E. M. Chetyrkin, I. L. Kalikhman

PROBABILITY AND STATISTICS

Moscow
«Finansy i Statistika»
1982

Е.М.Четыркин, И.Л.Калихман

ВЕРОЯТНОСТЬ И СТАТИСТИКА

Москва
«Финансы и статистика»
1982

ББК 22.17
Ч 54

Четыркин Е. М., Калихман И. Л.
Ч 54 Вероятность и статистика. — М.: Финансы и статистика, 1982. — 319 с., ил.
В пер.: 1 р. 50 к.

Излагаются общие теоретические основы теории вероятностей и математической статистики. Освещаются вопросы практического применения соответствующих методов в экономических исследованиях (анализ временных рядов, выборочных данных при оценке влияния факторов, определении качества продукции и т. д.).
Для экономистов, применяющих в своей работе вероятностные методы и методы математической статистики, а также для преподавателей и аспирантов вузов.

Ч $\frac{1702060000-096}{010(01)-82}$ 20—82

ББК 22.17
517.8

Оглавление

РАЗДЕЛ III

Математическая статистика

Введение

Задача любой науки и в том числе экономической состоит в конечном счете в выявлении и исследовании закономерностей, которым подчиняются реальные процессы. Найденные закономерности, относящиеся к экономике, имеют не только теоретическую, познавательную ценность, они широко применяются в практике — в планировании, управлении и прогнозировании. Как правило, экономические и социально-экономические процессы находятся под воздействием разнообразных факторов и многообразного переплетения взаимосвязей между отдельными элементами. В значительном числе случаев закономерности могут быть обнаружены при целенаправленном статистическом изучении массовых явлений, включающем сбор данных, их систематизацию и упорядочение и, наконец, статистический анализ. Поскольку наблюдаемый процесс или явление подвержено воздействию множества факторов, то каждое индивидуальное его проявление будет отличаться от другого (варьировать). Лишь в массовой совокупности объектов наблюдений проявляются общие закономерности, в формирование которых каждая единица совокупности вносит свой «вклад».

При достаточно большом объеме совокупности случайные воздействия в значительной мере взаимно погашаются (нейтрализуются) и результат (для совокупности в целом) становится мало зависящим от случая. Это явление лежит в основе образования статистических закономерностей. Статистические закономерности обладают известной *статистической устойчивостью*. Нетрудно понять, что устойчивость закономерности относительна и связана с неизменностью условий ее формирования. Существенное изменение этих условий неизбежно приводит к изменению самих закономерностей. Со статистической устойчивостью мы сталкиваемся в повседневной практике довольно часто, основывая на ней многие решения практического порядка. Например, характеризуя деятельность предприятия и принимая решения на дальнейший период, исходят из целого ряда средних характеристик: средней выработки одного работающего, процента брака, расхода сырья и материалов и т. д. Оперируя средними показателями, руководствуются их устойчивостью, хотя в индивидуальном проявлении (в определенные дни или у определенного рабочего) эти показатели будут колебаться в широких пределах.

Следует остановиться на весьма существенном различии между изучением закономерностей в естественных и социально-экономи-

ческих науках. Если в первых возможна постановка специальных целенаправленных экспериментов, позволяющих вычленить интересующие нас факторы и зависимости, то при изучении социально-экономических явлений возможности для постановки подобных экспериментов весьма ограниченны, а чаще всего такие эксперименты просто невозможны. Поэтому здесь основным источником информации оказывается непосредственное наблюдение реальных процессов во всем их многообразии. Лишь последующая обработка и анализ данных наблюдений позволяют выявить обособленное влияние исследуемого фактора, определить интересующую нас закономерность.

Приемы и способы научного анализа данных, относящихся к массовым явлениям, с целью определения некоторых обобщающих эти данные характеристик и выявления статистических закономерностей и составляют предмет *математической статистики.* В свою очередь математическая статистика опирается на *теорию вероятностей* — раздел математики, изучающий закономерности случайных явлений. Связь математической статистики и теории вероятностей обусловлена следующим. Методы, разрабатываемые в математической статистике, предполагают, что данные получают в ходе *выборочных наблюдений* или опытов. Таким образом, о свойствах совокупности приходится судить по результатам, которые в известном смысле случайны (попадание в выборку той или иной единицы совокупности — дело случая). Используя результаты, полученные в теории вероятностей, математическая статистика позволяет не только определить значения искомых характеристик, но и оценить степень точности получаемых при обработке выборочных данных выводов. Основным методом изучения в ней является построение абстрактных математических моделей этих закономерностей, получивших название *вероятностные модели.*

Модель представляет собой некоторое отображение (аналог) реальной действительности в основных, существенных для целей исследования чертах. Математическая модель описывает действительность на математическом языке, т. е. посредством различного рода математических выражений. Вероятностная модель — частный случай математической модели. Основным ее содержанием является соотношение между вероятностями отдельных случайных событий (отложим объяснение смысла понятий «вероятность» и «случайное событие» до гл. 3). Изучение вероятностных моделей позволяет понять различные свойства случайных событий на абстрактном и обобщенном уровне, *не прибегая к эксперименту.* В математической статистике исследование, наоборот, связано с конкретными данными и идет от практики (наблюдения) к гипотезе и ее проверке.

Цель настоящей книги заключается в описании методов математической статистики, которые имеют наибольшее применение в практике. Поскольку эти методы базируются на вероятностных моделях, то авторы сочли необходимым до изложения методов

математической статистики, чему собственно и посвящена книга, привести материал, который в сжатом виде дает представление об основных идеях теории вероятностей.

Книга делится на три раздела: статистические распределения, теория вероятностей и математическая статистика. В разделе I содержатся сведения о методах описания рядов распределения, полученных по данным массовых наблюдений. По существу, этот раздел представляет собой введение в математическую статистику, ее начальный раздел, который не опирается на абстрактные теоретико-вероятностные модели. Этот материал обычно относят к общей теории статистики, однако с целью сохранения целостности изложения и возможности содержательной интерпретации последующих теоретических построений мы сочли целесообразным предпослать его основным разделам книги.

Кроме того, формальный аппарат изучения статистических распределений в следующих разделах по аналогии распространяется на изучение распределений случайной величины. Читатель найдет здесь практические приемы обработки рядов распределения и получения простейших их характеристик.

В разделе II рассматриваются основные понятия теории вероятностей, необходимые для построения вероятностных моделей и последующего их исследования. Изложение здесь следует обычно принятому в математических курсах дедуктивному принципу изложения. Вначале формулируется ряд исходных определений и аксиом, а затем выводятся важные соотношения и теоремы. Без изучения этих вопросов нельзя обоснованно рассмотреть целый ряд методов математической статистики.

Раздел III посвящен изложению основных методов математической статистики, приложению вероятностных моделей к описанию конкретных ситуаций, статистическому оцениванию и выработке статистических решений. В основе этого раздела лежит концепция выборки, т. е. рассмотрение данных наблюдений, как результат выборки из некоторой конечной или гипотетической бесконечной генеральной совокупности. Таким образом, создается база для анализа статистических закономерностей по данным наблюдений и выработки статистических решений.

Особое место в книге занимает последняя глава этого раздела: «Введение в анализ временны́х рядов». В ней рассматривается проблема использования методов математической статистики при анализе статистической информации, заданной в виде временны́х (или динамических) рядов.

Раздел I

Статистические распределения

Глава 1

Распределение качественных и количественных признаков

1.1. Статистические совокупности

Множество однородных объектов, подлежащих статистическому изучению, называется *статистической совокупностью,* отдельные объекты — *элементами* (*единицами*) совокупности, а их число — *объемом совокупности.* Например, при статистическом изучении производительности труда может рассматриваться статистическая совокупность рабочих предприятия. В этом случае единицей совокупности является рабочий, а общая их численность — объемом совокупности. Можно рассматривать в качестве статистической совокупности множество предприятий данной отрасли, характеризуя каждое предприятие средней величиной производительности. В таком случае единицей совокупности будет каждое предприятие, а их число — объемом совокупности.

Элементы совокупности можно охарактеризовать одним или несколькими *признаками,* значения которых изменяются (варьируют) при переходе от одного элемента совокупности к другому. Признак может быть качественным (атрибутивным) или количественным. *Качественный* признак характеризует некоторое свойство или состояние наблюдаемой единицы совокупности (например, пол, профессия, сорт продукции, цвет и т. д.), а также наличие или отсутствие данного свойства (бракованное и небракованное изделие и др.). *Количественный* — это признак, отдельные значения которого (варианты), получаемые в результате измерения, наблюдения или счета, выражаются числами (например, рост, масса, объем, заработная плата, выпуск продукции, прибыль и т. д.).

К особой группе признаков следует отнести так называемые *ранговые* показатели. Их значения устанавливаются путем упорядочения (ранжирования) объектов совокупности в соответствии с некоторым признаком и присвоения каждому из объектов в таком упорядоченном ряду некоторого числа (обычно порядкового

номера) — *ранга*. Статистический анализ данных, указанных в ранговой шкале, может производиться лишь определенными, так называемыми *непараметрическими*, методами.

Количественные признаки могут быть *дискретными*, принимающими лишь отдельные, изолированные значения, и *непрерывными*, которые могут принимать в определенных пределах любые численные значения. Примерами дискретных признаков (обычно как результатов счета) могут служить: число рабочих, число машин, число отработанных дней и т. д. В этих примерах варианты могут выражаться лишь натуральными числами. Примерами непрерывных признаков (как результатов измерений) могут служить: масса или размер детали, процент выполнения нормы, длительность обработки и т. д.

Заметим, что указанное определение непрерывного признака носит условный характер. Так, при взвешивании с точностью до 1 г мы будем всегда получать массу в виде дискретных величин (число граммов), хотя сам признак является непрерывным.

Для изучения вариации признака производится *статистическое наблюдение* (измерение, регистрация, описание и т. д.), в результате которого определяются значения признака (варианты), соответствующие каждому элементу совокупности.

По охвату совокупности различают два вида статистических наблюдений: сплошное и выборочное. При *сплошном наблюдении* изучается каждый элемент совокупности. Однако часто такое наблюдение сопряжено со значительными затратами труда. Иногда оно может оказаться принципиально неосуществимым, например когда сам процесс наблюдения связан с уничтожением наблюдаемого объекта. В таких случаях прибегают к *выборочному наблюдению*, в основе которого лежит выделение из статистической совокупности некоторой ее части — *выборочной совокупности*, или кратко *выборки*. Исходная совокупность именуется в таком случае *генеральной*.

По результатам изучения вариации признака в выборочной совокупности делают заключение о свойствах признака в генеральной совокупности. Изучению может подлежать вариация одного или нескольких признаков. Вариация признака может изучаться либо без учета времени регистрации наблюдений, либо с учетом их хронологической последовательности. Методы анализа хронологических (временны́х) *рядов* рассматриваются в последней главе.

Для правильной организации статистического наблюдения должна быть прежде всего четко сформулирована задача, для решения которой оно проводится, определен признак, вариация которого изучается, и точно очерчена статистическая совокупность. Наконец, должен быть указан способ проведения наблюдений (сплошное или выборочное), порядок отбора элементов совокупности, способ регистрации данных и необходимая точность измерений. Одним из основных требований при организации статистических наблюдений является обеспечение *однородности* совокуп-

ности. Ограничимся на данном этапе лишь следующим общим указанием: единицы совокупности должны быть по всем другим, кроме изучаемого признака, существенно неразличимы, или, лучше сказать, *сопоставимыми*. Например, изучая вариацию производительности труда в совокупности предприятий, целесообразно включать в нее предприятия одного типа, одной отрасли и т. д., в противном случае ценность полученной информации снижается.

1.2. Распределение качественных признаков

Пусть элементы совокупности характеризуются одним качественным альтернативным признаком, относительно которого возможны лишь два суждения: «признак есть» (обозначим через A) или «признака нет» ($\bar{A}$). Так, при проверке годности изделий из некоторой партии (статистическая совокупность) изучается признак «брак». Любое изделие (элемент совокупности) может оказаться годным (A) или бракованным ($\bar{A}$). Результаты наблюдений могут быть представлены в виде табл. 1.1.

Таблица 1.1

Элемент совокупности	1	2	3	4	...	n
Результат наблюдения	A	$\bar{A}$	A	A	...	$\bar{A}$

Таблица 1.2

Варианта	A	$\bar{A}$	$\sum$
Частота	m_A	$m_{\bar{A}}$	n

Для удобства обозрения и дальнейшего анализа произведем группировку данных табл. 1.1, разбив совокупность на две группы: элементы, обладающие признаком A, и элементы, не обладающие этим признаком. Такая группировка называется *двухразрядной*. Каждую из двух групп достаточно охарактеризовать числом элементов в ней — *частотой* и обозначить соответственно через m_A и $m_{\bar{A}}$. Очевидно, что $m_A + m_{\bar{A}} = n$. В результате указанной группировки получим *распределение альтернативного признака* в виде табл. 1.2.

Разделив частоты m_A и $m_{\bar{A}}$ на n, получим *относительные частоты* $w_A = \frac{m_A}{a}$ и $w_{\bar{A}} = \frac{m_{\bar{A}}}{n}$, которые называются также *долями* признака.

Пусть теперь изучается вариация некоторого признака A, имеющего не две, а несколько вариант (A_1, ..., A_i, ..., A_s). Например, признак A — «сортность» может иметь три варианты: A_1 — I сорт, A_2 — II сорт, A_3 — III сорт. Произведя уже не двухразрядную, а *многоразрядную* (множественную) *группировку*, придем к распределению, указанному в табл. 1.3. Очевидно, что $\sum_{k=1}^{s} m_k = n$.

Таблица 1.3

Варианта	A_1	...	A_k	...	A_s	$\sum$
Частота	m_1	...	m_k	...	m_s	n

Табл. 1.3 также может быть перестроена в ряд относительных частот $w_k = \frac{m_k}{n}$. В этом случае $\sum w_k = 1$.

Рассмотрим теперь совокупность, элементы которой характеризуются двумя альтернативными признаками — A и B. Будем называть *группировкой нулевого порядка* рассмотрение всей совокупности как одной группы, *группировками первого порядка* — разбиение совокупности по каждому признаку на две группы и т. д. При изучении двух признаков более существенной является группировка совокупности по сочетанию вариант обоих признаков, т. е. образование четырех групп из элементов: 1) с вариантами AB; 2) с вариантами $A\bar{B}$; 3) с вариантами $\bar{A}B$; 4) с вариантами $\bar{A}\bar{B}$. Такое разбиение является *группировкой второго порядка*. Обозначив численности (частоты) указанных четырех групп через m_{AB}, $m_{A\bar{B}}$, $m_{\bar{A}B}$ и $m_{\bar{A}\bar{B}}$, запишем результат этой группировки в виде четырехклеточной табл. 1.4. Данные итоговых столбца и строки (обозначенных символом $\sum$) характеризуют результаты группировок первого порядка по каждому из признаков и нулевого порядка (число n).

Таблица 1.4

A \ B	B	$\bar{B}$	$\sum$
A	m_{AB}	$m_{A\bar{B}}$	m_A
$\bar{A}$	$m_{\bar{A}B}$	$m_{\bar{A}\bar{B}}$	$m_{\bar{A}}$
$\sum$	m_B	$m_{\bar{B}}$	n

Разделив все данные табл. 1.4 на n, получим аналогичную четырехклеточную таблицу в относительных частотах. Между величинами, содержащимися в табл. 1.4, справедливы следующие очевидные соотношения:

$$\left.\begin{aligned} &m_A + m_{\bar{A}} = n;\ m_B + m_{\bar{B}} = n;\ m_{AB} + m_{A\bar{B}} = m_A;\\ &m_{AB} + m_{\bar{A}B} = m_B;\ m_{\bar{A}B} + m_{\bar{A}\bar{B}} = m_{\bar{A}};\ m_{A\bar{B}} + m_{\bar{A}\bar{B}} = m_{\bar{B}}. \end{aligned}\right\} \quad (1.1)$$

Можно показать, что из этих шести уравнений одно является следствием остальных, т. е. линейно-независимых среди них пять. Поэтому задание любых четырех величин однозначно определяет всю табл. 1.4. Однако только частоты m_{AB}, $m_{A\bar{B}}$, $m_{\bar{A}B}$, $m_{\bar{A}\bar{B}}$ могут задаваться произвольно. Если же задать значения любых четырех частот, то при определении остальных из уравнений (1.1) можно получить для некоторых из них отрицательные значения. Это будет указывать на несовместимость исходных данных.

Подобно тому как строились группировки нулевого, первого и второго порядков по двум альтернативным признакам, можно образовывать многоступенчатые группировки по трем и более признакам.

В случае двух признаков A и B, каждый из которых характеризуется несколькими вариантами ($A_1, \ldots, A_i, \ldots, A_s, B_1, \ldots, B_k, \ldots, B_t$), результатом группировки первого порядка будет образование s групп по признаку A и t групп — по признаку B, а результатом группировки второго порядка (по сочетанию вариант двух признаков) — образование st групп. Данные этих группировок указываются в таблице, которая является обобщением четырехклеточной таблицы и получила название *таблицы взаимосопряженности.*

1.3. Распределение количественных признаков

Пусть результаты изучения вариации признака X в совокупности из n элементов представлены в виде перечня значений x_k.

Среди наблюденных вариант некоторые могут оказаться совпадающими.

Обилие и неупорядоченность информации, содержащейся в перечне, затрудняет ее использование для дальнейшего анализа. Поэтому данные наблюдений подвергают первичной обработке, состоящей в группировке совокупности по отдельным вариантам. Каждая группа состоит из элементов с одинаковым значением признака x_i и характеризуется численностью m_i, называемой *частотой* варианты x_i.

Расположив все различные варианты в возрастающем порядке ($x_1 < x_2 < \ldots < x_k < \ldots < x_s$) и указав против каждой варианты соответствующую ей частоту, получим *дискретный вариационный ряд*, или ряд распределения частот:

Таблица 1.5

X	x_1	x_2	$\ldots$	x_k	$\ldots$	x_s	$\sum$
m	m_1	m_2	$\ldots$	m_k	$\ldots$	m_s	n

Разделив все частоты на объем совокупности n, получим аналогичное распределение *относительных частот* $w_i = \frac{m_i}{n}$. Для ряда частот или относительных частот выполняется соотношение

$$\sum_{k=1}^{s} m_k = n \quad \text{или} \quad \sum_{k=1}^{s} w_k = 1. \tag{1.2}$$

Данные табл. 1.5 могут быть изображены графически в виде точечной диаграммы (рис. 1.1).

Введем понятие *накопленной*, или *кумулятивной*, частоты (относительной частоты), равной числу элементов совокупности (относительному числу элементов), имеющих значение признака X меньше данного числа. Обозначим накопленную частоту через $H(x)$ и относительную накопленную частоту — через $F(x)$. Метод их расчета одинаков. Так, по ряду относительных частот получим

$$F(x)=\begin{cases} 0 \quad \text{при} \quad x \leqslant x_1; \\ w_1 \quad \text{при} \quad x_1 < x \leqslant x_2; \\ w_1 + w_2 \quad \text{при} \quad x_2 < x \leqslant x_3; \\ \cdots\cdots\cdots\cdots\cdots \\ w_1 + \ldots + w_k \quad \text{при} \quad x_k < x \leqslant x_{k+1}; \\ \cdots\cdots\cdots\cdots\cdots \\ w_1 + \ldots + w_k + \ldots + w_s \quad \text{при} \quad x > x_s. \end{cases} \tag{1.3}$$

Поясним построение этих формул. Для любого $x \leqslant x_1$ нет элементов, у которых $X < x$, следовательно, $F(x) = 0$. Если x принимает значения в интервале $x_1 < x \leqslant x_2$, то неравенство $X < x$ выполняется только для элементов со значением признака $X = x_1$. Число таких элементов m_1, поэтому $F(x) = w_1$. Если $x_2 < x \leqslant x_3$, то неравенство $X < x$ будет выполняться для m_1 элементов со значением $X = x_1$ и для m_2 элементов со значением $X = x_2$, откуда $F(x) = w_1 + w_2$ и т. д.

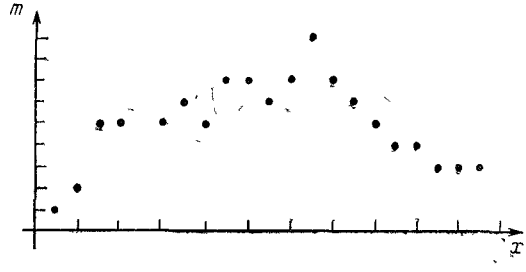


Рис. 1.1. Точечная диаграмма дискретного распределения

Аналогично рядам распределения частот строятся *кумулятивные вариационные ряды:*

Таблица 1.6

X	x_1	x_2	...	x_s
H	m_1	$m_1 + m_2$	...	$m_1 + m_2 + \ldots + m_s$

Таблица 1.7

X	x_1	x_2	...	x_s
F	w_1	$w_1 + w_2$	...	$w_1 + w_2 + \ldots + w_s$

Для графического их изображения по оси абсцисс откладываются значения признака, а по оси ординат — значения накопленной частоты $H(x)$ или относительной накопленной частоты $F(x)$. В результате получим (рис. 1.2) график функции $F(x)$, заданной

уравнениями (1.3), т. е. *кумуляту*. График представляет собой ступенчатую линию, имеющую разрывы (конечные скачки) в точках x_1, x_2 и т. д.

Дискретное распределение оказывается при большом числе различных вариант малоудобным для последующего анализа. В таком случае прибегают к интервальной группировке, сущность которой состоит в следующем. Весь диапазон изменения признака от наименьшего ($x_{\min}$) до наибольшего ($x_{\max}$) разбивают на

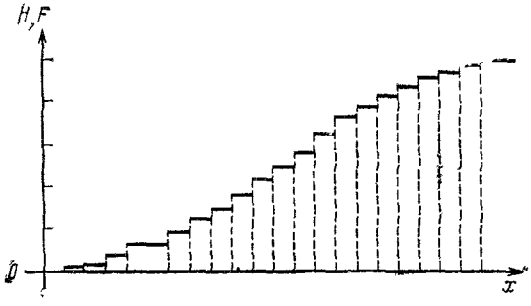


Рис. 1.2. Кумулята дискретного распределения

определенное число равных или неравных интервалов (x_0, x_1), (x_1, x_2), ..., (x_{t-1}, x_t) и собирают в отдельные группы элементы совокупности, признак которых принимает значения в каждом из интервалов. Обозначив, как и ранее, численности групп через $m_1, m_2, \ldots, m_t$ и сведя результат группировки в табл. 1.8, получим *интервальный вариационный ряд*, или *интервальное распределение*.

Таблица 1.8

Интервал	$x_0 - x_1$ $x_1 - x_2$ ... $x_{k-1} - x_k$ ... $x_{t-1} - x_t$	$\sum$
m	m_1 m_2 ... m_k ... m_t	n
w	w_1 w_2 ... w_k ... w_t	1

Как и ранее, данные таблицы можно перестроить в ряд относительных частот (третья строка).

Построение интервального ряда оказывается тем более целесообразным, когда речь идет об изучении вариации непрерывного признака. При этом надо интервалы выбирать так, чтобы признак был почти равномерно распределен в каждом из них. Кроме того, следует стремиться передать особенности вариации признака во всей области его изменения. Как правило, это удается осуществить (особенно если совокупность качественно неоднородна) выбором неравных интервалов, однако это существенно затрудняет дальнейшую обработку ряда. Поэтому в практике чаще всего данные наблюдений представляют в виде рядов с равными интервалами.

При использовании интервалов различной длины анализ интервального ряда становится затруднительным, так как численности групп m_k (или w_k) будут зависеть не только от свойств признака (*исследуемой вариации*), но и от выбранной длины интервала. Так, например, если интервалу в 10 единиц соответствует частота $m_1 = 30$ и другому интервалу длиной в 4 единицы соответствует частота $m_2 = 20$, то отсюда отнюдь не следует, что вблизи середины первого интервала значения признака встречаются чаще, чем вблизи середины второго интервала. Наоборот, на единицу длины первого интервала приходится $\frac{30}{10} = 3$ элемента совокупности, а на единицу длины второго — $\frac{20}{4} = 5$ элементов совокупности. Это приводит к целесообразности введения нового понятия *плотности частоты* $\frac{m_k}{\Delta x_k}$ или соответственно *плотности относительной частоты*

$$f_k = \frac{w_k}{\Delta x_k} = \frac{m_k}{n\,\Delta x_k}, \qquad (1.4)$$

характеризующей число элементов, или относительное их число, приходящееся на единицу длины интервала.

При изучении интервального ряда с неравными интервалами целесообразно либо перестроить его в кумулятивный ряд, либо пересчитать частоты в соответствующие плотности частоты.

Пусть, например, вариация признака среди 100 элементов совокупности задана первыми двумя строками табл 1.9. Наибольшая относительная частота 0,4 приходится на интервал 5—15, но и длина этого интервала наибольшая. После перехода к плотности относительной частоты f устанавливаем, что максимум ее $f_3 = 0{,}08$ соответствует интервалу 15—18, т. е. вблизи его середины группируется относительно наибольшее число элементов совокупности.

Таблица 19

Интервал	1—5	5—15	15—18	18—25	$\sum$
w	0,08	0,40	0,24	0,28	1
F	0,08	0,48	0,72	1,00	—
f	0,02	0,04	0,08	0,04	—

Интервальный ряд может быть условно перестроен в дискретный путем замены каждого интервала его серединой.

Для графического изображения интервальных рядов используются два вида графиков: *гистограмма* — для ряда частот (относительных частот) и *кумулята* — для кумулятивных рядов. В том и другом случае исходят из предположения, лежащего в основе

построения интервального ряда, о постоянстве плотности на каждом интервале.

Для построения гистограммы по оси абсцисс откладывают интервалы, а по оси ординат — соответствующие им плотности ча-

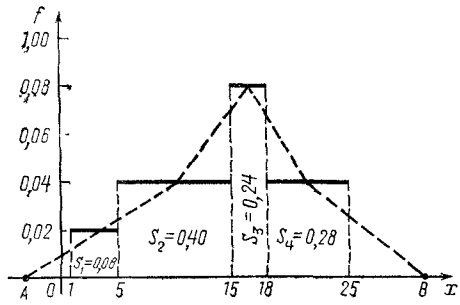


Рис. 1.3.
Гистограмма (сплошная линия) и полигон (пунктирная линия) интервального распределения

стоты (относительной частоты). В каждом интервале плотность предполагается постоянной. На рис. 1.3 изображена гистограмма, построенная по данным табл. 1.9.

При построении кумуляты по данным кумулятивного ряда исходят из той же предпосылки о постоянстве плотности частоты на

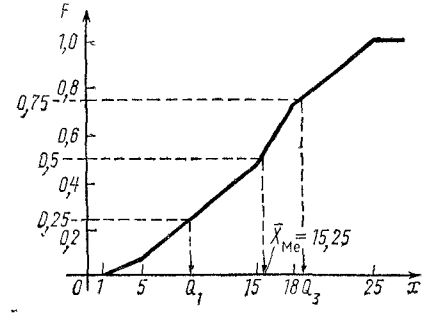


Рис. 1.4.
Кумулята интервального распределения

данном интервале. На рис. 1.4 построена кумулята, соответствующая данным табл. 1.9.

Используется еще один вид графического изображения ряда, получивший название *полигона распределения*. Полигон строится в тех же координатах, что и гистограмма, но в отличие от последней предполагается, что плотность частоты (относительной частоты), отнесенная в каждом интервале к его середине, меняется при переходе к соседнему интервалу по линейному закону. На рис. 1.3 показан пунктирной линией полигон распределения, полученный в результате перестройки гистограммы.

Можно показать, что площадь полигона, как и гистограммы, равна n, если построение велось для ряда частот, или 1 — для ряда относительных частот.

Условно полигон распределения используется и для изображения дискретных рядов. В этом случае последовательно соединяют отрезками прямых точки на точечной диаграмме.

Непосредственным обобщением статистического распределения одного признака являются *многомерные распределения*, получаемые в результате группировки элементов совокупности по сочетаниям значений двух или большего числа признаков (см. гл. 12).

Глава 2

Числовые характеристики статистических распределений

2.1. Общие замечания

Статистическое распределение содержит полную информацию о вариации признака. Однако обилие числовых данных, с помощью которых оно задается, усложняет их использование. Между тем при решении многих задач эта информация оказывается избыточной и может быть успешно заменена небольшим количеством сводных характеристик распределения. Более того, искусство статистического анализа как раз и определяется не непосредственным использованием статистического распределения, а тем, насколько успешно и полно удается такой анализ осуществить, опираясь на минимальное число показателей.

Для описания статистических распределений обычно используются три вида характеристик (показателей): 1) средние, или характеристики центральной тенденции; 2) характеристики изменчивости вариант (рассеяния); 3) характеристики, отражающие дополнительные особенности распределений, в частности их форму. Все они вычисляются по результатам наблюдений и построенных в результате их первичной обработки одномерных или многомерных распределений. Поэтому будем называть их *опытными* или *статистическими*. Расчет статистических характеристик представляет собой второй после группировки этап обработки данных наблюдений. Этим, конечно, не исчерпывается статистический анализ, основной целью которого является установление общих закономерностей вариации признака. Однако такой анализ опирается на построение теоретико-вероятностной модели и будет изложен в дальнейшем после подготовки необходимого математического аппарата.

В настоящей главе статистические характеристики рассматриваются лишь как средство компактного описания основных свойств распределений. Для каждой из характеристик предложены различные конкретные формы, выбор которых диктуется условиями задач.

2.2. Средние характеристики

Средняя характеризует типичный для совокупности размер признака, или, как иногда говорят, центральную тенденцию в распределении. Очевидно, практическое использование такой

характеристики целесообразно в том случае, когда отдельные варианты ряда распределения концентрируются вблизи некоторого значения. Если же совокупность крайне неоднородна, результаты наблюдений значительно отличаются друг от друга и не обнаруживают общей тенденции, ее использование становится чисто формальным.

Существуют различные формы средних. К выбору формы средней следует подходить, руководствуясь задачей исследования и определяющим свойством распределения, которое должно быть выражено этой характеристикой.

Основным видом средних является *средняя арифметическая*, определяемая формулой

$$\bar{X}_{\text{a}} = \frac{1}{n}\sum_k x_k \cdot m_k = \sum_k x_k w_k, \tag{2.1}$$

где x_k — значения признака в дискретном или середины интервалов в интервальном распределении; m_k — соответствующие им частоты; $n = \Sigma m_k$.

Величины m_k или w_k в этой формуле часто называют *весами*, а характеристику $\bar{X}_{\text{a}}$ — *средневзвешенной*.

Важнейшее свойство средней арифметической описывается следующей теоремой.

Теорема. Сумма отклонений вариант от средней равна нулю.

Действительно,

$$\sum (x_k - \bar{X}_{\text{a}})\, m_k = \sum x_k \cdot m_k - \bar{X}_{\text{a}} \sum m_k = n\bar{X}_{\text{a}} - \bar{X}_{\text{a}} \sum m_k = 0$$

так как $\Sigma m_k = n$.

Далеко не во всех случаях типичный размер признака можно охарактеризовать с помощью средней арифметической. Поэтому применяются и иные формы средней, которые легко могут быть получены из так называемой *средней степенной*, определяемой формулой

$$\bar{X}_{\nu} = \sqrt[\nu]{\frac{\sum x_k^{\nu} \cdot m_k}{\sum m_k}} = \sqrt[\nu]{\sum x_k^{\nu} w_k}, \quad \text{где} \quad x_k > 0. \tag{2.2}$$

Нетрудно установить, что при $\nu = 1$ формула (2.2) соответствует средней арифметической $\bar{X}_1 = \bar{X}_{\text{a}}$. При $\nu = -1$ получим *среднюю гармоническую:*

$$\bar{X}_{-1} = \left(\frac{\sum x_k^{-1} \cdot m_k}{\sum m_k}\right)^{-1} = \frac{\sum m_k}{\sum m_k / x_k}. \tag{2.3}$$

Как видно из формулы, эту среднюю целесообразно применять в том случае, когда усреднению подлежат величины, обратные значению признака.

При $\nu = 2$ получим *среднюю квадратическую*:

$$\bar{X}_2 = \frac{\sum x_k^2 \cdot m_k}{\sum m_k}. \tag{2.4}$$

При $\nu = 0$ и $x_k > 0$ формула (2.2) приводит к неопределенности, после раскрытия которой получаем

$$\bar{X}_0 = \sqrt[n]{x_1^{m_1} x_2^{m_2} \dots x_k^{m_k} \dots x_s^{m_s}}. \qquad (2.5)$$

Это *средняя геометрическая*, она применяется преимущественно при анализе временны́х рядов.

Логарифмируя обе части равенства (2.5), получим

$$\lg \bar{X}_0 = \frac{1}{n} \sum m_k \cdot \lg x_k, \qquad (2.6)$$

т. е. логарифм средней геометрической равен средней арифметической логарифмов вариант. Можно доказать, что $\bar{X}_\nu$ возрастает с ростом ν, т. е.

$$\bar{X}_{-1} \leqslant \bar{X}_0 \leqslant \bar{X}_1 < \bar{X}_2 \dots$$

Эти соотношения, получившие название свойства *мажорантности средних*, были впервые доказаны А. Я. Боярским.

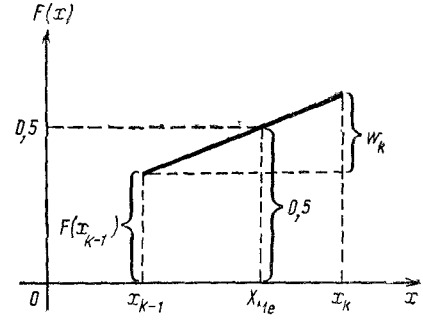


Рис. 2.1.
Нахождение медианы с помощью линейного интерполирования

Кроме рассмотренных средних, называемых *аналитическими*, в статистическом анализе применяют *структурные*, или *порядковые средние*. Из них наиболее широко применяются медиана и мода.

Медиана $\bar{X}_{Me}$ — *это значение признака*, разделяющее ранжированную совокупность на две равные по численности группы: первую, содержащую элементы со значениями признака $x < \bar{X}_{Me}$, и вторую — со значениями признака $x > \bar{X}_{Me}$. Из определения накопленной относительной частоты следует, что

$$F(\bar{X}_{Me}) = 0{,}5. \qquad (2.7)$$

Для интервального ряда находится *медианный интервал* (x_{k-1}, x_k) из условия $F(x_{k-1}) \leqslant 0{,}5$ и $F(x_k) \geqslant 0{,}5$. Значение медианы на этом интервале находят с помощью линейного интерполирования (рис. 2.1):

$$\bar{X}_{Me} = x_{k-1} + \frac{0{,}5 - F(x_{k-1})}{w_k} (x_k - x_{k-1}). \qquad (2.8)$$

Медиана может быть определена графически с помощью кумуляты (см. рис. 1.4).

Мода $\bar{X}_{\text{Мо}}$ — это значение признака, которому соответствует наибольшая частота. Во многих случаях эта величина является наиболее характерной для ряда распределения, и вокруг нее концентрируется бóльшая часть вариант. При изменении распределения в его «хвостах» мода не меняется, т. е. она обладает определенной устойчивостью к вариации признака. Поэтому ее особенно удобно применять при изучении рядов с неопределенными границами. Моду целесообразно применять также и в том случае, когда при изучении вариации признака важно выделить одну превалирующую над всеми другими частоту.

Для дискретного ряда мода находится непосредственно по определению. Для интервального распределения определяется модальный интервал $(x_{k-1},\ x_k)$, которому соответствует наибольшая плотность относительной частоты. Величина моды внутри модального интервала определяется по интерполяционной формуле:

$$\bar{X}_{\text{Мо}} = x_{k-1} + \frac{f(x_{k+1})}{f(x_{k+1}) + f(x_{k-1})}(x_k - x_{k-1}). \qquad (2.9)$$

Для строго симметричного распределения, у которого частоты вариант, равноотстоящих от моды, равны, значения средней арифметической, моды и медианы совпадают, т. е. $\bar{X}_{\text{а}} = \bar{X}_{\text{Ме}} = \bar{X}_{\text{Мо}}$. При нарушении симметрии эти три показателя расходятся. Однако если нарушение симметрии не очень сильно выражено, то между указанными тремя видами средней будет выполняться приближенное равенство

$$\bar{X}_{\text{а}} - \bar{X}_{\text{Мо}} \approx 3(\bar{X}_{\text{а}} - \bar{X}_{\text{Ме}}). \qquad (2.10)$$

В заключение вычислим $\bar{X}_{\text{а}}$, $\bar{X}_{\text{Ме}}$ и $\bar{X}_{\text{Мо}}$ по данным табл. 1.9:

$$\bar{X}_{\text{а}} = 3 \cdot 0{,}08 + 10 \cdot 0{,}40 + 16{,}5 \cdot 0{,}24 + 21{,}5 \cdot 0{,}28 = 14{,}22;$$

$$\bar{X}_{\text{Ме}} = 15 + \frac{0{,}5 - 0{,}48}{0{,}24}(18 - 15) = 15{,}25;$$

$$\bar{X}_{\text{Мо}} = 15 + \frac{0{,}04}{0{,}04 + 0{,}04}(18 - 15) = 16{,}5.$$

2.3. Характеристики рассеяния

Рассмотренные в предыдущем параграфе средние тем более характерны для данного распределения, чем теснее группируются отдельные варианты вокруг средней, т. е. чем менее они *рассеяны*. Поэтому средние характеристики должны быть дополнены измерением вариации признака относительно средней, т. е. характеристиками рассеяния. В практическом анализе оценка рассеяния может оказаться не менее важной, чем определение средней.

Самая грубая оценка рассеяния, легко определяемая по данным вариационного ряда, может быть произведена с помощью размаха варьирования

$$R = x_{\max} - x_{\min}, \qquad (2.11)$$

где $x_{\max}$ и $x_{\min}$ — наибольшая и наименьшая из вариант ряда.

Однако этот показатель не дает представления о характере вариационного ряда, расположении вариант вокруг средней и может сильно меняться от добавления или исключения крайней варианты даже с минимальной частотой.

Для оценки колеблемости значений признака относительно средней используются различные *характеристики рассеяния*, отличающиеся друг от друга выбранной формой средней и способами оценки отклонений от нее отдельных вариант. Перестроим дискретный вариационный ряд (табл. 1.5), заменив варианты x_k на их отклонения от средней $x_k - \bar{X}$.

Однако при усреднении отклонения, имеющие разные знаки, при суммировании будут взаимно погашаться, не давая истинного представления о величине отклонений. В частном случае, если $\bar{X} = \bar{X}_a$, то, как следует из доказанной в § 2.2 теоремы, сумма отклонений будет для любого исходного вариационного ряда равна нулю.

Для устранения влияния знака отклонений переходят либо к абсолютным величинам отклонений (табл. 2.1), либо к квадратам отклонений (табл. 2.2).

Таблица 2.1

$\lvert X-\bar{X}\rvert$	$\lvert x_1-\bar{X}\rvert \ldots \lvert x_k-\bar{X}\rvert \ldots \lvert x_s-\bar{X}\rvert$
m	$m_1 \quad \ldots \quad m_k \quad \ldots \quad m_s$

Таблица 2.2

$(X-\bar{X})^2$	$(x_1-\bar{X})^2 \ldots (x_k-\bar{X})^2 \ldots (x_s-\bar{X})^2$
m	$m_1 \quad \ldots \quad m_k \quad \ldots \quad m_s$

На основе данных табл. 2.1 могут быть построены две характеристики рассеяния:

1) *среднее линейное (среднее абсолютное) отклонение*, определяемое формулой

$$E_a = \frac{1}{n} \sum_k |x_k - \bar{X}| m_k; \qquad (2.12)$$

2) *срединное (среднее вероятное)* отклонение E, определяемое как медиана ряда табл. 2.1, т. е. абсолютная величина отклонения, которая не превышается у половины наблюденных вариант.

Первая является естественной мерой рассеяния, однако в силу отсутствия у нее некоторых математических свойств ею пользуются сравнительно редко. Вторая связана с определенной формой распределения, и поэтому для любых распределений ею пользоваться также нецелесообразно.

На основании данных табл. 2.2 может быть определен средний квадрат отклонения (*дисперсия*), вычисляемый по формуле

$$D = \sigma^2 = \frac{1}{n} \sum_k (x_k - \bar{X})^2 m_k. \qquad (2.13)$$

Эта характеристика, как мы увидим из дальнейшего изложения, обладает рядом важных математических свойств. Однако

вследствие суммирования *квадратов отклонений* дисперсия дает искаженное представление о самой величине отклонений, измеряя их в квадратных единицах. Поэтому на базе дисперсии вводятся две характеристики:

1) *среднее квадратическое отклонение* σ, равное корню квадратному из дисперсии, т. е.

$$\sigma = \sqrt{D} = \sqrt{\frac{1}{n}\sum_k (x_k - \bar{X})\, m_k}\ ; \qquad (2.14)$$

2) *коэффициент вариации* V, равный процентному отношению среднего квадратического отклонения к средней, т. е.

$$V = \frac{\sigma}{\bar{X}}\,100\%, \text{ где } \bar{X} \neq 0. \qquad (2.15)$$

Величина σ является наиболее удобной и чаще всего применяемой характеристикой рассеяния. Поэтому будем в дальнейшем производить оценку рассеяния с помощью σ.

Что касается коэффициента вариации, то он имеет ту же природу, что и σ, и оказывается более удобным для сравнительной оценки вариации в распределениях с разными значениями средней. Однако он теряет смысл при $\bar{X} = 0$ и становится малонадежным при близких к нулю значениях средней.

При расчете характеристик D, σ и V по формулам (2.13)—(2.15) используется среднее значение признака в форме средней арифметической.

Кроме перечисленных характеристик рассеяния, имеющих аналитическую форму, используются структурные характеристики, в основе определения которых лежит разделение совокупности на различные группы. Так, перестроив вариационный ряд в ряд накопленных относительных частот $F(x)$, можно определить по аналогии с медианой величины, именуемые *квартилями* и делящие совокупность на четыре равные части: первый — Q_1, второй — Q_2, третий — Q_3, которые находятся из равенств:

$$F(Q_1) = 0{,}25; \quad F(Q_2) = 0{,}5 \quad \text{и} \quad F(Q_3) = 0{,}75. \qquad (2.16)$$

Очевидно, что $Q_2 = \bar{X}_{\text{Ме}}$. Так как значения признака $x < Q_1$ и $x > Q_3$ будут встречаться у $25\% + 25\% = 50\%$ элементов совокупности, то интервал $Q_1 \leqslant x \leqslant Q_3$ также охватит половину элементов совокупности. Так как где-то на интервале (Q_1, Q_3) (не обязательно в его середине) будет располагаться медиана, характеризующая среднее значение признака, то естественно принять в качестве меры рассеяния, которую называют *квартильным отклонением* К, половину длины этого интервала, т. е.

$$\text{К} = \frac{Q_3 - Q_1}{2}. \qquad (2.17)$$

Квартили могут быть определены либо графически с помощью кривой $y = F(x)$ (см. рис 1.4), либо по интерполяционным формулам, аналогичным (2.6).

В заключение приведем пример расчета всех указанных характеристик рассеяния по данным табл. 19. Построив ряд абсолютных величин отклонений от средней арифметической, найденной в конце предыдущего параграфа ($\bar{X}_a = 14{,}22$), получим табл. 23. Используя формулу (2.12), получаем $E_a = 11{,}22 \times 0{,}08 + 4{,}22 \cdot 0{,}40 + 1{,}78\ 0{,}24 + 7{,}28\ 0{,}28 = 5{,}05$.

Для расчета срединного отклонения E необходимо прежде всего, расположив отклонения в возрастающем порядке, перейти к ряду накопленных частостей (табл. 24).

Таблица 23

$\lvert x - \bar{X}_a \rvert$	11,22	4,22	1,78	7,28
w	0,08	0,40	0,24	0,28

Таблица 24

$\lvert x - \bar{X}_a \rvert$	1,78	4,22	7,28	11,22
$F(x)$	0,24	0,64	0,92	1,0

Из таблицы видно, что медиана этого ряда заключена между 1,78 и 4,22, откуда с помощью интерполирования находим $E = 1{,}78 + \frac{0{,}5 - 0{,}24}{0{,}4}(4{,}22 - 1{,}78) = 3{,}37$. Построив ряд квадратов отклонений, определяем

$$D = 11{,}22^2 \cdot 0{,}08 + 4{,}22^2 \cdot 0{,}4 + 1{,}78^2 \cdot 0{,}24 + 7{,}28^2 \cdot 0{,}28 = 32{,}79,$$

откуда

$$\sigma = \sqrt{32{,}79} = 5{,}73 \quad \text{и} \quad V = \frac{5{,}73}{14{,}22} \cdot 100\% = 40\%.$$

Определив по рис. 14 $Q_1 = 9{,}5$ и $Q_3 = 19$, получим $K = \frac{19 - 9{,}5}{2} = 4{,}75$.

Как видим, мы получили сопоставимые значения показателей рассеяния:

$$E_a = 5{,}05,\ E = 3{,}37,\ \sigma = 5{,}73 \text{ и } K = 4{,}75.$$

2.4. Моменты распределения. Характеристики формы

Рассмотренных в предыдущих параграфах двух основных характеристик недостаточно для подробного описания свойств распределения, поэтому вводят дополнительные характеристики. Более удобной и логически обоснованной является система характеристик, построенных по тому же принципу, что и средняя арифметическая и дисперсия. Такие характеристики, впервые предложенные П. Л. Чебышевым, получили название *моментов распределения*.

Выберем произвольное число C. Моментом l-го порядка (где l — натуральное число) относительно постоянной C (обозначается через $M_l(C)$) является средняя арифметическая l-х степеней отклонений x_k от числа C, т. е.

$$M_l(C) = \frac{1}{n}\sum_k (x_k - C)^l m_k. \tag{2.18}$$

Если $C = 0$, то момент называется *начальным* и обозначается через $\mu_l(x)$ или μ_l, т. е.

$$\mu_l = \frac{1}{n}\sum_k x_k^l m_k. \tag{2.19}$$

Если $C = \bar{X}_a$, то момент называется *центральным* и будет обозначаться через $\nu_l(X)$ или ν_l, т. е.

$$\nu_l = \frac{1}{n}\sum_k (x_k - \bar{X}_a)^l m_k. \qquad (2.20)$$

Моменты распределения представляют собой систему числовых характеристик, с помощью которых можно описать все особенности вариации признака (среднюю тенденцию, рассеяние, форму распределения и др.). Чем больше моментов для данного распределения вычислено, тем, вообще говоря, точнее можно описать свойства распределения. Однако с ростом порядка l растет влияние случайных погрешностей в наблюдаемых или измеренных значениях вариант, поэтому на практике используются лишь моменты до четвертого порядка.

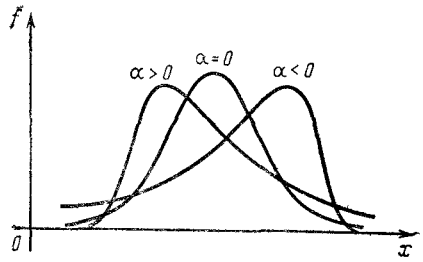


Рис. 2.2. Асимметия распределения

Характеристики распределения строятся на базе начальных или центральных моментов, что же касается моментов относительно произвольного начала C, то они в основном служат для упрощения вычислений этих характеристик.

Согласно определению начальный момент первого порядка есть средняя арифметическая

$$\mu_1 = \frac{1}{n}\sum_k x_k \cdot m_k = \bar{X}_a.$$

Центральный момент второго порядка есть дисперсия

$$\nu_2 = \frac{1}{n}\sum_k (x_k - \bar{X}_a)^2 m_k = D.$$

Центральный момент третьего порядка используется при построении показателя асимметрии распределения. Для того чтобы этот показатель асимметрии не зависел от выбранного при измерении вариант масштаба, вводят безразмерную характеристику

$$\alpha = \frac{\nu_3}{\sigma^3} = \frac{\frac{1}{n}\sum_k (x_k - \bar{X}_a)^3 m_k}{\left[\frac{1}{n}\sum_k (x_k - \bar{X}_a)^2\right]^{3/2}} \qquad (2.21)$$

— *коэффициент асимметрии* распределения. Для симметричного распределения варианты, равноудаленные от $\bar{X}_a$, имеют одинаковую частоту, и поэтому $\nu_3 = 0$, а следовательно, и $\alpha = 0$. Если $\alpha < 0$, то в вариационном ряду преобладают (имеют бóльшую частоту) варианты меньшие, чем средняя (рис. 2.2); такой ряд является *отрицательно асимметричным. Положительная асимметрия* будет характеризоваться значением $\alpha > 0$.

Еще одна особенность формы распределения измеряется центральным моментом четвертого порядка. Как и ранее, для образо-

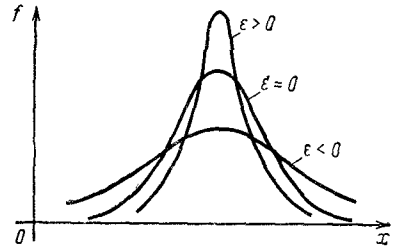


Рис. 2.3.
Эксцесс распределения

вания безразмерной характеристики определяется отношение $\frac{\nu_4}{\sigma^4}$. Оно характеризует крутизну (заостренность) графика распределения. При измерении асимметрии естественным эталоном являлось симметричное распределение, для которого $\alpha = 0$. Аналогично, при оценке крутизны в качестве эталонного выбирается так называемое *нормальное распределение,* соответствующая кривая * которого изображена на рис. 2.3 и которое будет подробно рассмотрено ниже (см. гл. 9).

Для этого распределения $\frac{\nu_4}{\sigma^4} = 3$, поэтому для оценки крутизны данного распределения в сравнении с нормальным вычисляется показатель

$$\varepsilon = \frac{\nu_4}{\sigma^4} - 3 \qquad (2.22)$$

крутизны, или *эксцесса,* распределения (см. рис. 2.3).

2.5. Свойства основных характеристик распределения. Формулы моментов

Рассмотрим линейное преобразование признака, т. е. переход от признака X к признаку

$$Y = aX + b. \qquad (2.23)$$

Теорема 1. Средние арифметические X и Y связаны тем же линейным выражением, т. е. $\bar{Y}_a = a\bar{X}_a + b$.

* Для наглядности на рис. 2.2 и 2.3 вместо полигонов (гистограмм) распределений изображены сглаженные кривые, которые хорошо их аппроксимируют при достаточно малых интервалах

Для доказательства запишем равенство (2.1) для $\bar{Y}_a$, заменив в нем значения y_k на $ax_k + b$. Получим

$$\bar{Y}_a = \frac{1}{n}\sum_k y_k m_k = \frac{1}{n}\sum_k (ax_k + b)\, m_k = \frac{a}{n}\sum_k x_k m_k + b = a\bar{X}_a + b.$$

Теорема 2. Центральные моменты любого порядка признаков X и Y связаны соотношениями $\nu_l(Y) = a^l \nu_l(X)$.

Доказательство вытекает из определения (2.20) и теоремы 1:

$$\nu_l(Y) = \frac{1}{n}\sum_k (y_k - \bar{Y}_a)^l m_k = \frac{1}{n}\sum_k [(ax_k + b - (a\bar{X}_a + b)]^l m_k =$$

$$= \frac{1}{n}\sum_k a^l (x_k - \bar{X}_a)^l m_k = a^l \nu_l(X).$$

В частности, для дисперсии получим $D(Y) = \nu_2(Y) = a^2 D(X)$.

Положив в равенстве (2.23) $a = 1$ или $b = 0$, получим следующие соотношения между моментами при изменении начала отсчета $(Y' = X + b)$ или изменении масштаба $(Y'' = aX)$:

$$\bar{Y}_a = \bar{X}_a + b; \quad \nu_l(\bar{Y}') = \nu_l(X); \quad \bar{Y}_a = a\bar{X}_a, \quad \nu_l(\bar{Y}'') = a^l \nu_l(X).$$

Положив в равенстве (2.23) $a = \frac{1}{\sigma}$ и $b = -\frac{\bar{X}_a}{\sigma}$ и обозначив в этом случае преобразованный признак, именуемый *нормированным отклонением*, через $X^0 = \frac{X - \bar{X}_a}{\sigma}$, получим из доказанных соотношений $\bar{X}_a^0 = 0$ и $D(X^0) = 1$.

В практике вычислений оказывается целесообразным переходить от признака X к преобразованному признаку $X' = \frac{X - C}{h}$, где C — начало отсчета и h — единица масштаба. Специальным подбором этих чисел можно, как правило, добиться, чтобы значения $x'_k = \frac{x_k - C}{h}$ выражались небольшими целыми числами. Для преобразованного признака X' вычисляют начальные моменты

$$\mu_l(X') = \frac{1}{n}\sum_k x'^l_k m_k,$$

а затем с их помощью рассчитывают основные центральные моменты по формулам моментов:

$$\nu_1(X) = h(\mu_1(X') + C - \bar{X}_a = 0, \quad \text{откуда} \quad \bar{X}_a = h\mu_1(X') + C;$$

$$\nu_2(X) = D(X) = h^2[\mu_2(X') - \mu_1^2(X')];$$

$$\nu_3(X) = h^3[\mu_3(X') - 3\mu_1(X')\mu_2(X') + 2\mu_1^3(X')],$$

$$\nu_4(X) = h^4[\mu_4(X') - 4\mu_3(X')\mu_1(X') + 6\mu_2(X')\mu_1^2(X') - 3\mu_1^4(X')].$$

Их применение становится особенно эффективным при расчете характеристик интервального ряда с равными интервалами. При этом в качестве C выбирают середину какого-либо интервала (обычно удобнее интервала, которому соответствует наибольшая частота), а в качестве h — ширину интервала. Тогда значения x'_k будут совпадать с порядковыми номерами интервалов, отсчитываемыми от 0-го номера интервала, середине которого соответствует значение C.

Найдем характеристики распределения, указанного в первых двух столбцах табл. 2 5 Необходимые подготовительные расчеты приведены также в таблице.

Таблица 2.5

Интервалы	m_k	x'_k	$x'_k m_k$	$(x'_k)^2 m_k$	$(x'_k)^3 \cdot m_k$	$(x'_k)^4 m_k$
5—10	12	—2	—24	48	—96	192
10—15	24	—1	—24	24	—24	24
15—20	35	0	0	0	0	0
20—25	22	1	22	22	22	22
25—30	15	2	30	60	120	240
Итого	108	0	4	154	22	478

При расчетах примем $c = 17{,}5$ и $h = 5$. По итогам столбцов получаем: $n = 108$;

$$\mu_1(X') = \frac{4}{108} = 0{,}037;\ \mu_2(X') = \frac{154}{108} = 1{,}426;\ \mu_3(X') = \frac{22}{108} = 0{,}204 \text{ и}$$

$$\mu_4(X') = \frac{478}{108} = 4{,}426.$$

Далее по формулам моментов находим:

$$\overline{X}_a = 5 \cdot 0{,}037 + 17{,}5 = 17{,}685;\ \nu_2(X) = D(X) = 5^2 (1{,}426 - 0{,}037^2) = 35{,}614;$$

$$\nu_3(X) = 5^3 (0{,}204 - 3 \cdot 0{,}037 \cdot 1\ 426 + 2 \cdot 0{,}037^3) = 5{,}727;$$

$$\nu_4(X) = 5^4 (4{,}426 - 4 \cdot 0{,}204 \cdot 0{,}037 + 6 \cdot 1{,}426 \cdot 0{,}037^2 - 3 \cdot 0{,}037^3) = 2754{,}6$$

Наконец вычисляем:

$$\sigma = \sqrt{D} = 5{,}968 \quad \alpha = \frac{\nu_3(X)}{\sigma^3} = 0{,}027 \text{ и } \varepsilon = \frac{\nu_4(X)}{\sigma^4} - 3 = -0{,}82.$$

Часто приходится изучать совокупности, образующиеся при объединении нескольких групп, или, наоборот, расчленять общую совокупность на отдельные группы. При этом возникает важный вопрос о соотношениях между характеристиками распределения объединенной совокупности и характеристиками распределений составляющих ее групп. Ограничимся рассмотрением соотношений только для первых двух характеристик: средней арифметической и дисперсии.

Пусть имеется l групп элементов ($i = 1, \ldots, l$) с численностями n_i, средними арифметическими $\overline{X}_i$ и дисперсиями D_i. Назовем

эти характеристики *групповыми средними* и *групповыми дисперсиями.* Объединим группы в общую совокупность, объем которой $n = \sum_{t=1}^{l} n_t$, общая средняя — $\bar{X}$ и общая дисперсия — D. Введем также понятие межгрупповой дисперсии $D(\bar{X}_t)$, определяемой из соотношения

$$D(\bar{X}_t) = \frac{1}{n} \sum_{t=1}^{l} (\bar{X}_t - \bar{X})^2 n_t$$

и характеризующей дисперсию групповых средних, если каждой из них приписать вес n_t. Нетрудно доказать, что общая средняя равна средней арифметической взвешенной групповых средних, а общая дисперсия — сумме средней арифметической взвешенной групповых дисперсий с весами, равными численностям групп n_t, и межгрупповой дисперсии, т. е.

$$\bar{X} = \frac{1}{n} \sum_{t=1}^{l} \bar{X}_t n_t; \quad D = \frac{1}{n} \sum_{t=1}^{l} D_t n_t + D(\bar{X}_t).$$

Раздел II

Теория вероятностей

Глава 3

Случайные события

3.1. События и вероятность

Теория вероятностей изучает закономерности случайных явлений, отражающие в абстрактной форме объективно существующие статистические закономерности массовых реальных явлений. Как и другие математические науки, теория вероятностей оперирует вероятностными моделями, в основе которых лежат формально-логические определения ряда основных понятий, таких, например, как «событие», «вероятность», «случайная величина» и т. д., и исходная система аксиом. На базе этой аксиоматики с помощью формально-логических операций («правил вывода») можно получить все многочисленные теоремы, формулы, правила и приемы анализа, составляющие содержание теории вероятностей. При этом основной задачей теории является не выяснение содержания исходных понятий, а установление связи между вероятностями различных событий или построение законов распределения случайных величин.

Однако применение теории к анализу реальных ситуаций, к решению практических задач требует перехода с абстрактного языка модели на описание изучаемого явления в содержательных терминах и понятиях. Этот переход представляет весьма сложную задачу, и для того, чтобы ее разъяснить, будем сопровождать теоретическое изложение многочисленными примерами. Кроме того, отступая от строгого формально-логического изложения, будем часто ему предпосылать описательное изложение, построенное на интуитивных представлениях.

Слово «событие» употребляется в обыденной речи наряду со словами «происшествие», «явление», «факт» и т. д. и связывается обычно с конкретным содержанием того, что при определенных условиях происходит или не происходит. При этом речь может идти как о событиях, носящих единичный, неповторяемый характер, так и о событиях, которые при определенных условиях могут

воспроизводиться многократно. Наконец, встречаются события, относительно которых суждение о появлении или непоявлении не может быть высказано однозначно.

В отличие от сказанного в теории вероятностей *событие* является формально-логическим понятием, абстрагированным от конкретного содержания. Событие характеризуется лишь тем, что при реализации определенного комплекса условий (*испытаний*) оно может произойти или не произойти, но в то же время при многократном повторении испытаний подчиняется некоторой статистической закономерности (проявляющейся в устойчивости относительной частоты его появления).

Приведем некоторые примеры:

1) из урны, содержащей 10 белых и 5 черных шаров извлекаются наудачу два шара (испытание); в качестве событий можно рассматривать: A_1 — извлечение двух белых шаров, A_2 — извлечение двух черных шаров и A_3 — извлечение двух шаров разного цвета; 2) извлечение номера из тиражной урны — испытание, выигрыш на данный билет — событие; 3) имеется генеральная совокупность с заданным распределением качественного признака A. Извлечение одного элемента в выборку — испытание, а наличие у этого элемента признака A — событие

Характерной особенностью всех перечисленных событий является то, что появление каждого из них при одном испытании непредсказуемо (может появиться или не появиться). Поэтому говорят, что наступление события зависит от случайных обстоятельств.

Особую разновидность представляют: *достоверное событие* (U), которое наступает при каждом испытании, и *невозможное событие* (V), которое не может появиться ни при одном испытании в данной серии испытаний. Так, извлечение красного шара из урны с белыми и черными шарами есть невозможное событие, а выпадание выигрыша в беспроигрышной лотерее — событие достоверное.

Пусть проводится несколько серий испытаний. Обозначим через n_i число испытаний в i-й серии и через m_i — частоту появления события A в этой серии. Тогда утверждение об устойчивости относительной частоты означает, что при достаточно больших n_i величины $\frac{m_i}{n_i}$ будут близки между собой и незначительно колебаться вокруг некоторого постоянного для данного события числа, называемого *вероятностью* этого события (обозначается $P(A)$). Указанное определение вероятности получило название *частотного*, или *статистического*.

Как видно, понятие вероятности тесно связано со свойством *устойчивости относительной частоты*, или, другими словами, с условием *статистической однородности* совокупности испытаний. Хотя многолетний опыт и подтверждает наличие устойчивости для многих событий, прямого способа ее проверки нет. А без этого введенное понятие вероятности оказывается необоснованным. В прикладных исследованиях вопрос о наличии статистической

устойчивости решается на основании интуитивных соображений или постулируется как обобщение прошлого опыта.

В приведенном определении можно усмотреть ряд нерешенных вопросов: какое количество испытаний следует считать достаточно большим, как должны выбираться серии испытаний, какие колебания относительной частоты считать незначительными и как по ряду их значений однозначно определить постоянное число $P(A)$? Следует заметить, что указанная неопределенность касается логических оснований теории вероятностей и не мешает практическому определению вероятности путем постановки соответствующего эксперимента. Например, регистрируя последовательные относительные частоты появления герба при бросании монеты достаточно большое число раз, можно установить, что они будут незначительно колебаться вокруг числа 0,5, которое и принимается за вероятность появления герба. С аналогичным положением приходится сталкиваться при проведении измерений длины, массы и т. д. Результат измерения можно получить достаточно точно при большом числе измерений, не задаваясь вопросом о логическом определении понятий длины, массы и т. д. Более того, исходя из данного выше определения, можно установить основные свойства вероятности:

1) $0 \leqslant P(A) \leqslant 1$;
2) вероятность невозможного события равна нулю;
3) вероятность достоверного события равна единице.

Продолжая далее интуитивно-логическое построение теории, можно рассмотреть некоторые операции над событиями, позволяющие в дальнейшем сводить изучение одних событий к изучению других, в некотором смысле более простых.

Если при каждом испытании, при котором происходит событие A, происходит и событие B, то будем говорить, что B *включает* A или что A *входит* в B, обозначая это в виде $A \subset B$.

Например, бросается игральная кость (испытание). Обозначим через A_6 появление шести очков и через $A_ч$ — появление четного числа очков. Очевидно, что $A_6 \subset A_ч$.

Если всякий раз, когда происходит A, происходит и B и, наоборот, когда происходит B, происходит и A, т. е. одновременно $A \subset B$ и $B \subset A$, то события A и B называются *равными* (точнее, *эквивалентными*), что обозначается в виде $A = B$.

Суммой двух событий A и B, обозначаемой $A + B$, является событие, наступающее всякий раз, когда наступает хотя бы одно из событий A или B. Это определение может быть распространено на любое число слагаемых.

Так, если в предыдущем примере через A_2, A_4 и A_6 обозначены события, состоящие в появлении соответственно 2, 4 или 6 очков, то событие $A_2 + A_4 + A_6$ означает появление четного числа очков.

Произведением событий A и B, обозначаемым в виде $A \cdot B$, является событие, наступающее всякий раз, когда наступает и событие A, и событие B. Это определение может быть распространено на любое число сомножителей.

Пусть в произведенной продукции имеются изделия I, II сортов и бракованные. Наудачу выбирается изделие (испытание). Обозначим через A_1, A_2 и $A_б$ события, состоящие в том, что выбранное изделие окажется соответственно I сорта, II сорта или бракованное. Тогда $A_1 + A_2$ означает событие, состоящее в том, что выбранное изделие не бракованное.

Рассмотрим последовательное извлечение двух изделий. Будем соответственно одним и двумя штрихами обозначать результат первого и второго извлечений. Тогда событие $A_1' \cdot A_1''$ означает, что оба раза выбраны изделия I сорта, событие $A_1' \cdot A_б''$ — что первое изделие оказалось I сорта, а второе — бракованное и т. д.

Событие $\bar{A}$, наступающее всякий раз, когда не появляется событие A, называется *противоположным* A.

Так, в примере с игральной костью событие $\bar{A}_6$ означает выпадание любого числа очков, кроме шестерки. В предыдущем примере событие $\bar{A}_б$ означает извлечение годного изделия (т. е. событие $A_1 + A_2$), поэтому можно записать $\bar{A}_б = A_1 + A_2$.

Мы не случайно при рассмотрении операций над событиями использовали алгебраическую терминологию и обозначения. Обусловлено это тем, что указанные операции обладают многими свойствами действий над числами, позволяющими говорить о так называемой *алгебре событий*. Однако, прежде чем обратиться к ее подробному изучению, необходимо преодолеть отмеченную выше логическую незавершенность исходных понятий «событие» и «вероятность».

3.2. Алгебра событий

Математическим дисциплинам свойственно дедуктивное построение, при котором в основу кладется некоторая исходная система определений и аксиом, в абстрактной форме отражающих реальные понятия и отношения, из которых, следуя правилам формальной логики, выводятся остальные теоретические положения. Образцом такого дедуктивного построения может служить геометрия. Впервые аксиоматическое построение теории вероятностей, основанное на теоретико-множественном определении события и его вероятности, было предложено А. Н. Колмогоровым и в настоящее время получило всеобщее признание*. Не имея возможности в рамках настоящей книги следовать строгому аксиоматическому изложению, опирающемуся на ряд сложных теоретико-множественных понятий, ограничимся рассмотрением частного случая.

Пусть имеется некоторое множество $\Omega = \{e_1, \ldots, e_i, \ldots\}$ взаимно исключающих исходов испытания e_i, которые будем называть *элементарными событиями* или *исходами*. Множество Ω, которое может быть конечным или бесконечным, будем называть *пространством элементарных событий* (исходов). Элементы множества Ω являются первичными понятиями и не подлежат формальному определению. Однако, для того чтобы аксиоматика приводила к со-

* Ранее идея аксиоматического построения теории вероятностей, использующая понятие равновозможности событий, была высказана С. Н. Бернштейном.

держательным теоретическим построениям, формально-логические определения и аксиомы должны отражать реальные понятия и отношения между ними. Поэтому проиллюстрируем их на некоторых примерах.

1. При бросании игральной кости (испытание) Ω состоит из шести элементарных событий $\Omega = \{e_1, e_2, e_3, e_4, e_5, e_6\}$, где e_i означает выпадание i очков.

2. В урне находятся два белых и три красных шара. Извлекается наудачу один шар. В данном случае можно рассматривать Ω как состоящее из двух элементарных событий: $e_б$ — извлечение белого шара, $e_к$ — извлечение красного шара. Можно, однако, мысленно перенумеровав все пять шаров, рассматривать $\Omega_1 = \{e_1, e_2, e_3, e_4, e_5\}$, где e_i — извлечение i-го шара.

3. Производится многократное бросание игральной кости до появления шести очков. Обозначим через e_1 появление шести очков при первом бросании, через e_2 — появление шести очков при втором бросании, через e_i — появление шести очков при i-м бросании и т. д. Тогда $\Omega = \{e_1, e_2, \ldots, e_i, \ldots\}$. Как видим, в этом случае Ω есть бесконечное множество, элементы которого могут быть однако перенумерованы. Такое бесконечное множество называется *счетным*.

Рассмотрим конечное, или счетное, пространство элементарных событий Ω, которое будем называть *дискретным* (см. п. 1—3). Событием A будем называть любое подмножество этого пространства ($A \subseteq \Omega$), а элементы $e_i \in A$, его составляющие,— *входящими* (благоприятными) в событие A. Так, в п. 1 событие «четное число очков» есть подмножество $\{e_2, e_4, e_6\}$, в п. 2 событие «белый шар» есть подмножество $\{e_1, e_2\} \subset \Omega_1$, в п. 3 событие «появление шести очков не более чем за четыре бросания» есть подмножество $\{e_1, e_2, e_3, e_4\}$ и т. д.

Достоверному событию (U) соответствует все множество Ω и невозможному (V) — пустое множество $\varnothing$.

Событие A есть *следствие* события B («частный случай B», «входит в B»), что обозначается в виде $A \subset B$, если множество A есть подмножество множества B. Если одновременно $A \subseteq B$ и $B \subseteq A$, то события A и B называются *равными* (равносильными), что обозначается в виде $A = B$.

Суммой событий $A + B$ называется объединение множеств $A \cup B$.

Произведением событий AB (или $A \times B$) называется пересечение множеств $A \cap B$.

Событием $\bar{A}$, *противоположным* A (отрицание A), является множество элементов из Ω, не входящих в A (дополнение A), т. е. $\Omega - A$.

Эти основные операции над событиями наглядно иллюстрируются диаграммами Венна (рис. 3.1), на которых множество Ω есть множество точек квадрата, а события A и B — множества точек круга и прямоугольника. Результаты операций над событиями показаны заштрихованной фигурой.

Хотя диаграммы Венна предполагают несчетное множество Ω, однако они очень удобны для наглядной иллюстрации и обоснования свойств операций над событиями, поэтому мы их используем и в случае дискретного пространства Ω.

Приведем перечень основных свойств операций над событиями (или, что то же, над множествами), часть которых следует непо-

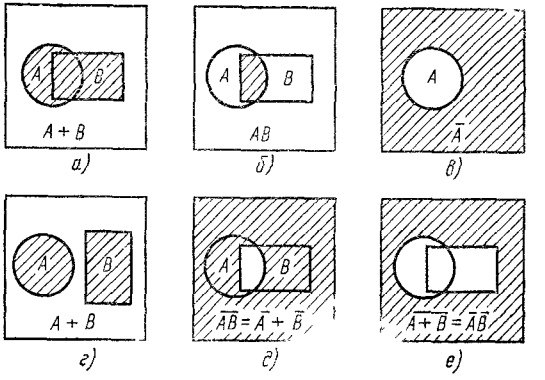


Рис 3.1. Диаграммы Венна. Операции над событиями

средственно из определений, а часть — легко обосновывается с помощью диаграмм Венна.

1) $A + B = B + A$; 2) $A + V = A$; 3) $A + U = U$; 4) $A + A = A$;
5) $A + B = A$, если $B \subset A$; 6) $AB = BA$; 7) $AV = V$; 8) $AU = A$;
9) $AA = A$; 10) $AB = B$, если $B \subset A$; 11) $A(B + C) = AB + AC$;
12) $AB + C = (A + C)(A + B)$.

Как видим, ряд свойств аналогичны соответствующим свойствам алгебры чисел, при этом невозможное V и достоверное U события играют как бы роль нуля и единицы, но некоторые свойства (3, 4, 5, 9, 10 и 12) не имеют аналога в обычной алгебре. Все это позволяет говорить о своеобразной *алгебре событий*. Еще более разительное отличие от обычной алгебры связано с новой операцией, не имеющей аналога в действиях над числами — образование противоположного события:

13) $\bar{V} = U$; 14) $\bar{U} = V$; 15) $\bar{\bar{A}} = A$; 16) $A + \bar{A} = U$;
17) $A\bar{A} = V$; 18) $\overline{A + B} = \bar{A}\bar{B}$; 19) $\overline{AB} = \bar{A} + \bar{B}$.

Свойства 13—17 вытекают непосредственно из определения противоположного события. Обоснование последних двух свойств можно получить из диаграмм Венна.

События A и B являются *несовместными*, если $AB = V$ (т. е. пересечение соответствующих им подмножеств есть пустое множество). В противном случае события *совместны* (см. рис 3.1). Различают попарную несовместность и несовместность в совокуп-

ности. Так, на рис. 3.2 показаны события *попарно совместные*, но *несовместные в совокупности*

В следующем примере дана содержательная интерпретация некоторых операций над событиями.

В группе из шести перенумерованных изделий первое и второе I сорта; третье, четвертое и пятое — II сорта и шестое — бракованное. При этом изделия второе, четвертое и шестое маркированы. Испытание состоит в извлечении наудачу одного изделия. Обозначим через e_i (элементарное событие) появление изделия i-го номера и через A_1, A_2, $Б$ и M — появление соответственно изделия I сорта, II сорта, бракованного и маркированного. В качестве пространства элементарных событий можно рассмотреть множество $\Omega = \{e_1, e_2, e_3, e_4, e_5, e_6\}$. События $A_1 = \{e_1, e_2\}$, $A_2 = \{e_3, e_4, e_5\}$, $Б = \{e_6\}$ и $M = \{e_2, e_4, e_5, e_6\}$. Очевидно, что $A_1A_2 = V$, $A_1Б = V$ и $A_2Б = V$, т. е. эти пары событий несовместные. В отличие от этого событие M совместно с любым из событий A_1, A_2 или $Б$. Далее очевидно, что $Б \subset M$ и $A \subset M$. Рассмотрим событие $A_1 + A_2$, означающее, что выбранное изделие годное (I и II сорта). Это событие можно записать также как $\overline{Б}$, т. е. $A_1 + A_2 = \overline{Б}$. Произведением событий A_2 M будет событие $\{e_4, e_5\}$, означающее, что извлеченное изделие II сорта и маркированное.

Суммой событии $A_2 + M$ будет событие $\{e_2, e_3, e_4, e_5, e_6\}$, которое означает, что изделие оказалось или II сорта, или маркированное. Противоположным ему $\overline{(A_2 + M)}$ будет событие, которое можно охарактеризовать, как не II сорта и не маркированное, т. е. $\overline{A_2 + M} = \bar{A}_2\bar{M}$.

Событие $(A_1 + A_2)M = \{e_3, e_4, e_5\}$ означает, что изделие оказалось I или II сорта и маркированное. Его можно записать на основании 11-го свойства как сумму $A_1M + A_2M$. Сумма $A_1 + A_2 + Б = U$ есть достоверное событие.

В заключение отметим, что мы рассматривали только дискретное пространство элементарных событий Ω, в котором любое подмножество есть событие. В этом случае результат операции сложения, умножения и образования противоположного события есть подмножества из Ω, т. е. представляют собой событие.

В общем случае, если Ω есть несчетное множество, указанные операции над любыми его подмножествами не всегда выполнимы

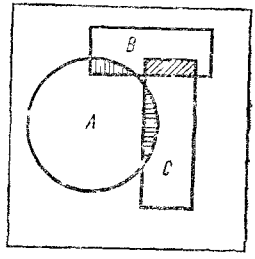


Рис. 3.2.
События *попарно совместные, но несовместные в совокупности*

(т. е. не всегда образуется подмножество Ω). Поэтому для произвольного множества элементарных событий Ω в качестве событий выделяют специальный класс *подмножеств, входящих в* Ω, в котором указанные операции всегда выполнимы, т. е. приводят вновь к подмножествам этого класса. Такой класс подмножеств называют *полем событий*. В силу ряда математических осложнений при изучении произвольного множества Ω ограничимся в дальнейшем рассмотрением лишь дискретного пространства элементарных событий.

3.3. Вероятностное пространство

Обратимся теперь к введению понятия вероятности. Так как в основе определения любого события лежит пространство элементарных событий Ω (которое мы будем предполагать дискретным), то понятие вероятности следует прежде всего определить для элементарных событий. При этом вероятность, являясь абстрактным отображением относительной частоты события, должна сохранять ее свойства, а именно, во-первых, выражаться неотрицательным числом и, во-вторых, если событие A есть сумма s несовместных событий $A_1, A_2, \ldots, A_i, \ldots, A_s$, появляющихся при n испытаниях с частотами $m_1, \ldots, m_i, \ldots, m_s$, то относительная частота появления события A равна сумме относительных частот этих слагаемых. Эти соображения приводят к следующим определениям.

Определение 1. Каждому элементарному событию e_i соответствует неотрицательное число $P(e_i)$, называемое *вероятностью элементарного события.*

Определение 2. *Вероятностью события* $A = \{e_1, \ldots, e_i, \ldots\}$ называется число $P(A)$, равное сумме вероятностей входящих в это событие элементарных событий, т. е.

$$P(A) = \sum_{e_i \in A} P(e_i). \tag{3.1}$$

Дискретное пространство элементарных событий с определенной на нем вероятностной мерой называется *дискретным вероятностным пространством.*

Из определений 1 и 2 можно получить ряд свойств вероятности:

$$1)\ P(V) = 0;\quad 2)\ P(U) = \sum_{e_i \in A} P(e_i) = 1;\quad 3)\ 0 \leqslant P(A) \leqslant 1;$$

$$4)\ \text{если } AB = V, \text{ то } P(A + B) = P(A) + P(B).$$

Докажем, например, последнее свойство, известное под названием *теоремы сложения для несовместных событий.* Пусть события A и B определяются подмножествами элементарных событий $\{e_i\}$ и $\{e'_i\}$, т. е. $e_i \in A$ и $e'_i \in B$. Тогда событие $A + B$ есть объединение этих подмножеств $\{e_i, e'_i\}$. Согласно определению 2 получаем

$$P(A + B) = \sum_{\delta_k \in (A+B)} P(\delta_k) = \sum_{e_i \in A} P(e_i) + \\ + \sum_{e'_i \in B} P(e'_i) = P(A) + P(B), \tag{3.2}$$

где через δ_k обозначены элементы объединенного множества $A + B$. Используя метод доказательств по индукции, можно теорему сложения обобщить на любое число слагаемых событий.

Теорема 1 (*сложения несовместных событий*). Вероятность суммы несовместных событий равна сумме вероятностей этих со-

бытий:

$$P\left(\sum_i A_i\right) = \sum_i P(A_i). \qquad (3.2')$$

Пусть события $A_1, \ldots, A_i, \ldots, A_s$ попарно несовместны и сумма их есть достоверное событие (в последнем случае говорят, что они образуют *полную группу событий*). Тогда, используя теорему и определение 2, получим

$$P\left(\sum_{i=1}^{s} A_i\right) = \sum_{i=1}^{s} P(A_i) = P(U) = 1. \qquad (3.3)$$

В частности, для пары событий A и $\bar{A}$ из последней формулы получаем еще одно важное свойство вероятности:

$$P(A) + P(\bar{A}) = 1. \qquad (3.4)$$

Если события A и B совместны, то приведенное выше доказательство теоремы сложения осложняется тем, что часть элементарных событий (обозначим их через e_j'') входит и в событие A, и в событие B, т. е. принадлежит их пересечению $e_j'' \in AB$. Для того чтобы исключить их двойной счет, необходимо в разложении (3.2) общей суммы по $\delta_k \in (A+B)$ на суммы по $e_i \in A$ и $e_i' \in B$ исключить слагаемые, соответствующие $e_j'' \in AB$. В результате придем к выражению

$$P(A+B) = \sum_{\delta_k \in (A+B)} P(\delta_k) = \sum_{e_i \in A} P(e_i) + \sum_{e_i' \in B} P(e_i') - \sum_{e_j'' \in AB} P(e_j'').$$

Заменив согласно определению 2 полученные суммы вероятностями соответствующих событий, придем к следующей теореме.

Теорема 2 (*сложения двух совместных событий*). Вероятность суммы двух совместных событий равна сумме вероятностей этих событий минус вероятность их совместного наступления, т. е.

$$P(A+B) = P(A) + P(B) - P(AB). \qquad (3.5)$$

Два стрелка производят по одному выстрелу в мишень, при этом заданы вероятности попадания первого стрелка ($P(A_1) = 0{,}8$), второго стрелка ($P(A_2) = 0{,}5$) и обоих стрелков ($P(A_1A_2) = 0{,}4$). Найти вероятность поражения мишени.

Здесь A_1 и A_2 — совместные события. Поэтому, используя (3.5), получим

$$P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1A_2) = 0{,}9.$$

Обобщение теоремы 2 на случай более чем двух слагаемых приводит к громоздкому выражению, поэтому используется окольный путь для вычисления вероятности суммы совместных событий.

Теорема 3 (*сложения для произвольных событий*). Вероятность суммы произвольных событий равна разности между единицей и вероятностью произведения противоположных им событий.

Пусть $A_1, A_2, \ldots, A_i, \ldots, A_n$ являются совместными событиями. Используя свойство 18 (см. с. 40), запишем соотношение

$$\overline{A_1 + \ldots + A_i + \ldots + A_n} = \bar{A}_1\bar{A}_2 \ldots \bar{A}_i \ldots \bar{A}_n.$$

Переходя к вероятностям и используя (3.4), получим окончательно

$$P(A_1 + \ldots + A_i + \ldots + A_n) = 1 - P(\bar{A}_1 \bar{A}_2 \ldots \bar{A}_i \ldots \bar{A}_n). \quad (3.6)$$

При переходе к произвольному полю событий (если Ω есть несчетное множество) вероятность любого события A из этого поля определяется как неотрицательное число $P(A) \geqslant 0$, удовлетворяющее двум аксиомам: 1) $P(U) = 1$ и 2) $P\left(\sum_i A_i\right) = \sum_i P(A_i)$, если события A_i попарно несовместны.

Поле событий, на котором введена таким образом вероятностная мера, называется *произвольным вероятностным пространством.*

Как видим, свойства 2 и 4, доказываемые для дискретного вероятностного пространства, здесь выступают как аксиомы. Что касается остальных свойств и теорем, то справедливость их доказывается, как и в случае дискретного пространства. Поэтому мы в дальнейшем часто будем говорить о вероятностном пространстве, не уточняя, является ли оно дискретным или произвольным.

3.4. Классическое определение вероятности

В предыдущем параграфе при установлении вероятностной меры совершенно не затрагивался вопрос о том, чему конкретно численно равна вероятность $P(e_i)$ элементарного события в дискретном или $P(A)$ любого события в произвольном вероятностном пространстве. Да это и не требовалось при аксиоматическом построении теории, подобно тому как в геометрии достаточно постулировать существование длины отрезка или площади фигуры, не касаясь вопросов их измерения. Однако, как только мы переходим к вычислению вероятностей, нам необходимы, кроме ряда теорем, устанавливающих соотношения между вероятностями различных событий, еще какие-то численные значения вероятностей некоторых простейших событий. В общем случае для этой цели может быть привлечено статистическое определение вероятности, позволяющее опытным путем найти вероятность через относительную частоту при многократном повторении испытаний.

Однако имеется случай, когда, опираясь лишь на аксиоматическое определение вероятностной меры, можно однозначно определить ее численное значение. Это имеет место, если испытания обладают особым свойством симметрии, выражающимся в том, что возможные исходы можно полагать равновероятными. Говорят при этом, что испытания сводятся к схеме «случаев» или «шансов».

Рассмотрим вначале конечное множество $\Omega = \{e_i\}$ $(i = 1, \ldots, n)$ возможных исходов, которые будем предполагать равновероятными Достоверное событие U есть множество Ω и потому $P(U) = \sum_{e_i \in \Omega} P(e_i) = 1$. Поскольку все слагаемые $P(e_i)$ равны между собой

и число их равно n, то получим из этого равенства $P(e_i) = \frac{1}{n}$. Таким образом, мы однозначно определили величину вероятности элементарного события.

Пусть теперь событию A благоприятствуют m исходов из общего числа n взаимоисключающих равновероятных исходов. Тогда из (3.1) получаем

$$P(A) = \sum_{e_i \in A} P(e_i) = m\frac{1}{n} = \frac{m}{n}. \qquad (3.7)$$

Определение вероятности по данной формуле как отношения числа m благоприятных исходов к общему числу n всевозможных взаимоисключающих и равновероятных исходов получило название *классического определения вероятности.*

При возникновении теории вероятности оно служило определением понятия вероятности (откуда и термин «классическое»). Действительно, для задач, которые в тот период рассматривались (бросание монеты или игральной кости, извлечение шара из урны и др.), исходы испытаний обладали свойством симметрии и результат испытаний мог быть описан конечным числом равновероятных исходов (при бросании монеты — герб или число, при бросании кости — выпадание любой цифры и т. д.).

В современном изложении формула (3.7) должна рассматриваться не как определение понятия вероятности, а как частный прием ее вычисления для событий, сводящихся к схеме случаев. Этот прием получил название *способа непосредственного подсчета* вероятности.

Несмотря на ограниченную возможность использования формулы (3.7) на практике, рассмотрение событий, сводящихся к схеме случаев, позволяет яснее представить себе ряд фундаментальных понятий теории вероятностей, наглядно иллюстрировать многие ее абстрактные построения. К таким иллюстрациям мы будем неоднократно прибегать в дальнейшем.

Применение формулы (3.7) сопряжено с определенными затруднениями, связанными с необходимостью описания пространства равновероятных исходов и подсчета чисел n и m. Некоторую помощь в преодолении этого затруднения может оказать использование формул комбинаторики. Проиллюстрируем способ непосредственного подсчета на примерах.

1. Определим вероятность выигрыша при выпадании герба, если монета бросается не более двух раз

Обозначим событие, состоящее в выпадании герба через «г» и выпадании числа — через «ч». Предполагая, что с появлением герба отпадает надобность во втором бросании, можно было бы рассмотреть в качестве элементарных исходов e_1 = «г», e_2 = «ч, г» и e_3 = «ч, ч». В таком случае имели бы $n = 3$, $m = 2$ и $P = \frac{2}{3}$.

Однако приведенные рассуждения неверны с точки зрения условий применения формулы (3.7), так как исходы e_1, e_2 и e_3 неравновероятны. Следует рассмотреть всевозможные исходы двукратного бросания: e_1 = «г, г», e_2 = «г, ч»,

e_3 = «ч, г» и e_4 = «ч, ч». Все они равновероятны. Выигрыш наступит при первых трех исходах, т. е. $m = 3$, откуда $P = \frac{3}{4}$.

2. Имеется 20 деталей, среди которых пять бракованных. Какова вероятность того, что в случайной выборке из восьми деталей окажутся две бракованные

Будем считать все 20 деталей различными и выборки различать не только составом отобранных деталей, но и порядком их расположения *, т. е. под элементарным событием (исходом) понимать выбор любых восьми деталей из 20 Возможны два принципа отбора, которые приведут к двум различным результатам

1. *Отбор с возвращением* (повторная выборка). Каждая последовательно отобранная деталь вновь возвращается и может быть отобрана вторично. Таким образом, каждая деталь может повторяться в выборке от 1 до 8 раз. Всего число возможных выборок (элементарных событий) будет равно числу размещений из 20 по 8 с повторениями, т. е. $n = \tilde{A}^8_{20} = 20^8$. Благоприятные выборки, т. е. такие, в которых окажутся две бракованные детали, образуются, если шесть деталей отбираются из 15 годных (это можно сделать $\tilde{A}^6_{15} = 15^6$ способами) и две отбираются из пяти бракованных ($A^2_5 = 5^2$ способами). Следовательно, всего число благоприятных исходов будет $m = 15^6 \cdot 5^2$. Отсюда искомая вероятность

$$P = \frac{15^6 \cdot 5^2}{20^8} = 0{,}0111.$$

2. *Отбор без возвращения* (бесповторная выборка) Отобранные детали не возвращаются и потому повторно в выборке оказаться не могут. Следовательно, в отличие от предыдущего подсчет чисел n и m должен производиться с помощью числа размещений без повторения. Итак, получим $n = A^8_{20} = 20 \cdot 19 \ldots 13$ и $m = A^5_{16} \cdot A^2_5 = 15 \cdot 14 \cdot 13 \cdot 12 \cdot 11 \cdot 10 \cdot 5 \cdot 4$, откуда

$$P = \frac{A^6_{15} \cdot A^2_5}{A^8_{20}} = 0{,}0142.$$

Заметим, что можно было бы пространство элементарных событий строить и по-другому, различая выборки (возможные исходы) только по составу входящих в них деталей, а не по порядку, т. е. как сочетание восьми деталей. Важно лишь, чтобы при подсчете числа всевозможных и благоприятных исходов пользоваться одним и тем же определением элементарного события. Такой способ решения для отбора без возвращения дал бы $n = C^8_{20}$, $m = C^6_{15} \cdot C^2_5$, откуда $P = \frac{C^6_{15} \cdot C^2_5}{C^8_{20}} = 0{,}0142$, т. е. результат, совпадающий с предыдущим.

Классическое определение вероятностей может быть обобщено на случай бесконечного счетного или несчетного пространства равновероятных исходов. Однако в этом случае числа n и m в формуле (3.7) должны быть заменены на соответствующие *меры множеств* всевозможных и благоприятных исходов.

3.5. Условная вероятность. Зависимые и независимые события. Теорема умножения вероятностей

Пусть события A и B определены в одном и том же поле событий. Обратимся к частотной интерпретации вероятности. Если

* Образование такой выборки можно себе представить как последовательный отбор восьми деталей по одной.

через m_A, $m_{\bar{A}}$, m_B, $m_{\bar{B}}$, m_{AB} и т. д. обозначены частоты появления соответствующих событий A, $\bar{A}$, B, $\bar{B}$, AB, указанных в индексе, при n испытаниях, то относительные частоты $\frac{m_A}{n}$, $\frac{m_{\bar{A}}}{n}$, $\frac{m_B}{n}$, $\frac{m_{\bar{B}}}{n}$, $\frac{m_{AB}}{n}$, ... являются прообразом соответствующих вероятностей $P(A)$, $P(\bar{A})$, $P(B)$, $P(\bar{B})$, $P(AB)$... Тогда отношение $\frac{m_{AB}}{m_B}$ может рассматриваться как условная относительная частота события A при наступлении вместе с ним события B. Разделив числитель и знаменатель последней дроби на n, представим ее в виде

$$\frac{m_{AB}}{m_B} = \frac{m_{AB}}{n} : \frac{m_B}{n}.$$

Абстрактной моделью этого соотношения между относительными частотами будет аналогичное соотношение между вероятностями событий:

$$P_B(A) = \frac{P(AB)}{P(B)}, \tag{3.8}$$

которое примем в качестве определения *условной вероятности* $P_B(A)$ *события* A *при наступлении события* B.

Аналогично может быть определена *условная вероятность* $P_A(B)$ *события* B *при наступлении события* A:

$$P_A(B) = \frac{P(AB)}{P(A)}. \tag{3.9}$$

Естественно, что в этих определениях предполагается, что $P(A)$ и $P(B)$, которые в этом случае называются *безусловными* * *вероятностями*, отличны от нуля.

Если относительная частота $\frac{m_A}{n}$ события A будет равна относительной частоте появления A и B в общем числе испытаний, при которых появилось B, т. е. $\frac{m_A}{n} = \frac{m_{AB}}{m_B}$, то событие A не зависит от события B; если же $\frac{m_A}{n} \neq \frac{m_{AB}}{m_B}$, то событие A является зависимым от B. Переход к вероятностям приводит к следующему важному определению: событие A является *независимым* от B, если $P(A) = P_B(A)$, и *зависимым*, если $P(A) \neq P_B(A)$. При этом возможны два соотношения: $P_B(A) > P(A)$ и $P_B(A) < P(A)$. Нетрудно убедиться в том, что если A зависимо (независимо) от B, то и B зависимо (независимо) от A. Можно также показать, что свойства зависимости или независимости сохраняются при

---

* Термин «безусловная» вероятность употребляется только при сопоставлении с «условной». Обычно этот термин опускается, так как любая вероятность определяется при определенных условиях и потому является «условной».

переходе к противоположным событиям. Для иллюстрации изложенных важных понятий рассмотрим пример.

Производится двукратное последовательное извлечение одного шара из урны, содержащей 10 белых и 5 черных шаров. Обозначим через A извлечение белого шара при первом извлечении и через B — при втором извлечении. Возможны две схемы извлечения: «с возвращением» и «без возвращения». В обеих схемах вероятность события A одна и та же и равна $P(A) = \frac{10}{15} = \frac{2}{3}$. Но вероятность события B зависит от принятой схемы.

1. *Извлечение с возвращением.* В этом случае $P(B) = \frac{10}{15} = \frac{2}{3} = P(A)$ вне зависимости от результата первого извлечения, т. е. от того, наступило или нет событие A. Поэтому $P_A(B) = P_{\overline{A}}(B) = P(B) = \frac{2}{3}$, т. е. события B и A независимы.

2. *Извлечение без возвращения.* Перед вторым извлечением в урне остается только 14 шаров. Если первый шар оказался белым, то $P_A(B) = \frac{9}{14} > P(B)$, откуда следует зависимость событий B и A. Заметим, что если бы первый шар оказался черным, то получили бы аналогично $P_{\overline{A}}(B) = \frac{10}{14} < P(B)$.

Из (3.8) и (3.9) после умножения на знаменатели можно получить соотношение

$$P(AB) = P(B)\,P_B(A) = P(A)\,P_A(B), \qquad (3.10)$$

составляющее содержание следующей *теоремы умножения вероятностей.*

Теорема 1. Вероятность произведения двух событий ($и$ A, $и$ B) равна произведению вероятности одного из этих событий на условную вероятность другого.

Хотя равенство (3.10) является тривиальным следствием определения условной вероятности, его практическое значение состоит в том, что оно позволяет определить вероятность сложного события AB по известным вероятностям событий сомножителей, минуя построение для события AB пространства Ω.

Пусть, например, необходимо определить вероятность появления двух белых шаров при применении схемы извлечения «без возвращения».

Применяя формулу (3.10) и ранее найденные вероятности $P(A) = \frac{2}{3}$ и $P_A(B) = \frac{9}{14}$, получим $P(AB) = \frac{2}{3} \cdot \frac{9}{14} = \frac{3}{7}$.

Решение этой же задачи с помощью построения пространства исходов Ω выглядело бы следующим образом. Всего возможных и равновероятных исходов двукратного извлечения будет $n = A_{15}^2 = 15 \cdot 14$. Среди них благоприятных исходов, при которых появляются два белых шара, будет $m = A_{10}^2 = 10 \cdot 9$. Отсюда искомая вероятность $P(AB) = \frac{10 \cdot 9}{15 \cdot 14} = \frac{3}{7}$.

Определение вероятности произведения произвольного числа событий приводит к следующему обобщению теоремы умножения.

Теорема 2.

$$P(A_1A_2A_3 \ldots A_n) = P(A_1)\,P_{A_1}(A_2) \ldots P_{A_1A_2 \ldots A_{n-1}}(A_n). \qquad (3.11)$$

Это равенство можно получить из (3.10), рассматривая последовательно очевидные равенства. $A_1A_2A_3\ldots A_n = (A_1A_2A_3\ldots A_{n-1})A_n$, $A_1A_2A_3\ldots A_{n-1} = (A_1A_2\ldots A_{n-2})A_{n-1}$ и т. д., применяя к ним формулу (3.10).

В случае, когда события-сомножители оказываются независимыми, все условные вероятности в равенстве (3.11) могут быть заменены на безусловные, в результате чего приходим к следующей теореме умножения для независимых событий.

Теорема 3. Вероятность произведения нескольких независимых событий равна произведению вероятностей этих событий, т. е.

$$P(A_1A_2\ldots A_n) = P(A_1)\cdot P(A_2)\ldots P(A_n). \qquad (3.12)$$

Общий прием определения вероятностей сложных событий заключается в представлении их в виде суммы или произведения более простых событий и последующего применения теорем сложения или умножения вероятностей. Обычно затруднение вызывает первый этап решения — построение так называемых структурных формул, описывающих соотношения между событиями. На втором этапе необходимо следить за правильностью применения теорем, связанных с учетом совместности или несовместности слагаемых и зависимости или независимости сомножителей.

В заключение отметим, что при рассмотрении нескольких (более двух) событий следует различать *попарную независимость* и *независимость в совокупности*, предполагающую независимость каждого из событий от любой совокупности остальных (в том числе попарную).

3.6. Формула полной вероятности. Теорема гипотез

Пусть событие A может появиться вместе с одним из следующих попарно несовместных событий H_1, H_2, ..., H_i, ..., H_s, образующих полную группу событий *. Будем называть события H_i *гипотезами* для события A.

На рис. 3.3 большой прямоугольник изображает пространство исходов Ω, маленькие прямоугольники соответствуют гипотезам H_i, а заштрихованная фигура — событию A. Как видно из рисунка, событие A может быть представлено как сумма попарно несовместных событий $A = AH_1 + AH_2 + \ldots + AH_i + \ldots + AH_s$. Применяя теорему сложения, получим

$$P(A) = P(AH_1) + P(AH_2) + \ldots + P(AH_i) + \ldots + P(AH_s).$$

Используя теорему умножения для зависимых событий (зависимость A от H_i на рис. 3.3 проявляется в том, что площади фигур, соответствующих появлению A при различных гипотезах H_i, различны), получим окончательно соотношение, или *формулу*

* Группа событий называется полной, если их сумма есть достоверное событие.

полной вероятности:

$$P(A) = P(H_1) P_{H_1}(A) + \ldots + P(H_i) P_{H_i}(A) + \ldots + P(H_s) P_{H_s}(A). \quad (3.13)$$

Эта формула, как бы объединяющая в одно целое теоремы сложения и умножения вероятностей, позволяет определить вероятность события A (полную вероятность), если известны вероятности гипотез $P(H_i)$ и условные вероятности события A при наступлении каждой гипотезы.

Предположим, что из партии деталей 20 изготовлено на первом станке, 25 — на втором станке и 5 — на третьем станке. Известно, что вероятности выпуска бракованной детали на первом, втором и третьем станках соответственно равны 0,02, 0,01 и 0,05. Какова вероятность, что взятая наудачу деталь окажется бракованной?

Под событием A будем понимать появление бракованной детали, а под гипотезами H_1, H_2, H_3 — события, состоящие в том, что взятая наудачу деталь была изготовлена соответственно на первом, втором и третьем станке. Из условий задачи имеем:

$$P(H_1) = \frac{20}{50} = \frac{2}{5}, \; P(H_2) = \frac{25}{50} = \frac{1}{2} \text{ и } P(H_3) = \frac{5}{50} = \frac{1}{10}.$$

Далее, известно, что $P_{H_1}(A) = 0{,}02$, $P_{H_2}(A) = 0{,}01$ и $P_{H_3}(A) = 0{,}05$. Применив формулу (3.13), получим

$$P(A) = \frac{2}{5} \cdot 0{,}02 + \frac{1}{2} \cdot 0{,}01 + \frac{1}{10} \cdot 0{,}05 = 0{,}018.$$

Мы предполагали, что A зависит от H_i. Но в таком случае и H_i будут зависеть от A, т. е. существуют условные вероятности $P_A(H_i)$.

Пусть известны предварительные вероятности гипотез $P(H_i)$ и условные вероятности $P_{H_i}(A)$. В результате проведения испытания наступило событие A. Это позволяет переоценить гипотезы,

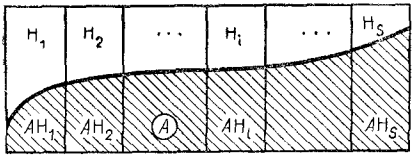


Рис. 3.3.
События A и гипотезы H_i

т. е. найти их новые условные вероятности $P_A(H_i)$. Если вероятность наступления события A при некоторой гипотезе H_i невелика и все же событие A наступило, то это может служить основанием для вывода о малой вероятности данной гипотезы. Так, если в рассматриваемом примере извлеченная деталь оказалась бракованной, то маловероятно, что она была изготовлена на втором станке, брак на котором наименее вероятен, несмотря на относительно большую долю продукции второго станка.

Для вывода формулы по переоценке гипотез обратимся к равенству (3.10). Разделив обе его части на $P(A) \neq 0$, получим

$$P_A(H_i) = \frac{P(H_i)\, P_{H_i}(A)}{P(A)} = \frac{P(H_i)\, P_{H_i}(A)}{\sum_i P(H_i)\, P_{H_i}(A)}. \qquad (3.14)$$

Это равенство получило название *формулы Байеса* или *теоремы гипотез.*

Определим по данным предыдущего примера вероятности всех трех гипотез при условии, что взятая наудачу деталь оказалась бракованной

Используя данные примера на с. 50 и найденную при его решении вероятность $P(A) = 0{,}018$, получим из (3.14):

$$P_A(H_1) = \frac{0{,}4 \cdot 0{,}02}{0{,}018} = 0{,}44;\ P_A(H_2) = \frac{0{,}5 \cdot 0{,}01}{0{,}018} = 0{,}28 \text{ и}$$

$$P_A(H_3) = \frac{0{,}1 \cdot 0{,}05}{0{,}018} = 0{,}28.$$

Естественно, что сумма условных вероятностей всех трех гипотез оказалась равной единице, как и сумма первоначальных вероятностей этих гипотез, так как достоверно, что одна из гипотез имела место в действительности При этом наиболее вероятной оказалась гипотеза H_1 Если бы эта максимальная вероятность оказалась равной единице, то вывод был бы однозначен — деталь была изготовлена на первом станке. Однако мы получили $P_A(H_1) = 0{,}44 < 1$, поэтому наряду с гипотезой H_1 нельзя отвергать и остальные.

Вероятности гипотез до опыта $P(H_i)$ иногда называют *априорными*, а вероятности $P_A(H_i)$ — *апостериорными.*

Формула Байеса может служить основанием для принятия решений после проведения эксперимента, т. е. для выработки статистического решения. Но для того чтобы выбор правдоподобной гипотезы имел достаточно оснований, необходимо, чтобы в результате эксперимента ее апостериорная вероятность была достаточно близкой к единице.

Если эксперимент проводить повторно достаточно большое число раз, то апостериорные вероятности оказываются в конечном счете мало зависящими от априорных, которые можно принять, если нет других достаточно обоснованных предложений, одинаковыми, т. е. положить $P(H_1) = \ldots = P(H_i) = \ldots = P(H_s)$. В таком случае формула (3.14) упрощается:

$$P_A(H_i) = \frac{P_{H_i}(A)}{\sum_i P_{H_i}(A)}. \qquad (3.15)$$

Рассмотрим на примере схему выработки статистического решения на основании формулы Байеса Имеются два суждения о надежности прибора, характеризуемой долей безотказно работающих приборов: завода изготовителя — 0,95 и испытательной лаборатории — 0,8. Определим на основании эксперимента, проведенного потребителем, какой из указанных характеристик следует отдать предпочтение.

Обозначим через H_1 и H_2 гипотезы, состоящие в том, что верны данные завода и лаборатории. Будем исходить из априорного предположения, что обе гипотезы равновероятны, т. е. $P(H_1) = P(H_2) = 0{,}5$

а) Проводится испытание одного прибора Обозначим через A и $\bar{A}$ события, состоящие в том, что прибор выдержал и, соответственно, не выдержал испыта-

ние. Тогда по условию задачи

$$P_{H_1}(A) = 0{,}95 \quad P_H(A) = 0{,}8 \quad P_{H_1}(\overline{A}) = 0{,}05 \text{ и } P_H(\overline{A}) = 0\,2$$

По формуле (3 15) получим:

$$P_A(H_1) = \frac{0{,}95}{0\,95 + 0{,}8} = 0.54; \; P_A(H_2) = \frac{0{,}8}{0{,}95 + 0{,}8} = 0{,}46;$$

$$P_{\overline{A}}(H_1) = \frac{0{,}05}{0{,}05 + 0{,}2} = 0\,2; \; P_{\overline{A}}(H_2) = \frac{0{,}2}{0.05 + 0{,}2} = 0{,}8.$$

Как видим, только в случае неудачи (событие $\overline{A}$) можно относительно уверенно (с вероятностью 0,8) отдать предпочтение гипотезе H_2.

б) Проводится испытание двух приборов В таком случае возможны три варианта результатов эксперимента: AA, $A\overline{A}$, и $\overline{A}\overline{A}$, при этом результат $A\overline{A}$ означает, что один прибор выдержал, а другой — не выдержал испытания, безразлично в каком порядке Найдем их вероятности при каждой из гипотез:

$$P_{H_1}(AA) = 0{,}95^2 = 0{,}9085; \; P_{H_1}(A\overline{A}) = 2 \cdot 0{,}95 \cdot 0{,}05 = 0{,}095 \text{ и}$$

$$P_{H_1}(\overline{A}\overline{A}) = 0{,}05^2 = 0{,}0025.$$

Аналогично при гипотезе H_2 получим:

$$P_{H_2}(AA) = 0{,}64; \; P_{H_2}(A\overline{A}) = 0{,}32 \text{ и } P_{H_2}(\overline{A}\overline{A}) = 0{,}04$$

Результаты расчетов по формуле (3.15) вероятностей обеих гипотез при различных результатах эксперимента приведены в табл 3.1 Как видим, в данном случае суждения о правдоподобии гипотез оказались более надежными, особенно при результате $\overline{A}\overline{A}$, оба испытуемых прибора не выдержали испытания), подтверждающем данные лаборатории, а не завода. Очевидно, что при испытании трех приборов результаты окажутся еще более надежными.

Таблица 3.1

Результаты эксперимента	Полная вероятность	Вероятности гипотез	
		H_1	H_2
A	0,7710	0,585	0,415
$\overline{A}\overline{A}$ или $\overline{A}A$	0,2075	0,229	0,771
$\overline{A}\overline{A}$	0,0212	0,059	0,941

Возможна и другая схема применения формулы Байеса, когда предполагается последовательное проведение эксперимента (последовательное испытание приборов по одному) и апостериорные вероятности гипотез после первого испытания принимаются в качестве априорных вероятностей к началу второго испытания и т д

Глава 4

Случайные величины

4.1. Основные понятия

Мы уже отмечали ранее, что понятие события было введено как абстрактная модель качественного признака, о котором имеет смысл лишь альтернативное суждение «есть» признак или «нет»

его. Дальнейшее развитие теории вероятностей требует введения нового понятия — случайной величины, представляющей собой абстрактную модель количественного признака. Случайная величина характеризует количественный результат испытания. Примерами величин, принимающих различные числовые значения под влиянием многих случайных обстоятельств, могут служить: а) остаток вклада по выбранному наудачу лицевому счету; б) число вызовов, поступающих на телефонную станцию в течение определенного промежутка времени; в) число попаданий при 5 выстрелах; г) продолжительность обслуживания покупателя.

Приведенные примеры, взятые из различных областей, объединяет одно общее положение: результат «испытания» (выбор наудачу лицевого счета, поступление телефонных вызовов, число попаданий и т. д.) под влиянием случайных обстоятельств принимает то или иное числовое значение из некоторого множества. Таким образом, принятие каждого из возможных значения есть событие, которому соответствует некоторая вероятность. Такие величины называются *случайными величинами.*

В приведенных примерах встречались два вида случайных величин: *дискретные*, множества возможных значений которых дискретны, и величины, множества возможных значений которых сплошь заполняют некоторый интервал.

Для описания случайной величины необходимо указать не только множество ее возможных значений (что было бы достаточно при описании обычной переменной величины), но и охарактеризовать вероятности, с которыми принимаются те или иные значения. Такое полное описание случайной величины называется ее *законом распределения.*

Пусть рассматривается дискретная случайная величина X, которая может принимать значения $x_1, \ldots, x_k, \ldots, x_n$, с вероятностями $P(X = x_1) = p_1, \ldots, P(X = x_k) = p_k, \ldots, P(X = x_n) = p_n$. Соответствие между двумя последовательностями $\{x_k\}$ и $\{p_k\}$ может быть в этом случае задано в форме таблицы, в которой принято x_k располагать в возрастающем порядке (табл. 4.1).

Таблица 4.1

X	x_1	$\ldots$	x_k	$\ldots$	x_n
p_k	p_1	$\ldots$	p_k	$\ldots$	p_n

Очевидно, что в любом распределении совокупность событий $X = x_1$, $X = x_k$ образует полную группу несовместных событий и потому сумма их вероятностей равна единице. Таким образом, получаем основное свойство любого дискретного распределения:

$$\sum_k p_k = 1, \tag{4.1}$$

представляющее собой теоретический аналог соответствующего свойства статистического ряда.

Таблица 42

X	0	5	100	1000
p_k	0,64	0,25	0,1	0,01

Распределение случайной величины X, обозначающей выигрыш на лотерейный билет, если на каждые 100 билетов разыгрываются 1 выигрыш в 1 000 руб., 10 — по 100 руб. и 25 — по 5 руб, задается табл. 4.2.

Пусть распределение признака в генеральной совокупности характеризуется табл. 4.3, где через p_k обозначена относительная частота (доля) значения признака $X = x_k$.

Производится отбор наудачу одного элемента. Значение признака у отобранного элемента может оказаться равным любому из чисел $x_1, \ldots, x_s$, поэтому его можно рассматривать как случайную величину X. При этом отбор элемента со значением признака $X = x_k$ есть событие, вероятность которого $P(X = x_k) = p_k$. В результате получаем распределение X, совпадающее с генеральным распределением (см. табл. 4.3).

Таблица 4.3

Значение признака X	x_1	...	x_k	...	x_s
Доля признака p	p_1	...	p_k	...	p_s

Поэтому признак, варьирующий в генеральной совокупности (обычно весьма большого или гипотетически бесконечного объема) может рассматриваться как случайная величина, а его распределение — как распределение этой случайной величины (иногда его называют *теоретическим распределением*).

Для получения информации об этом распределении производится выборочное наблюдение (опыт) некоторого ограниченного объема n. Результат выборки фиксируется в виде *выборочного* (*опытного*, или *статистического*) распределения, по форме аналогичного теоретическому распределению. Отличие будет лишь в том, что вместо вероятностей отдельных значений признака (табл. 4.3) будут указываться относительные частоты $\frac{m_k}{n}$ соответствующих значений признака в выборке. Если объем выборки n достаточно большой, то относительные частоты события $X = x_n$ будут близки к вероятностям этого события p_k, т. е. все выборочное распределение будет достаточно близким к теоретическому.

Сходство между статистическим распределением, подробно рассмотренным в первых двух главах, и теоретическим распределением дискретной случайной величины позволяет по аналогии распространить на последнее приемы описания и графического изображения, а также определение числовых характеристик, которые подробно излагались в гл. 1 и 2. Будем называть эти характеристики *теоретическими*, в отличие от ранее рассмотренных статистических.

Если для тех и других используются одинаковые буквенные обозначения ($\bar{X}$, D, σ и т. д.), то статистические характеристики будут дополнительно отмечаться звездочкой.

В табл. 4.4 приведены для сопоставления функции распределения и основные характеристики статистического и дискретного теоретического распределений.

Таблица 44

Характеристика	Распределение	
	статистическое	теоретическое
Доля признака Вероятность	$\frac{m_k}{n} = p_k^*$	$P(X = x_k) = p_k$
Средняя	$\overline{X}^* = \sum_k x_k p_k^*$	$M(X) = \sum_k x_k p_k$
Дисперсия	$D^* = \sum_k (x_k - \overline{X}^*)^2 p_k^*$	$D(X) = \sum_k (x_k - \overline{X})^2 p_k$
Среднее квадратическое отклонение	$\sigma^* = \sqrt{D^*}$	$\sigma(X) = \sqrt{D(X)}$
Моменты: центральные начальные	$\nu_l^* = \sum_k (x_k - \overline{X}^*)^l p_k^*$ $\mu_l^* = \sum_k x_k^l p_k^*$	$\nu_l = \sum_k (x_k - \overline{X})^l p_k$ $\mu_l = \sum_k x_k^l p_k$
Функция распределения	$F^*(x_k) = \sum_{i=1}^{k-1} p_i^*$	$F(x) = P(X < x_k)$

Остановимся подробнее на двух основных характеристиках дискретной случайной величины.

Математическое ожидание (среднее значение) *дискретной случайной величины* равно сумме попарных произведений всех ее значений на соответствующие им вероятности:

$$\overline{X} = M(X) = \sum_k x_k p_k. \qquad (4.2)$$

Если множество возможных значений случайной величины бесконечно, но счетно, то вместо конечной суммы получим бесконечный ряд, который может *сходиться* (математическое ожидание существует) или *расходиться* (математического ожидания не существует).

Укажем основные свойства математического ожидания, доказательство которых можно получить из (4.2) так же, как доказы-

вались аналогичные свойства средней арифметической (см. § 2.2)

$$1)\ M(c)=c;\quad 2)\ M(X+c)=M(X)+c;\quad 3)\ M(cX)=cM(X);$$
$$4)\ M(aX+b)=aM(X)+b,$$

где a, b, c — неслучайные величины.

Дисперсия дискретной случайной величины равна математическому ожиданию квадрата отклонения случайной величины от ее математического ожидания:

$$D(X)=M(X-\bar{X})^2=\sum_k (x_k-\bar{X})^2 p_k. \qquad (4.3)$$

И здесь для бесконечного (но счетного) множества возможных значений случайной величины получим ряд, который может оказаться сходящимся или расходящимся

Основные свойства дисперсии также легко получаются из определения:

$$1)\ D(c)=0;\quad 2)\ D(X+c)=D(X);\quad 3)\ D(cX)=c^2D(X):$$
$$4)\ D(aX+b)=a^2D(X):\quad 5)\ D(X)=M(X^2)-\overline{X^2}.$$

Последним соотношением удобно пользоваться при подсчете $D(X)$. Корень квадратный из дисперсии есть среднее квадратическое отклонение (стандарт) случайной величины $\sigma(X)=\sqrt{D(X)}$.

4.2. Аксиоматическое определение случайной величины. Функция распределения

Рассмотренное в предыдущем параграфе понятие случайной величины было введено чисто описательно. При этом рассматривалась лишь дискретная случайная величина. Поскольку случайная величина характеризует количественный результат (исход) испытания, а множество взаимоисключающих исходов образуют пространство Ω элементарных событий, то естественно связать это новое фундаментальное понятие с пространством Ω.

Случайной величиной X называется вещественная функция $X=X(e)$, отображающая множество $\Omega=\{e\}$ в множество действительных чисел. В случае дискретного множества Ω каждому элементарному событию $e_i\in\Omega$ соответствует некоторое действительное число x_i — *значение* (*реализация*) случайной величины X. В этом случае множество значений X также будет дискретным. В случае несчетного множества Ω приведенное определение нуждается в уточнении, которое мы, однако, не приводим, в связи с необходимостью привлечения ряда сложных понятий, выходящих за рамки книги. Отметим лишь, что в этом случае множество значений случайной величины также может оказаться несчетным *, в частности заполнять некоторый интервал или всю дей-

* Заметим, что и для бесконечного несчетного множества Ω множество значений X может оказаться дискретным или даже конечным.

...ительную ось. В дальнейшем будем рассматривать и такие случайные величины, опираясь на полуинтуитивное их описание.

Вернемся к рассмотрению дискретной случайной величины X, заданной на дискретном вероятностном пространстве. Пусть определенное ее значение $X = x_k$ соответствует подмножеству элементарных исходов $\{e_1^k, e_2^k, \ldots\} = A$, образующих некоторое событие. Вероятность этого события, по определению, есть

$$P(X = x_k) = p_k = \sum_{e_i^k \in A} p(e_i^k).$$

Определив таким образом значения p_k для всех значений x_k, получим две последовательности $\{x_k\}$ и $\{p_k\}$, которые и задают *закон распределения* дискретной случайной величины.

Однако для случайной величины с несчетным множеством значений такое задание распределения непригодно, хотя бы потому,

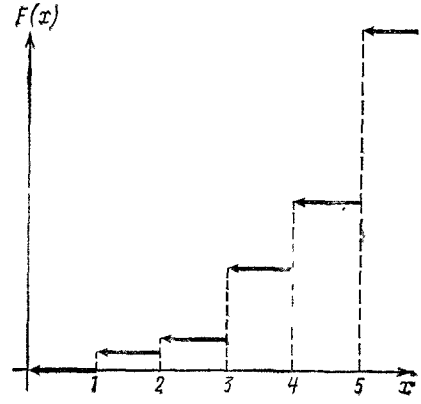


Рис. 4.1.
Функция распределения дискретной случайной величины

что вероятности отдельных числовых значений (т. е. вероятности попадания в заданную точку интервала) будут равны нулю. Поэтому необходимо обратиться к другому способу задания распределения, применимому к произвольной случайной величине (в том числе и дискретной).

Функцией распределения $F(x)$ случайной величины X называется вероятность того, что X примет значение меньше данного числа x, т. е.

$$F(x) = P(X < x). \tag{4.4}$$

Для дискретной случайной величины это определение представляет собой полную аналогию с определением накопленной относительной частоты $F^*(x)$. График $F(x)$ в этом случае является «ступенчатой» линией (рис. 4.1), меняющейся скачком и непрерывной слева в каждой точке x_k:

$$F(x_k + 0) - F(x_k) = P(X = x_k).$$

Для произвольной случайной величины эта функция также определена на всей числовой оси, и график ее будет на некоторых

интервалах изображаться монотонно возрастающей кривой (рис. 4.2).

Рассмотрим основные свойства функции распределения.

1. Вероятность того, что случайная величина принимает значения в промежутке $[x', x'']$, равна разности значений $F(x)$ на концах этого промежутка, т. е.

$$P\{x' \leqslant X < x''\} = F(x'') - F(x'). \tag{4.5}$$

Доказательство. Рассмотрим три события, состоящие в выполнении следующих неравенств: $A - (X < x')$,

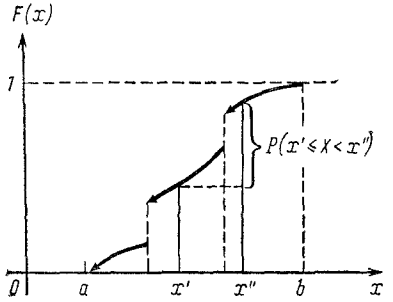


Рис. 4.2.
Функция распределения произвольной случайной величины

$B - (x' \leqslant X < x'')$ и $C - (X < x'')$ (рис. 4.3). Очевидно, что $C = A + B$, откуда, так как события A и B несовместны, $P(C) = P(A) + P(B)$. Выразив вероятности $P(A)$ и $P(C)$ из (4.4), получим важное для многочисленных приложений равенство (4.5).

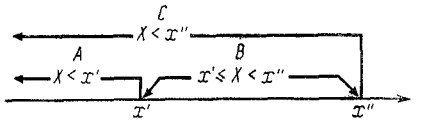


Рис. 4.3.
К доказательству свойства функции распределения

2. Значения функции $F(x)$ заключены в границах $0 \leqslant F(x) \leqslant 1$, что следует непосредственно из определения $F(x)$, как вероятности.

3. $F(x)$ — монотонно возрастающая функция.

Доказательство. Пусть $x'' > x'$. Тогда из (4.5), учитывая неотрицательность второго слагаемого, получим

$$F(x'') = F(x') + P\{x' \leqslant X < x''\} \geqslant F(x').$$

4. $\lim\limits_{x \to -\infty} F(x) = 0$ и $\lim\limits_{x \to \infty} F(x) = 1$. Это свойство следует из определения $F(x)$ и того, что событие $X < -\infty$ невозможное, а $X < \infty$ — достоверное.

5. Если $a \leqslant X \leqslant b$, то $F(x) = 0$ при $x \leqslant a$ и $F(x) = 1$ при $x \geqslant b$.

Рассмотрим приращение $F(x)$ на интервале $(x, x+\Delta x)$:

$$\Delta F(x) = F(x+\Delta x) - F(x) = P\{x \leqslant X < x+\Delta x\}. \qquad (4.6)$$

Если $\lim\limits_{\Delta x \to 0} \Delta F(x) = 0$ при любом x, т. е. $F(x)$ — непрерывная функция, то вероятность того, что случайная величина принимает данное значение x, равна нулю. Поэтому в дальнейшем не будем под знаком вероятности различать строгие и нестрогие неравенства, так как

$$P(X \leqslant x) = P(X < x) + P(X = x) = P(X < x).$$

Если функция $F(x)$ к тому же и дифференцируемая при любом x (за исключением, быть может, счетного числа точек), т. е. существует

$$\frac{dF}{dx} = \lim_{\Delta x \to 0} \frac{\Delta F(x)}{\Delta x} = f(x), \qquad (4.7)$$

то случайная величина X называется *непрерывной*, а функция $f(x)$ — *дифференциальной функцией распределения.*

Кроме рассмотренных двух типов (дискретные и непрерывные), существуют и более общие типы случайных величин, например дискретные на одном участке и непрерывные на другом, однако мы ограничимся в дальнейшем рассмотрением только указанных двух типов.

Подставив в последнее равенство выражение для $\Delta F(x)$ из (4.6), получим

$$f(x) = \lim_{\Delta x \to 0} \frac{P\{x \leqslant X < x+\Delta x\}}{\Delta x}, \qquad (4.8)$$

откуда следует вероятностная интерпретация дифференциальной функции распределения: функция $f(x)$ *есть плотность распределения вероятностей* в точке x или, короче, плотность вероятности. Заметим, что, таким образом, $f(x)$ есть теоретический аналог статистического понятия плотности относительной частоты. График некоторой функции $f(x)$ изображен на рис. 4.4. Интегрируя (4.7) в пределах от $-\infty$ до x, получим соотношение

$$F(x) = \int_{-\infty}^{x} f(u)\, du, \qquad (4.9)$$

геометрически означающее, что $F(x)$ численно равна площади $S(x)$ криволинейной трапеции (см. рис. 4.4), ограниченной сверху кривой $f(x)$ и справа — ординатой в точке x. Далее, вероятность попадания значений случайной величины в заданный интервал (x', x'') определится из выражения

$$P\{x' \leqslant X < x''\} = \int_{x'}^{x''} f(u)\, du \qquad (4.10)$$

и численно равна площади S_1 трапеции, заключенной между ординатами в точках x' и x'' (см. рис. 4.4).

Из определения $f(x)$ вытекают следующие ее свойства:

1. $f(x) \geqslant 0$, что следует из монотонного возрастания $F(x)$.

2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$, что вытекает из равенства (4.9) и 4-го свойства $F(x)$. Данное свойство, получившее название условия *нормируемости* $f(x)$, представляет собой обобщение основного свойства ряда распределения (4.1). Геометрически оно означает, что пло-

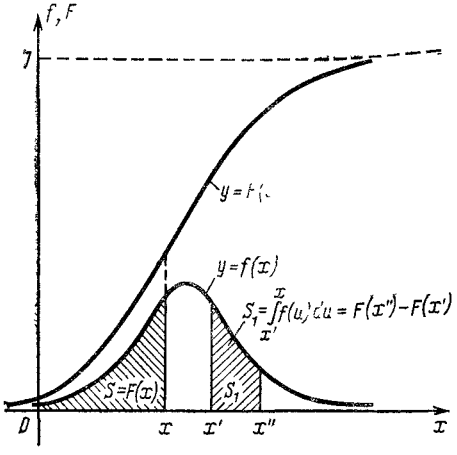


Рис. 4.4.
Графики функции распределения и плотности вероятности

щадь фигуры, ограниченной сверху кривой плотности вероятности, равна единице.

3. Существует хотя бы одно значение $\xi \in (x,\ x + \Delta x)$, для которого выполняется равенство

$$P\{x \leqslant X < x + \Delta x\} = \Delta F(x) = f(\xi)\,\Delta x, \tag{4.11}$$

вытекающее из теоремы «о среднем» для интеграла $\Delta F(x) = \int_{x}^{x+\Delta x} f(u)\,du$. Полагая $\Delta x = dx \to 0$, откуда $\xi \to x$, получим из (4.11) выражение для вероятности попадания случайной величины в бесконечно малый интервал

$$P\{x \leqslant X < x + dx\} = f(x)\,dx, \tag{4.12}$$

получившее название *элемента вероятности.*

Описание распределения непрерывной случайной величины с помощью плотности вероятности $f(x)$ более удобно, чем с помощью функции распределения $F(x)$, так как позволяет наглядно передать его особенности на различных участках изменения случайной величины. Однако вероятности попадания случайной величины в заданный интервал предпочтительнее вычислять по формуле (4.5), а не (4.11), так как при использовании последней ве-

личина ξ является неопределенной. Ее практически выбирают равной $x + \frac{\Delta x}{2}$.

Приведем пример определения $f(x)$ и $F(x)$ для непрерывной случайной величины, распределенной равномерно (рис. 4.5) на интервале (a, b).

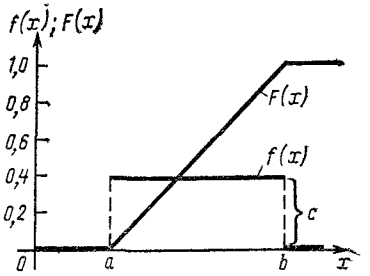


Рис. 4.5.
Равномерное распределение на интервале (a, b)

Из условия следует, что $f(x) = 0$ для $-\infty < x < a$ и для $b \leqslant x < \infty$ и $f(x) = c = \text{const}$ для $a \leqslant x < b$. Величина c может быть определена из условия нормируемости $f(x)$:

$$1 = \int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{a} f(x)\,dx + \int_{a}^{b} f(x)\,dx + \int_{b}^{\infty} f(x)\,dx.$$

Подставляя указанные выражения для $f(x)$, получим

$$1 = \int_{-\infty}^{a} 0\,dx + \int_{a}^{b} c\,dx + \int_{b}^{\infty} 0\,dx = c\,(b-a),$$

$$\text{откуда } c = \frac{1}{b-a}.$$

Окончательно получим выражение для плотности вероятности *равномерного распределения* (закон равномерной плотности):

$$f(x) = \begin{cases} 0 \text{ для } x < a; \\ \frac{1}{b-a} \text{ для } a \leqslant x \leqslant b; \\ 0 \text{ для } x > b. \end{cases} \tag{4.13}$$

Используя соотношение (4.8), получим интегральную функцию для равномерного распределения:

$$F(x) = \begin{cases} \int_{-\infty}^{x} 0 \cdot du = 0 \text{ для } x \leqslant a; \\ \int_{-\infty}^{x} f(u)\,du = \int_{-\infty}^{a} 0\,du + \int_{a}^{x} c\,du = \frac{x-a}{b-a} \text{ для } a < x \leqslant b; \\ \int_{-\infty}^{x} f(u)\,du = \int_{-\infty}^{a} 0\,du + \int_{a}^{b} c\,du + \int_{b}^{x} 0 \cdot du = 1 \text{ для } x > b. \end{cases} \tag{4.13'}$$

Графики этих функций изображены на рис. 4.5.

Определим вероятность того, что случайная величина, равномерно распределенная на интервале (2, 7), принимает значения в промежутке (5, 6).

С помощью формулы (4, 5) получаем

$$P\{5 \leqslant X \leqslant 6\} = F(6) - F(5) = \frac{1}{7-2}[(6-2) - (5-2)] = 0{,}2.$$

Тот же результат получим при расчете по формуле (4.11):

$$P\{5 \leqslant X \leqslant 6\} \approx \frac{1}{7-2} \cdot 1 = 0{,}2,$$

так как $f(x) = \text{const} = \frac{1}{b-a}$ при любом $\xi \in (a, b)$.

Часто возникает необходимость в рассмотрении зависимости между случайными величинами и определении закона распределения функции $Y = \varphi(X)$ по известной плотности распределения $f(x)$. Пусть зависимость между значениями этих случайных величин $y = \varphi(x)$ изображена на рис. 4.6. Обозначим функцию распределения Y через $\theta(y)$, т. е. $\theta(y) = P(Y < y)$. Проведем прямую, параллельную оси абсцисс, на расстоянии y. Тогда на оси абсцисс выделятся участки $\Delta_1, \ldots, \Delta_i$ изменения аргумента x, на которых выполняется неравенство $Y < y$. Таким образом, событие $Y < y$ равносильно сумме попарно несовместных событий

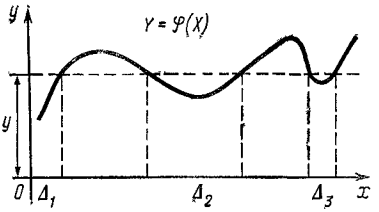


Рис. 4.6.
Произвольная функция от случайной величины

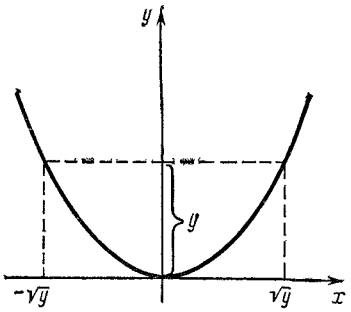


Рис. 4.7.
К определению плотности распределения $Y = X^2$

$$(X \in \Delta_1) + (X \in \Delta_2) + \ldots + (X \in \Delta_i) + \ldots$$

Согласно (4.10) вероятность попадания X в каждый из интервалов равна интегралу от $f(x)$ по данному интервалу, откуда

$$\theta(y) = \int\limits_{(\Delta_1)} f(x)\,dx + \int\limits_{(\Delta_2)} f(x)\,dx + \ldots + \int\limits_{(\Delta_i)} f(x)\,dx + \ldots \tag{4.14}$$

Дифференцируя эту функцию по y, получим плотность распределения $P(y) = \frac{d\theta(y)}{dy}$.

Найдем плотность распределения случайной величины $Y = X^2$ по заданной плотности $f(x)$.

График функции $\varphi(x) = x^2$ изображен на рис. 4.7. Проведем прямую, параллельную оси абсцисс, на расстоянии y от нее. На оси абсцисс образуется интер-

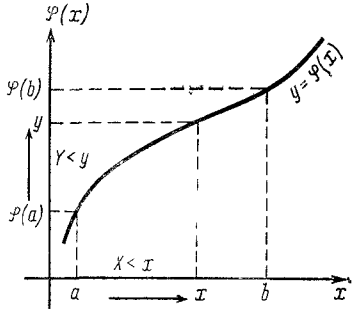


Рис. 4.8.
К доказательству теоремы 1

вал $(-\sqrt{y}, \sqrt{y})$, на котором выполняется неравенство $Y < y$. Поэтому, используя определение функции распределения $\theta(y) = P(Y < y)$ и (4.10), получим

$$\theta(y) = P(Y<y) = \int_{-\sqrt{y}}^{\sqrt{y}} f(u)\,du = \int_{-\sqrt{y}}^{0} f(u)\,du + \int_{0}^{\sqrt{y}} f(u)\,du =$$

$$= -\int_{0}^{-\sqrt{y}} f(u)\,du + \int_{0}^{\sqrt{y}} f(u)\,du.$$

Дифференцируя по y и используя правило дифференцирования интеграла по верхнему пределу интегрирования, найдем

$$P(y) = -f(-\sqrt{y})\left(\frac{-1}{2\sqrt{y}}\right) + f(\sqrt{y})\frac{1}{2\sqrt{y}} = [f(-\sqrt{y}) + f(\sqrt{y})]\frac{1}{2\sqrt{y}}. \tag{4.15}$$

Особенно простым становится вывод формулы для $P(y)$, если функция $\varphi(x)$ монотонная.

Теорема 1. Если $Y = \varphi(X)$ есть монотонная функция случайной величины X, то плотность распределения Y определится из равенства

$$P(y) = f(x)\left|\frac{dx}{dy}\right|, \tag{4.16}$$

где $x = \psi(y)$ есть обратная функция к заданной.

Рассмотрим вначале монотонно возрастающую (рис. 4.8) функцию $\varphi(x)$ в заданном промежутке (a, b) изменения X. Когда X возрастает от a до b, то Y также будет возрастать от $\varphi(a)$ до $\varphi(b)$.

Пусть $\theta(y) = P(Y < y)$. Так как $\varphi(x)$ монотонно возрастающая, то неравенство $Y < y$ будет выполняться всякий раз, когда

выполняется неравенство $X < x$, и наоборот. Следовательно, будут равны и их вероятности, т. е.

$$\theta(y) = P(Y < y) = P(X < x) = \int_{-\infty}^{x} f(u)\,du.$$

Дифференцируя по y, получим

$$\frac{d\theta(y)}{dy} = P(y) = \frac{d}{dx}\int_{-\infty}^{x} f(u)\,du\,\frac{dx}{dy} = f(x)\,\frac{dx}{dy}. \qquad (4.16')$$

Выполнив аналогичные преобразования для монотонно убывающей функции $\varphi(x)$ и учитывая, что в этом случае неравенство $Y < y$ равносильно противоположному неравенству $X \geqslant x$, т. е. $P(Y < y) = P(X \geqslant x) = 1 - P(X < x)$, получим

$$\theta(y) = 1 - \int_{-\infty}^{x} f(u)\,du.$$

Дифференцируя по y, получим выражение для плотности распределения $P(y)$ монотонно убывающей функции:

$$P(y) = -f(x)\,\frac{dx}{dy}. \qquad (4.16'')$$

Равенства (4.16′) и (4.16″) объединяются в формулу (4.16).

Найдем плотность распределения $P(y)$ линейной функции $Y = ax + b$ по заданной плотности $f(x)$.

Очевидно, что линейная функция монотонно возрастает при $a > 0$ и монотонно убывает при $a < 0$ Поэтому, используя (4 16′) и (4 16″) и учитывая, что $x = \dfrac{y-b}{a}$, откуда $\dfrac{dx}{dy} = \dfrac{1}{a}$, получим

$$P(y) = f\left(\frac{y-b}{a}\right)\frac{1}{|a|}. \qquad (4.17)$$

4.4. Числовые характеристики непрерывной случайной величины и их основные свойства

Пусть X — непрерывная случайная величина с плотностью распределения $f(x)$. По аналогии с определением математического ожидания дискретной случайной величины, заменив лишь в формуле (4.2) вероятности p_k на элемент вероятности $f(x)\,dx$ и сумму на интеграл, получим основное определение.

Математическим ожиданием непрерывной случайной величины называется число, определяемое из равенства

$$M(X) = \int_{-\infty}^{\infty} xf(x)\,dx. \qquad (4.18)$$

Несобственный интеграл может оказаться расходящимся, тогда математического ожидания не существует. Рассмотрим, напри-

мер, определение математического ожидания случайной величины x, равномерно распределенной на интервале (a, b) Согласно определению и найденной ранее плотности вероятности (4.13) получим

$$M(X)=\int_{-\infty}^{\infty} xf(x)\,dx=\int_{-\infty}^{a} x\cdot 0\cdot dx+\int_{a}^{b} x\cdot\frac{1}{b-a}\,dx+$$

$$+\int_{b}^{\infty} x\cdot 0\cdot dx=\int_{a}^{b}\frac{x}{b-a}\,dx.$$

После интегрирования и подстановки пределов получим

$$M(X)=\frac{1}{b-a}\,\frac{x^2}{2}\Big|_a^b=\frac{b^2-a^2}{2(b-a)}=\frac{b+a}{2}.$$

Дисперсия и моменты любого порядка определяются через математическое ожидание, одинаково как для дискретной, так и для непрерывной случайной величины, а именно

$$D(X)=M(X-\bar{X})^2;\ \nu_l=M(X-\bar{X})^l;\ \mu_l=M(X^l). \qquad (4.19)$$

Для их вычисления необходимо раскрыть лишь выражение для математического ожидания. Обозначим $(X-\bar{X})^2=Y$.
Тогда, по определению получим

$$D(X)=M(Y)=\int_{-\infty}^{\infty} yP(y)\,dy=\int_{0}^{\infty} yP(y)\,dy,$$

так как $y=(X-\bar{X})^2>0$ при любом x. Подставляя выражение для $P(y)$ из (4.15) и разбивая на два интеграла, придем к выражению

$$D(X)=\int_{0}^{\infty} yf(-\sqrt{y})\frac{dy}{2\sqrt{y}}+\int_{0}^{\infty} yf(\sqrt{y})\frac{dy}{2\sqrt{y}}.$$

Произведя в первом интеграле замену $-\sqrt{y}=x-\bar{X}$, откуда $\frac{-dy}{2\sqrt{y}}=dx$, и во втором интеграле $\sqrt{y}=x-\bar{X}$, откуда $\frac{dy}{2\sqrt{y}}=dx$, и переходя соответственно к новым пределам интегрирования в первом интеграле, получим

$$D(X)=-\int_{0}^{-\infty}(x-\bar{X})^2 f(x)\,dx+\int_{0}^{\infty}(x-\bar{X})^2 f(x)\,dx=$$

$$=\int_{-\infty}^{0}(x-\bar{X})^2 f(x)\,dx+\int_{0}^{\infty}(x-\bar{X})^2 f(x)\,dx.$$

Наконец, объединяя полученные два интеграла в один, находим

$$D(X)=\int_{-\infty}^{\infty}(x-\bar{X})^2\ f(x)\,dx. \tag{4.20}$$

Нетрудно увидеть аналогию между этой формулой и соответствующей формулой (4.3) для дискретной случайной величины.

Аналогичным образом могут быть получены формулы для моментов любого порядка:

$$\nu_l=\int_{-\infty}^{\infty}(x-\bar{X})^l\ f(x)\,dx;\ \nu_l(c)=\int_{-\infty}^{\infty}(x-C)^l\ f(x)\,dx;$$
$$\mu_l=\int_{-\infty}^{\infty}x^l\ f(x)\,dx. \tag{4.21}$$

Для иллюстрации полученных формул найдем дисперсию и центральные моменты 3-го и 4-го порядков для случайной величины, равномерно распределенной на интервале (a, b):

$$D(X)=\int_a^b(x-\bar{X})^2\ c\,dx=c\,\frac{(x-\bar{X})^3}{3}\Bigg|_a^b=\frac{c}{3}((b-\bar{X})^3-(a-\bar{X})^3);$$

$$\nu_l=\int_a^b(x-\bar{X})^l\ c\,dx=\frac{c}{l+1}(x-\bar{X})^{l+1}\Bigg|_a^b=$$
$$=\frac{c}{l+1}\left((b-\bar{X})^{l+1}-(a-\bar{X})^{l+1}\right).$$

Подставив $\bar{X}=\frac{b+a}{2}$ и $c=\frac{1}{b-a}$, найдем:

$$D(X)=\frac{(b-a)^2}{12};\ \nu_3=0;\ \nu_4=\frac{(b-a)^4}{5\cdot 2^4}.$$

Теперь остановимся на основных свойствах числовых характеристик непрерывной случайной величины, которые естественно совпадают со свойствами характеристик дискретной случайной величины, но отличаются обоснованием.

Теорема 2 Математическое ожидание линейной функции $Y=aX+b$ равно той же функции от математического ожидания X.

Имеем $y=ax+b$, откуда $dy=a\,dx$ и $P(y)=\frac{1}{|a|}f(x)$ (см. пример на с. 64). На основании (4.18) получим

$$M(Y)=\int_{-\infty}^{\infty}yP(y)\,dy.$$

Переходим к переменной x, отдельно для случая $a > 0$ и $a < 0$. Так, при $a > 0$ получаем

$$M(aX + b) =$$

$$= \int_{-\infty}^{\infty} (ax + b) \frac{f(x)}{a} a \, dx = a \int_{-\infty}^{\infty} x f(x) \, dx + b \int_{-\infty}^{\infty} f(x) \, dx = a\bar{X} + b.$$

При $a < 0$ имеем $|a| = -a$ и при изменении y от $-\infty$ до $+\infty$ x будет меняться от ∞ до $-\infty$, откуда

$$M(aX + b) =$$

$$= \int_{\infty}^{-\infty} (ax + b) \frac{f(x)}{-a} a \, dx = \int_{-\infty}^{\infty} (ax + b) f(x) \, dx = a\bar{X} + b.$$

Таким образом, в обоих случаях получаем

$$M(aX + b) = a\bar{X} + b. \tag{4.22}$$

При $a = 0$ получаем $M(b) = b$. Опираясь на (4.22), можно получить выражения для дисперсии и моментов линейной функции.

Теорема 3. Дисперсия и центральные моменты любого порядка линейной функции $Y = aX + b$ выражаются через соответствующие характеристики X по формулам

$$D(aX + b) = a^2 D(X); \quad \nu_l(aX + b) = a^l \nu_l(X). \tag{4.23}$$

Из определения дисперсии имеем

$$D(aX + b) = M[(aX + b)^2 - (a\bar{X} + b)^2] = M(a^2 (X - \bar{X})^2) =$$

$$= a^2 M(X - \bar{X})^2 = a^2 D(X).$$

Аналогично доказываются и остальные формулы.

Особое значение приобретают выведенные формулы для случайной величины

$$X^0 = \frac{X - \bar{X}}{\sigma}, \tag{4.24}$$

получившей название *нормированного отклонения*. Из (4.22) имеем

$$M(X^0) = \frac{1}{\sigma} M(X - \bar{X}) = \frac{1}{\sigma} (\bar{X} - \bar{X}) = 0. \tag{4.25}$$

Далее из (4.23) получаем последовательно

$$D(X^0) = \frac{1}{\sigma^2} D(X) = 1 \quad \text{и} \quad \nu_l(X^0) = \frac{1}{\sigma^l} \nu_l(X). \tag{4.26}$$

Можно показать, что все остальные свойства числовых характеристик и формулы моментов, о которых мы подробно говорили при рассмотрении статистических распределений (см. гл. 2) и распространили по аналогии на характеристики дискретной случайной величины, остаются справедливыми и для непрерывной случайной величины (табл. 4.5).

Таблица 45

	Дискретная случайная величина	Непрерывная случайная величина
Математическое ожидание	$M(X) = \sum_k x_k p_k$	$M(X) = \int_{-\infty}^{\infty} x f(x)\,dx$
	$M(aX + b) = aM(X) + b$	
	1) $M(b) = b$ 2) $M(aX) = aM(X)$ 3) $M(X - \overline{X}) = 0$ 4) $M\left(\frac{X - \overline{X}}{\sigma}\right) = 0$	
Дисперсия	$D(X) = \sum_k (x_k - \overline{X})^2 p_k$	$D(X) = \int_{-\infty}^{\infty} (x - \overline{X})^2 f(x)\,dx$
	$D(aX + b) = a^2 D(X)$	
	1) $D(b) = 0$, 2) $D(X + b) = D(X)$	3) $D\left(\frac{X - \overline{X}}{\sigma}\right) = 1$ 4) $D(X) = M(X^2) - \overline{X}^2$
Моменты	$\nu_l(X) = \sum_k (x_k - \overline{X})^l p_k$	$\nu_l(X) = \int_{-\infty}^{\infty} (x - \overline{X})^l f(x)\,dx$
	$\mu_l(X) = \sum_k x_k^l p_k$	$\mu_l(X) = \int_{-\infty}^{\infty} x^l f(x)\,dx$
	$\nu_l(aX + b) = a^l \nu_l(X)$	
Формулы моментов	$X' = \frac{X - C}{h}$	
	1) $\overline{X} = \mu_1(X) = h\mu_1(X') + C$; 2) $D(X) = \nu_2(X) = h^2(\mu_2(X') - \mu_1^2(X'))$;	
	3) $\nu_3(X) = h^3\{\mu_3(X') - 3\mu_1(X')\mu_2(X') + 2\mu_1^3(X')\}$,	
	4) $\nu_4(X) = h^4\{\mu_4(X') - 4\mu_3(X')\mu_1'(X') + 6\mu_1^2(X')\mu_2(X') - 3\mu_1^4(X')\}$.	

Глава 5

Модели повторных испытаний

5.1. Повторные испытания

Мы уже познакомились с такими фундаментальными понятиями теории вероятностей, как случайная величина и ее закон распределения. Однако при этом не рассматривался сам механизм испы-

таний, результат которых характеризовался данной случайной величиной. Между тем при изучении реальных процессов, сопровождаемых воздействием случайных факторов, именно раскрытие этого механизма позволяет дать полное описание процесса, построить его вероятностную модель

Рассмотрим серию повторных испытаний, каждое из которых есть отбор одного элемента из генеральной совокупности. Тогда такая серия из n испытаний может рассматриваться как процесс образования выборки объемом n Пусть результат каждого испытания есть появление одного из двух альтернативных исходов наступление некоторого события A или противоположного ему события $\bar{A}$. Испытания могут проводиться при условиях, когда результат каждого испытания не зависит от предшествующих испытаний (*независимые* испытания) или зависит от них (*зависимые* испытания). Если обратиться к интерпретации повторных испытаний как последовательному отбору, то независимые испытания соответствуют возвратной выборке или выборке из бесконечной генеральной совокупности, а зависимые испытания соответствуют безвозвратной выборке или выборке из совокупности ограниченного объема

Пусть например, производится выборка объемом $n = 10$ из генеральной совокупности содержащей 30% изделии высшего сорта и 70% прочих сортов Отбор одного изделия есть испытание, исходом которого может быть появление изделия высшего сорта (событие A) или изделия не высшего сорта (событие $\bar{A}$)

Рассмотрим вначале *возвратную* выборку (отобранное изделие возвращается перед выбором следующего) В этом случае $P(A) = 0{,}3$ и $P(\bar{A}) = 0{,}7$ вне зависимости от порядкового номера производимого отбора

Пусть теперь производится *безвозвратная* выборка из совокупности, содержащей $N = 100$ изделий Если первый раз было отобрано изделие высшего сорта, то вероятность во второй раз отобрать такое же изделие будет равна $P_A(A) = \frac{29}{99}$ и т д Если же в первый раз было отобрано изделие не первого сорта, то $P_{\bar{A}}(A) = \frac{30}{99}$ и т д Однако, если бы имели $N = 1000$, то соответствующие вероятности $P_A(A) = \frac{299}{999}$ и $P_{\bar{A}}(A) = \frac{300}{999}$ оказались бы практически неразличимыми Очевидно, в пределе при $N \to \infty$, т е при отборе из гипотетической генеральной совокупности бесконечно большого объема любой способ отбора можно рассматривать как повторные независимые испытания

Наиболее простой из приведенных схем является модель повторных независимых испытаний с альтернативными исходами каждого испытания, получившая название *схемы* (модели) *Бернулли.*

5.2. Биномиальное распределение

Рассмотрим серию из n повторных независимых испытаний, результатом каждого из которых может быть или появление события A с вероятностью $P(A) = p$ или события $\bar{A}$ с вероятностью $P(\bar{A}) = 1 - p = q$ Результат всей серии из n испытаний можно характеризовать некоторой последовательностью событий A и $\bar{A}$, например $e_i = A\bar{A}A \ldots A\bar{A} \ldots A$.

Рассмотрим конечное множество Ω всех взаимоисключающих исходов e_i, или *пространство выборок*. Оно содержит всего $\tilde{A}_2^n = 2^n$ элементарных исходов e_i; среди них C_n^m исходов, в каждом из которых событие A появляется m раз и не появляется $n-m$ раз *. Такие исходы отличаются друг от друга лишь порядком чередования событий A и $\bar{A}$, например $e_1 = \underbrace{AA\ldots A}_{m\text{ раз}}\underbrace{\bar{A}\bar{A}\ldots\bar{A}}_{n\text{ раз}}$, $e_2 = \underbrace{\bar{A}\bar{A}\ldots\bar{A}}_{n-m\text{ раз}}\underbrace{AA\ldots A}_{m\text{ раз}}$ и т. д., и потому вероятность каждого из них по теореме умножения вероятностей $P(e_1) = P(e_2) = \ldots = p^m q^{n-m}$.

Определим на множестве Ω дискретную случайную величину X_n, равную числу появления события A в серии из n испытаний. Эта случайная величина может принимать значения 0, 1, ..., m, ..., n. При этом событию «$X_n = m$» соответствует C_n^m равновероятных исходов, вероятность каждого из которых равна $p^m q^{n-m}$, откуда, на основании определения вероятности,

$$P\{X_n = m\} = C_n^m p^m q^{n-m}. \tag{5.1}$$

Вероятность $P(X_n = m)$ будем обозначать через $P_n(m)$ (*вероятность появления события m раз из n испытаний*).

Равенство (5.1) получило название *формулы Бернулли*, а распределение вероятностей, задаваемое функцией

$$f_{\text{Б}}(m,\ n,\ p) = P_n(m) = C_n^m p^m q^{n-m}, \tag{5.2}$$

где $m = 0, 1, \ldots, n$ — текущая переменная, а n и p — параметры, *биномиального распределения*. Оно приведено в следующей таблице:

Таблица 5.1

X_n	0	1	2	...	m	...	$n-1$	n	$\sum$
$P_n(m)$	$C_n^0 q^n$	$C_n^1 p q^{n-1}$	$C_n^2 p^2 q^{n-2}$	...	$C_n^m p^m q^{n-m}$	...	$C_n^{n-1} p^{n-1} q$	$C_n^n p^n$	1

Термин «биномиальное» связан с тем, что этот ряд вероятностей можно получить как последовательные члены разложения бинома Ньютона:

$$(q+p)^n = \sum_{m=0}^{n} C_n^m p^m q^{n-m}. \tag{5.3}$$

Так как $q + p = 1$, то из (5.3) следует основное свойство ряда:

$$\sum_{m=0}^{n} P_n(m) = 1.$$

* Число таких исходов можно определить как число подмножеств n-элементного множества.

Пусть в партии готовой продукции 2/3 изделий высшего сорта Производится возвратная выборка 5 изделий. Построить ряд вероятностей, характеризующих всевозможные результаты отбора

Обозначим через A появление изделия высшего сорта при отборе одного изделия Тогда $P(A) = p = 2/3$, $q = 1 - p = 1/3$ и $n = 5$ По формуле (5 2) рассчитаем вероятности, соответствующие $m = 0, 1, \ldots, 5$ и сведем результаты расчетов в таблицу:

Таблица 5 2

X_5	0	1	2	3	4	5	Σ
$P_5(m)$	0,0041	0,0412	0,1646	0,3292	0,3292	0,1317	1

Биномиальное распределение при различных значениях параметров n и p показано на рис. 5.1. Поскольку распределение дис-

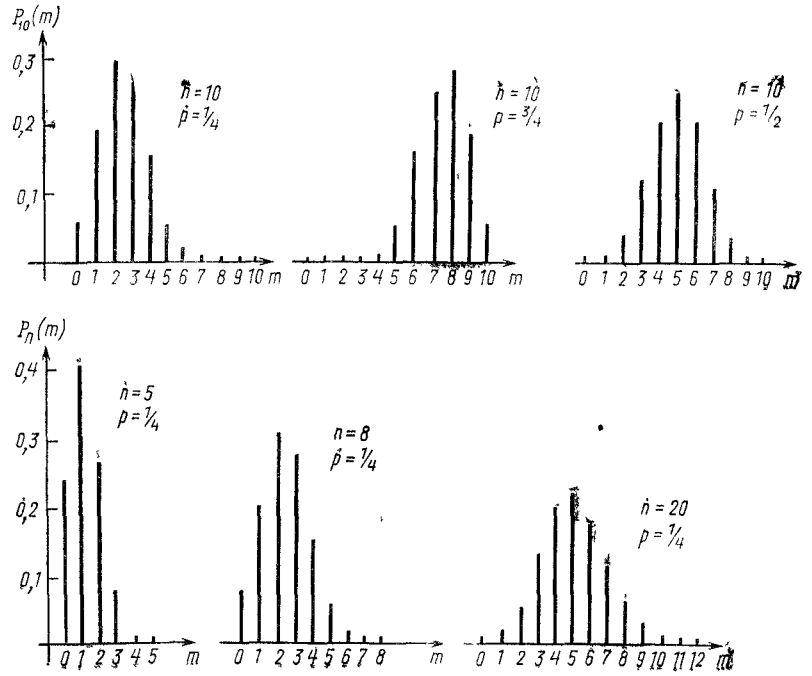


Рис. 5.1.
Графики биномиального распределения при различных значениях n и p

кретное, его графическим изображением служит совокупность ординат, называемая иногда пилообразным графиком. Условно соединяя концы ординат отрезками прямых и концы крайних ординат с точками $(-1, 0)$ и $(n + 1,0)$, получим замкнутый многоугольник, или *полигон биномиального распределения.* На рис. 5.2 изображен полигон, соответствующий данным табл. 5.2, а на рис. 5.3 — изме-

нение полигона с изменением p и n. Нетрудно убедиться на основании (5.3), что площадь S полигона равна 1. Действительно,

$$S = \frac{1}{2} P_n(0) + \frac{1}{2}[P_n(0) + P_n(1)] + \frac{1}{2}[P_n(1) + P_n(2)] +$$

$$\ldots + \frac{1}{2}[P_n(n-1) + P_n(n)] + \frac{1}{2} P_n(n) = \sum_{m=0}^{n} P_n(m) = 1.$$

Определим функцию *биномиального распределения*:

$$F_{\text{Б}}(m,\ n,\ p) = P\{X_n < m\} = \sum_{k=0}^{m-1} C_n^k p^k q^{n-k}. \qquad (5.4)$$

Функции (5.2) и (5.4) могут быть табулированы для разных значений p и параметров m и n. Заметим, что достаточно иметь

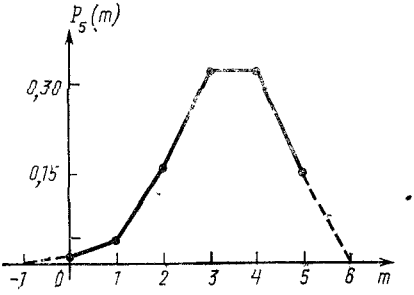


Рис. 5.2.
Полигон биномиального распределения при $n = 5$ и $p = 2/3$

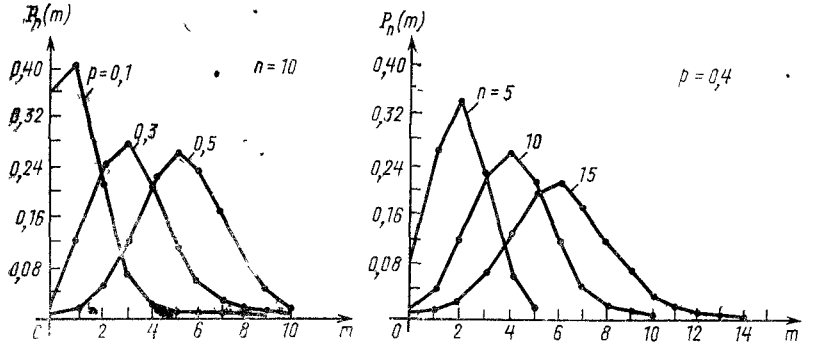


Рис. 5.3.
Изменение полигона биномиального распределения с изменением p и n

таблицу для $F_{\text{Б}}(m,\ n,\ p)$, так как значение функции (5.2) можно получить из нее простым вычитанием:

$$f_{\text{Б}}(m,\ n,\ p) = F_{\text{Б}}(m+1,\ n,\ p) - F_{\text{Б}}(m,\ n,\ p). \qquad (5.5)$$

С помощью функции распределения (5.4) можно, например, определить следующие вероятности:

$$P\{X_n \geqslant m\} = 1 - P\{X_n < m\} = 1 - F_{\text{Б}}(m,\ n,\ p); \qquad (5.6)$$

$$P\{m_1 \leqslant X < m_2\} = P\{X_n < m_2\} - P\{X_n < m_1\} = \\ = F_{\text{Б}}(m_2,\ n,\ p) - F_{\text{Б}}(m_1,\ n,\ p). \qquad (5.7)$$

Определим по данным предыдущего примера вероятности того, что число изделий высшего сорта в выборке окажется «не менее 3» и «в границах от 1 до 4 исключительно»

По данным таблицы $F_{\text{Б}}(m, n, p)$ получим:

$$P\{X_n \geqslant 3\} = 1 - F_{\text{Б}}(3;\ 5;\ 2/3) = 1 - 0{,}2099 = 0{,}7901,$$

$$P\{1 \leqslant X_n < 4\} = F_{\text{Б}}(4;\ 5;\ 2/3) - F_{\text{Б}}(1,\ 5;\ 2/3) = 0{,}5391 - 0{,}0041 = 0{,}5350.$$

5.3. Наивероятнейшая частота. Основные характеристики биномиального распределения

Как видно из табл. 5.2, члены биномиального ряда вначале возрастают, достигая наибольшего значения (0,3292) для некоторой частоты $X_n = m_0$ (в табл. $m_0 = 3$ или 4), и затем убывают до конца ряда. Это свойство биномиального распределения может быть доказано в общем виде. Рассмотрим отношение двух соседних членов ряда

$$\frac{P_n(m)}{P_n(m-1)} = \frac{C_n^m p^m q^{n-m}}{C_n^{m-1} p^{m-1} q^{n-m+1}} = \frac{n!\,(m-1)!\,(n-m+1)!}{m!\,(n-m)!\,n!}\,\frac{p}{q} = \\ = \left(\frac{n+1}{m} - 1\right)\frac{p}{q}.$$

При данных n и p это отношение убывает с ростом m. Поэтому если $\frac{P_n(1)}{P_n(0)} < 1$, т. е. $P_n(1) < P_n(0)$, то и для любого m будет $\frac{P_n(m)}{P_n(m-1)} < \frac{P_n(1)}{P_n(0)} < 1$, откуда следует, что члены ряда будут монотонно убывать, начиная с первого члена $P_n(0)$. Если же $\frac{P_n(1)}{P_n(0)} > 1$, то $P_n(1) > P_n(0)$, т. е. члены ряда вначале возрастают. Однако так как с ростом m отношение $\frac{P_n(m)}{P_n(m-1)}$ убывает, то рост будет продолжаться до некоторой частоты m_0, после которой вероятности биномиального распределения начнут убывать. Значение частоты m_0, которой соответствует наибольшая вероятность $P_n(m_0)$, называется *наивероятнейшей частотой*. Величину m_0 можно определить, используя выше найденное отношение двух последовательных членов ряда:

$$\frac{P_n(m_0)}{P_n(m_0-1)} = \left(\frac{n+1}{m_0} - 1\right)\frac{p}{q} \geqslant 1;$$

$$\frac{P_n(m_0+1)}{P_n(m_0)} = \left(\frac{n+1}{m_0+1} - 1\right)\frac{p}{q} \leqslant 1,$$

откуда после несложных преобразований получим

$$np + p - 1 \leqslant m_0 \leqslant np + p. \tag{5.8}$$

Так как m_0 по определению есть число целое, то получим

$$m_0 = \begin{cases} np + p \text{ или } np + p - 1, \text{ если } np + p - \text{целое число;} \\ [np + p], \text{ если } np + p - \text{дробное число}^*. \end{cases} \tag{5.9}$$

Разделив (5.8) почленно на n, получим аналогичные неравенства для наивероятнейшей относительной частоты:

$$p + \frac{p-1}{n} \leqslant \frac{m_0}{n} \leqslant p + \frac{p}{n}. \tag{5.10}$$

При больших n значения $\frac{p-1}{n}$ и $\frac{p}{n}$ будут значительно меньше, чем p, откуда следует приближенное выражение

$$\frac{m_0}{n} \approx p \text{ или } m_0 \approx np. \tag{5.11}$$

Последнее равенство оказывается точным, если np — целое число.

Соответствующую вероятность наивероятнейшей частоты найдем из формулы (5.2), подставив в нее значение m_0.

Определим наивероятнейшую частоту и соответствующую ей вероятность при следующих исходных данных: 1) $n = 10$, $p = 1/3$; 2) $n = 8$, $p = 1/3$; 3) $n = 80$, $p = 0{,}01$.

1) $np + p = 10 \cdot {}^1/_3 + {}^1/_3 = 3\frac{2}{3}$, следовательно, $m_0 = \left[3\frac{2}{3}\right] = 3$ и $P_{10}(m_0) = P_{10}(3) = C_{10}^3 \left(\frac{1}{3}\right)^3 \left(\frac{2}{7}\right)^7 = 0{,}26$,

2) $np + p = 8 \cdot {}^1/_3 + {}^1/_3 = 3$, следовательно, $m_0 = 3$ или $m_0 = 2$ и $P_8(3) = P_8(2) = C_8^3 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^5 = C_8^2 \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^6 = 0{,}35$;

3) $np + p = 80 \cdot 0{,}01 + 0{,}01 = 0{,}81$, откуда $m_0 = [0{,}81] = 0$ и $P_{80}(0) = C_{80}^0 0{,}01^0 0{,}99^{80} = 0{,}447$.

Найдем основные характеристики случайной величины X_n. Согласно определению математического ожидания получаем из табл. 5.1

$$M(X_n) = 0C_n^0 q^n + 1C_n^1 pq^{n-1} + 2p^2 q^{n-2} + \ldots + mC_n^m p^m q^{n-m} + \ldots + (n-1)C_n^{n-1} p^{n-1} q + nC_n^n p^n,$$

откуда

$$M(X_n) = np\{C_{n-1}^0 p^0 q^{n-1} + C_{n-1}^1 pq^{n-2} + \ldots + C_{n-1}^{m-1} p^{m-1} q^{n-m} + \ldots + C_{n-1}^{n-2} p^{n-2} q + C_{n-1}^{n-1} p^{n-1}\}.$$

* Символом $[a]$ обозначена целая часть числа a.

Но сумма, содержащаяся в фигурных скобках, есть бином $(q+p)^{n-1}=1$. Следовательно, получим окончательно

$$M(X_n)=np. \tag{5.12}$$

Разделив обе части на n, найдем выражение для математического ожидания относительной частоты:

$$M\left(\frac{X_n}{n}\right)=p. \tag{5.13}$$

Аналогичным образом* могут быть получены выражения для дисперсии:

$$D(X_n)=npq \text{ и } D\left(\frac{X_n}{n}\right)=\frac{pq}{n}. \tag{5.14}$$

Извлекая квадратный корень, найдем стандарт частоты и относительной частоты:

$$\sigma(X_n)=\sqrt{npq} \text{ и } \sigma\left(\frac{X_n}{n}\right)=\sqrt{\frac{pq}{n}}. \tag{5.15}$$

5.4. Гипергеометрическое распределение

Рассмотрим безвозвратную выборку объемом n из генеральной совокупности, состоящей из N элементов, среди которых признак A встречается N_1 раз. При этом предполагается, что $n<N$. Обозначим, как и ранее, через A событие, состоящее в том, что отобранный элемент обладает данным признаком. Так как отобранный элемент обратно не возвращается, то последовательный отбор n элементов равносилен одновременному их выбору из генеральной совокупности.

Определим, как и в модели Бернулли, вероятность появления события «$X_n=m$» при n испытаниях, используя способ непосредственного подсчета вероятностей. Всего различных выборок объемом n можно образовать C_N^n. Среди них выборок, у которых данный признак встречается m раз, будет $C_{N_1}^m \cdot C_{N-N_1}^{n-m}$. Следовательно, искомая вероятность

$$P\{X_n=m\}=\frac{C_{N_1}^m C_{N-N_1}^{n-m}}{C_N^n}=f(m,\ n,\ N,\ N_1), \tag{5.16}$$

где $m\leqslant N_1$ и $m\leqslant n$. Задаваясь теперь разными значениями частоты $m=0,\ 1,\ \ldots$ и подсчитав по формуле (5.16) соответствующие им вероятности, получим ряд вероятностей (табл. 5.3), получивший название *гипергеометрического распределения:*

* В дальнейшем (см § 6.9) будет показан иной вывод этих формул, опирающийся на теоремы о математическом ожидании и дисперсии.

Таблица 5.3

X_n	0	1	...	m	...	n	$\sum$
$p \cdot C_N^n$	$C_{N-N_1}^n$	$N_1 C_{N-N_1}^{n-1}$	...	$C_{N_1}^m C_{N-N_1}^{n-m}$	...	$C_{N_1}^n$	1

Можно показать, что для этого ряда также будет выполняться равенство, характеризующее основное свойство распределения:

$$\sum_{m=0}^{n} \frac{C_{N_1}^n C_{N-N_1}^{n-m}}{C_N^n} = 1.$$

Пусть в партии из 12 изделий находится 8 изделий I сорта. Производится случайная безвозвратная выборка 5 изделий Построим ряд вероятностей, характеризующих всевозможные результаты отбора

По условию имеем: $N = 12$, $N_1 = 8$, $N - N_1 = 4$, $n = 5$. Обозначив через X_n число изделий I сорта, получим из формулы (5.16)

$$P\{X_n = m\} = \frac{C_8^m C_4^{5-m}}{C_{12}^5} = \frac{C_8^m C_4^{5-m}}{210}.$$

Задаваясь значениями $m = 0, 1, \ldots, 5$, рассчитаем ряд вероятностей, указанных в табл. 5.4:

Таблица 5.4

X_n	0	1	2	3	4	5
$P_5(m)$	0	0,0101	0,1414	0,4242	0,3535	0,0707

Первый член $P(X_n = 0) = 0$ получен из того очевидного факта, что из $N - N_1 = 4$ изделий не I сорта выбрать 5 таких изделий невозможно.

Для того чтобы формула (5.16) автоматически давала необходимый результат при любом m, условно полагают, что $C_n^m = 0$ при $m > n$.

Можно показать, что математическое ожидание и дисперсия относительной частоты для гипергеометрического распределения определяются следующими равенствами:

$$\mathrm{M}\left(\frac{X_n}{n}\right) = p; \quad \mathrm{D}\left(\frac{X_n}{n}\right) = \frac{pq}{n} \cdot \frac{N-n}{N-1}, \qquad (5.17)$$

где $p = N_1/N$.

Очевидно, что гипергеометрическое распределение асимптотически приближается к биномиальному при неограниченном увеличении N. Это следует и из приведенных формул. Поэтому мы и говорили ранее о возвратном отборе как об отборе из бесконечной генеральной совокупности. В свою очередь гипергеометрическое распределение представляет собой статистическую модель выборки из ограниченной генеральной совокупности, или безвоз-

вратной выборки. Практически указанные два распределения малоразличимы при относительно малом значении n в сравнении с N. Учитывая, что расчеты по формуле (5.16) оказываются значительно более трудоемкими, чем по формуле (5.2), мы в дальнейшем неоднократно воспользуемся этим замечанием.

5.5. Другие распределения, основанные на схеме Бернулли

Пусть рассматривается схема независимых испытаний Бернулли с постоянной вероятностью $P(A)=p$. Определим вероятность того, что A появится на m-м испытании, т. е. после $m-1$ раз появления $\bar{A}$. Обозначив эту вероятность через $\Gamma(m, p)$, получим

$$\Gamma(m, p)=(1-p)^{m-1}p. \tag{5.18}$$

Меняя значения m от 0 до n, получим *геометрическое распределение.*

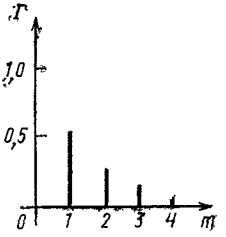


Рис. 5.4.
График геометрического распределения

Аналогично можно получить вероятность первого непоявления A после $m-1$ раз его наступления:

$$\Gamma'(m, p)=p^{m-1}(1-p). \tag{5.19}$$

Эти распределения зависят только от одного параметра p. На рис. 5.4 построен график геометрического распределения при $p=1/2$.

Обобщением геометрического распределения является *распределение Паскаля*, связанное с определением вероятности того, что до появления k раз события A оно не появится $m-1$ раз:

$$\Pi(m, p, k)=C_{k+m-2}^{m-1}(1-p)^{m-1}p^k. \tag{5.20}$$

Как видно, геометрическое распределение является частным случаем распределения Паскаля при $k=1$.

Наконец, обобщим модель испытаний Бернулли, предполагая, что в каждом испытании могут появиться не два альтернативных исхода (A или $\bar{A}$), а s исходов ($s>2$): A_1 — с вероятностью $P(A_1)=p_1$; A_2 — с вероятностью $P(A_2)=p_2$ и т. д., A_s — с вероятностью $P(A_s)=p_s$. Очевидно, что $p_1+\ldots+p_s=1$.

Обозначим вероятность того, что частоты появления этих событий при n испытаниях будут соответственно $m_1, m_2, \ldots, m_s$,

через $P_n(m_1, m_2, \ldots, m_s)$. Тогда

$$P_n(m_1, m_2, \ldots, m_s) = \frac{n!}{m_1!\, m_2! \ldots m_s!}\, p_1^{m_1} p_2^{m_2} \ldots p_s^{m_s}. \qquad (5.21)$$

Распределение вероятностей, задаваемых функцией (5.21) при любых значениях m_1, m_2, ..., m_s, где $m_1 + m_2 + \ldots + m_s = n$, называется *полиномиальным распределением*. Оно представляет собой пример многомерного распределения (см. гл. 6).

Глава 6

Многомерные случайные величины

6.1. Определение и закон распределения многомерной случайной величины

Очень часто результат испытания характеризуется не одной случайной величиной, а некоторой системой случайных величин X_1, X_2, ..., X_j, ..., X_s, которую будем называть *многомерной случайной величиной* или *случайным вектором* $\mathbf{Z} = (X_1, \ldots, X_j, \ldots, X_s)$; X_j в этом случае составляющие этого вектора.

Например, поступающая после обработки деталь характеризуется несколькими размерами; погода в данном месте и в определенное время характеризуется температурой, влажностью, скоростью ветра, давлением и т. д. В дальнейшем значения составляющих случайного вектора $\mathbf{Z}$ будем обозначать строчными буквами с двумя индексами. Так x_{j1}, ..., x_{jt}, ..., x_{jn} есть n значений составляющей X_j.

Многомерные случайные величины служат для построения моделей статистических многомерных распределений.

Пусть дано пространство $\Omega = \{e\}$ элементарных событий. Многомерную случайную величину, или случайный вектор, определим как векторную функцию $z(e)$, ставящую в соответствие каждому элементарному исходу e упорядоченную совокупность s чисел x_{je} $(j = 1, \ldots, s)$ или вектор $\mathbf{z} = (x_{1e}, \ldots, x_{je}, \ldots, x_{se})$, называемый *реализацией* случайного вектора $\mathbf{Z}$

Если Ω — дискретное множество, то $\mathbf{Z}$ будет дискретным случайным вектором. Его закон распределения при конечном множестве всех возможных значений может быть описан в форме таблицы *многомерного распределения*, содержащей всевозможные сочетания значений каждой из одномерных случайных величин, входящих в данную систему (составляющих случайного вектора), и соответствующие этим сочетаниям вероятности. Построение многомерных таблиц распределения необычайно сложно. Так, если рассматривается система из пяти случайных величин (пятимерный случайный вектор) и каждая из одномерных составляющих может принимать по три значения, то таблица должна будет содержать $3^5 = 243$ реализации и соответствующие им вероят-

ности. Если Ω — несчетное множество, то определенный на нем случайный вектор также может иметь несчетное множество реализаций, т. е. соответствующие им точки $(x_{1e}, \ldots, x_{je}, \ldots, x_{se})$ могут заполнять сплошь некоторую область s-мерного пространства.

Таблица 6.1

X \ Y	y_1	$\ldots$	y_k	$\ldots$	y_n	$\sum$
x_1	p_{11}	$\ldots$	p_{1k}	$\ldots$	p_{1n}	$p_{1\cdot}$
$\cdot$	$\cdot$		$\cdot$		$\cdot$	$\cdot$
$\cdot$	$\cdot$		$\cdot$		$\cdot$	$\cdot$
$\cdot$	$\cdot$		$\cdot$		$\cdot$	$\cdot$
x_i	p_{i1}	$\ldots$	p_{ik}	$\ldots$	p_{in}	$p_{i\cdot}$
$\cdot$	$\cdot$		$\cdot$		$\cdot$	$\cdot$
$\cdot$	$\cdot$		$\cdot$		$\cdot$	$\cdot$
$\cdot$	$\cdot$		$\cdot$		$\cdot$	$\cdot$
x_m	p_{m1}	$\ldots$	p_{mk}	$\ldots$	p_{mn}	$p_{m\cdot}$
$\sum$	$p_{\cdot 1}$	$\ldots$	$p_{\cdot k}$	$\ldots$	$p_{\cdot n}$	1

Пусть рассматривается *двумерная* дискретная величина $\mathbf{Z} = (X, Y)$. *Ее двумерное распределение* можно представить в виде шахматной табл. 6.1, содержащей в соответствующих клетках вероятности произведения событий $P\{(X = x_i) \cdot (Y = y_k)\} = p_{ik}$. Итоговые столбец или строка такой таблицы задают распределения одномерных составляющих $\{x_i;\ p_{i\cdot}\}$ или $\{y_k;\ p_{\cdot k}\}$.

Действительно, распределение одномерной случайной величины X можно получить, вычислив по данным табл. 6.1 вероятности $P(X = x_i)$, которые обозначим через $p_{i\cdot}$. Но случайное событие $(X = x_i)$ можно представить как сумму попарно несовместных событий:

$$(X = x_i) = (X = x_i)(Y = y_1) + \ldots + (X = x_i)(Y = y_k) + \\ + \ldots + (X = x_i)(Y = y_n),$$

откуда по формуле полной вероятности получим

$$P(X = x_i) = p_{i\cdot} = P\{(X = x_i)(Y = y_1)\} + \ldots + \\ + P\{(X = x_i)(Y = y_k)\} + \ldots + P\{(X = x_i)(Y = y_n)\} = \sum_{k=1}^{n} p_{ik}. \tag{6.1}$$

Аналогично можно построить распределение одномерной случайной величины Y, вычислив вероятности

$$p_{\cdot k} = \sum_{i=1}^{m} p_{ik}. \tag{6.1'}$$

В табл 6 2 в виде примера приведены данные о возможных сочетаниях отклонений длины валика (X) и диаметра (Y) от номинальных размеров и соответствующие им вероятности

Таблица 6.2

X \ Y	-1	0	1	$p_{i\cdot}$
-2	0,15	0,35	0,05	0,55
3	0,10	0,25	0,10	0,45
$p_{\cdot k}$	0,25	0,60	0,15	1,0

Рассмотрим распределение одной из случайных величин (например, X) при определенном значении другой ($Y = y_k$). Для этого необходимо вычис-

лить условные вероятности событий $(X = x_1)$, ..., $(X = x_i)$, ..., $(X = x_m)$ при наступлении события $(Y = y_k)$. Обозначив эти условные вероятности через $p_k(x_i)$, получим из определения условных вероятностей

$$p_k(x_i) = \frac{P\{(X = x_i)(Y = y_k)\}}{P(Y = y_k)} = \frac{p_{ik}}{p_{\cdot k}}. \qquad (6.2)$$

Таким образом, разделив все данные k-го столбца табл. 6.1 на итоговую вероятность по этому столбцу $p_{\cdot k}$, получим *условное распределение* X при $Y = y_k$. По аналогичной формуле

$$p_i(y_k) = \frac{p_{ik}}{p_{i\cdot}} \qquad (6.2')$$

определим *условные вероятности значений* Y при данном значении $X = x_i$. Так, по данным табл. 6.2, разделив все числа первой строки на 0,55, получим условное распределение Y при $X = x_1 = -2$, приведенное в табл. 6.3.

Таблица 6.3

Y	-1	0	1	Σ
$p_1(y_k)$	0,273	0,636	0,091	1

6.2. Функция распределения многомерной случайной величины

Для описания закона распределения многомерных случайных величин обобщим рассмотренное в предыдущей главе понятие функции распределения. *Функцией распределения случайного вектора* $\mathbf{Z}$ является функция $F(z)$ векторного аргумента z, равная вероятности одновременного выполнения s неравенств, т. е.

$$F(z) = P\{(X_1 < x_1)(X_2 < x_2) \ldots (X_j < x_j) \ldots (X_s < x_s)\}. \qquad (6.3)$$

В фигурных скобках содержится произведение s событий $(X_j < x_j)$, которые в общем случае могут быть зависимы, и потому вероятность их произведения не может быть выражена через произведение вероятностей этих событий.

Функция распределения является универсальным средством описания закона распределения любой случайной величины, в том числе и дискретной. При $s = 1$ получаем рассмотренное ранее определение функции распределения одномерной случайной величины.

Графическое построение функции $F(x, y)$ оказывается сложным. Отметим лишь, что $F(x, y)$ будет изображаться некоторой ступенчатой поверхностью, «ступени» которой параллельны плоскости XOY (рис. 6.1). Переход со «ступени» на «ступень» соответствует разрывам (конечным скачкам) функции $F(x, y)$. Если X и Y могут принимать любые значения в некоторой области, то $F(x, y)$ будет изображаться криволинейной поверхностью.

Ограничимся в дальнейшем рассмотрением двумерной случайной величины $\mathbf{Z} = (X, Y)$, которая геометрически может интер-

претироваться как случайная точка (X, Y) в плоскости XOY. Из определения (6.3) получаем

$$F(z) = F(x, y) = P\{(X < x)(Y < y)\}, \qquad (6.3')$$

т. е. $F(x, y)$ равна вероятности попадания случайной точки (X, Y) в заштрихованную на рис. 6.2 область, лежащую левее и ниже точки $A(x, y)$.

Соответственно вероятность попадания в бесконечную прямоугольную полосу, заштрихованную на рис. 6.3, будет равна раз-

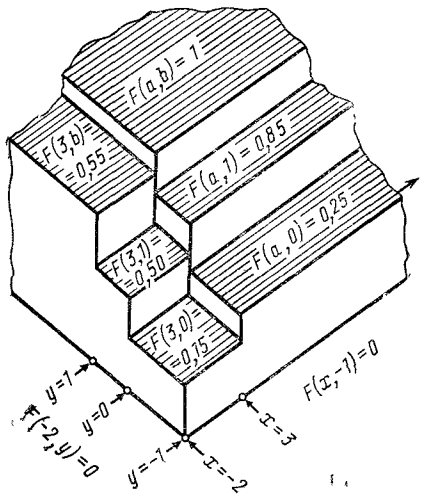


Рис. 6.1.
График функции распределения двумерной дискретной случайной величины, заданной табл. 6.2

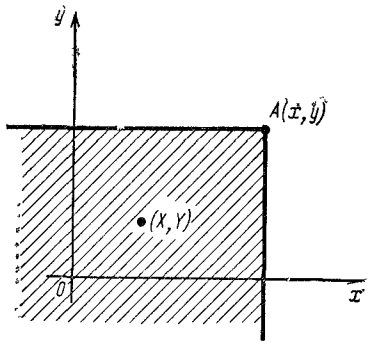


Рис. 6.2.
К определению функции распределения двумерной случайной величины

ности между вероятностями попадания в область, лежащую левее и ниже точки B, и в область, лежащую левее и ниже точки A, т. е.

$$P\{(x_1 \leqslant X < x_2)(Y < y)\} = P\{(X < x_2)(Y < y)\} - \\ - P\{(X < x_1)(Y < y)\} = F(x_2, y) - F(x_1, y). \qquad (6.4)$$

Наконец, вероятность попадания случайной точки в прямоугольник $ABCД$ (рис. 6.4) равна разности между вероятностями

попадания в полосу под AB и в полосу под $CД$, т. е.

$$P\{(x_1 \leqslant X < x_2)(y_1 \leqslant Y < y_2)\} = P\{(x_1 \leqslant X < x_2) \times (Y < y_2)\} - P\{(x_1 \leqslant X < x_2)(Y < y_1)\} = [F(x_2, y_2) - F(x_1, y_2)] - [F(x_2, y_1) - F(x_1, y_1)]. \quad (6.5)$$

Функция распределения обладает свойствами, обобщающими аналогичные свойства функции распределения одномерной случайной величины (см. § 4.2).

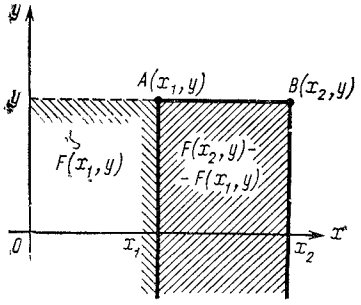


Рис. 6.3.
К определению вероятности попадания в полосу

1. Значения $F(\mathbf{z})$ заключены в промежутке от 0 до 1

$$0 \leqslant F(x_1, \ldots, x_j, \ldots, x_s) \leqslant 1,$$

что вытекает из определения $F(\mathbf{z})$ как вероятности

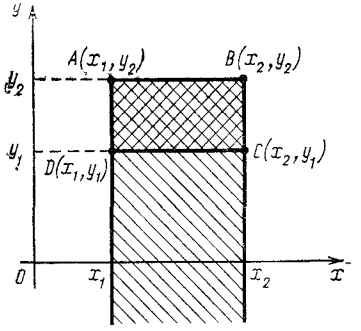


Рис. 6.4.
К определению вероятности попадания в прямоугольник

2. $F(\mathbf{z})$ — монотонно возрастающая функция по любому из аргументов при фиксированных значениях остальных аргументов, т. е.

$$F(x_1, \ldots, x_j', \ldots, x_s) \leqslant F(x_1, \ldots, x_j'', \ldots, x_s)$$

для любых $x_j' < x_j''$.
Это свойство легко получить в частном случае из (6.4), записав это равенство в виде

$$F(x'', y) = F(x', y) + P\{(x' \leqslant X < x'')(Y < y)\} \geqslant F(x', y) \text{ при } x' < x'',$$

3. При неограниченном увеличении всех аргументов

$$\lim_{\substack{x_1\to\infty \\ \cdots\cdots \\ x_s\to\infty}} F(x_1, \ldots, x_j, \ldots, x_s) = 1.$$

Свойство вытекает непосредственно из определения (6.3), так как события $(X_1 < \infty)$, ..., $(X_j < \infty)$, ..., $(X_s < \infty)$ достоверны.

4. $\lim\limits_{x_j\to-\infty} F(x_1, \ldots, x_j, \ldots, x_s) = 0$ при любых значениях остальных аргументов. Действительно, событие $(X_j < -\infty)$ — невозможное, и потому его совмещение с любыми другими событиями также будет невозможным событием.

6.3. Плотность распределения

Если функция $F(z)$ допускает s-кратное дифференцирование по всем переменным, то можно ввести понятие плотности вероятности, которое обобщает аналогичное понятие для одномерной случайной величины. Ограничимся рассмотрением двумерной случайной величины. Найдем вероятность попадания случайной точки (X, Y) в прямоугольник со сторонами Δx и Δy, показанный на рис. 6.4. Используя формулу (6.5), положив в ней $x_1 = x$, $x_2 = x + \Delta x$, $y_1 = y$ и $y_2 = y + \Delta y$, получим

$$P\{(x \leqslant X < x + \Delta x)(y \leqslant Y < y + \Delta y)\} = [F(x + \Delta x, y + \Delta y) - F(x, y + \Delta y)] - [F(x + \Delta x, y) - F(x, y)]. \tag{6.6}$$

Разделив обе части равенства на площадь прямоугольника $\Delta x \cdot \Delta y$, получим среднюю плотность вероятности в данном прямоугольнике:

$$f_{\text{ср}} = \frac{P\{(x \leqslant X < x + \Delta x)(y \leqslant Y < y + \Delta y)\}}{\Delta x\, \Delta y}.$$

Переходя к пределу при $\Delta x \to 0$ и $\Delta y \to 0$ (предполагая, что он существует), получим плотность вероятности $f(x, y)$ в точке (x, y):

$$f(x, y) = \lim_{\substack{\Delta x\to 0 \\ \Delta y\to 0}} \frac{P\{(x \leqslant X < x + \Delta x)(y \leqslant Y < y + \Delta y)\}}{\Delta x\, \Delta y}.$$

Подставив в последнее выражение (6.6) и перейдя к пределу последовательно, вначале по $\Delta x \to 0$, а затем по $\Delta y \to 0$, получим

$$f(x, y) = \lim_{\Delta y\to 0} \frac{1}{\Delta y} \left\{ \lim_{\Delta x\to 0} \frac{F(x + \Delta x, y + \Delta y) - F(x, y + \Delta y)}{\Delta x} - \lim_{\Delta x\to 0} \frac{F(x + \Delta x, y) - F(x, y)}{\Delta x} \right\}.$$

Но каждый из пределов, указанных в фигурных скобках, есть соответствующие частные производные $F'_x(x, y + \Delta y)$ и $F'_x(x, y)$

Следовательно,

$$f(x,\ y)=\lim_{\Delta y\to 0}\frac{F'_x(x,\ y+\Delta y)-F'_x(x,\ y)}{\Delta y}=F''_{xy}(x,\ y). \qquad (6.7)$$

Таким образом, приходим к следующему утверждению.

Т е о р е м а. Если в данной точке существует плотность вероятности многомерной случайной величины, то она численно равна значению в данной точке смешанной частной производной функции распределения по всем аргументам.

Мы умышленно не указали размерность случайной величины, так как теорема оказывается справедливой при любом числе измерений.

Многомерная случайная величина, для которой существует плотность вероятности $f(x_1, \ldots, x_j, \ldots, x_s)$ в каждой точке области ее возможных значений (за исключением, быть может, счетного множества точек), называется *непрерывной.*

Из равенства (6.7), применяя двукратное интегрирование, получим выражение функции распределения через заданную плотность:

$$F(x,\ y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,\ v)\,\mathrm{d}u\,\mathrm{d}v. \qquad (6.8)$$

При числе измерений $s>2$ формулы (6.7) и (6.8) заменяются на аналогичные:

$$f(x_1, \ldots, x_j, \ldots, x_s)=\frac{\partial^s F(x_1, \ldots, x_s)}{\partial x_1 \ldots \partial x_j \ldots \partial x_s};\ F(x_1, \ldots, x_s)=$$

$$=\int_{-\infty}^{x_1}\ldots\int_{-\infty}^{x_s} f(u,\ v, \ldots)\,\mathrm{d}u\,\mathrm{d}v.$$

Эти формулы послужили основанием для введения терминов *дифференциальная* (для f) и *интегральная* (для F) *функции распределения.*

Величина $f(x,\ y)\,\mathrm{d}x\mathrm{d}y$ — *элемент вероятности двумерной случайной величины* — приближенно выражает собой вероятность попадания случайной точки в квадрат со сторонами $\mathrm{d}x$ и $\mathrm{d}y$. Укажем основные свойства дифференциальной функции распределения:

1. $f(x_1, \ldots, x_j, \ldots, x_s)\geqslant 0;$

2. $\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(x_1, \ldots, x_j, \ldots, x_s)\,\mathrm{d}x_1 \ldots \mathrm{d}x_j \ldots \mathrm{d}x_s=1.$

Первое свойство вытекает из определения плотности вероятности как предела отношения двух неотрицательных величин. Второе является следствием соотношения между $F(x_1, \ldots, x_s)$ и $f(x_1, \ldots, x_s)$ и свойства 3 функции распределения. Оно часто используется для так называемого нормирования функции $f(x_1, \ldots, x_s)$ с помощью постоянного множителя.

Для двумерной случайной величины плотность вероятности изображается графически некоторой поверхностью распределения (рис. 6.5).

Найдем плотность вероятности двумерной величины (X, Y), распределенной равномерно в прямоугольнике $a \leqslant x \leqslant b$ и $c \leqslant y \leqslant d$.

По условию $f(x, y) = h = \text{const}$ для любой точки внутри прямоугольника и $f(x, y) = 0$ — для любой точки вне этого прямо-

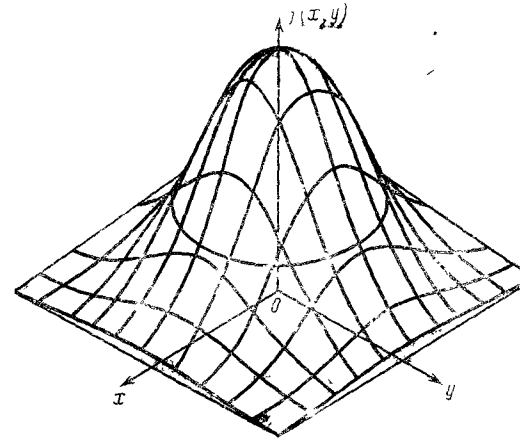


Рис. 6.5.
Плотность распределения вероятностей двумерной случайной величины

угольника. Величину h можно найти из второго свойства плотности вероятности, заменив бесконечные пределы интегрирования границами прямоугольника:

$$\int_a^b \int_c^d h\,dx\,dy = h(b-a)(d-c) = 1, \text{ откуда } h = 1/(b-a)(d-c).$$

Предположим, что поставка сырья и поступление требования на него из цеха могут произойти в любой день месяца (30 дней) с равной вероятностью. Определить вероятность следующих событий: 1) поставка сырья произойдет до поступления требования на него (событие A); 2) поставка произойдет после поступления требования (событие B)

Обозначим через X время поставки и через Y — время поступления требования. Согласно условию $0 \leqslant X \leqslant 30$; $0 \leqslant Y \leqslant 30$ и $f(x, y) = \frac{1}{h} = \frac{1}{30^2}$. Тогда

1) $P(A) = P\{(0 \leqslant X \leqslant y)\ (0 \leqslant Y \leqslant 30)\} = h \int_0^{30} \int_a^y dx\,dy =$

$$= h \int_0^{30} y\,dy = \frac{1}{30^2} \cdot \frac{30^2}{2} = 0{,}5;$$

2) $P(B) = P\{(y \leqslant X \leqslant 30)\ (0 \leqslant Y \leqslant 30)\} = h \int_0^{30} \int_y^{30} dx\,dy =$

$$= h \int_0^{30} (30 - y)\,dy = \frac{1}{30^2} \cdot \frac{30^2}{2} = 0{,}5.$$

6.4. Распределения составляющих случайного вектора

Каждая составляющая случайного вектора есть одномерная случайная величина, функцию распределения которой (эти распределения будем называть в дальнейшем безусловными) можно получить из (6.3), если все неравенства, кроме одного, заменить достоверными неравенствами $X_j < \infty$. Так, если дана функция распределения $F(x, y)$ двумерной случайной величины, то функции распределения составляющих $F_1(x) = P(X < x)$ и $F_2(y) = P(Y < y)$ можно получить из соотношений:

$$F_1(x) = P\{(X < x)(Y < \infty)\} = F(x, \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, v)\, du\, dv; \quad (6.9)$$

$$F_2(y) = P\{(X < \infty)(Y < y)\} = F(\infty, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y} f(u, v)\, du\, dv. \quad (6.9')$$

Геометрически величины $F_1(x)$ и $F_2(y)$ равны соответственно вероятностям попадания случайной точки в полуплоскости $X < x$ и $Y < y$. Для дискретной случайной величины эти одномерные распределения задаются итоговыми столбцами $(p_{i\cdot})$ и строкой $(p_{\cdot k})$ табл. 6.1.

Для многомерных $(s > 2)$ распределений аналогично вводятся понятия распределений одномерных составляющих, задаваемых функциями $F_1(x_1) = F(x_1, \infty, \ldots, \infty)$; $f_1(x_1) = \frac{dF_1(x_1)}{dx_1}$ и т. д.

Дифференцируя (6.9) и (6.9′) соответственно по x и y, получим дифференциальные функции распределений:

$$f_1(x) = \frac{dF_1(x)}{dx} = \int_{-\infty}^{\infty} f(x, v)\, dv; \quad f_2(y) = \frac{dF_2(y)}{dy} = \int_{-\infty}^{\infty} f(u, y)\, du. \quad (6.10)$$

Найдем функции распределения двумерной случайной величины, равномерно распределенной в прямоугольнике $a \leqslant x \leqslant b$ и $c \leqslant y \leqslant d$.

Так как $f(x, y) = h$ внутри прямоугольника и $f(x, y) = 0$ вне его, то в формулах (6.9) и (6.9′) бесконечные пределы интегрирования можно заменить на конечные. В результате получим

$$F_1(x) = \int_{a}^{x} \int_{c}^{d} h\, dx\, dy = h(x - a)(d - c) = \frac{x - a}{b - a}; \quad F_2(y) = \frac{y - c}{d - c}.$$

Полученные функции характеризуют закон равномерной плотности для составляющей X на интервале (a, b) и составляющей Y на интервале (c, d). Соответствующие дифференциальные функции распределения будут:

$$f_1(x) = \frac{dF_1(x)}{dx} = \frac{1}{b - a}; \quad f_2(y) = \frac{dF_2(y)}{dy} = \frac{1}{d - c}.$$

Рассмотрим теперь условные распределения составляющей X при $Y = y_k$ и составляющей Y при $X = x_i$ (в дискретном случае они характеризовались соответствующими столбцом и строкой

табл. 6.1). Используя выражение (6.2) для вероятностей условного распределения X при $Y = y_k$ и заменив вероятности на элементы вероятностей, получим $f_y(x)\,dx = \frac{f(x,\ y\ dx\,dy}{f_2(y)\,dy}$, откуда после сокращения на dx и dy найдем

$$f_y(x) = \frac{f(x,\ y)}{f_2(y)}. \tag{6.11}$$

Аналогично из (6.2′) получим

$$f_x(y) = \frac{f(x\ \ y)}{f_1(x)}. \tag{6.11′}$$

Эти равенства и принимаются в качестве определения *плотности вероятности условных распределений*. Равенства (6.11) и (6.11′) можно представить в виде

$$f(x,\ y) = f_1(x)\,f_x(y) = f_2(y)\,f_y(x), \tag{6.12}$$

т. е. плотность распределения вероятностей двумерной случайной величины равна произведению плотности вероятности одной составляющей на плотность условного распределения вероятностей другой. Последнее равенство аналогично выражению для вероятности произведения двух событий, поэтому его называют также *теоремой умножения законов распределения*.

Введенные выше понятия условных распределений можно распространить на многомерные ($s > 2$) случайные величины как распределения одной из составляющих при определенных значениях, принимаемых остальными. Наконец, можно ввести более общее понятие условного многомерного распределения r составляющих при определенных значениях остальных $s - r$ составляющих.

6.5. Числовые характеристики многомерной случайной величины

Мы уже убедились в том, насколько сложным оказывается изучение многомерных распределений. Тем более важным представляется описание основных свойств такого распределения с помощью числовых характеристик, которые вводятся аналогично характеристикам одномерной случайной величины. Пусть, например, рассматривается система двух случайных величин $(X,\ Y)$. Математическое ожидание и дисперсия составляющих определяются по безусловным распределениям так же, как и для одномерной случайной величины (табл. 6.4), и называются *безусловными математическими ожиданиями* и *безусловными дисперсиями*. При этом соответствующие формулы могут строиться на базе распределений составляющих, а могут — на базе исходного двумерного распределения.

Точка $(\bar{X},\bar{Y})$ есть центр группирования двумерного распределения. Дисперсии $D(X)$ и $D(Y)$ характеризуют рассеяние значений случайных величин X и Y относительно их математических

Таблица 6.4

Характеристика	Дискретные случайные величины	Непрерывные случайные величины
$M(X)=\overline{X}$	$\sum_i x_i p_{i\cdot} = \sum_i \sum_k x_i p_{ik}$	$\int_{-\infty}^{\infty} x f_1(x)\,dx = \iint_{-\infty}^{\infty} x f(x,y)\,dx\,dy$
$M(Y)=\overline{Y}$	$\sum_k y_k p_{\cdot k} = \sum_k \sum_i y_k p_{ik}$	$\int_{-\infty}^{\infty} y f_2(y)\,dy = \iint_{-\infty}^{\infty} y f(x,y)\,dx\,dy$
$D(X)=M(X-\overline{X})^2$	$\sum_i (x_i-\overline{X})^2 p_{i\cdot} = \sum_i \sum_k (x_i-\overline{X})^2 p_{ik}$	$\int_{-\infty}^{\infty} (x-\overline{X})^2 f_1(x)\,dx = \iint_{-\infty}^{\infty} (x-\overline{X})^2 f(x,y)\,dx\,dy$
$D(Y)=M(Y-\overline{Y})$	$\sum_k (y_k-\overline{Y})^2 p_{\cdot k} = \sum_k \sum_i (y_k-\overline{Y})^2 p_{ik}$	$\int_{-\infty}^{\infty} (y-\overline{Y})^2 f_2(y)\,dy = \iint_{-\infty}^{\infty} (y-\overline{Y})^2 f(x,y)\,dx\,dy$

Таблица 6.5

Дискретные случайные величины	Непрерывные случайные величины
$\overline{X}_k = \sum_i x_i p_k(x_i) = \sum_i x_i p_{ik}/p_{\cdot k}$	$\overline{X}_y = \int_{-\infty}^{\infty} x f_y(x)\,dx = \int_{-\infty}^{\infty} x \frac{f(x,y)}{f_2(y)}\,dx$
$\overline{Y}_i = \sum_k y_k p_i(y_k) = \sum_k y_k p_{ik}/p_{i\cdot}$	$\overline{Y}_x = \int_{-\infty}^{\infty} y f_x(y)\,dy = \int_{-\infty}^{\infty} y \frac{f(x,y)}{f_1(x)}\,dy$
$D_k(X) = \sum_i (x_i-\overline{X}_k)^2 p_k(x_i) = \sum_i (x_i-\overline{X}_k)^2 p_{ik}/p_{\cdot k}$	$D_y(X) = \int_{-\infty}^{\infty} (x-\overline{X}_y) f_y(x)\,dx$
$D_i(Y) = \sum_k (y_k-\overline{Y}_i)^2 p_i(y_k) = \sum_k (y_k-\overline{Y}_i)^2 p_{ik}/p_{i\cdot}$	$D_x(Y) = \int_{-\infty}^{\infty} (y-\overline{Y}_x) f_x(y)\,dy$

ожиданий. Свойства этих характеристик и техника их вычисления не отличаются от свойств и способов расчета характеристик одномерной случайной величины.

При рассмотрении дискретной случайной величины нетрудно убедиться в аналогии формул для соответствующих статистических характеристик двумерного статистического распределения. Продолжив эту аналогию, можно теперь ввести характеристики условных распределений: $\bar{X}_k$ — *условное математическое ожидание* и $D_k(X)$ — *условную дисперсию* X при $Y = y_k$, а также $\bar{Y}_i$ и $D_i(Y)$ при $X = x_i$. Аналогичные характеристики можно построить и для непрерывной двумерной случайной величины. Соответствующие формулы приведены в табл. 6.5.

Условные математические ожидания и условные дисперсии могут, в свою очередь, рассматриваться как случайные величины, значения которых зависят от условного распределения, а следовательно, от значений, принимаемых другой величиной. Так, для дискретной случайной величины $\bar{X}_k$ и $D_k(X)$, рассчитанные при $Y = y_k$ (т. е. по условному распределению при $Y = y_k$), будут характеризоваться вероятностью $P(Y = y_k) = p_{\cdot k}$. В результате получаем для них распределения, приведенные в табл. 6.6.

Таблица 6.6

$\bar{X}_k$	$\bar{X}_1$	...	$\bar{X}_n$
$D_k(X)$	$D_1(X)$	...	$D_n(X)$
$p_{\cdot k}$	$p_{\cdot 1}$	...	$p_{\cdot n}$

Таблица 6.7

$\bar{Y}_i$	$\bar{Y}_1$	...	$\bar{Y}_m$
$D_i(Y)$	$D_1(Y)$	...	$D_m(Y)$
$p_{i\cdot}$	$p_{1\cdot}$	...	$p_{m\cdot}$

Аналогичные распределения получаем для $\bar{Y}_i$ и $D_i(Y)$ (табл. 6.7). Для непрерывной двумерной случайной величины аналогичные распределения величин $\bar{X}_y$, $D_y(X)$ и $\bar{Y}_x$, $D_x(Y)$ будут задаваться соответственно плотностями $f_2(y)$ и $f_1(x)$. Эти распределения, в свою очередь, могут характеризоваться математическими ожиданиями и дисперсиями. Всего получим таким образом восемь характеристик. Однако из них две характеристики ($M(\bar{X}_y)$ и $M(\bar{Y}_x)$) не несут новой информации, так как можно доказать, что они совпадают с общими математическими ожиданиями составляющих, т. е.

$$M(\bar{X}_y) = M(X) = \bar{X} \text{ и } M(\bar{Y}_x) = M(Y) = \bar{Y}. \qquad (6.13)$$

Еще две характеристики — дисперсии условных дисперсий — практически не используются. Таким образом, остаются четыре характеристики, расчетные формулы для которых приведены в табл. 6.8.

Часто, употребляя статистическую терминологию, характеристики условных распределений ($\bar{X}_y$, $\bar{Y}_x$, $D_y(X)$, $D_x(Y)$) называют

Таблица 68

Дискретные случайные величины	Непрерывные случайные величины
$D(\overline{X}_k)=\sum_k(\overline{X}_k-\overline{X})^2 p_{\cdot k}$	$D(\overline{X}_y)=\int_{-\infty}^{\infty}(\overline{X}_y-\overline{X})^2 f_2(y)\,dy$
$D(\overline{Y}_i)=\sum_i(\overline{Y}_i-\overline{Y})^2 p_{i\cdot}$	$D(\overline{Y}_x)=\int_{-\infty}^{\infty}(\overline{Y}_x-\overline{Y})^2 f_1(x)\,dx$
$M(D_k(X))=\sum_k D_k(X)\,p_{\cdot k}$	$M(D_y(X))=\int_{-\infty}^{\infty} D_y(x)\,f_2(y)\,dy$
$M(D_i(Y))=\sum_i D_i(Y)\,p_{i\cdot}$	$M(D_x(Y))=\int_{-\infty}^{\infty} D_x(Y)\,f_1(x)\,dx$

групповыми. Тогда математические ожидания и дисперсии групповых средних естественно назвать *межгрупповыми средними* и *межгрупповыми дисперсиями.* Между групповыми, межгрупповыми и общими дисперсиями существуют важные соотношения, которые составляют содержание следующей теоремы.

Теорема 1. Общая дисперсия равна сумме средней групповых дисперсий и дисперсии групповых средних, т. е.

$$\begin{aligned} D(X) &= M(D_y(X)) + D(\overline{X}_y);\\ D(Y) &= M(D_x(Y)) + D(\overline{Y}_x). \end{aligned} \qquad (6.14)$$

Эти важные соотношения — формулы *разложения дисперсии* — будут неоднократно использоваться в дальнейшем. Заметим, что в дискретном случае теорема не требует нового доказательства, поскольку оно было проведено для статистического распределения (см. § 2.5). Для непрерывного случая доказательство опирается на использование определения математического ожидания и дисперсии, однако из-за громоздких выкладок мы его опускаем.

Рассчитаем основные характеристики двумерной непрерывной случайной величины, заданной плотностью распределения

$$f(x, y)=\begin{cases} e^{-x-y} & \text{при } x\geqslant 0 \text{ и } y\geqslant 0,\\ 0 & \text{при } x<0 \text{ или } y<0. \end{cases}$$

Поскольку функция $f(x, y)$ симметрична относительно x и y, ограничимся расчетом характеристик для X. Так как при $x<0$ и $y<0$ имеем $f(x, y)=0$, то можно нижний предел интегрирования в формулах для расчета $\overline{X}$, $\overline{X}_y$, $D(X)$,

$D_y(X)$ и $M(D_y(X))$ заменить нулем. Кроме того, для данной дифференциальной функции имеем по формуле (6.10)':

$$f(x, y) = e^{-x-y} = e^{-x} \cdot e^{-y}; \quad f_1(x) = \int_0^\infty e^{-x} \cdot e^{-y}\, dy = e^{-x} \int_0^\infty e^{-y} dy = e^{-x};$$

$$f_2(y) = e^{-y} \int_0^\infty e^{-x} dx = e^{-y}.$$

Далее, по формулам (6.11) и (6.11)' получим:

$$f_y(x) = \frac{e^{-x}e^{-y}}{e^{-y}} = e^{-x}, \quad f_x(y) = \frac{e^{-x}e^{-y}}{e^{-x}} = e^{-y}$$

Теперь перейдем к расчету характеристик, используя формулы из табл. 6.4, 6.5 и 6.8. Имеем:

$$M(X) = \int_0^\infty x e^{-x} dx = 1, \quad D(X) = \int_0^\infty (x-1)^2 e^{-x} dx =$$

$$= \int_0^\infty x^2 e^{-x} dx - 2\int_0^\infty x e^{-x} dx + \int_0^\infty e^{-x} dx = 1;$$

$$\overline{X}_y = \int_0^\infty x e^{-x} dx = 1, \quad D_y(X) = \int_0^\infty (x-1)^2 e^{-x} dx = 1, \quad M(\overline{X}_y) = M(1) = 1;$$

$$D(\overline{X}_y) = D(1) = 0; \quad M(D_y(X)) = M(1) = 1.$$

Теперь легко убедиться в справедливости равенств (6.13) и (6.14).

Остановимся еще на одной важной теореме — *теореме сложения математических ожиданий*, к которой будем неоднократно обращаться в дальнейшем.

Теорема 2. Математическое ожидание суммы случайных величин $\{X_1, \ldots, X_j, \ldots, X_s\}$ равно сумме их математических ожиданий, т. е.

$$M\sum_{j=1}^{s} X_j = \sum_{j=1}^{s} M(X_j). \qquad (6.15)$$

По распределению, приведенному в табл. 6.2, найдем $M(X+Y)$ и проверим справедливость формулы (6.15) путем непосредственного построения распределения для $Z = X + Y$.

По данным табл. 6.2 имеем: $\bar{X} = 0{,}25$ и $\bar{Y} = -0{,}1$; вычисляем $M(X+Y) = 0{,}25 + (-0{,}1) = 0{,}15$.

Построим распределение Z по данным табл. 6.2. Имеем $z_{11} = x_1 + y_1 = -2 + (-1) = -3$ с вероятностью $p_{11} = 0{,}15$, $z_{12} = x_1 + y_2 = -2 + 0 = -2$ с вероятностью $p_{12} = 0{,}35$ и т. д. В результате получаем распределение, приведенное в табл. 6.9.

Таблица 6.9

Z	-3	-2	-1	2	3	4
P	0,15	0,35	0,05	0,10	0,25	0,10

По этой таблице находим $M(Z) = -3 \cdot 0{,}15 - 2 \cdot 0{,}35 - 1 \cdot 0{,}05 + 2 \cdot 0{,}1 + 3 \cdot 0{,}25 + 4 \cdot 0{,}1 = 0{,}15$, что совпадает с результатами предыдущих расчетов.

6.6. Стохастическая зависимость

В гл. 3 было дано определение зависимости между событиями. Распространим эти понятия на систему случайных величин (X, Y), понимая под событиями A и B выполнение неравенств $X < x$ и $Y < y$ Две случайные величины X и Y *независимы*, если вероятность совместного наступления двух событий $(X < x)$ и $(Y < y)$ для любых x и y равна произведению вероятностей этих событий, т. е.

$$P\{X < x)(Y < y)\} = P(X < x) \cdot P(Y < y). \qquad (6.16)$$

Заменив вероятности функциями распределения, получим определение независимости в виде равенства

$$F(x, y) = F_1(x) \cdot F_2(y). \qquad (6.17)$$

Если функция распределения $F(x, y)$ не может быть представлена как произведение $F_1(x)$ на $F_2(y)$, то величины X и Y *зависимы.*

Для непрерывных случайных величин, дифференцируя обе части (6.16) по x и y, выразим условие независимости X и Y через дифференциальные функции распределения:

$$f(x, y) = f_1(x) f_2(y). \qquad (6.18)$$

Заметим, что для рассмотренных в примерах § 6.4 и 6.5 случайных величин эти условия выполняются, следовательно, X и Y независимы.

Для дискретных случайных величин условием независимости будет выполнение системы равенств:

$$P_{ik} = P(X = x_i) \cdot P(Y = y_k) = p_i \cdot p_{\cdot k} \text{ для всех } i \text{ и } k. \qquad (6.19)$$

Например, в распределении из табл. 6 2 $P_{11} = 0{,}15$, $P(X = x_1) = 0{,}55$, $P(Y = y_1) = 0{,}25$, поэтому $P_{11} \neq P(X = x_1)\ P(Y = y_1)$ и случайные величины X и Y зависимые.

Наоборот, непрерывные случайные величины X и Y, совместное распределение которых характеризуется дифференциальной функцией

$$f(x, y) = \begin{cases} \alpha\beta e^{-\alpha x - \beta y} & \text{при } x \geqslant 0 \text{ и } y \geqslant 0, \\ 0 & \text{при } x < 0 \text{ или } y < 0 \end{cases}$$

независимы, так как эту функцию можно представить в виде произведения $\alpha e^{-\alpha x} \cdot \beta e^{-\beta y}$.

Если случайные величины X и Y независимы, то по их безусловным распределениям можно составить, пользуясь равенствами (6.18) или (6.19), соответствующее двумерное распределение. Например, по заданным в табл. 6.10 и 6 11

Таблица 6 10

X	1	2	3
$P(x)$	0,3	0,5	0,2

Таблица 6.11

Y	4	6	8	10
$P(y)$	0,1	0,4	0,3	0,2

распределениям независимых случайных величин X и Y получим двумерное распределение, приведенное в табл 6 12

Условию независимости (6.18) можно придать и другую форму, если вместо $f(x, y)$ подставить выражения из (6 12): $f(x, y) = f_x(y) f_1(x) = f_y(x) f_2(y)$. После сокращения на $f_1(x)$ или $f_2(y)$ придем к двум равносильным равенствам:

$$f_x(y) = f_2(y) \text{ или } f_y(x) = f_1(x), \tag{6.20}$$

означающим, что независимость двух случайных величин равносильна совпадению плотности условного распределения каждой из них с соответствующими плотностями безусловных распределений.

Таблица 6 12

X \ Y	4	6	8	10	P_i
1	0,03	0,12	0,09	0,06	0,3
2	0,05	0,20	0,15	0,10	0,5
3	0,02	0,08	0,06	0,04	0,2
P_k	0,1	0,4	0,3	0,2	1

Зависимость случайных величин в некотором смысле обобщает понятие функциональной зависимости между переменными $y = f(x)$. Так, если последняя означает, что каждому значению x соответствует определенное значение y, то при зависимости между случайными величинами каждому значению $X = x$ может соответствовать множество значений Y, характеризуемых условным распределением с плотностью $f_x(y)$. Следовательно, зависимость случайных величин означает аналитическую зависимость плотности условного распределения одной из них от значений, принимаемых другой. Поэтому такую зависимость называют *вероятностной* или *стохастической*.

Поясним это понятие графически. На рис. 6.6 показана область изменения системы случайных величин (X, Y). В данном случае величины независимы, что можно заключить из независимости графиков условных распределений одной из составляющих от изменения значений другой. На рис. 6.7 и 6.8 показаны зависимые случайные величины. При этом на рис. 6.7 зависимость проявляется не только в изменении условных законов распределения, но и в изменении условных средних — центра каждого из этих распределений, а на рис. 6.8 изменения условных распределений характеризуются изменением условных дисперсий при неизменности центров распределений.

Линия, проходящая через точки $(x, \bar{Y})$ есть *линия регрессии* Y на X, а проходящая через точки $(y, \bar{X}_y)$ — *линия регрессии* X на Y. В случае независимости условные средние постоянны и равны общим средним (см. рис. 6.6):

$$\bar{Y}_x = \bar{Y};\ \bar{X}_y = \bar{X}, \tag{6.21}$$

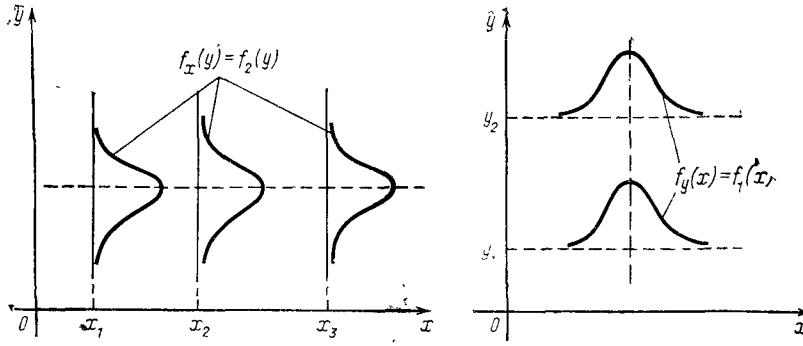


Рис. 6.6.
Иллюстрация независимости случайных величин

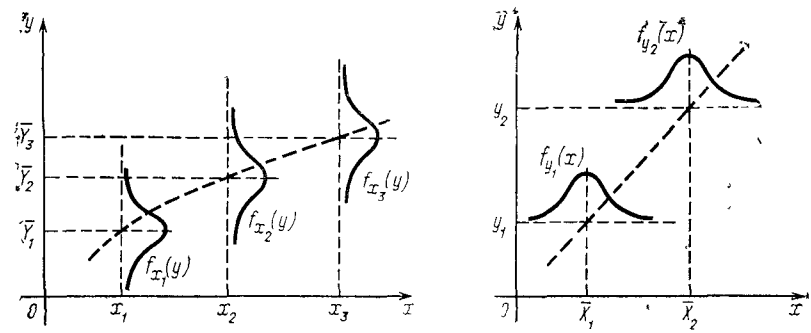


Рис. 6.7.
Иллюстрация зависимости случайных величин (изменение условных средних)

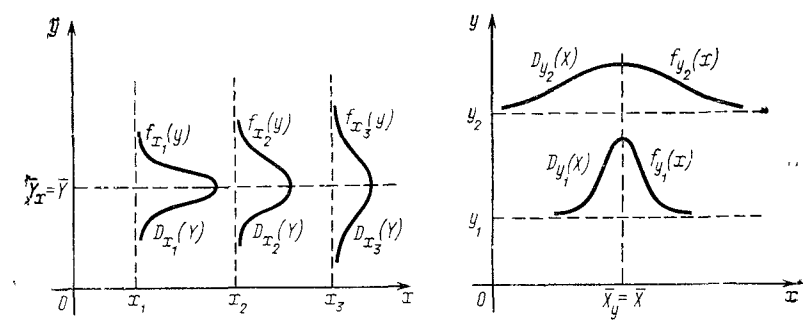


Рис. 6.8.
Иллюстрация зависимости случайных величин (изменение условных дисперсий)

в чем можно убедиться из преобразований $\overline{Y}_x = \int_{-\infty}^{\infty} y f_x(y)\,dy =$

$= \int_{-\infty}^{\infty} y f_1(y)\,dy = \overline{Y}$. Аналогично доказывается второе равенство.

Следовательно, линии регрессии в этом случае представляют собой две взаимно перпендикулярные прямые, параллельные осям координат и проходящие через центр группирования $(\overline{X}, \overline{Y})$. Однако обратный вывод неверен.

Пусть, например, системы двух величин (X, Y) распределены равномерно в круге (рис. 6.9). Из условия симметрии очевидно,

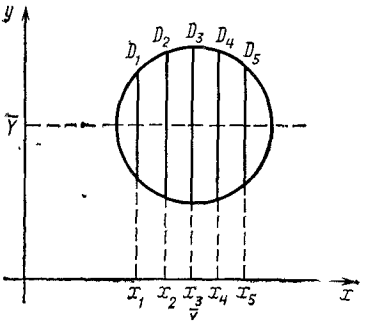


Рис. 6.9.
Демонстрация **зависимости** между X и Y при постоянстве **условных** средних

что условные средние будут располагаться на двух взаимно перпендикулярных диаметрах. А между тем величины X и $У$ стохастически зависимы. Действительно, по мере удаления от центра круга условные распределения $f_x(y)$ при данном x и $f_y(x)$ при данном y характеризуются все уменьшающейся дисперсией.

Понятие *стохастической зависимости* может быть распространено на многомерные (более двух) распределения. При этом можно говорить о попарной независимости в изложенном выше смысле и более общем понятии независимости в совокупности.

6.7. Корреляционные моменты. Коэффициент корреляции

Для анализа зависимости между двумя случайными величинами X_j и X_l введем еще одну числовую характеристику K_{jl} — *смешанный момент второго порядка*, *ковариацию* или *корреляционный момент*, определяемую как математическое ожидание произведения отклонений этих случайных величин от их математических ожиданий, т. е.

$$k_{jl} = M(X_j - \overline{X}_j)(X_l - \overline{X}_l). \qquad (6.22)$$

Используя определение математического ожидания, получим: для дискретного распределения

$$k_{jl} = \sum_{i=1}^{m} \sum_{k=1}^{n} (x_{ji} - \overline{X}_j)(x_{lk} - \overline{X}_l) p_{ik}; \qquad (6.23)$$

для непрерывного распределения

$$k_{jl} = \int_{-\infty}^{\infty}\int (x_j - \overline{X}_j)(x_l - \overline{X}_l) f(x_j, x_l)\, dx_j\, dx_l. \qquad (6.24)$$

При $l = j$ корреляционный момент совпадает с дисперсией:

$$k_{jl} = M(X_j - \overline{X}_j)^2 = D(X_j). \qquad (6.25)$$

Из (6 22) следует симметричность корреляционного момента относительно случайных величин X_j и X_l, т. е.

$$k_{jl} = k_{lj}. \qquad (6.26)$$

Всего для s-мерной случайной величины могут быть определены s^2 корреляционных моментов, из которых, исключая s дисперсий (при $j = l$), различных будет $\frac{s(s-1)}{2}$ Их удобно располагать в виде квадратной симметричной матрицы s-го порядка

$$\mathbf{K} = \begin{bmatrix} k_{11} & \dots & k_{1l} & \dots & k_{1s} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ k_{j1} & \dots & k_{jl} & \dots & k_{js} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ k_{s1} & \dots & k_{sl} & \dots & k_{ss} \end{bmatrix}, \qquad (6.27)$$

т е *корреляционной матрицы*

Далее ограничимся рассмотрением двумерной случайной величины (X, Y), для которой

$$k_{xy} = M(X - \overline{X})(Y - \overline{Y}). \qquad (6.28)$$

Для облегчения расчетов корреляционного момента можно использовать формулы, аналогичные формулам моментов, которые применялись для вычисления математического ожидания и дисперсии Раскрывая скобки в (6 28) и используя теорему о математическом ожидании суммы, получим

$$k_{xy} = M(XY - \overline{X}Y - \overline{Y}X + \overline{X}\overline{Y}) = M(XY) - \overline{X}\overline{Y}. \qquad (6.29)$$

Введя новые случайные величины $X' = \frac{X - C_1}{h_1}$ и $Y' = \frac{Y - C_2}{h_2}$, получим после несложных преобразований

$$k_{xy} = h_1 h_2 [M(X'Y') - \overline{X}'\overline{Y}']. \qquad (6.30)$$

Для иллюстрации последних формул вычислим по данным табл 6 2 корреляционный момент

$$M(XY) = (-2)\cdot(-1)\cdot 0{,}15 + (-2)\cdot 1\cdot 0{,}05 + 3\cdot(-1)\cdot 0{,}1 + 3\cdot 1\cdot 0{,}1 = 0{,}20$$

откуда $k_{xy} = 0{,}20 - 0{,}25\cdot(-0{,}1) = 0{,}225.$

В частности, для нормированных отклонений $X^0 = \frac{X - \overline{X}}{\sigma(X)}$ и $Y^0 = \frac{Y - \overline{Y}}{\sigma(Y)}$ получим $k_{xy} = \sigma(X)\sigma(Y) M(X^0 Y^0)$, так как $M(X^0) =$

$= M(Y^0) = 0$. Разделив обе части (6 28) на $\sigma(X)\sigma(Y)$, получим безразмерную характеристику — коэффициент корреляции.

$$\rho_{xy} = \frac{k_{xy}}{\sigma(x)\sigma(y)} = M(X^0 Y^0). \qquad (6.31)$$

Соответственно корреляционная матрица, при переходе к ρ_{xy} получившая название *нормированной корреляционной матрицы*, приобретает следующий вид

$$\begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix}.$$

6.8. Корреляционная зависимость

Перейдем к изучению зависимости между случайными величинами с помощью корреляционного момента. Прежде всего заметим, что проверка условий независимости с помощью (6 17)—(6.19) представляет собой весьма сложную задачу. Поэтому естественно обратиться к решению этой проблемы с помощью числовых характеристик. Оказывается, что имеет место следующий необходимый признак независимости.

Теорема Необходимым условием независимости случайных величин X и Y является равенство нулю их корреляционного момента (или коэффициента корреляции).

Рассмотрим вначале дискретное распределение Из независимости X и Y следует $p_{ik} = p_{i\cdot} p_{\cdot k}$ Подставив в (6 23), получим

$$k_{xy} = \sum_k \sum_i (x_i - \overline{X})(y_k - \overline{Y}) p_{i\cdot} p_{\cdot k} =$$
$$= \sum_k (y_k - \overline{Y}) p_{\cdot k} \left(\sum_i (x_i - \overline{X}) p_{i\cdot} \right),$$

но каждая из последних сумм равна нулю как математическое ожидание отклонений от математического ожидания. Следовательно, $k_{xy} = 0$. Аналогично для непрерывных случайных величин из их независимости следует $f(x, y) = f_1(x) f_2(y)$, откуда по определению (6.28) получим

$$k_{xy} = \int_{-\infty}^{\infty}\int (x - \overline{X})(y - \overline{Y}) f_1(x) f_2(y)\, dx\, dy =$$
$$= \int_{-\infty}^{\infty} (y - \overline{Y}) f_2(y)\, dy \cdot \int_{-\infty}^{\infty} (x - \overline{X}) f_1(x)\, dx.$$

Каждый из полученных интегралов равен нулю как математическое ожидание отклонения от математического ожидания, откуда следует утверждение теоремы.

Однако равенство $k_{xy} = 0$ (или $\rho_{xy} = 0$) есть только необходимое, но недостаточное условие независимости.

Этот вывод можно сделать и из рис 6.10, на котором изображены точки (x_i, y_k), лежащие на параболе $x-2=(y-4)^2$. Между тем здесь корреляционный момент $k_{xy}=0$.

В связи с этим вводится более узкое понятие *некоррелированности* (если $k_{xy}=0$) или *коррелированности* (если $k_{xy}\neq 0$) случайных величин.

Из доказанной теоремы и приведенного примера следует, что *независимость* случайных величин означает их *некоррелированность* ($k_{xy}=0$) и, наоборот, *коррелированность* ($k_{xy}\neq 0$) — *зависимость*. В случае коррелированности возникает необходимость

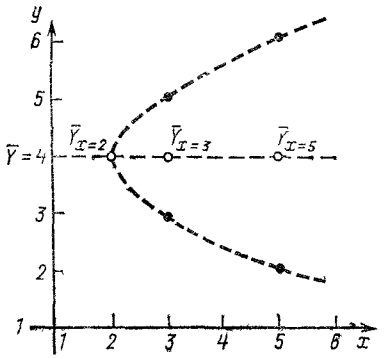


Рис. 6.10.
График зависимости $x-2=(y-4)^2$

в измерении степени зависимости X и Y. Для этого удобнее использовать безразмерную характеристику — коэффициент корреляции ρ_{xy}. Переход к оценке зависимости с помощью коэффициента корреляции диктуется еще и следующими соображениями. Величина k_{xy}, как видно из (6.28), оценивает не только зависимость между X и Y, но и рассеивание каждой из этих величин. Так, если, например, значения X мало разбросаны вокруг своей средней $\bar{X}$, то при любой степени зависимости между X и Y множитель $(X-\bar{X})$ в выражении (6.28), а следовательно, и величина k_{xy}, может оказаться сколь угодно малой. Этот недостаток исключается при переходе к ρ_{xy}.

Можно показать, что ρ_{xy} по абсолютной величине не превосходит 1, т. е.

$$-1\leqslant\rho_{xy}\leqslant 1. \tag{6.32}$$

Если случайные величины X и Y связаны точной линейной зависимостью, т. е. $Y=aX+b$, то $\bar{Y}=a\bar{X}+b$ и $k_{xy}=M(X-\bar{X})(aX+b-a\bar{X}-b)=aM(X-\bar{X})^2=aD(X)$. Далее, $D(Y)=a^2D(X)$, откуда $\sigma(Y)=|a|\sigma(X)$. Окончательно получим

$$\rho_{xy}=\frac{aD(X)}{\sigma(X)|a|\sigma(X)}=\frac{a}{|a|}=\begin{cases}-1 \text{ при } a<0,\\ 1 \text{ при } a>0.\end{cases}$$

В общем случае, когда $-1<\rho_{xy}<1$, точки (X, Y) будут разбросаны вокруг прямой тем более тесно, чем больше величина $|\rho_{xy}|$. Таким образом, коэффициент корреляции характеризует не

любую зависимость между X и Y, а *степень тесноты линейной зависимости* между ними.

Так, в частности, даже при $\rho_{xy} = 0$, т. е. при полном отсутствии линейной зависимости, между X и Y может существовать сколь угодно сильная статистическая и даже нелинейная функциональная зависимость.

При значениях $\rho_{xy} > 0$ говорят о *положительной корреляции* между X и Y, означающей, что с возрастанием одной величины другая также имеет тенденцию к возрастанию. При $\rho_{xy} < 0$ говорят об *отрицательной корреляции*, означающей противоположную тенденцию в изменении случайных величин X и Y.

6.9. Теоремы о числовых характеристиках системы случайных величин

С одной такой теоремой мы уже ознакомились в § 6.5, когда рассматривали математическое ожидание суммы случайных величин. Продолжим изучение теорем, с помощью которых можно определять математическое ожидание и дисперсию суммы и произведения случайных величин.

Теорема 1. Математическое ожидание произведения двух случайных величин равно сумме произведения математических ожиданий этих величин и корреляционного момента:

$$M(XY) = M(X)\,M(Y) + k_{xy}. \tag{6.33}$$

Доказательство вытекает непосредственно из (6.29).

В частном случае, если X и Y некоррелированны, т. е. $k_{xy} = 0$, теорема приобретает более простой вид.

Теорема 1′. Математическое ожидание произведения двух некоррелированных случайных величин равно произведению их математических ожиданий.

Из равенства (6.33), положив в нем $Y = X$ и учитывая, что $k_{xx} = M(X - \bar{X})^2 = D(X)$, получим ранее указанную формулу для дисперсии:

$$D(X) = M(X^2) - \bar{X}^2.$$

Теорема 2. Дисперсия суммы нескольких случайных величин $X_1, \ldots, X_j, \ldots, X_s$ равна сумме всех возможных корреляционных моментов, т. е.

$$D(X_1 + \ldots + X_j + \ldots + X_s) = \sum_{j=1}^{s} \sum_{l=1}^{s} k_{jl}. \tag{6.34}$$

Если учесть, что $k_{jl} = D(X_j)$ при $j = l$ и что $k_{jl} = k_{lj}$, то последнее равенство может быть представлено в виде

$$D \sum_{j=1}^{s} X_j = \sum_{j=1}^{s} D(X_j) + 2 \sum_{\substack{j=1 \\ j<l}}^{s} \sum_{l=1}^{s} k_{jl}. \tag{6.35}$$

Приведем для упрощения выкладок доказательство этой теоремы для $s = 2$. Имеем по определению дисперсии

$$D(X+Y) = M[(X+Y) - M(X+Y)]^2 = M[(X-\bar{X})^2 + (Y-\bar{Y})^2]^2 = \\ = M[(X-\bar{X})^2 + 2(X-\bar{X})(Y-\bar{Y}) + (Y-\bar{Y})^2].$$

Используя теорему сложения математических ожиданий, получим

$$D(X+Y) = M(X-\bar{X})^2 + 2M(X-\bar{X})(Y-\bar{Y}) + M(Y-\bar{Y})^2 = \\ = D(X) + D(Y) + 2k_{xy}. \qquad (6.36)$$

Для некоррелированных случайных величин получаем более простую формулировку теоремы.

Теорема 2′. Дисперсия суммы некоррелированных случайных величин равна сумме их дисперсий:

$$D\left(\sum_{j=1}^{s} X_j\right) = \sum_{j=1}^{s} D(X_j). \qquad (6.37)$$

Рассмотрим применение этих теорем к определению математического ожидания и дисперсии частоты $X_{(n)}$ появления события при n повторных независимых испытаниях.

Обозначим через X_j частоту появления события A при одном j-м испытании. Случайная величина X_j может принимать только два значения: 0 (что означает непоявление A) с вероятностью $P(\bar{A}) = 1-P(A) = 1-p$ и 1 (появление A) с вероятностью $P(A) = p$. Таким образом, для X_j получаем предельно простое распределение, откуда следует

$$M(X_j) = p \text{ и } D(X_j) = p(1-p).$$

Рассмотрим теперь частоту $X_{(n)}$ при n повторных независимых испытаниях. Очевидно, что $X_{(n)}$ можно представить в виде суммы частот при каждом единичном испытании, т. е. $X_{(n)} = \sum_{j=1}^{n} X_j$. Слагаемые X_j и X_l попарно некоррелированны, так как

$$M(X_j X_l) = 0^2(1-p)^2 + 2 \cdot 0 \cdot 1 (1-p)p + 1^2 p^2 = p^2,$$

откуда

$$k_{jl} = M(X_j - p)(X_l - p) = M(X_j X_l) - p^2 = 0.$$

Поэтому, применяя к $X_{(n)}$ теоремы 1′ и 2′, получим

$$M(X_{(n)}) = \sum_{j=1}^{n} M(X_j) = np; \; D(X_{(n)}) = \sum_{j=1}^{n} D(X_j) = np(1-p),$$

что совпадает с ранее полученными формулами (5.12) и (5.14).

Глава 7

Основные виды распределений

7.1. Вводные замечания

Мы уже познакомились в гл. 5 с одной группой дискретных распределений, связанных со схемой повторных испытаний. Между тем в различных как теоретических, так и практических исследованиях возникает необходимость использовать другие виды распределений, с помощью которых могут строиться теоретико-вероятностные модели статистических закономерностей. Каждое распределение характеризует некоторую случайную величину — результат определенного вида испытаний. Как уже указывалось, такое распределение может рассматриваться как теоретическое (генеральное), а результат опыта — как статистическое (выборочное) распределение. Последнее, в силу ограниченности числа наблюдений, будет лишь приближенно (тем более точно, чем больше число наблюдений) характеризовать теоретическое распределение.

В настоящей главе рассматриваются некоторые виды распределений случайных величин, которые найдут применение в дальнейшем. Все они могут быть разделены на две группы, концентрируемые вокруг двух основных распределений: дискретного распределения Пуассона и непрерывного нормального распределения Муавра — Лапласа.

При моделировании реальных процессов можно столкнуться с двумя ситуациями. В первой удается на основании общих теоретических исследований механизма образования статистической закономерности определить соответствующий закон распределения. В этом случае по опытным данным остается лишь найти его параметры. Во второй теоретическое исследование не позволяет установить закон распределения. Тогда по статистическому ряду подбирается подходящая кривая (теоретическое распределение), которая в некотором смысле наилучшим образом сглаживает данный ряд.

Такой подбор теоретического распределения, получивший название *выравнивания* статистического ряда, вообще говоря, имеет неоднозначное решение. Поэтому весьма важным является установление некоторых объективных признаков, при помощи которых оно может быть произведено. К ним относится сопоставление общестатистических свойств опытного и теоретического распределений (значений коэффициентов асимметрии и эксцесса, общих особенностей формы распределения и его графика, соотношений между отдельными числовыми характеристиками и т. д.).

7.2. Распределение Пуассона

Пусть событие A появляется некоторое число раз в фиксированном участке пространства (отрезке, площади, объеме) или промежутке времени с постоянной интенсивностью. Для определенности

рассмотрим последовательное появление событий во времени, называемое *потоком событий*. Графически поток событий можно иллюстрировать множеством точек, расположенных на оси времени (рис. 7.1).

Допустим, что поток событий обладает следующими свойствами:

1. Вероятность появления некоторого числа событий в данном промежутке времени зависит только от длины этого промежутка, а не от его положения на временно́й оси. Это — свойство *стационарности.*

2. Появление более одного события в достаточно малом промежутке времени Δt практически невозможно, т. е. имеет вероятность порядка $O(\Delta t)$*. Это — свойство *ординарности.*

3. Вероятность появления данного числа событий на фиксированном промежутке времени не зависит от числа событий, появ-

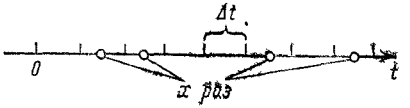


Рис. 7.1.
Поток событий на оси времени

ляющихся в другие промежутки времени. Это — свойство *отсутствия последействия.*

Поток событий, удовлетворяющий перечисленным трем свойствам, называется *простейшим потоком.*

Рассмотрим достаточно малый промежуток времени Δt. На основании свойства 2 событие может появиться на этом промежутке один раз или совсем не появиться. Обозначим вероятность появления событий через p. На основании свойства 3 эта вероятность постоянна для любого промежутка Δt, а на основании свойства 1 она зависит только от величины Δt. Вероятность непоявления события на этом промежутке $q = 1 - p$. Так как в промежутке времени Δt может появиться или 0, или одно событие, то математическое ожидание числа появлений события в промежутке Δt будет равно $0 \cdot q + 1 \cdot p = p$. Отношение этой величины к Δt, представляющее собой математическое ожидание числа появлений событий в единицу времени, является *интенсивностью* потока и обозначается через λ, т. е. $\lambda = \frac{p}{\Delta t}$.

Рассмотрим конечный отрезок времени t и разделим его на n равных частей $\Delta t = \frac{t}{n}$ (см. рис. 7.1). Появления события в каждом из этих промежутков на основании свойства 2 независимы. Определим вероятность $P_t(x)$ того, что в отрезке времени t при постоянной интенсивности потока λ событие появится x раз. Так как событие может в каждом из n промежутков Δt появиться не более чем 1 раз, то для появления его x раз на отрезке длительностью t оно должно появиться в любых x промежутках из об-

* Символ $O(\Delta t)$ означает более высокий порядок малости в сравнении с Δt.

щего числа n. Всего таких комбинаций будет C_n^x, а вероятность каждой равна $p^x q^{n-x}$. Следовательно, по теореме сложения вероятностей получаем для искомой вероятности

$$P_t(x) \approx C_n^x p^x q^{n-x}. \tag{7.1}$$

Равенство (7.1) записано как приближенное, так как исходной посылкой при его выводе служило свойство 2, выполняемое тем точнее, чем меньше Δt. Поэтому для получения точного равенства перейдем к пределу при $\Delta t \to 0$, или, что то же, $n \to \infty$. Получим после замены $p = \lambda \Delta t = \frac{\lambda t}{n}$ и $q = 1 - \frac{\lambda t}{n}$.

$$P_t(x) = \lim_{n\to\infty} C_n^x p^x q^{n-x} = \lim_{n\to\infty} \left[\frac{n(n-1)\ldots(n-x+1)}{1\cdot 2\ldots x}\left(\frac{\lambda t}{n}\right)^x \times \right.$$

$$\left. \times \left(1 - \frac{\lambda t}{n}\right)^{n-x}\right]. \tag{7.2}$$

Введем новый параметр $a = \lambda t$, означающий среднее число появлений события в отрезке t. После несложных преобразований и перехода к пределу в сомножителях получим

$$P_t(x) = \lim_{n\to\infty} \frac{n(n-1)\ldots(n-x+1)}{n^x} \lim_{n\to\infty} \frac{a^x}{x!} \lim_{n\to\infty} \left(1 - \frac{a}{n}\right)^n \times$$

$$\times \lim_{n\to\infty} \left(1 - \frac{a}{n}\right)^{-x}.$$

Учитывая, что

$$\lim_{n\to\infty} \frac{n(n-1)\ldots(n-x+1)}{n^x} = 1, \quad \lim_{n\to\infty} \frac{a^x}{x!} = \frac{a^x}{x!},$$

$$\lim_{n\to\infty} \left(1 - \frac{a}{n}\right)^n = e^{-a} \text{ и } \lim_{n\to\infty} \left(1 - \frac{a}{n}\right)^{-x} = 1,$$

получим окончательно

$$P_t(x) = \frac{a^x e^{-a}}{x!}, \text{ где } a = \lambda t. \tag{7.3}$$

Это равенство получило название формулы Пуассона.

Теорема. Вероятность того, что в данном промежутке времени длительностью t появится ровно x событий из простейшего потока событий с постоянной интенсивностью λ, определяется равенством (7.3).

Мы рассматривали поток событий во времени. Но тот же вывод можно распространить на события, распределенные с постоянной плотностью в некоторой пространственной области. Естественно, что в этом случае параметр λ есть среднеожидаемое (математическое ожидание) число событий в данной области. Этим объясняется широкий круг случайных явлений, для описания которых может быть использована формула (7.3). Поэтому, отвлекаясь от происхождения этой формулы, введем следующее определение: распределение вероятностей дискретной случайной

величины X, принимающей значения $x = 0, 1, 2, \ldots, m, \ldots$, с вероятностями

$$P(X = x) = \frac{a^x e^{-a}}{x!} = f_п(x, a), \tag{7.4}$$

называется *законом распределения Пуассона* (табл. 7.1).

Таблица 7.1

X	0	1	2	...	m	...	$\sum$
P	e^{-a}	ae^{-a}	$\frac{a^2 e^{-a}}{2!}$	...	$\frac{a^m e^{-a}}{m!}$	...	1

Формула (7.4) охватывает и случай $x = 0$, если учесть принятое ранее условие, что $0! = 1$.

Из (7.4) следует справедливость рекуррентного соотношения

$$f_п(x + 1, a) = \frac{a}{x + 1} f_п(x, a), \tag{7.5}$$

которое может быть использовано для расчета последовательных членов ряда (7.4). Суммируя все вероятности распределения Пуассона и используя разложение $e^a = \sum_{x=0}^{\infty} \frac{a^x}{x!}$, получим подтверждение основного свойства распределения:

$$\sum_{x=0}^{\infty} \frac{a^x e^{-a}}{x!} = e^{-a} \sum_{x=0}^{\infty} \frac{a^x}{x!} = e^{-a} \cdot e^a = 1.$$

Как видно из (7.4), распределение Пуассона зависит от одного параметра a, равного математическому ожиданию случайной величины. Действительно, согласно определению

$$M(X) = \sum_{0}^{\infty} x f_п(x, a) = \sum_{0}^{\infty} x \frac{a^x}{x!} e^{-a} = e^{-a} \cdot a \sum_{1}^{\infty} \frac{a^{x-1}}{(x-1)!}.$$

Но $\sum_{1}^{\infty} \frac{a^{x-1}}{(x-1)!} = e^a$, поэтому получим

$$M(X) = a. \tag{7.6}$$

Из (7.5) получаем отношение l любого члена ряда к предыдущему:

$$l = \frac{f_п(x + 1, a)}{f_п(x, a)} = \frac{a}{x + 1}.$$

Пока $x < a - 1$, отношение $l > 1$, т. е. последующий член ряда больше предыдущего и члены ряда возрастают, принимая наибольшее значение при $x = [a - 1]$. Далее, при $x > a - 1$ получим $l < 1$ и члены ряда убывают. Подобное свойство мы уже наблюдали у биномиального распределения.

Графики распределения Пуассона для разных значений a показаны на рис. 7.2, а значения $f_{\text{п}}(x, a)$ табулированы (см. [12]). Распределение, как видно из рисунка, несимметричное. С ростом параметра a значения вероятностей ряда убывают и распределение становится все более симметричным.

Найдем дисперсию случайной величины

$$D(X) = M(X^2) - [M(X)]^2 = \sum x^2 \frac{a^x e^{-a}}{x!} - a^2.$$

Выполнив преобразования аналогично примененным при выводе формулы (7.6), получим

$$D(X) = a = M(X). \tag{7.7}$$

Выполнение этого равенства для характеристик статистического ряда ($D^* = \bar{X}^*$) может быть использовано как признак при-

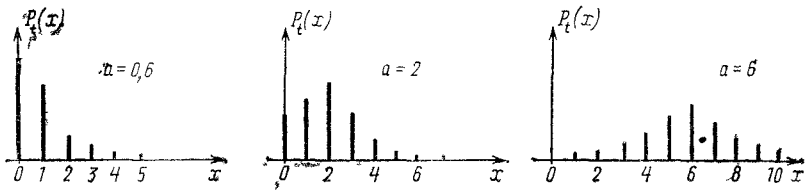


Рис. 7.2.
График распределения Пуассона

менимости распределения Пуассона для выравнивания статистического распределения.

Через параметр a могут быть также выражены еще две характеристики распределения Пуассона: коэффициент асимметрии $\alpha = \frac{1}{\sqrt{a}}$ и эксцесс $\varepsilon = \frac{1}{a} - 3$.

Интегральная функция распределения $F_{\text{п}}(x, a)$ для закона Пуассона

$$F_{\text{п}}(x,a) = P(X < x) = \sum_{m=0}^{x-1} \frac{a^m e^{-a}}{m!} \tag{7.8}$$

табулирована (см. [12]). С ее помощью удобно вычислять, например, такие вероятности, как

$$P\{x_1 \leqslant X < x_2\} = F_{\text{п}}(x_2, a) - F_{\text{п}}(x_1, a),$$

а также значения плотности вероятности

$$f_{\text{п}}(x, a) = F_{\text{п}}(x, a) - F_{\text{п}}(x - 1, a). \tag{7.8'}$$

Для случая, когда a велико и $x = a$, справедлива асимптотическая формула

$$f_{\text{п}}(x, a) \approx \frac{0{,}4}{\sqrt{a}}$$

Одним из существенных свойств распределения Пуассона является его *устойчивость* относительно линейных операций над случайными величинами. Так, может быть доказано следующее утверждение.

Теорема. Если имеется n попарно независимых случайных величин $X_1, \ldots, X_j, \ldots, X_n$, распределенных по закону Пуассона, с математическими ожиданиями $M(X_j) = a_j$ для $j = 1, \ldots, n$, то их линейная комбинация

$$X = \sum_{j=1}^{n} C_j X_j,$$

где C_j — целые числа, также подчиняется закону распределения Пуассона с параметром $a = \sum_{j=1}^{n} C_j a_j$.

Распределение Пуассона используется для вычисления вероятностей появления данного числа относительно редких (т. е. характеризуемых малой вероятностью) событий в данном промежутке времени или участке пространства.

При выводе закона распределения Пуассона существенным было предположение о постоянной интенсивности λ. Однако формула (7.3) может быть использована и при нарушении этого условия. Так, например, при описании нестационарного потока событий, для которого $\lambda = \lambda(t)$, можно воспользоваться той же формулой (7.3), определяя a из выражения $a = \int_{t}^{t+\tau} \lambda(u)\, du$.

Укажем некоторые случаи, когда может быть использован закон распределения Пуассона: 1) распределение числа появлений событий в промежутке времени: поступление вызовов на телефонную станцию, приход в магазин покупателя, поступление требования на ремонт или переналадку оборудования, прибытие в порт на разгрузку судов и т. д.; 2) распределение числа появлений событий в данной области пространства: дефекты на заданном участке проводной связи, аварии автомашин в данном районе города, ошибки на странице текста, брак изделия в партии, задержка в пути на данной железнодорожной ветке и т. д.

В заключение рассмотрим числовой пример.

Пусть на 10 однотипных станках зафиксировано за месяц их работы 20 повреждений. Считая, что появление повреждений на каждом станке — независимые события и число их появлений на каждом станке подчиняется закону распределения Пуассона, определим: 1) вероятность p_1 появления трех повреждений на одном станке за такой же период; 2) вероятность p_2 того, что число повреждений окажется не менее четырех

Принимаем в качестве параметра a среднее число повреждений, приходящихся на один станок: $a = 20 : 10 = 2$. Следовательно, можно исходить из распределения $f_{\text{п}}(x, 2) = \frac{2^x e^{-2}}{x!}$. С помощью таблицы при $a = 2$ находим:

$p_1 = f_{\text{п}}(3, 2) = 0{,}1804$; $p_2 = P(X \geqslant 4) = 1 - P(X < 4) = 1 - F_{\text{п}}(4, 2) = 0{,}0527$.

7.3. Аппроксимация биномиального распределения распределением Пуассона

При изучении биномиального распределения указывалось на целесообразность использования асимптотических формул, облегчающих расчет $P_n(m)$ при большом n. При выводе (7.3) было получено

$$\lim_{n\to\infty} C_n^x p^x q^{n-x} = \frac{a^x e^{-a}}{x!}, \quad \text{где} \quad a = \lambda t, \quad \text{и} \quad p = \frac{\lambda t}{n} = \frac{a}{n}.$$

Таким образом, биномиальное распределение при $n \to \infty$ стремится к распределению Пуассона с параметром $a = np$. Поскольку параметр a — постоянное число (среднее число появления события), то с ростом n: $p = \frac{a}{n} \to 0$, т. е. предельное равенство предполагает неограниченное уменьшение вероятности p.

Переходя от предельного равенства к приближенному, при конечном n получим асимптотическую формулу Пуассона для биномиального распределения:

$$P_n(m) \approx \frac{(np)^m e^{-np}}{m!}. \tag{7.9}$$

Строго говоря, предпосылка о постоянстве a, лежащая в основе вывода закона Пуассона (откуда следует переменная вероятность $p = \frac{a}{n}$), противоречит исходным условиям биномиального распределения ($p = \text{const}$). Однако приближенное равенство (7.9) дает достаточно точные результаты при a непостоянном, т. е. при постоянном p. Важно лишь сохранить условие малости p и достаточно большого n так, чтобы произведение np было невелико ($np < 10$).

В следующей таблице приведены сравнительные результаты расчетов $P_n(m)$ по точной формуле Бернулли и асимптотической формуле Пуассона. Как видно, результаты вычислений оказываются достаточно близкими даже при небольших значениях n.

Таблица 7.2

Формула	$p = 0{,}1;\ m = 1;$ $n = 10$	$p = 0{,}1;\ m = 5;$ $n = 50$	$p = 0{,}01;$ $m = 1;\ n = 10$	$p = 0{,}95;\ m = 9;$ $n = 10$
Бернулли	0,3874	0,1809	0,0913	0,3132
Пуассона	0,3679	0,1755	0,0905	0,3030

Можно руководствоваться следующей ориентировочной рекомендацией по использованию асимптотической формулы (7.9): $n \geqslant 100$ и $0 < np < 10$. Асимптотическая формула дает хорошие результаты и при переменной вероятности p, т. е. при зависимости между результатами отдельных испытаний (например, для гипергеометрического распределения).

7.4. Показательное распределение

Вернемся к рассмотрению простейшего потока событий с интенсивностью λ и введем непрерывную случайную величину T — промежуток времени между двумя появлениями события. Очевидно, что по смыслу T принимает только неотрицательные значения. Определим для нее функцию распределения. $F(t) = P(T < t)$. Вероятность противоположного события $(T \geqslant t)$ равна вероятности того, что в промежутке времени $(0, t)$ не наступит ни одно событие потока, т. е.

$$P(T \geqslant t) = P_t(0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t} = H(t, \lambda). \qquad (7.10)$$

Следовательно, вероятность искомого события

$$P(T < t) = 1 - e^{-\lambda t} = F(t, \lambda) \quad \text{для} \quad t \geqslant 0. \qquad (7.11)$$

Закон распределения непрерывной случайной величины T, заданный функцией распределения (7.11), получил название *показательного*. Дифференциальная функция для этого распределения найдется по общему правилу:

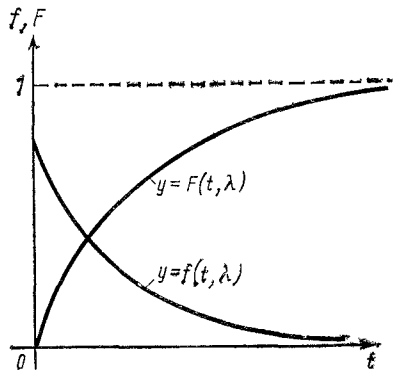


Рис. 7.3.
Плотность и функция показательного распределения

$$f(t, \lambda) = \begin{cases} \lambda e^{-\lambda t}, & \text{при } t \geqslant 0, \\ 0, & \text{при } t < 0. \end{cases} \qquad (7.11')$$

Графики функций $f(t, \lambda)$ и $F(t, \lambda)$ показаны на рис. 7.3, а значения $F(t, \lambda)$ табулированы (см. [12]). Показательное распределение также оказывается зависящим от одного параметра λ. Найдем основные характеристики случайной величины T. Согласно определению получим $M(T) = \int_0^\infty t\lambda e^{-\lambda t}\, dt$. Интегрируя по частям и учитывая, что $\lim_{t \to \infty} te^{-\lambda t} = 0$, получим

$$M(T) = \frac{-1}{\lambda} e^{-\lambda t} \Big|_0^\infty = \frac{1}{\lambda}. \qquad (7.12)$$

Аналогичным образом, не останавливаясь подробно на аналитических преобразованиях, получим

$$D(T)=\frac{1}{\lambda^2}, \quad \alpha=2{,}0 \quad \text{и} \quad \varepsilon=0. \tag{7.12'}$$

Соотношение (7.12) может быть использовано для оценки параметра λ по опытным данным. В качестве значения для λ может быть принята величина $\lambda^*=\frac{1}{\overline{T}^*}$, т. е. обратная среднему промежутку времени $\overline{T}^*$ между двумя событиями. Если поток наблюдается в течение времени τ и при этом n раз появилось событие, то $\lambda^*=\frac{n}{\tau}$.

Важным свойством показательного распределения является отсутствие последействия, т. е. независимость вероятности данного промежутка времени $T=t$ между появлениями событий простейшего потока от длительности предыдущих промежутков времени. Собственно, это свойство является перефразировкой свойства 2 простейшего потока, но его можно доказать и исходя из определения показательного распределения.

Показательное распределение нами получено как распределение промежутка времени между двумя событиями простейшего потока. Можно доказать обратное положение: если исходить из показательного закона распределения длительности промежутка времени между некоторыми событиями, то события образуют простейший поток с интенсивностью $\lambda=\frac{1}{\overline{T}}$.

Из этого обстоятельства вытекает самостоятельное значение показательного распределения. Оно часто используется для описания распределения времени безотказной работы прибора или системы, если интенсивность отказов можно считать постоянной, длительности ремонта или другого вида обслуживания и т. д.

Наконец, для практического применения показательного распределения важна предельная теорема. Согласно этой теореме при объединении достаточно большого числа потоков с любыми законами распределения промежутков времени между появлениями событий промежуток времени между событиями в объединенном потоке будет в пределе при неограниченном увеличении числа составляющих потоков и равномерной их малости подчиняться показательному закону распределения с параметром $\lambda=\frac{n}{\tau}$, где τ — общее время наблюдения за всеми потоками и n — общее число появлений событий за время τ.

Показательное распределение широко используется в теории надежности, изучающей условия безотказной работы некоторой системы, если отказы в ее работе образуют простейший поток.

Так, полученная в (7.10) функция $H(t, \lambda)$, характеризующая вероятность того, что система будет работать надежно (без отказов) в течение промежутка времени t, называется *функцией*

надежности. Пользуясь ею, можно определить, например, вероятность того, что время надежной работы не выйдет из заданного интервала:

$$P(t_1 \leqslant T \leqslant t_2) = e^{-\lambda t_1} - e^{-\lambda t_2}.$$

Наоборот, задаваясь вероятностью $P(T \geqslant t) = p_1$, можно определить время надежной работы, гарантируемой с этой вероятностью:

$$t = \frac{-\ln p_1}{\lambda}.$$

Пусть в результате наблюдений за работой системы в течение 100 ч зарегистрировано пять отказов. Определим: 1) функцию надежности системы; 2) вероятность безотказной работы в течение 50 ч; 3) время безотказной работы, гарантированное с вероятностью $P = 0{,}99$.

Значит, имеем $\tau = 100$ ч, $n = 5$, откуда $\lambda^* = 5 : 100 = 0{,}05$ отказа в 1 ч. Следовательно: 1) функция надежности $1 - F(t;\ 0{,}05) = e^{-0{,}05t}$; 2) $P(T \geqslant 50) = e^{-0{,}05 \cdot 50} = 0{,}61$; 3) задаваясь $p_1 = 0{,}99$, получим

$$t = \frac{-\ln 0{,}99}{0{,}05} = 0{,}2 \text{ ч}.$$

7.5. Распределения, связанные с показательным распределением

Несмотря на довольно широкую область приложения, показательное распределение далеко не всегда может быть использовано при построении теоретико-вероятностных моделей. Кроме того, оно включает только один параметр λ и потому трудно подобрать его значение, чтобы при выравнивании передать особенности опытного распределения.

В данном параграфе рассматриваются распределения, сходные в каком-то смысле с показательным, но содержащие большее число параметров.

Гамма-распределение. Моделью образования гамма-распределения является поток событий с постоянной интенсивностью λ. При этом в качестве случайной величины T рассматривается время, необходимое для появления данного числа β событий. Таким образом, как и в показательном законе, величина T меняется в интервале $(0, \infty)$. Можно показать, что это распределение будет задаваться плотностью вероятности

$$f_\Gamma(t, \beta, \lambda) = \begin{cases} \dfrac{\lambda^\beta}{\Gamma(\beta)} t^{\beta-1} e^{-\lambda t} & \text{при } t \geqslant 0, \\ 0 & \text{при } t < 0 \end{cases} \tag{7.13}$$

или функцией распределения

$$P(T < t) = F_\Gamma(t, \beta, \lambda) = \begin{cases} \dfrac{\lambda^\beta}{\Gamma(\beta)} \displaystyle\int_0^t u^{\beta-1} e^{-\lambda u}\, du & \text{при } t \geqslant 0, \\ 0 & \text{при } t < 0, \end{cases} \tag{7.14}$$

где $\Gamma(\beta)$ — некоторая зависящая от β постоянная величина *. Распределение (7.13) или (7.14) называется гамма-распределением. Оно зависит от двух параметров β и λ, принимающих неотрицательные значения.

Табулирование функции (7.14) для разных t сложно (зависимость от трех входов β, λ, t), поэтому используются таблицы неполной гамма-функции:

$$\gamma(z, \beta) = \frac{1}{\Gamma(\beta)} \int_0^z v^{\beta-1} e^{-v}\, dv,$$

которая табулируется в зависимости уже от двух величин $z = \lambda t$ и β. При этом $F(t, \beta, \lambda) = \gamma(z, \beta)$.

Для определения $P(T < t)$ по заданным значениям λ, t и β находим $z = \lambda t$, затем по таблице значения $\gamma(z, \beta)$ и, наконец, $P(T < t) = F(t, \beta, \lambda) = \gamma(z, \beta)$.

Предположим, расход материала на производство носит случайный характер со средней интенсивностью 20 единиц в день. Для покрытия расхода производятся регулярные ежемесячные поставки объемом в 640 ед. Определим: 1) вероятность образования дефицита материалов; 2) объем поставок, при котором вероятность дефицита не превысит 0,01.

Обозначим через T промежуток времени, в течение которого суммарный расход окажется равным объему поставки β Величина T является случайной, подчиняющейся гамма-распределению с параметрами β и λ. Дефицит образуется, если T окажется меньше заданного интервала между поставками, т. е. при $T < 30$. Из определения функции распределения имеем

$$P(T < 30) = F_\Gamma(30, \beta, \lambda) = \gamma(z, \beta).$$

1. Задаваясь $\beta = 640$, $\lambda = 20$ и $t = 30$, получим

$$P(T < 30) = F_\Gamma(30, 640, 20) = \gamma(600, 640) = 0{,}057.$$

2. По заданной величине $p_1 = P(T < 30) = 0{,}01 = F_\Gamma(30, \beta, 20)$ находим $\beta = 660$.

Графики плотности вероятности гамма-распределения при различных значениях β показаны на рис. 7.4.

Моменты первого и второго порядков определяются из равенств:

$$M(T) = \frac{\beta}{\lambda}; \quad D(T) = \frac{\beta}{\lambda^2}. \tag{7.15}$$

Из этих двух уравнений можно получить оценки параметров гамма-распределения по опытным данным:

$$\lambda^* = \frac{\overline{T}^*}{D^*(T)}; \quad \beta^* = \frac{(\overline{T}^*)^2}{D^*(T)} = \lambda^* \overline{T}^*. \tag{7.16}$$

---

* $\Gamma(\beta)$ — гамма-функция, определяемая несобственным интегралом $\Gamma(\beta) = \int_0^\infty x^{\beta-1} e^{-x}\, dx$. Если β – целое положительное число, то $\Gamma(\beta) = (\beta - 1)!$

При отдельных значениях параметра β получаем частные случаи гамма-распределения. Так, при $\beta = 1$ гамма-распределение совпадает с показательным При β — положительном целом числе

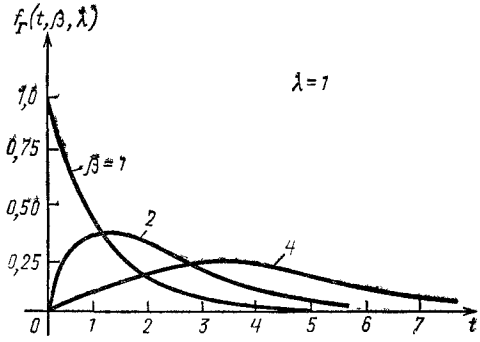


Рис. 7.4.
Графики плотности вероятности гамма-распределения

получаем распределение Эрланга, широко используемое в теории массового обслуживания.

Бета-распределение. Бета-распределение используется в качестве статистической модели для величины, принимающей любые значения в некотором промежутке $[a, b]$. Изменением масштаба можно этот интервал привести к стандартному промежутку $(0, 1)$. Плотность вероятности бета-распределения имеет следующий вид:

$$f_\beta(x, \beta, \gamma) = \begin{cases} \dfrac{\Gamma(\beta+\gamma)}{\Gamma(\beta)\,\Gamma(\gamma)}\, x^{\beta-1}(1-x)^{\gamma-1} & \text{при} \quad 0 \leqslant x \leqslant 1, \\ 0 & \text{при} \quad x < 0 \quad \text{или} \quad x > 1, \end{cases} \tag{7.17}$$

и функция распределения

$$F_\beta(x, \beta, \gamma) = \begin{cases} 0 & \text{при} \quad x < 0, \\ \dfrac{\Gamma(\beta+\gamma)}{\Gamma(\beta)\,\Gamma(\gamma)} \displaystyle\int_0^x u^{\beta-1}(1-u)^{\gamma-1}\,du & \text{при} \quad 0 \leqslant x \leqslant 1. \end{cases} \tag{7.18}$$

На рис. 7.5 показаны графики плотности вероятности (7.17).

Наличие двух параметров β и γ и разнообразие форм бета-распределения при изменении этих параметров делают его удобным при построении различных статистических моделей. Так, бета-распределение можно использовать для описания распределения доли брака на производственном участке за рабочую смену, доли элементов генеральной совокупности, лежащих между наименьшим и наибольшим значениями в выборке, продолжительности времени завершения работы между оптимистической $t_{оп}$ и пессимистической $t_п$ оценками этого времени в системе сетевого планирования и управления (СПУ) и т. д.

Оценки параметров β и γ по опытным данным могут быть получены из выражений:

$$\gamma^* = \frac{1-\overline{X}^*}{D^*}\left[\overline{X}^*(1-\overline{X}^*) - D^*\right]; \quad \beta^* = \frac{\overline{X}^*}{1-\overline{X}^*}\gamma^*, \qquad (7.19)$$

где $\overline{X}^*$ и D^* — математическое ожидание и дисперсия, найденные по опытным данным.

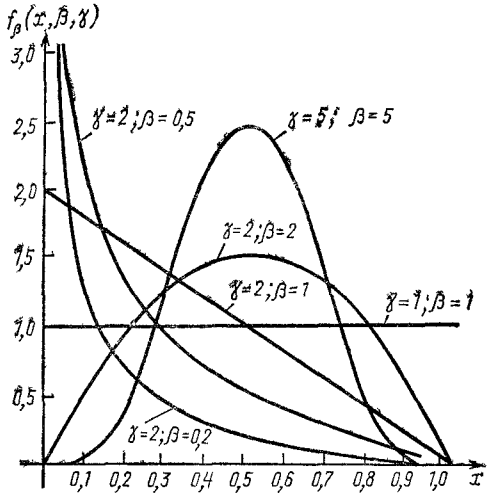


Рис. 7.5.
Графики плотности вероятности бета-распределения

При $\beta = \gamma = 1$ получаем из (7.17) равномерное распределение на отрезке (0, 1). При $\beta = 2$ и $\gamma = 1$ получаем треугольное распределение, задаваемое плотностью вероятности

$$f_\beta(x, 2, 1) = \begin{cases} 2x & \text{при} \quad 0 \leqslant x \leqslant 1, \\ 0 & \text{при} \quad x < 0 \quad \text{или} \quad x > 1. \end{cases}$$

7.6. Нормальное распределение

При построении различных статистических моделей наиболее широко применяется нормальное распределение. Теоретическим основанием к его применению служит центральная предельная теорема Ляпунова, рассматриваемая в следующей главе. Согласно этой теореме распределение суммы n попарно независимых и произвольно распределенных случайных величин при некоторых дополнительных условиях и неограниченном возрастании n стремится к нормальному закону распределения.

Широкому использованию нормального распределения способствует также ряд простых, но важных его свойств.

Нормальное распределение — это распределение непрерывной случайной величины X, характеризуемое плотностью вероятности

$$f_N(x, a, b) = \frac{1}{b\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2b^2}} \quad \text{при} \quad -\infty < x < \infty, \qquad (7.20)$$

где a и b — параметры.

Найдем математическое ожидание случайной величины X, подчиняющейся распределению (7.20):

$$M(X) = \int_{-\infty}^{\infty} x f_N(x, a, b)\,dx = \frac{1}{b\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{(x-a)^2}{2b^2}}\,dx.$$

Введя подстановку $z = \frac{x-a}{b}$, откуда $dz = \frac{1}{b}dx$, получим

$$M(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (bz + a) e^{-\frac{z^2}{2}}\,dz = \frac{b}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}}\,dz + \frac{a}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz.$$

Но первый интеграл от нечетной функции в симметричных границах интегрирования равен нулю, а второй есть интеграл Пуассона, равный $\sqrt{2\pi}$. Поэтому получим окончательно

$$M(X) = a. \qquad (7.21)$$

Выполнив аналогичные преобразования при расчете дисперсии, получим

$$D(X) = b^2 \quad \text{или} \quad b = \sigma(X). \qquad (7.22)$$

Таким образом, два параметра нормального распределения a и b^2 равны соответственно математическому ожиданию и дисперсии распределения. На рис. 7.6 показан график функции (7.20).

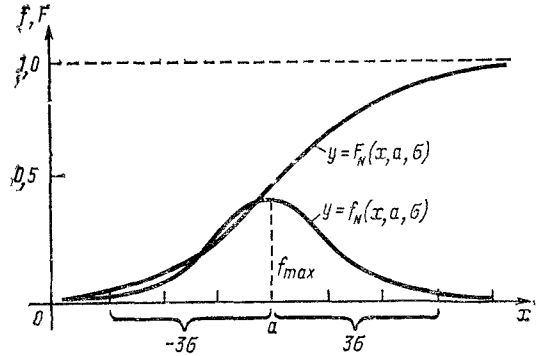


Рис. 7.6. Нормальное распределение. Графики $f_N(x, a, \sigma)$ и $F_N(x, a, \sigma)$

Кривая симметрична относительно прямой $x = a$; в точке $x = a$ достигается максимум, равный

$$f_{\max} = \frac{1}{\sigma\sqrt{2\pi}} = \frac{0{,}3989}{\sigma}.$$

При $x \to \pm\infty$ кривая асимптотически приближается к оси абсцисс. Действительно,

$$\lim_{x\to\pm\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\overline{X})}{2\sigma^2}} = 0.$$

Площадь, ограниченная кривой, равна 1. Можно показать, что если ограничить ее ординатами $x = a - 3\sigma$ и $x = a + 3\sigma$, то площадь окажется почти равной 1 (точнее, 0,9973). При уменьшении

σ растет f_{max}, а так как площадь должна оставаться равной 1, то кривая будет стягиваться к своей оси симметрии $x = \bar{X}$, что соответствует смыслу σ как меры рассеяния. На рис. 7.7 показаны три кривые, соответствующие значениям $\sigma_1 < \sigma_2 < \sigma_3$.

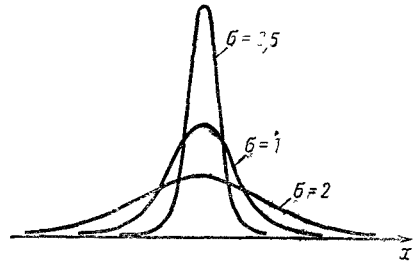


Рис. 7.7.
Зависимость формы кривой нормального распределения от параметра σ

Вычислив для нормального распределения центральные моменты третьего и четвертого порядков, получим $M(X - \bar{X})^3 = 0$ и $M(X - \bar{X})^4 = 3\sigma^4$. Следовательно, асимметрия распределения

$$\alpha = \frac{M(X-\bar{X})^3}{\sigma^3} = 0 \qquad (7.23)$$

и эксцесс распределения

$$\varepsilon = \frac{M(X-\bar{X})^4}{\sigma^4} - 3 = 0. \qquad (7.24)$$

Функция распределения для нормального закона распределения определяется из выражения

$$P(X < x) = F_N(x, \bar{X}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\bar{X})^2}{2\sigma^2}}\, du. \qquad (7.25)$$

График ее показан на рис. 7.6. Нетрудно убедиться, что эта функция будет обладать свойствами, общими для любой функции распределения:

1) $\lim\limits_{x\to-\infty} F_N(x, \bar{X}, \sigma) = 0$; 2) $\lim\limits_{x\to\infty} F_N(x, \bar{X}, \sigma) = 1$;

3) $0 \leqslant F_N(x, \bar{X}, \sigma) \leqslant 1$ для любого x.

С помощью функции распределения может быть определена вероятность попадания случайной величины в заданный интервал:

$$P(x_1 \leqslant X < x_2) = F_N(x_2, \bar{X}, \sigma) - F_N(x_1, \bar{X}, \sigma). \qquad (7.26)$$

Одним из важнейших свойств нормального распределения оказывается его своеобразная «устойчивость», проявляющаяся в том, что сумма n нормально распределенных случайных величин X_j, т. е. $\sum\limits_{j=1}^{n} X_j$ также будет подчиняться нормальному закону распределения. При этом если X_j попарно независимые, то параметрами этого объединенного нормального распределения будут:

$$a = M(\Sigma X_j) = \Sigma M(X_j); \quad b^2 = D(\Sigma X_j) = \Sigma D(X_j). \qquad (7.27)$$

Оказывается, что и при произвольных распределениях X_i, но при выполнении некоторых дополнительных условий (гл. 8), распределение $\sum X_i$ будет стремиться к нормальному при неограниченном увеличении числа слагаемых n.

7.7. Стандартное нормальное распределение. Функции Гаусса и Лапласа

Табулировать функции f_N и F_N сложно, так как они зависят от двух параметров a и b. Поэтому перейдем к нормированному отклонению

$$Z = \frac{X - \overline{X}}{\sigma}. \tag{7.28}$$

Если X есть нормально распределенная случайная величина, заданная функцией (7.20), то соответствующую плотность вероятности для Z можно получить из (4.16):

$$\varphi(z) = f_N(\sigma z + \overline{X}) \cdot \sigma = \frac{\sigma}{\sigma\sqrt{2\pi}} e^{-\frac{(\sigma z + \overline{X} - \overline{X})^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{7.29}$$

Но это есть нормальный закон распределения с параметрами $a = 0$ и $b = 1$, которые соответствуют математическому ожиданию

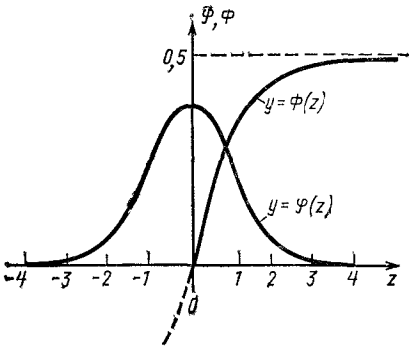


Рис. 7.8.
Кривая Гаусса и функция Лапласа

и дисперсии случайной величины Z, или *стандартный нормальный закон распределения*. График функции $\varphi(z)$, изображенный на рис. 7.8, называется *кривой вероятностей* или *кривой Гаусса*. Функция $\varphi(z)$ обладает теми же свойствами, что и любая функция плотности вероятности, т. е. 1) $\varphi(z) \geqslant 0$; 2) $\int_{-\infty}^{\infty} \varphi(z)\,dz = 1$.

Кроме того, из (7.29) можно получить и ряд специальных свойств, которые отражены на ее графике:

3) $\varphi(-z) = \varphi(z)$; 4) $\lim\limits_{z \to \pm\infty} \varphi(z) = 0$.

Последнее предельное равенство практически выполняется почти точно уже начиная с $z = 4$ (так, $\varphi(4) = 0{,}0001 \approx 0$) Поэтому при табулировании $\varphi(z)$ указываются значения z лишь в интервале от 0 до 4.

По значению $\varphi(z)$ можно найти значение $f_N(x, \overline{X}, \sigma)$ из соотношения

$$f_N(x, \overline{X}, \sigma) = \varphi\left(\frac{x - \overline{X}}{\sigma}\right) \cdot \frac{1}{\sigma}. \tag{7.30}$$

Преобразуем выражение для элемента вероятности:

$$P(x < X < x + \mathrm{d}x) \approx f_N(x, \overline{X}, \sigma)\,\mathrm{d}x.$$

Переходя к стандартной функции плотности, получим

$$P(x < X < x + \mathrm{d}x) = \varphi(z)\,\mathrm{d}z.$$

Интегральная функция для стандартного нормального закона распределения найдется из выражения

$$P(Z < z) = F_N(z, 0, 1) = \int_{-\infty}^{z} \varphi(u)\,\mathrm{d}u = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-u^2/2}\,\mathrm{d}u. \tag{7.31}$$

Так как события $Z < z$ и $X < x$ равновероятны, то получим

$$F_N(x, \overline{X}, \sigma) = F_N(z, 0, 1), \quad \text{где } z = \frac{x - \overline{X}}{\sigma}. \tag{7.32}$$

Это соотношение может быть использовано для определения функции распределения нормального закона по значениям функции $F_N(z, 0, 1)$. Используя (7.32), можем формулу (7.26) записать в виде

$$P(x_1 < X < x_2) = F_N(z_2, 0, 1) - F(z_1, 0, 1), \tag{7.33}$$

где

$$z_1 = \frac{x_1 - \overline{X}}{\sigma} \text{ и } z_2 = \frac{x_2 - \overline{X}}{\sigma}.$$

Функция распределения $F_N(z, 0, 1)$, определенная равенством (7.32), зависит только от одной переменной z и поэтому удобна для табулирования. Однако интеграл, через который она определена, не выражается через элементарные функции, поэтому введем функцию Лапласа

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{0}^{z} \mathrm{e}^{-u^2/2}\,\mathrm{d}u, \tag{7.34}$$

численно равную площади фигуры, ограниченной сверху кривой Гаусса $y = \varphi(z)$, снизу — осью абсцисс, слева — осью ординат и справа — прямой $y = z$.

Из определения $\Phi(z)$ следуют основные свойства, которые необходимо учитывать при ее табулировании и практическом ее применении:

1) $\Phi(z)$ — монотонно возрастающая, что следует из положительности производной

$$\Phi'(z)=\frac{1}{\sqrt{2\pi}}e^{-z^2/2}>0 \text{ при любом } z;$$

2) $\Phi(z)$ — нечетная, т. е. $\Phi(-z)=-\Phi(z)$ (график симметричен относительно начала координат). В частности, $\Phi(0)=0$;

3) $\lim\limits_{z\to\infty}\Phi(z)=0{,}5$, что следует из геометрической интерпретации $\Phi(z)$ как площади (или из интеграла Пуассона). Так как $\Phi(4)=0{,}49997\approx 0{,}5$ и функция монотонно возрастает, то с достаточной для практики точностью можно полагать $\Phi(z)=0{,}5$ для $z\geqslant 4$. С учетом указанных свойств она табулируется для значений z от 0 до 4.

На рис. 7.8 построен график $y=\Phi(z)$. Ветви такого графика асимптотически приближаются к прямым $y=+0{,}5$ и $y=-0{,}5$. Выразим функцию $F_N(z, 0, 1)$ через функцию Лапласа:

$$F_N(z,0,1)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{0}e^{-\frac{u^2}{2}}du+\frac{1}{\sqrt{2\pi}}\int\limits_{0}^{z}e^{-u^2/2}du=$$

$$=\frac{-1}{\sqrt{2\pi}}\int\limits_{0}^{-\infty}e^{-u^2/2}du+\frac{1}{\sqrt{2\pi}}\int\limits_{0}^{z}e^{-\frac{u^2}{2}}du.$$

Подставляя вместо полученных интегралов функцию Лапласа и учитывая, что $\int\limits_{0}^{-\infty}e^{-u^2/2}du=-\frac{\sqrt{2\pi}}{2}$, получим окончательно

$$F_N(z,0,1)=0{,}5+\Phi(z). \tag{7.35}$$

На основании (7.31) получаем

$$P(X<x)=F_N(z,0,1)=0{,}5+\Phi\left(\frac{x-\overline{X}}{\sigma}\right). \tag{7.36}$$

Соответственно (7.33) запишется в виде

$$P(x_1<X<x_2)=\Phi\left(\frac{x_2-\overline{X}}{\sigma}\right)-\Phi\left(\frac{x_1-\overline{X}}{\sigma}\right). \tag{7.37}$$

Если границы x_1 и x_2 симметрично располагаются относительно $\overline{X}$, т. е. $x_1=\overline{X}-\varepsilon$ и $x_2=\overline{X}+\varepsilon$, то неравенство $\overline{X}-\varepsilon<X<\overline{X}+\varepsilon$ можно записать более компактно в виде $|X-\overline{X}|<\varepsilon$. Заменив в правой части (7.37) $\Phi\left(\frac{x_1-\overline{X}}{\sigma}\right)=\Phi\left(\frac{\varepsilon}{\sigma}\right)$ и $\Phi\left(\frac{x_1-\overline{X}}{\sigma}\right)=\Phi\left(\frac{-\varepsilon}{\sigma}\right)=-\Phi\left(\frac{\varepsilon}{\sigma}\right)$, получим

$$P\left(|X-\overline{X}|<\varepsilon\right)=2\Phi\left(\frac{\varepsilon}{\sigma}\right). \tag{7.38}$$

Последнему равенству можно придать другой вид, выразив величину отклонения ε в единицах, кратных σ, т. е. подставив $\varepsilon=z\sigma$:

$$P\left(|X-\overline{X}|<z\sigma\right)=2\Phi(z).$$

При $z = 3$, учитывая, что $\Phi(3) = 0{,}4986$, найдем

$$P(|X - \overline{X}| < 3\sigma) = 0{,}9972 \approx 1.$$

Это равенство может быть сформулировано в виде *правила трех сигм*: практически достоверно (т. е. с вероятностью $P = 0{,}9972 \approx 1$), что отклонение нормально распределенной случайной величины от ее математического ожидания не превысит по абсолютной величине трех сигм.

Соответственно отклонения, не превышающие 2σ, гарантируются с вероятностью $P = 0{,}9544$ и не превышающие σ — с вероятностью $P = 0{,}6826$. Наконец, можно определить вероятность того, что случайная величина примет значение, превышающее данное:

$$P(X \geqslant x) = 1 - P(X < x) = 0{,}5 - \Phi\left(\frac{x - \overline{X}}{\sigma}\right). \qquad (7.39)$$

Пусть размер изготовленной детали есть случайная величина, распределенная нормально со значениями $\overline{X} = 10$ см и $\sigma = 0{,}01$ см. Определим: 1) вероятность того, что взятая наудачу деталь окажется годной, если допускаются отклонения от 10 см не более чем на 0,025 см; 2) вероятность того, что размер детали не превысит 10,03 см; 3) какова должна быть точность изготовления детали, характеризуемая величиной σ, чтобы при допуске $\pm 0{,}02$ см годность детали гарантировалась с вероятностью 0,99?

1. Определим $P(|X - \overline{X}| < 0{,}025)$. Используя (7.39), при $\varepsilon = 0{,}025$ получим

$$P(|X - \overline{X}| < 0{,}025) = 2\Phi\left(\frac{0{,}025}{0{,}01}\right) = 2\Phi(2{,}5) = 0{,}9876.$$

2. На основании (7.37) найдем

$$P(X < 10{,}03) = 0{,}5 + \Phi\left(\frac{10{,}03 - 10}{0{,}01}\right) = 0{,}9984$$

3. Согласно условию

$$P(|X - 10| < 0{,}02) = 2\Phi\left(\frac{0{,}02}{\sigma}\right) = 0{,}99.$$

Из таблицы $\Phi(z)$ находим $z = \dfrac{0{,}02}{\sigma} = 2{,}58$, откуда $\sigma = \dfrac{0{,}02}{2{,}58} = 0{,}0077$.

7.8. Распределения, связанные с нормальным распределением

Несмотря на широкое распространение нормального распределения, в некоторых случаях при построении статистических моделей возникает необходимость в использовании других распределений. В настоящем параграфе рассмотрены распределения, связанные с нормальным и представляющие собой распределение некоторых функций от нормально распределенной величины.

Логарифмически-нормальное распределение. Пусть X — нормально распределенная случайная величина с плотностью распределения $f_N(x, \overline{X}, \sigma) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\overline{X})^2}{2\sigma^2}}$. Рассмотрим новую неотрицательную величину $Y = e^X$, откуда $X = \ln Y$. Из

(4.16) можно получить плотность распределения величины Y:

$$f_l(y, \overline{X}, \sigma) = \frac{1}{y\sigma\sqrt{2\pi}} e^{\frac{-(\ln y - \overline{X})^2}{2\sigma^2}}. \qquad (7.40)$$

Это распределение получило название *логарифмически-нормального* или сокращенно *логнормального*. График логнормального распределения при различных значениях $\overline{X}$ и σ показан на рис. 7.9. Как видно из рисунка, это распределение существенно асимметричное (коэффициент асимметрии α растет с ростом σ).

В основе применения логнормального распределения лежит предельная теорема, согласно которой распределение произведения n независимых положительных случайных величин при условии

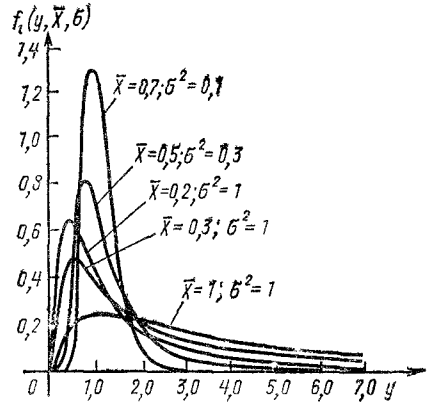


Рис. 7.9.
Логнормальное распределение

их равномерной малости в пределе стремится к логнормальному распределению. Практически логнормальное распределение используется для описания совокупного мультипликативного действия многих случайных факторов, когда их влияние на изменение конечного результата примерно пропорционально изменению самих факторов. Свидетельством близости распределения к логнормальному является значительная асимметрия, обусловленная ограничением на случайную величину слева от нуля. После логарифмирования левая часть распределения значительно растягивается, приближаясь к нормальному.

Логнормальное распределение используется, в частности, для описания распределения доходов, банковских вкладов, месячной заработной платы, посевных площадей на разные культуры и т. д. Поскольку функция распределения

$$F_l(y, \overline{X}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^y e^{\frac{-(\ln u - \overline{X})^2}{2\sigma^2}} \frac{du}{u} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(v-\overline{X})^2}{2\sigma^2}} dv =$$

$$= F_N(x, \overline{X}, \sigma)$$

совпадает с функцией нормального распределения для $X = \ln Y$, то расчеты на базе логнормального распределения могут выполняться с помощью той же таблицы функции Лапласа.

Иногда используется некоторая модификация логнормального распределения:

$$f(y, \overline{X}, \sigma, \alpha) = \frac{1}{(y-\alpha)\,\sigma\sqrt{2\pi}}\, e^{\frac{-(\ln(y-\alpha)-\overline{X})^2}{2\sigma^2}}, \qquad (7.41)$$

включающая еще один параметр α и благодаря этому представляющая собой более гибкую модель для описания случайного про-

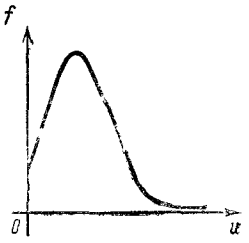


Рис. 7.10.
Полунормальное распределение

цесса. Так, в частности, здесь уже y изменяется не от 0 до ∞, а от α до ∞.

Полунормальное распределение. Если X—нормально распределенная случайная величина, то распределение абсолютной величины ее отклонения от средней, т. е. величины $u = |X - \overline{X}| \geqslant 0$, будет задаваться плотностью

$$f(u, \sigma) = \frac{2}{\sigma\sqrt{2\pi}}\, e^{-\frac{u^2}{2\sigma^2}} \qquad (7.42)$$

и называться *полунормальным распределением* (рис. 7.10).

χ^2-распределение. Это распределение также связано с нормальным и широко используется при решении различных задач статистического анализа. В основе образования распределения лежит рассмотрение случайной выборки из нормально распределенной совокупности. Пусть имеется некоторая нормально распределенная совокупность, характеризуемая плотностью $f_N(x, \overline{X}, \sigma)$. Образуем из нее возвратную выборку $(X_1, \ldots, X_j, \ldots, X_n)$, которую можно рассматривать как совокупность n одинаково распределенных независимых случайных величин с плотностью распределения для каждой $f_N(x, \overline{X}, \sigma)$.

Можно показать, что случайная величина

$$\chi^2 = \sum_{j=1}^{n} \frac{(X_j - \overline{X})^2}{\sigma^2} \qquad (7.43)$$

подчиняется закону распределения, задаваемому плотностью вероятности

$$f_{\chi^2}(x, n) = Ke^{-x/2} x^{\frac{n}{2}-1}, \quad x > 0, \qquad (7,44)$$

получившему название χ^2-распределения (хи-квадрат распределения).

Коэффициент K определяется из условия нормируемости:

$$K = 1 : \int_0^\infty e^{-u/2} u^{n/2-1}\, du.$$

Заменив в (7.43) $\Sigma (X_j - \bar{X})^2$ на $n\sigma^{*2}$, где σ^{*2} — выборочная дисперсия, получим выражение

$$\chi^2 = \frac{\sigma^{*2}}{\sigma^2} n. \qquad (7.45)$$

Оказывается, что то же распределение, зависящее от одного параметра n, будет иметь сумма $\sum \frac{(X_j - \bar{X}_j)^2}{\sigma_j^2}$, где X_j — попарно независимые нормально распределенные случайные величины, имеющие различные математические ожидания $\bar{X}_j$ и дисперсии σ_j^2. Сравнив

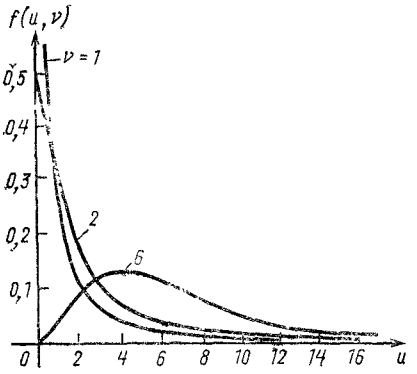


Рис. 7.11.
χ^2-распределение

плотность распределения (7.44) с ранее приведенной плотностью (7.13), заключаем, что данное распределение представляет собой частный случай гамма-распределения при $\beta = \frac{n}{2}$ и $\lambda = \frac{1}{2}$.

Наконец, если случайные величины X_j удовлетворяют s линейным зависимостям, то распределение суммы квадратов их нормированных отклонений будет подчиняться аналогичному закону распределения:

$$f_{\chi^2}(x, \nu) = K e^{x/2} x^{\nu/2-1}, \qquad (7.46)$$

где параметр $\nu = n - s$ есть *число степеней свободы* системы случайных величин X_j.

Рассчитав первые два момента χ^2-распределения, получим:

$$M(\chi^2) = \nu; \quad D(\chi^2) = 2\nu.$$

Как видно из графика χ^2-распределения, показанного на рис. 7.11, оно асимметрично, имеет правый удлиненный «хвост». При неограниченном возрастании n (или ν) асимметричность рас-

пределения уменьшается и оно стремится к нормальному. Как и нормальное, оно является аддитивным, т. е. если имеются две величины χ_1^2 и χ_2^2 с плотностями (7.46) и степенями свободы ν_1 и ν_2, то величина $\chi_1^2+\chi_2^2$ имеет то же распределение с параметром $\nu=\nu_1+\nu_2$. При $\nu\geqslant 30$ распределение χ^2 близко к стандартному нормальному распределению величины $z=2\chi^2-\sqrt{2\nu}$. Таблица функции распределения χ^2

$$F_{\chi^2}(x,\nu)=P(\chi^2<x)=\int\limits_{-\infty}^{x} f(u,\nu)\,du \tag{7.47}$$

приведена в [12].

Чаще всего это распределение используется для определения критического значения $x=\chi_\alpha^2$ по уровню значимости α из равенства

$$P(\chi^2\geqslant\chi_\alpha^2)=1-F_{\chi^2}(\chi_\alpha^2,\nu)=\alpha. \tag{7.48}$$

Значения $\chi_{0,05}^2$ и $\chi_{0,01}^2$ для 5- и 1%-ных уровней значимости α и значений ν от 1 до 43 приведены в табл. 7.3.

Таблица 7.3

ν	1	3	5	7	9	11	13	15	17	19	21
$\chi_{0,05}^2$	3,8	7,8	11,0	14,1	16,9	19,7	22,4	25,0	27,6	30,1	32,7
$\chi_{0,01}^2$	6,6	11,4	15,1	18,5	21,7	24,7	27,7	30,6	33,4	36,2	38,9
ν	23	25	27	29	31	33	35	37	39	41	43
$\chi_{0,05}^2$	35,2	37,7	40,1	42,6	45,0	47,4	49,8	52,2	54,6	56,9	59,3
$\chi_{0,01}^2$	41,6	44,3	47,0	49,6	52,2	54,8	57,3	59,9	62,4	64,9	67,5

Распределение Стьюдента. Пусть Z — стандартная нормально распределенная случайная величина и U — независимая от нее случайная величина, имеющая χ^2-распределение с ν степенями свободы. Тогда величина

$$T=\frac{Z\sqrt{\nu}}{U} \tag{7.49}$$

подчиняется *t-распределению*, или *распределению Стьюдента*, с плотностью вероятности [см. 3]:

$$f_s(t,\nu)=B\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{7.50}$$

где B, как обычно, определяется из условия нормируемости *.

Пользуясь определениями математического ожидания и дисперсии, можно получить [3]:

$$M(T)=0; \quad D(T)=\frac{\nu}{\nu-2}. \tag{7.51}$$

Распределение Стьюдента зависит от одного параметра ν — числа степеней свободы. Оно симметрично, как и нормальное рас-

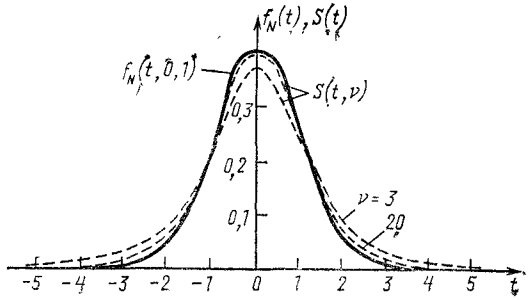


Рис. 7.12.
Распределение Стьюдента

пределение (рис. 7.12), но более пологое (т. е. с более значительными «хвостами»). С ростом числа степеней свободы ν распределение Стьюдента быстро приближается к нормальному и потому при больших ν (практически начиная с $\nu=20$) можно определять вероятность $P(T<t)$ с помощью функции Лапласа:

$$P(T<t)=S(t,\nu)=B\int_{-\infty}^{t}\left(1+\frac{v^2}{\nu}\right)^{-\frac{\nu+1}{2}}dv\approx 0{,}5+\Phi(t). \tag{7.52}$$

Если $\bar{X}^*$ и σ^{*2} — средняя и дисперсия выборки объемом n из нормально распределенной генеральной совокупности со средней $\bar{X}$, то величина

$$T=\frac{\bar{X}^*-\bar{X}}{\sigma^*}\sqrt{n-1} \tag{7.53}$$

имеет t-распределение с $\nu=n-1$ числом степеней свободы.

Распределение Фишера—Снедекора. Пусть U и V — независимые случайные величины, имеющие χ^2-распределения со степенями свободы ν_1 и ν_2. Тогда распределение величины

$$F=\frac{U}{\nu_1}:\frac{V}{\nu_2} \tag{7.54}$$

* Величина B определяется из равенства

$$\int_{-\infty}^{\infty}f_S(v,\nu)\,dv=B\int_{-\infty}^{\infty}\left(1+\frac{v^2}{\nu}\right)^{-\frac{\nu+1}{2}}dv=1.$$

зависит только от двух параметров — степеней свободы ν_1 и ν_2 и задается плотностью вероятности

$$f_F(x, \nu_1, \nu_2) = Cx^{\frac{\nu_1}{2}-1}\left(1 + \frac{\nu_1 x}{\nu_2}\right)^{-\frac{\nu_1+\nu_2}{2}}, \qquad (7.55)$$

где коэффициент C определяется из условия нормируемости. Это распределение получило название *F-распределения* или *распределения Фишера — Снедекора.* В частности, F-распределению подчиняется отношение дисперсий двух выборок объемом n_1 и n_2 из

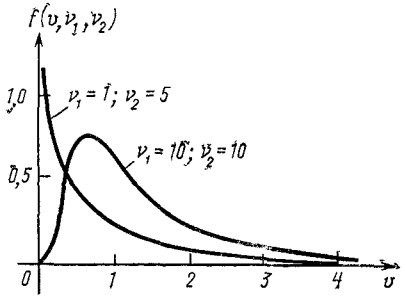


Рис. 7.13.
F-распределение

двух нормально распределенных генеральных совокупностей с равными дисперсиями. В этом случае параметры $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$. Значения степеней свободы будем указывать в индексе величины F в виде $F_{\alpha,\, \nu_1,\, \nu_2}$.

Таблица значений $F_{\alpha,\, \nu_1,\, \nu_2}$ дана например в [12]. На рис. 7.13 приведен график F-распределения при $\nu_1 = \nu_2 = 10$. График несимметричный; удлиненный вправо. Для определения значений F_α при α, близком к 1, можно воспользоваться соотношением

$$F_{\alpha,\, \nu_1,\, \nu_2} = 1 : F_{1-\alpha,\, \nu_1, \nu_2}. \qquad (7.56)$$

При больших значениях ν_1 и ν_2 распределение приближается к нормальному. Это позволяет находить значения F_α по заданному уровню значимости α с помощью приближенной формулы:

$$F_\alpha \approx e^{z_\alpha \sqrt{\frac{2(\nu_1+\nu_2)}{\nu_1\nu_2}}}, \qquad (7.57)$$

где z_α — нормированная нормально распределенная величина, определяемая из равенства

$$P(Z > z_\alpha) = 0{,}5 - \Phi(z_\alpha) = \alpha. \qquad (7.58)$$

7.9. Двумерное нормальное распределение

Распределение двумерной случайной величины (X, Y), задаваемое плотностью вероятности

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}}\, e^{-\frac{\bar{x}^2+\bar{y}^2-2r\bar{x}\bar{y}}{2(1-r^2)}}, \qquad (7.59)$$

где σ_x и σ_y — стандарты случайных величин X и Y; r — коэффициент корреляции; $\tilde{x} = \frac{x - \overline{X}}{\sigma_x}$, $\tilde{y} = \frac{y - \overline{Y}}{\sigma_y}$ — нормированные отклонения, называется *двумерным нормальным распределением*. Это распределение зависит от пяти параметров $\overline{X}$, $\overline{X}$, σ_x, σ_y, r. Графически оно изображается поверхностью, называемой иногда «палаткой Гаусса» (см. рис. 6.5).

Сечениями этой поверхности плоскостями, параллельными плоскости XOY, будут эллипсы $\tilde{x}^2 - 2r\tilde{x}\tilde{y} + \tilde{y}^2 = c$, во всех точках $(\tilde{x}, \tilde{y})$ которых плотность вероятности постоянна и равна $f(\tilde{x}, \tilde{y}) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} e^{-\frac{c}{2(1-r^2)}}$. Эти эллипсы — эллипсы рассеяния подобны и имеют общий центр $(\overline{X}, \overline{Y})$, называемый *центром распределения*.

Безусловные распределения одномерных составляющих

$$f_1(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{\tilde{x}^2}{2}}; \quad f_2(y) = \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{\tilde{y}^2}{2}} \tag{7.60}$$

представляют собой одномерные нормальные распределения.

Ранее отмечалось (см. § 6.7), что некоррелированность случайных величин ($r_{xy} = 0$) является только необходимым, но недостаточным условием их независимости. Однако для нормального распределения эти понятия совпадают, т. е. имеет место следующая теорема.

Теорема. Если величины X и Y, подчиняющиеся двумерному нормальному распределению, некоррелированны, т. е. $r = 0$, то они независимы.

Действительно, положив в (7.59) $r = 0$, получим

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{\tilde{x}^2}{2} - \frac{\tilde{y}^2}{2}} = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\tilde{x}^2} \cdot \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{\tilde{y}^2}{2}} = f_1(x) \cdot f_2(y). \tag{7.61}$$

Но это соотношение, как следует из определения (см. § 6.7), является условием стохастической независимости. В данном случае оси эллипсов рассеяния будут параллельны осям координат.

Вернемся к общему двумерному нормальному распределению. Условные распределения X при $Y = y$ и Y при $X = x$ будут задаваться плотностями вероятностей

$$f_y(x) = \frac{1}{\sigma_x\sqrt{2\pi}\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}(\tilde{x} - r\tilde{y})^2};$$
$$f_x(y) = \frac{1}{\sigma_y\sqrt{2\pi}\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}(\tilde{y} - r\tilde{x})^2}. \tag{7.62}$$

Записав первое из этих распределений в виде

$$f_y(x) = \frac{1}{\sigma_x\sqrt{1-r^2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2(1-r^2)\sigma_x^2}\left[x - \overline{X} - \frac{\sigma_x}{\sigma_y} r(y - \overline{Y})\right]^2},$$

заключаем, что оно представляет собой нормальное распределение со средней

$$\overline{X}_y = \overline{X} + \frac{\sigma_x}{\sigma_y} r (y - \overline{Y}) \tag{7.63}$$

и дисперсией

$$\sigma_y^2 (X) = \sigma_x^2 (1 - r^2). \tag{7.64}$$

Аналогично получим:

$$\overline{Y}_x = \overline{Y} + \frac{\sigma_y}{\sigma_x} r (x - \overline{X}); \tag{7.63'}$$

$$\sigma_x^2 (Y) = \sigma_y^2 (1 - r^2). \tag{7.64'}$$

Таким образом, если система двух случайных величин (X, Y) подчиняется двумерному нормальному распределению, то между X и Y имеется линейная корреляционная зависимость, задаваемая уравнениями прямой регрессии (7.63) и (7.63').

Глава 8

Закон больших чисел

8.1. Сущность закона больших чисел

Говоря о статистических закономерностях и о возможностях их теоретико-вероятностного моделирования, мы ссылались на установленный многолетним опытом факт объективного существования этих закономерностей, проявляющихся, в частности, в устойчивости относительной частоты при многократном повторении испытаний. На этой основе было введено фундаментальное понятие вероятности. Теперь можно проверить, как согласуется исходная предпосылка об устойчивости относительной частоты с выводами, которые можно получить из теоретико-вероятностной модели. Рассмотрим ряд теорем, устанавливающих условия возникновения соответствующей статистической закономерности и подтверждающих адекватность модели исследуемому явлению. Статистическая закономерность проявляется при многократном повторении опыта или в математической форме в виде предельных свойств при неограниченном увеличении числа испытаний. Поэтому соответствующие утверждения и получили название теорем *закона больших чисел*.

Мы упомянули простейший вид статистической закономерности — *устойчивость относительной частоты*. Другой вид статистической закономерности — *устойчивость средней*. Логическим основанием для нее является следующий ход рассуждений. При каждом отдельном испытании его результат, характеризуемый некоторым показателем, под влиянием случайных обстоятельств будет отклоняться от некоторого постоянного значения в ту и другую сторону. Поэтому среднее значение показателя при достаточно

большом числе испытаний из-за взаимного погашения индивидуальных отклонений приобретает определенную устойчивость, стремясь к этому постоянному числу, т. е. практически не зависит от случая.

Однако значение теорем закона больших чисел заключается не только в возможности проверки согласованности теоретико-вероятностной модели с исследуемым явлением. Установление факта статистической устойчивости лежит в основе статистических выводов и обобщений, позволяет на базе изучения всегда ограниченного числа опытных данных судить о поведении исследуемого явления в других, пока не наблюдаемых, но сходных ситуациях, сформулировать статистические закономерности, которым эти явления подчиняются. Так, регистрируя за достаточно длительный промежуток времени относительную частоту брака на определенной операции, установив ее устойчивость и переходя к соответствующей вероятностной модели — вероятности брака p, можно на основании соответствующих теорем (см. § 8.4) предсказать, если сохранятся неизменными условия выполнения этой операции, относительную частоту брака в будущем, среднее ожидаемое число бракованных изделий в партии из n единиц, возможные отклонения от средней и т. д.

Таким образом, закон больших чисел приводит к установлению детерминированных закономерностей в поведении относительной частоты, средней или других итоговых показателей, характеризующих результат достаточно большого числа испытаний при случайности, а следовательно, и непредсказуемости результата каждого испытания в отдельности.

Разные теоремы закона больших чисел отличаются исходными вероятностными моделями и различными формами статистической устойчивости.

Исторически первой формой закона больших чисел была теорема Бернулли об устойчивости относительной частоты в модели повторных независимых испытаний. В последующем она была обобщена на случай испытаний с изменяющейся вероятностью p_i, где i — номер испытания. Значительно более общие формы закона больших чисел были доказаны П. Л. Чебышевым, А. А. Марковым, А. Я. Хинчиным и др.

Кроме исследования такой формы статистических закономерностей, как устойчивость относительной частоты или средней, возникает задача изучения закономерностей, которым подчиняются суммы большого числа случайных слагаемых. Оказывается, что при некоторых условиях и неограниченном возрастании числа слагаемых их сумма подчиняется в пределе определенному закону распределения. Такая форма закона больших чисел получила название *центральной предельной теоремы*. Исследования в этом направлении, начатые еще П. Лапласом, были значительно продвинуты в работах А. М. Ляпунова (ему принадлежит одна из наиболее общих предельных теорем), А. А. Маркова, С. Н. Бернштейна, А. Н. Колмогорова и др.

8.2. Неравенства Чебышева

Мы знаем из предыдущего, что по заданному закону распределения случайной величины могут быть определены ее характеристики $M(X)$, $D(X)$ и др. С другой стороны, неоднократно подчеркивалось, что знание числовых характеристик если не полностью, то в некотором смысле характеризует закон распределения. Целью рассматриваемых далее двух неравенств и является оценка закона распределения по заданным числовым характеристикам.

Таблица 8.1

X	x_1	x_2	...	x_k	x_{k+1}	...	$\sum$
P	p_1	p_2	...	p_k	p_{k+1}	...	1

Лемма. Если случайная величина X принимает неотрицательные значения ($X \geqslant 0$) и имеет математическое ожидание * $M(X) = \overline{X}$, то выполняется неравенство

$$P\{X < \alpha\} \geqslant 1 - \frac{\overline{X}}{\alpha}, \tag{8.1}$$

где $\alpha > 0$ — любое заданное положительное число.

Пусть X — дискретная случайная величина, заданная распределением, указанным в табл. 8.1. Выберем некоторое число $\alpha > 0$ и допустим, что в данном ряду $x_1 < x_2 < \ldots < x_k < \alpha$ и $x_{k+1} \geqslant \alpha$.

Согласно определению

$$\overline{X} = \sum_{i=1}^{k} x_i p_i + \sum_{i=k+1} x_i p_i.$$

Так как по условию все $x_i \geqslant 0$ и $p_i \geqslant 0$, то, отбросив первую сумму и заменив во второй все x_i на меньшую величину α, получим неравенство

$$\overline{X} \geqslant \alpha \sum_{i=k+1} p_i \quad \text{или} \quad \sum_{i=k+1} p_i \leqslant \frac{\overline{X}}{\alpha}.$$

Но $\sum\limits_{i=k+1} p_i = 1 - \sum\limits_{i=1}^{k} p_i$, откуда $\sum\limits_{i=1}^{k} p_i \geqslant 1 - \frac{\overline{X}}{\alpha}$.

С другой стороны, по условию

$$P(X < \alpha) = P(X = x_1) + \ldots + P(X = x_k) = \sum_{i=1}^{k} p_i$$

---

* Условие существования математического ожидания не тривиальное, так как дискретная случайная величина с бесконечным множеством значений или непрерывная случайная величина могут и не иметь математического ожидания, если соответственно ряд $\sum\limits_{i=1}^{\infty} x_i p_i$ или $\int\limits_{-\infty}^{\infty} x f(x)\,dx$ расходятся.

Используя только что полученное неравенство, получим

$$P(X<\alpha)\geqslant 1-\frac{\overline{X}}{\alpha}.$$

Неравенство (8.1) доказано в предположении, что α — некоторое число, относительно которого ряд делится на две части: $x_1<\alpha, \ldots, x_k<\alpha$ и $x_{k+1}\geqslant\alpha, \ldots$ Если α лежит левее начала ряда, т. е. уже $x_1\geqslant\alpha$, то $P(X<\alpha)=0$. В этом случае неравенство (8.1) становится тривиальным, так как $\overline{X}\geqslant\alpha$, $1-\frac{\overline{X}}{\alpha}\leqslant 0$ и, очевидно, $P\geqslant 1-\frac{\overline{X}}{\alpha}$. Аналогично если $x_k<\alpha$ при любом k, то $\overline{X}<\alpha$ и $P(X<\alpha)=1>1-\frac{\overline{X}}{\alpha}$. Пусть теперь X — непрерывная случайная величина с плотностью вероятности $f(x)$. Выполняя преобразования, аналогичные предыдущим, получим последовательно

$$\overline{X}=\int_0^\infty xf(x)\,dx=\int_0^\alpha xf(x)\,dx+\int_\alpha^\infty xf(x)\,dx\geqslant\alpha\int_\alpha^\infty f(x)\,dx,$$

откуда

$$\int_\alpha^\infty f(x)\,dx\leqslant\frac{\overline{X}}{\alpha}\quad\text{или}\quad 1-\int_0^\alpha f(x)\,dx\leqslant\frac{\overline{X}}{\alpha}.$$

Так как $P(X<\alpha)=\int_0^\alpha f(x)\,dx$, то приходим к тому же неравенству (8.1).

Заметим, что вероятность $P(X<\alpha)$ по определению равна функции распределения $F(x)$. Таким образом, из (8.1) получаем

$$F(x)\geqslant 1-\frac{\overline{X}}{\alpha}.$$

С другой стороны, по определению $F(x)$ имеем $F(\alpha)\leqslant 1$, в результате чего получаем двустороннее неравенство

$$1-\frac{\overline{X}}{\alpha}\leqslant F(x)\leqslant 1. \tag{8.2}$$

Это неравенство можно рассматривать как оценку неизвестной функции распределения $F(\alpha)$. Сказанное иллюстрируется рис. 8.1, на котором построены графики произвольной функции распределения $y=F(\alpha)$ и функции $y=1-\frac{\overline{X}}{\alpha}$. При $\alpha<\overline{X}$ значение $1-\frac{\overline{X}}{\alpha}<0$ и неравенство (8.2) дает тривиальную оценку, так как $F(\alpha)\geqslant 0$. Однако с ростом α значение $1-\frac{\overline{X}}{\alpha}\to 1$ и оценка (8.2) становится достаточно хорошей.

Известно, что средний размер вклада в сберкассу равен 200 руб. Определить вероятность того, что взятый наудачу лицевой счет будет иметь размер вклада, меньший 1000 руб.

Обозначим размер вклада на взятом наудачу лицевом счете (случайную величину) через X. По условию $X > 0$, $\bar{X} = 200$ руб. и $\alpha = 1000$. Тогда, применяя оценку (8.1), получим $P(X < 1000) \geqslant 1 - \frac{200}{1000} = 0{,}8$.

Теорема. Если X — произвольная случайная величина, имеющая математическое ожидание $\bar{X}$ и дисперсию $D(X)$, то для любого положительного числа ε справедливо соотношение, известное под названием неравенства Чебышева:

$$P\{|X - \bar{X}| < \varepsilon\} \geqslant 1 - \frac{D(X)}{\varepsilon^2}. \qquad (8.3)$$

Неравенство, заключенное в скобках, равносильно, а следовательно, и равновероятно неравенству $(X - \bar{X})^2 < \varepsilon^2$, т. е. $P\{|X - \bar{X}| < \varepsilon\} = P\{|X - \bar{X}|^2 < \varepsilon^2\}$. Так как случайная величина $(X - \bar{X})^2 > 0$ и $M(X - \bar{X})^2 = D(X)$, то применяя к ней неравенство (8.1), получим неравенство (8.3).

Это неравенство подтверждает выбор дисперсии в качестве меры рассеяния значений случайной величины относительно ее

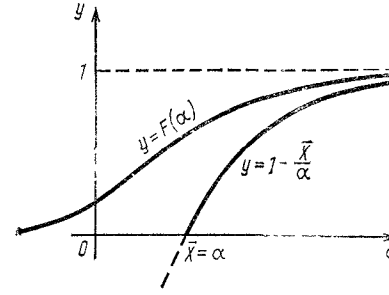


Рис. 8.1.
Ограничение (оценка сверху) функции распределения, даваемое леммой Чебышева

математического ожидания. Действительно, как видно из (8.3), чем меньше величина $D(X)$, тем при меньшем значении ε можно с той же степенью достоверности (оценкой вероятности P) утверждать выполнение неравенства $|X - \bar{X}| < \varepsilon$, т. е. значения X будут группироваться ближе к $\bar{X}$.

Если бы была известна функция распределения $F(x) = P(X < x)$, то оцениваемую неравенством (8.3) вероятность можно было бы найти точно. Действительно, неравенство $|X - \bar{X}| < \varepsilon$ равносильно двустороннему неравенству $\bar{X} - \varepsilon < X < \bar{X} + \varepsilon$. Используя свойство 1 функции распределения (см. § 4.2) получим

$$P\{|X - \bar{X}| < \varepsilon\} = F(\bar{X} + \varepsilon) - F(\bar{X} - \varepsilon).$$

Из неравенства (8.3), положив $\varepsilon = 3\sigma$, получим *

$$P\{|X - \bar{X}| < 3\sigma\} \geqslant 1 - \frac{\sigma^2}{9\sigma^2} = 0{,}8889.$$

* Заметим, что для нормального распределения мы получили $P\{|X - \bar{X}| < 3\sigma\} = 0{,}9973$, т. е. число, отличное от 1 на 0,003.

Таким образом, с достаточно близкой к единице вероятностью $P = 0{,}89$ можно утверждать, что отклонение случайной величины от ее математического ожидания будет по абсолютной величине меньше, чем 3σ.

При изготовлении партий одинаковых деталей должен быть выдержан некоторый размер $a = 10$ мм с допуском $\pm 0{,}1$ мм. Оценить вероятность того, что взятая наудачу деталь окажется бракованной, если разброс размеров характеризуется дисперсией $\sigma^2 = 0{,}0025$.

Обозначим размер взятой наудачу детали через X. По условию деталь окажется бракованной, если $|X - 10| > 0{,}1$. Следовательно, необходимо определить $P\{|X - 10| > 0{,}1\}$. Из (8.3) находим $P\{|X - 10| < 0{,}1\} \geqslant 1 - \dfrac{0{,}0025}{0{,}01} = 0{,}75$.

Теперь можно оценить искомую вероятность как вероятность противоположного события $P\{|X - 10| \geqslant 0{,}1\} \leqslant 1 - 0{,}75 = 0{,}25$. Следовательно, искомая вероятность не превысит 0,25.

8.3. Теорема Чебышева

Пусть дана система попарно независимых случайных величин $X_1, X_2, \ldots, X_j, \ldots, X_n$, имеющих математические ожидания $M(X_j) = \bar{X}_j$ и дисперсии $D(X_j) = D_j$. Образуем новую случайную величину, равную средней арифметической данных случайных величин:

$$\overline{X}_a = \frac{X_1 + X_2 + \ldots + X_j + \ldots + X_n}{n}.$$

В соответствии с теоремами о математическом ожидании и дисперсии суммы будем иметь:

$$M(\overline{X}_a) = \frac{1}{n}\sum_{j=1}^{n} \bar{X}_j; \quad D(\overline{X}_a) = \frac{1}{n^2}\sum_{j=1}^{n} D_j.$$

Применяя неравенство (8.3), получим

$$P\left\{\left|\frac{\sum_j X_j}{n} - \frac{\sum_j \bar{X}_j}{n}\right| < \varepsilon\right\} \geqslant 1 - \frac{\sum_j D_j}{n^2\varepsilon^2}. \tag{8.4}$$

Теперь допустим, что все дисперсии равномерно ограничены, т. е. существует некоторое постоянное число C, для которого выполняются неравенства $D_j < C$ при $j = 1, 2, \ldots, n$. Заменив в неравенстве D_j на C, мы этим самым только усилим его. Переходя после этого к пределу при $n \to \infty$ и фиксированном значении $\varepsilon > 0$, получим

$$\lim_{n\to\infty} P\left\{\left|\frac{\sum_j X_j}{n} - \frac{\sum_j \bar{X}_j}{n}\right| < \varepsilon\right\} \geqslant \lim_{n\to\infty}\left(1 - \frac{Cn}{n^2\varepsilon^2}\right) = 1.$$

Но вероятность $P \leqslant 1$, следовательно, и $\lim P \leqslant 1$. Сопоставив это неравенство с предыдущим, получим окончательно предельное равенство

$$\lim_{n\to\infty} P\left\{\left|\frac{\sum X_j}{n} - \frac{\sum \bar{X}_j}{n}\right| < \varepsilon\right\} = 1, \tag{8.5}$$

составляющее содержание следующей фундаментальной теоремы Чебышева.

Теорема 1. Закон больших чисел в форме Чебышева. При неограниченном увеличении числа n попарно независимых случайных величин, имеющих математические ожидания и равномерно ограниченные дисперсии, их средняя арифметическая стремится по вероятности к средней арифметической их математических ожиданий.

Это утверждение теоремы условно записывают в виде

$$\frac{1}{n}\sum_j X_j \xrightarrow{P} \frac{1}{n}\sum_j \overline{X}_j. \qquad (8.5')$$

Поясним смысл формулировки теоремы и соответственно запись ее содержания в виде (8.5)′. Понятие предела переменной величины ($\lim X = a$ или $X \to a$) означает выполнение неравенства $|X - a| < \varepsilon$ для любого $\varepsilon > 0$ начиная с некоторого момента изменения X. В фигурных скобках (8.5) содержится аналогичное неравенство $\left|\frac{\sum X_j}{n} - \frac{\sum \overline{X}_j}{n}\right| < \varepsilon$, где $\frac{\sum X_j}{n}$ является случайной переменной и $\frac{\sum \overline{X}_j}{n}$ — постоянное число. Однако из (8.5) отнюдь не следует, что это неравенство будет выполняться всегда начиная с некоторого момента изменения $\frac{\sum X_j}{n}$. Так как переменная $\frac{\sum X_j}{n}$ случайная величина, то не исключено, что в отдельных случаях может наблюдаться и неравенство противоположного смысла. Однако с увеличением числа n вероятность неравенства, указанного в скобках, стремится к 1, т. е. будет выполняться в *подавляющем числе случаев*, и, наоборот, вероятность неравенства противоположного смысла стремится к нулю, т. е. оно может считаться *практически невозможным.*

Таким образом, стремление $\frac{\sum X_j}{n}$ к $\frac{\sum \overline{X}_j}{n}$ следует понимать не как категорическое утверждение, а как утверждение, верность которого гарантируется с вероятностью, сколь угодно близкой к 1 при $n \to \infty$. Это обстоятельство и отражается в тексте теоремы словами «*стремится по вероятности*», а в записи (8.5′) — буквой «P», поставленной над стрелкой.

Пусть X_j — случайный результат j-го опыта. Тогда $\frac{\sum X_j}{n} = \overline{X}_a$ — средняя арифметическая результатов n опытов, а $\frac{\sum \overline{X}_j}{n}$ — некоторая постоянная (не случайная) величина. Рассмотрим несколько серий по n опытов в каждой. Утверждение теоремы означает, что средняя арифметическая результатов серии из n опытов при $n \to \infty$ *стремится по вероятности* к постоянному числу $\frac{\sum \overline{X}_j}{n}$,

т. е обладает *статистической устойчивостью.* При этом стремление по вероятности не исключает того, что в отдельных сериях могут отмечаться значительные отклонения $\bar{X}_a$ от этой постоянной, но число таких серий незначительно и, чем больше число n, тем относительное число этих серий будет меньшим.

Читателю, привыкшему к категоричности математических утверждений, может показаться несколько необычной подобная формулировка, однако следует иметь в виду, что речь идет о случайной величине, поведение которой не может быть однозначно предсказано в каждом отдельном опыте. Лишь в массовой совокупности опытов оно проявляется в форме статистической закономерности. Именно эта закономерность в виде *устойчивости средней арифметической* и утверждается со всей определенностью в рассматриваемой теореме.

Важное значение для дальнейшего анализа имеет частный случай теоремы Чебышева.

Теорема 2. При неограниченном увеличении числа n попарно независимых случайных величин $X_j (i = 1, \ldots, n, \ldots)$, имеющих равные математические ожидания $M(X_j) = a$ и равномерно ограниченные дисперсии $D(X_j) < C$ для $j = 1, 2, \ldots$, их средняя арифметическая стремится по вероятности к числу a:

$$\lim P\left\{\left|\frac{1}{n}\sum_j X_j - a\right| < \varepsilon\right\} = 1, \text{ или } \frac{1}{n}\sum_j X_j \xrightarrow{P} a. \qquad (8.6)$$

Соотношение (8.6) получается непосредственно из (8.5), если в последнем заменить

$$\frac{1}{n}\sum M(X_j) = \frac{1}{n} \cdot na = a.$$

Доказанная теорема имеет многочисленные практические приложения, из которых приведем лишь пример из теории измерений. Пусть производится n раз измерение некоторой величины, истинное значение которой равно a. Обозначим через X_j случайный результат j-го измерения. Если измерения производятся без систематической погрешности, то естественно предположить, что $M(X_j) = a$ при любом j. Можно также полагать, что разброс результатов каждого измерения, характеризуемый величиной $D(X_j)$, не может неограниченно возрастать с ростом числа измерений, т. е. $D(X_j)$ ограничены некоторым постоянным числом C. Если измерения производятся при постоянных условиях, то все величины X_j могут рассматриваться как попарно независимые случайные величины.

Таким образом, X_j удовлетворяют условиям теоремы 2, и, следовательно, средняя арифметическая результатов n измерений стремится по вероятности к истинному значению измеряемой величины. Этим обосновывается выбор средней арифметической в качестве меры истинного значения a.

Если все измерения производятся с одной и той же точностью, характеризуемой $D(X_l) = \sigma^2$, то

$$D(\bar{X}_a) = \frac{1}{n^2} D\left(\sum X_l\right) = \frac{1}{n^2} \sum D(X_l) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n},$$

откуда

$$\sigma(\bar{X}_a) = \frac{\sigma}{\sqrt{n}}.$$

Последнее соотношение, известное под названием «правила корня из n», говорит о том, что средний ожидаемый разброс средней n измерений в $\sqrt{n}$ раз меньше разброса каждого измерения. Таким образом, можно в сколь угодное число раз уменьшить влияние случайных погрешностей, увеличивая число измерений.

Заметим, что при наличии систематической погрешности ее нельзя уменьшить путем увеличения числа измерений и $\bar{X}_a$ будет обладать той же погрешностью.

Большое значение закона больших чисел для теоретического обоснования методов математической статистики, ее приложений обусловило проведение целого ряда исследований, направленных на изучение общих условий статистической устойчивости, т. е. основного вывода закона больших чисел. Так, учеником П. Л. Чебышева А. А. Марковым была доказана теорема, утверждающая справедливость предельного равенства (8.6) для зависимых случайных величин X_l при условии $\lim\limits_{n\to\infty} \frac{1}{n^2} D\left(\sum\limits_l X_l\right) = 0$. Другое обобщение теоремы Чебышева было установлено А. Я. Хинчиным

Теорема 3. Средняя арифметическая n одинаково распределенных независимых случайных величин X_l, имеющих математические ожидания $M(X_l) = a$, при неограниченном возрастании n стремится по вероятности к числу a

Здесь условие равномерной ограниченности дисперсий заменено условием одинакового распределения всех X_l. Заметим, что если предположить существование конечных дисперсий $D(X_l) = \sigma^2$, то теорема Хинчина окажется прямым следствием теоремы Чебышева.

Все упомянутые выше теоремы содержали лишь достаточные условия, при выполнении которых последовательность случайных величин $\{X_l\}$ подчиняется закону больших чисел. Некоторый итог этим многочисленным исследованиям подведен в теоремах о необходимых и достаточных условиях подчинения закону больших чисел последовательности независимых случайных величин (теорема Колмогорова) и зависимых (теорема Гнеденко).

8.4. Закон больших чисел в модели повторных испытаний

Рассмотрим частоту $X_{(n)}$ появления события A при n повторных независимых испытаниях (модель Бернулли). Как указывалось (§ 6.9) ее можно представить в виде суммы n одинаково

распределенных и независимых случайных величин X_j, каждая из которых может принимать только значения 0 или 1 с соответствующими вероятностями $1-p=q$ и p, при этом $M(X_j)=p$ и $D(X_j)=pq$. Следовательно, относительная частота $X_{(n)}/n$ есть средняя арифметическая n одинаково распределенных величин X_j и к ней полностью применима теорема 3 предыдущего параграфа. В результате получаем следующую теорему.

Теорема 1. При неограниченном увеличении числа повторных независимых испытаний относительная частота появления события стремится по вероятности к вероятности события при одном испытании, т. е.

$$\frac{X_{(n)}}{n} \xrightarrow{P} p. \tag{8.7}$$

Эта теорема была доказана Я. Бернулли, отчего ее называют также *законом больших чисел в форме Бернулли*. По этой же причине предельное равенство (8.7) записывают также в виде

$$\frac{X_{(n)}}{n} \to p\,(m.B),$$

где $(m\,B)$ означает «modo Bernulliano», т. е. «в смысле Бернулли».

Доказанная теорема как бы подтверждает (в схеме повторных независимых испытаний) согласованность исходной посылки об устойчивости относительной частоты и о вероятности p как постоянном числе, вокруг которого она колеблется, с дедуктивными выводами, вытекающими из принятой модели. Остановимся несколько подробнее на этой теореме, проиллюстрировав еще раз на ее примере смысл закона больших чисел.

Пусть производится возвратная выборка объемом n из генеральной совокупности, доля некоторого признака A в которой равна p. Такая выборка может рассматриваться как n повторных независимых испытаний. Назовем событием A отбор элемента с признаком А. Тогда $P(A)=p$. Очевидно, что относительная частота $\frac{X_n}{n}$ элементов с признаком A в выборке (т. е. относительная частота события A) может принимать любые значения от 0 до 1, т. е. является случайной величиной, значения которой зависят от того, какие элементы случайно оказались отобранными в выборку. Например, может оказаться $\frac{X_{(n)}}{n}=1$, если каждый из отобранных элементов обладает признаком A, $\frac{X_n}{n}=0$, если отобранные в выборку элементы признаком A не обладают, и т. д. Как указывалось в § 5.3, наиболее вероятным результатом выборки является $\frac{X_{(n)}}{n}=\frac{m_0}{n}=p$. Таково значение математического ожидания относительной частоты $M\left(\frac{X_{(n)}}{n}\right)=p$. Однако для раз-

личных выборок того же объема n возможны различные отклонения $\frac{X_{(n)}}{n}$ от p в ту или другую сторону (рис. 8.2). Эти отклонения в среднем оцениваются величиной $\sigma\left(\frac{X_{(n)}}{n}\right)=\sqrt{\frac{pq}{n}}$.

Теорема утверждает, что с практической достоверностью эти отклонения при достаточно большом n будут сколь угодно ма-

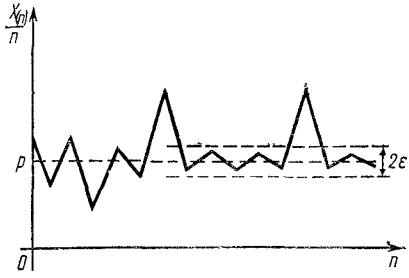


Рис. 8.2.
Колебания относительной частоты в последовательности серий из n испытаний

лыми, т. е. в серии выборок большого объема *почти всегда* (с *практической достоверностью*) будет наблюдаться сколь угодно малое отклонение $\frac{X_{(n)}}{n}$ от p.

Непосредственным обобщением теоремы Бернулли является теорема Пуассона (*закон больших чисел в форме Пуассона*), допускающая в отличие от теоремы Бернулли зависимость вероятности наступления события $P(A)=p_j$ от номера испытания j.

Теорема 2. При неограниченном увеличении числа испытаний относительная частота $\frac{X_{(n)}}{n}$ появления события с практической достоверностью сколь угодно мало отличается от средней арифметической вероятностей p_j, т. е.

$$\frac{X_{(n)}}{n} \to \frac{\sum_j p_j}{n} (m.B) . \tag{8.8}$$

Доказательство этой теоремы может быть получено как следствие теоремы Маркова.

8.5. Предельные теоремы

Мы уже указывали, что при некоторых условиях совокупное действие множества случайных факторов (слагаемых, сомножителей и т. д.) приводит в пределе к определенному закону распределения суммарного результата. Проявляющаяся в таком виде статистическая закономерность, которая также может быть отнесена к закону больших чисел, играет большую роль в статистическом анализе, поскольку устанавливает не только свойства устойчивости отдельных характеристик, но и всего закона распределения.

Подобные математические теоремы получили название *предельных*. Мы уже упоминали о подобных предельных теоремах при изучении законов распределения Пуассона и показательного. Очевидно, что переход к рассмотрению более общих случайных величин, чем фигурирующие в этих теоремах, требует введения дополнительных ограничений на них.

Среди цикла предельных теорем наиболее важной является так называемая *центральная предельная теорема*, устанавливающая условия образования в пределе нормального закона распределения. Впервые такая теорема для повторных независимых испытаний была доказана П. Лапласом и для ошибок измерений — К. Гауссом. В значительно более общей форме центральная предельная теорема была доказана А. М. Ляпуновым. В последующем рядом ученых (А. А. Марков, С. Н. Бернштейн, А. Я. Хинчин и др.) были указаны более общие условия, приводящие в пределе к образованию нормального распределения. Фундаментальное значение, которое имеет для математической статистики центральная предельная теорема и вместе с ней нормальный закон распределения, обусловлено весьма общим характером этих условий, которые на практике выполняются довольно часто. Рассмотрим общую часть центральной предельной теоремы в любой ее форме. Пусть $X_1, \ldots, X_j, \ldots, X_n \ldots$ — последовательность независимых случайных величин с математическими ожиданиями $\overline{X}_j$ и дисперсиями σ_j^2 для $j = 1, 2, \ldots n, \ldots$. Введем новые случайные величины:

$$Y_n = X_1 + \ldots + X_j + \ldots + X_n, \tag{8.9}$$

для которых

$$M(Y_n) = \sum_{j=1}^{n} \overline{X}_j; \quad \sigma^2(Y_n) = \sum_{j=1}^{n} \sigma_j^2. \tag{8.10}$$

Тогда при определенных условиях справедливо предельное равенство

$$\lim_{n\to\infty} P\left\{\left|\frac{Y_n - M(Y_n)}{\sigma(Y_n)}\right| < z\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}}\, du = 0{,}5 + \Phi(z), \tag{8.11}$$

утверждающее, что закон распределения нормированных отклонений суммы Y_n при $n \to \infty$ стремится к стандартному нормальному закону распределения. В таком случае говорят, что Y_n подчиняется *асимптотически-нормальному распределению*.

Различные формы центральной предельной теоремы отличаются друг от друга ограничениями, налагаемыми на последовательность X_j, при которых выполняется утверждение теоремы. Так в теореме Муавра — Лапласа (см. § 8.6) в качестве X_j рассматриваются частоты при j-м испытании в модели Бернулли. Значительно более общим является условие одинакового распределения всех X_j при существовании конечного математического ожидания $M(X_j) = \overline{X}$ и дисперсии $D(X_j) = \sigma^2$. А. М. Ляпунов, создав специальный метод характеристических функций, используемый в до-

казательстве различных предельных теорем, показал, что требование одинакового распределения X_j можно заменить условием их равномерной малости в образовании общей суммы. Это условие Ляпунова математически выражается так:

$$\lim_{n\to\infty} \frac{\sum_{j=1}^{n} a_j}{\left(\sum_{j=1}^{n} \sigma_j^2\right)^{3/2}} = 0, \text{ где } a_j = \mathrm{M}|X_j - \bar{X}_j|^3. \tag{8.12}$$

В последующем Ляпунов доказал справедливость предельного нормального распределения при более общих условиях: существует $\delta > 0$, для которого

$$\lim_{n\to\infty} \frac{1}{\sigma(Y_n)} \sum_{j=1}^{n} \mathrm{M}(X_j - \bar{X}_j)^{2+\delta} = 0. \tag{8.13}$$

Это предельное равенство получило название *условия Ляпунова.*

В дальнейшем многими исследователями предлагались различные расширения условий Ляпунова, позволяющие центральную предельную теорему применить к более общим видам последовательностей случайных величин. Так, было показано, что достаточно условие равномерной ограниченности X_j, в ряде работ доказывалась возможность ослабления требования независимости X_j и т. д.

Наиболее общим необходимым и достаточным условием справедливости центральной предельной теоремы является условие Линдберга.

Не имея возможности в рамках настоящей книги привести доказательство * центральной предельной теоремы, ограничимся иллюстрацией и примерами ее практического использования.

1. *Ошибки измерения.* Случайная ошибка измерения образуется под воздействием достаточно большого числа факторов, каждый из которых вызывает как бы часть (слагаемое) суммарной ошибки. Если при этом какой-то из факторов оказывает превалирующее воздействие, то и распределение суммарной ошибки будет в основном формироваться его распределением. Однако если все факторы оказываются примерно равноценными, то при достаточно большом их числе можно ожидать почти нормального распределения суммарной ошибки измерения.

2. *Рассеивание при стрельбе.* Исторически одним из первых примеров нормального распределения явилось распределение отклонений точки падения при одних и тех же условиях стрельбы. Причину следует искать в проявлении все той же центральной предельной теоремы: отклонение точки падения снаряда обусловлено действием множества факторов (ветер, колебания ствола орудия, неоднообразие формы и поверхности снаряда, неоднообразие

* С указанным доказательством можно познакомиться в [1].

заряда и т. д.) и является как бы равнодействующей (суммой) всех таких частичных отклонений. В предположении относительной равноценности воздействия этих факторов и достаточно большого их числа можно ожидать почти нормального распределения суммарного отклонения.

3. *Выборочные наблюдения.* Проявления центральной предельной теоремы в формировании распределений выборочной средней будет предметом подробного рассмотрения в специальной главе. Здесь же ограничимся лишь указанием, что выборочная средняя $\bar{X}_в$ как средняя арифметическая n одинаково распределенных случайных величин X_j, где X_j — значение признака у j-го отобранного элемента, в случае повторной выборки (в этом случае все X_j независимы) удовлетворяет условиям центральной предельной теоремы и потому может рассматриваться как асимптотически нормально распределенная случайная величина.

Из изложенного становится понятным широкое распространение нормального распределения как предельного для нормированных сумм произвольно распределенных слагаемых. Вместе с тем оно является не единственным предельным распределением. Так, при изучении распределения Пуассона было указано на возможность его использования в качестве предельного для сумм независимых слагаемых.

Это связано с общим для обоих распределений свойством устойчивости: если слагаемые подчиняются одному из этих законов, то и сумма будет подчиняться тому же закону.

8.6. Асимптотические формулы Лапласа для биномиального распределения

Мы уже указывали на сложность расчета вероятностей для биномиального распределения при большом числе испытаний n и на необходимость построения для этой цели приближенных формул. В § 7.2 были выведены асимптотические формулы Пуассона, однако при этом указывалось на ограниченные условия ее применения (большое n и малое значение p). Между тем часто приходится проводить расчеты при большом n и не малом p.

Приведем асимптотические формулы для биномиального распределения, основанные на нормальном распределении и практически пригодные при большом n и не малом p (точнее, если $np \geqslant 10$). Пусть проводится n повторных независимых испытаний при постоянной вероятности события $P(A) = p$. Обозначим частоту появления события при j-м испытании через X_j, где $j = 1, \ldots, n$. Все величины X_j представляют собой n одинаково распределенных случайных величин, принимающих каждая значения 0 или 1 с вероятностями $q = 1 - p$ и p, имеющих математическое ожидание $M(X_j) = p$ и дисперсию $D(X_j) = pq$. Тогда частота $X_{(n)}$ при n испытаниях, выражающаяся в виде $X_{(n)} = \sum_j m_j$, будет иметь $M(X_{(n)}) = np$ и $D(X_{(n)}) = npq$ или $\sigma(X_{(n)}) = \sqrt{npq}$.

Величины X_j полностью удовлетворяют условиям теоремы Ляпунова (они попарно независимы, имеют конечные математические ожидания и дисперсии и как одинаково распределенные удовлетворяют условию (8.12)). Следовательно, при $n \to \infty$ их нормированная сумма, т. е. случайная величина

$$Z = \frac{X_{(n)} - \mathrm{M}(X_{(n)})}{\sigma(X_{(n)})} = \frac{X_{(n)} - np}{\sqrt{npq}}, \tag{8.14}$$

подчиняется нормальному распределению с параметрами $\mathrm{M}(z) = 0$ и $\sigma^2(z) = 1$, задаваемому плотностью вероятности

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{8.15}$$

Переходя к элементу вероятности, запишем последнее утверждение в виде предельного равенства

$$\lim_{n\to\infty} P(z \leqslant Z < z + \Delta z) = \varphi(z)\,\Delta z. \tag{8.16}$$

Рассмотрим два соседних значения: $X_{(n)} = m$ и $X'_{(n)} = m + 1$. Тогда

$$z = \frac{m - np}{\sqrt{npq}}; \qquad z' = \frac{m + 1 - np}{\sqrt{npq}},$$

откуда

$$\Delta z = z' - z = \frac{1}{\sqrt{npq}}.$$

При $n \to \infty$ получим $\Delta z \to 0$ и

$$P(z \leqslant Z < z + \Delta z) = P(m \leqslant X_{(n)} < m + 1) = P(X_{(n)} = m) = P_n(m).$$

Подставляя в (8.16), получим

$$\lim_{n\to\infty} P_n(m) = \frac{1}{\sqrt{npq}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(m-np)^2}{2npq}}. \tag{8.17}$$

Это предельное равенство составляет содержание локальной теоремы Муавра — Лапласа, доказанной независимо от теоремы Ляпунова.

Теорема 1. При неограниченном возрастании числа n повторных независимых испытаний вероятность $P_n(m)$ стремится к выражению

$$\frac{1}{\sqrt{npq}} \frac{1}{\sqrt{2\pi}} e^{\frac{-(m-np)^2}{2npq}}.$$

На рис. 8.3 наглядно показано постепенное приближение полигона биномиального распределения к кривой Гаусса $y = \varphi(z)$ с ростом числа n. Причем это приближение происходит достаточно быстро, уже начиная с небольших значений n ($n \approx 10$). На рис. 8.3, б по оси абсцисс откладываются значения z, а по оси ординат — значения $y = \sqrt{npq}P_n(m)$. При этом наивероятнейшая частота будет находиться в начале координат, сжатие вдоль оси

абсцисс в $\sqrt{npq}$ раз компенсируется таким же растяжением вдоль оси ординат и площадь полигона остается равной единице.

Переходя от плотности распределения к интегральной функции нормального распределения, получим предельное равенство, аналогичное (8.17):

$$\lim_{n\to\infty} P\left\{\frac{X_{(n)}-np}{\sqrt{npq}} < z\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}}\, du = 0{,}5 + \Phi(z), \quad (8.18)$$

где $\Phi(z)$ — функция Лапласа (см. § 7.7).

Последнее равенство составляет содержание интегральной теоремы Лапласа, которая тоже получена в данном случае как след-

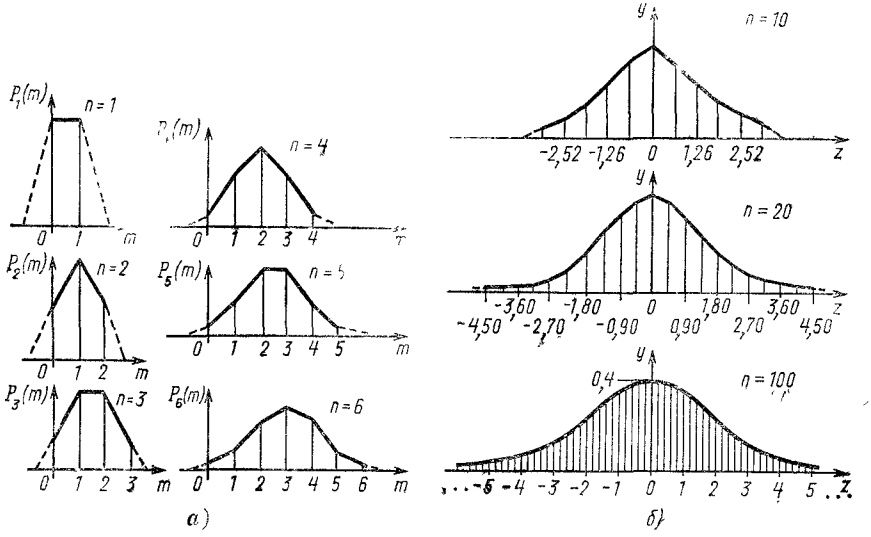


Рис. 8.3.
Трансформация полигона биномиального распределения в кривую вероятностей

ствие теоремы Ляпунова. Для практических расчетов при конечном n предельные равенства (8.17) и (8.18) можно заменить приближенными асимптотическими формулами:

$$P(X_{(n)} = m) = P_n(m) \approx \frac{1}{\sqrt{npq}} \frac{1}{\sqrt{2\pi}} e^{\frac{-(m-np)^2}{2npq}} = \frac{\varphi(z)}{\sqrt{npq}}; \quad (8.19)$$

$$P\left\{\frac{X_{(n)}-np}{\sqrt{npq}} < z\right\} \approx 0{,}5 + \Phi(z), \quad \text{где} \quad z = \frac{m-np}{\sqrt{npq}}, \quad (8.20)$$

тем более точными, чем больше величина n. Практически эти формулы дают хорошее приближение уже начиная с $np \geqslant 10$.

В следующей таблице приведены сравнительные результаты расчета по точной формуле Бернулли и по асимптотическим формулам Пуассона и Лапласа для значения $n = 20$ при разных значениях p и $\bar{m}$:

Таблица 82

m	Формула	Значения p							
		0,005	0,01	0,05	0,10	0,20	0,30	0,40	0,50
1	Бернулли	0,0944	0,1652	0,3773	0,2701	0,0577	0,0068	0,0005	0,0000
	Пуассона	0,0948	0,1637	0,3679	0,2707	0,0733	0,1490	0,0027	0,0000
	Лапласа	0,0219	0,1782	0,4091	0,2253	0,0547	0,0099	0,0011	0,0000
5	Бернулли	0	0	0,0023	0,0319	0,1746	0,1789	0,0746	0,0146
	Пуассона	0	0	0,0031	0,0361	0,1563	0,1606	0,0916	0,0378
	Лапласа	0	0	0	0,0244	0,1909	0,1728	0,0712	0,0146

Порядок выполнения расчетов по формулам (8.19), (8.20) иллюстрируется следующим примером.

Предположим, что вероятность изготовления бракованного изделия $p = 0{,}05$. Какова вероятность обнаружения среди 200 изделий 7 бракованных? Какова вероятность того, что число бракованных изделий окажется меньше 7?

Дано: $n = 200$; $p = 0{,}05$; $q = 0{,}95$; $m = 7$. Определим $P_{200}(7)$ и $P\{X_{(n)} < 7\}$.

Вычисляем

$$\sqrt{npq} = \sqrt{200 \cdot 0{,}05 \cdot 0{,}95} = 3{,}0822;$$

$$z = \frac{m - np}{\sqrt{npq}} = \frac{7 - 200 \cdot 0{,}05}{3{,}0822} = -0{,}9733.$$

Находим по таблице $\varphi(z) = \varphi(-0{,}9733) = 0{,}2484$. По формуле (8.19) вычисляем

$$P_{200}(7) = \frac{\varphi(z)}{\sqrt{npq}} = \frac{0{,}2484}{3{,}0822} = 0{,}0806.$$

Расчет по формуле Пуассона дал бы 0,09, а точное значение (по формуле Бернулли) составляет 0,089562. Как видим, хорошее приближение дает формула Пуассона (ошибка 0,5%), хуже — формула Лапласа (ошибка около 10%), так как $n \cdot p = 9{,}5$, т. е. близко к границе применимости формулы.

Ответ на второй вопрос найдем с помощью (8.20). Неравенство $X_{(n)} < 7$ равносильно неравенству $\dfrac{X_{(n)} - np}{\sqrt{npq}} < \dfrac{7 - np}{\sqrt{npq}} = z$, поэтому из (8.20) получаем

$$P(X_{(n)} < 7) = P\left(\frac{X_{(n)} - np}{\sqrt{npq}} < z\right) = 0{,}5 + \Phi(z).$$

Подставив найденное выше значение $z = -0{,}9733$, *получим*

$$P(X_{(n)} < 7) = 0{,}5 + \Phi(-0{,}9733) = 0{,}5 - \Phi(0{,}9733) = 0{,}16519.$$

Из (8.20), применяя ее к частоте $X_{(n)}$ или к относительной частоте $\dfrac{X_{(n)}}{n}$, можно получить ряд рабочих формул, (табл. 8.3), используемых при решении различных задач.

Таблица 83

Частота	Относительная частота
$z = \dfrac{m - np}{\sqrt{npq}}$	$z = \dfrac{m/n - p}{\sqrt{pq/n}}$
$P\left(X_{(n)} < m\right) \approx 0{,}5 + \Phi\left(z\right)$	$P\left(\dfrac{X_{(n)}}{n} < p\right) \approx 0{,}5 + \Phi\left(z\right)$
$P\left(X_{(n)} \geqslant m\right) \approx 0{,}5 - \Phi\left(z\right)$	$P\left(\dfrac{X_{(n)}}{n} \geqslant p\right) \approx 0{,}5 - \Phi\left(z\right)$
$P\left(m_1 \leqslant X_{(n)} < m_2\right) \approx \Phi\left(z_2\right) - \Phi\left(z_1\right)$	$P\left(\dfrac{m_1}{n} \leqslant \dfrac{X_{(n)}}{n} < \dfrac{m_2}{n}\right) \approx \Phi\left(z_2\right) - \Phi\left(z_1\right)$
$P\left(\left\| X_n - np \right\| < \varepsilon\right) \approx 2\Phi\left(\dfrac{\varepsilon}{\sqrt{npq}}\right)$	$P\left(\left\| \dfrac{X_{(n)}}{n} - p \right\| < \varepsilon\right) \approx 2\Phi\left(\dfrac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right).$

Раздел III

Математическая статистика

Глава 9

Математическая теория выборки

9.1. Сплошное и выборочное наблюдения

Методы систематизации, обработки и использования статистических данных, выявление статистических закономерностей составляют основное содержание математической статистики. Так как исходной базой для всех построений математической статистики является рассмотрение результатов опыта или наблюдений как выборки из некоторой генеральной совокупности, то математическая теория выборки является центральным разделом математической статистики. Основу ее составляют методы статистической оценки распределения и характеристик генеральной совокупности. Кроме непосредственного практического приложения теория выборки имеет фундаментальное значение для обоснования многих вопросов статистического анализа.

Статистическое наблюдение может быть *сплошным* или *выборочным*. Первое предполагает наблюдение (измерение, исследование и т. д.) всех изучаемых объектов. Однако по ряду причин оно может оказаться принципиально неосуществимым или практически нецелесообразным. Так, если изучаемая совокупность содержит труднообозримое множество объектов, то, очевидно, организация сплошного наблюдения либо просто невозможна, либо связана со значительными затратами труда и времени (например, при изучении урожайности, производительности труда, годности партии продукции, бюджета доходов и расходов семей и т. п.). Лишено оно смысла и в том случае, если в процессе наблюдения объекты уничтожаются (например, исследование крепости пряжи на разрыв, качества снарядов отстрелом и т. д.).

Во всех таких случаях прибегают к наблюдению части изучаемых объектов и по его результатам делают выводы о свойствах всей совокупности. Такой метод наблюдения получил название *выборочного*, отобранная для изучения часть объектов называется *выборкой*, а вся исходная совокупность объектов —

генеральной совокупностью. Практически выборочное наблюдение оказывается одним из основных методов получения статистической информации. Заметим, что подобный принцип логического анализа от частного к общему характерен для любого научного эксперимента. Однако если в последнем на основании наблюдения делаются выводы по поводу всех мыслимых опытов, проводимых при одних и тех же условиях, т. е. относительно некоторой гипотетической бесконечной генеральной совокупности, то в теории выборки мы всегда имеем дело с определенной генеральной совокупностью конечного объема.

Информация о генеральной совокупности, полученная на основании выборочного наблюдения, всегда будет обладать некоторой погрешностью, так как основывается на изучении только части элементов. Это выдвигает сразу же две взаимосвязанные проблемы, составляющие содержание математической теории выборки: 1) как организовать выборочное наблюдение, чтобы указанная информация была наиболее полной (*проблема репрезентативности выборки*); 2) как использовать результаты выборки для суждения по ним с наибольшей надежностью о свойствах и параметрах генеральной совокупности (*проблема оценки*).

В теоретическом обосновании выборочного метода построение соответствующей вероятностной модели исходит из заданного генерального распределения, на основании которого строится выборочное распределение или распределение некоторых функций от элементов выборки. Но на практике именно генеральное распределение неизвестно. Поэтому необходимо решить как бы обратную задачу: по результатам выборочного наблюдения высказать определенные суждения о генеральном распределении. Такие выводы формулируются в виде вероятностных утверждений. При этом по поводу генерального распределения делаются некоторые априорные предположения, начиная от самых общих, например о существовании конечных математического ожидания и дисперсии, до предположения о виде распределения (нормальное, Пуассона и т. д.). В последнем случае по выборке оцениваются параметры этого распределения.

Различают два принципиально различных способа отбора элементов генеральной совокупности в выборку: *случайный* и *неслучайный*. Случайным называется отбор, при котором каждый элемент совокупности случайным образом может попасть в выборку. На языке теории вероятностей такой отбор есть последовательность повторных испытаний. Случайный механизм отбора позволяет положить в основу изучения выборки вероятностную модель, закон больших чисел, методы изучения закономерностей случайных явлений. Случайный отбор может производиться путем обычной жеребьевки или с помощью специальных таблиц случайных чисел. В последнем случае элементы совокупности перенумеровываются и из таблицы случайных чисел, открытой на любой взятой наудачу странице, выписываются подряд номера элементов, которые должны быть взяты в выборку. Существуют различные кон-

струкции таких таблиц, основным свойством которых является отсутствие какой-либо закономерности в расположении чисел.

Может показаться, что случайный отбор можно осуществить выбором наудачу первых попавшихся «под руку» элементов. Однако практически установлено, что при таком способе в выбор элементов непроизвольно вносится упорядоченность, нарушающая механизм чисто случайного выбора. Если не придать этому значения, то в результаты выборки можно внести систематическую неустранимую погрешность.

Однако строгое соблюдение правил случайного отбора не всегда осуществимо, так как оно требует четко очерченной базы статистического анализа, каковой является генеральная совокупность, возможности перенумерации всех ее элементов или непосредственного их извлечения при жеребьевке. Так, при проведении бюджетных обследований населения в масштабах страны, республики или района, очевидно, практически невозможно составить список всех жителей или семей с последующей организацией выборки с помощью таблицы случайных чисел. Аналогично невозможно организовать всевозможные опросы по изучению покупательского спроса, потребностей населения и т. д. путем образования строго случайной выборки.

Поэтому прибегают к различным приемам *неслучайного отбора*, стремясь, однако, приблизиться к условиям случайного отбора. К ним относится *серийный* (*гнездовой*) отбор, заключающийся в том, что случайным образом отбираются не отдельные элементы, а целые их группы (серии, гнезда), на которые предварительно делится совокупность, или *механический* отбор, при котором элементы совокупности, предварительно упорядоченные, отбираются по заранее установленному правилу, не связанному с вариацией исследуемого признака. Такие методы отбора должны применяться с крайней осторожностью, и к обработке результатов таких выборок не всегда можно применять вероятностные методы. Поэтому мы будем рассматривать выборку, образованную методом *случайного отбора*.

Случайный отбор может производиться по схеме *возвращенного* (*возвратная* выборка) или *невозвращенного* шара (*безвозвратная* выборка). На практике выборка производится обычно как безвозвратная, однако в теоретическом плане проще возвратная выборка, моделью которой служит схема повторных независимых испытаний. Отличие этих двух выборок тем меньше, чем меньше отношение объема выборки к объему генеральной совокупности. Практически, если отношение составляет менее 5—10%, этим отличием можно пренебречь и пользоваться более простыми соотношениями, предполагающими возвратную выборку и в том случае, когда производится безвозвратный отбор.

Как мы увидим далее, точность результатов выборки существенно повышается с ростом ее объема. Но при этом растет и трудоемкость, а следовательно, и стоимость проведения выборочного обследования. Поэтому, естественно, возникла проблема, как уве

личить точность выборки, не увеличивая ее объема? Оказывается, что этого можно добиться, переходя от простого случайного отбора к более сложным способам организации отбора, как-то: *типический* (стратифицированный), *многоступенчатый*, *последовательный*. Однако основой математической теории выборки является простой случайный отбор (возвратный или безвозвратный), или *простая выборка*. Особое место в теории выборки занимает так называемый *способ моментных наблюдений*, рассмотрению которого посвящен далее отдельный параграф.

В заключение отметим, что выборочный метод, несмотря на весьма широкое распространение, все же на практике мог бы найти еще большее применение, если бы не бытующее против него предубеждение. Связано оно с опасением, что выводы, полученные на основании изучения не всей совокупности, а только ее части, могут привести к ошибкам. При этом не учитывают, что эти ошибки могут быть заранее оценены и посредством правильной организации выборки сведены к практически незначимым величинам. Между тем использование сплошного наблюдения, даже там, где это принципиально возможно, не говоря уже о росте трудоемкости, стоимости и увеличении необходимого времени, часто приводит к тому, что каждое отдельное наблюдение поневоле проводится с меньшей точностью. А это уже сопряжено с неустранимыми ошибками и в конечном счете приводит к снижению точности по сравнению с выборочным наблюдением.

Очевидно, в частности, что многие задачи оперативного экономического анализа могли бы с успехом базироваться на данных, полученных выборочным методом, подобно тому, как это реализуется в системе производственного контроля.

9.2. Статистические оценки

Как уже указывалось, задача заключается в том, чтобы по результатам одной случайной выборки оценить достаточно точно значения характеристик генерального распределения, как, например, долю признака, среднюю, дисперсию и т. д. Задачу об оценке можно разделить на две части: какую величину, подсчитанную по выборке, принять в качестве приближенного значения характеристики генерального распределения (*точечная оценка*), и в каком интервале вокруг этой величины будет заключена с заданной надежностью искомая характеристика (*интервальная оценка*).

Пусть генеральное распределение задается некоторой функцией $F(x, \xi_1, \ldots, \xi_k)$, где $\xi_1, \ldots, \xi_k$ — его параметры. Если распределение задается одним параметром ξ, то он обычно характеризует среднее значение (*центр распределения*), если имеются два параметра — ξ_1 и ξ_2, то ξ_2 характеризует рассеяние вокруг центра, и т. д.

Случайный отбор позволяет выборку объема n рассматривать как n повторных испытаний. Результат каждого испытания (j-го единичного отбора) есть случайная величина X_j, а вся выборка —

совокупность n случайных величин $\{X_1, \ldots, X_j, \ldots, X_n\}$. В случае возвратного отбора (или отбора из гипотетической генеральной совокупности бесконечного объема) эти величины независимы и имеют распределения, совпадающие с генеральным. Любая конкретная выборка $x_1, \ldots, x_j, \ldots, x_n$ есть реализация этой совокупности случайных величин. Для пояснения рассмотрим следующую упрощенную схему образования выборки.

Таблица 9.1

X	5	6	7	$\sum$
P	$\frac{2}{7}$	$\frac{4}{7}$	$\frac{1}{7}$	1

Из генеральной совокупности, заданной дискретным распределением табл. 9.1, образуется возвратная выборка из двух элементов ($n = 2$). Обозначим через X_1 и X_2 значение признака первого и второго из отобранных элементов. Очевидно, что X_1 и X_2 будут иметь одинаковые распределения, задаваемые табл. 9.1.

Малый объем выборки и простота табл. 9.1 позволяют в данном случае реально описать всевозможные выборки и вычислить вероятность каждой из них (табл. 9.2)

Таблица 9.2

	Номер выборки								
	1	2	3	4	5	6	7	8	9
Результат выборки $(X_1; X_2)$	5; 5	5; 6	5; 7	6; 5	6; 6	6; 7	7; 5	7; 6	7; 7
Вероятность выборки	4/49	8/49	2/49	8/49	16/49	4/49	2/49	4/49	1/49
Выборочная средняя $\bar{X}^*$	5	5,5	6	5,5	6	6,5	6	6,5	7
Выборочная дисперсия D^*	0	0,25	1	0,25	0	0,25	1	0,25	0

Вычисление вероятностей производится по теореме умножения. Так, например, для второй выборки имеем

$$P = P\{(X_1 = 5)(X_2 = 6)\} = P(X_1 = 5)\, P(X_2 = 6) = \frac{2}{7} \cdot \frac{4}{7} = \frac{8}{49}.$$

Для оценки неизвестного параметра ξ генеральной совокупности введем некоторую величину θ, вычисляемую по результатам выборки, т. е. $\theta = \theta(X_1, \ldots, X_j, \ldots, X_n)$, называемую *статистикой*. Статистика θ есть случайная величина, распределение которой зависит от генерального распределения и, в частности, от параметра ξ. Так, если для оценки генеральной средней $\xi = \bar{X}$ выбрана статистика $\theta = \bar{X}^*$ (выборочная средняя), то ее значения могут быть подсчитаны по результатам выборки как $\bar{X}^* = \frac{1}{n}\sum_{j=1}^{n} X_j$.

Сгруппировав в предыдущем примере одинаковые значения $\bar{X}^*$, суммируя их вероятности и располагая $\bar{X}^*$ в возрастающем порядке, получим распределение $\bar{X}^*$ (табл. 9.3).

Таблица 9.3

$\overline{X}^*$	5	5,5	6	6,5	7
P	$\frac{4}{49}$	$\frac{16}{49}$	$\frac{20}{49}$	$\frac{8}{49}$	$\frac{1}{49}$

Таблица 9.4

D^*	0	0,25	1
P	$\frac{21}{49}$	$\frac{24}{49}$	$\frac{4}{49}$

Если для оценки генеральной дисперсии выбрана статистика $\theta = D^*$, то ее значения для разных выборок могут быть рассчитаны по формуле $D^* = \frac{1}{n}\sum_j (x_j - \overline{X}^*)^2$ (см. табл 9.2). Используя данные табл. 9.2, получим подобно предыдущему распределению величины D^* (табл. 9.4).

Пусть в общем случае генеральное распределение задается функцией распределения $F(x, \xi_1, \ldots, \xi_k)$ или плотностью вероятности $f(x, \xi_1, \ldots, \xi_k)$, где $\xi_1, \ldots, \xi_k$ — параметры генерального распределения. Тогда для возвратной выборки закон распределения выборки (т. е. совокупности n независимых случайных величин X_j) может быть определен по георему умножения функций распределения (см. § 6.7), с помощью одного из следующих выражений:

$$\begin{aligned} F(x_1, \ldots, x_n, \xi_1, \ldots, \xi_k) &= F(x_1, \xi_1, \ldots, \xi_k) \cdot \ldots \cdot F(x_n, \xi_1, \ldots, \xi_k); \\ f(x_1, \ldots, x_n, \xi_1, \ldots, \xi_k) &= f(x_1, \xi_1, \ldots, \xi_k) \cdot \ldots \cdot f(x_n, \xi_1, \ldots, \xi_k). \end{aligned} \tag{9.1}$$

Именно эти соотношения использовались для дискретного случая при расчете вероятностей выборок в табл. 9.2.

По распределению выборки можно найти распределение статистики $\theta = \theta(x_1, \ldots, x_n)$, которое будет зависеть от вида генерального распределения и его параметров.

Пусть предстоит оценить один параметр генеральной совокупности ξ. Для этого необходимо, чтобы значение θ, подсчитанное по данным выборки, оказалось с достаточно высокой вероятностью близким к истинному значению параметра ξ. Очевидно, достичь этого при малом объеме выборки невозможно, так как слишком велика роль случайности отбора, поэтому естественно предположить, что объем выборки достаточно велик. Вначале рассмотрим предельный случай — $n \to \infty$, потребовав при этом, чтобы статистика θ стремилась по вероятности к оцениваемому параметру, т. е. чтобы для любого заданного $\varepsilon > 0$ выполнялось предельное равенство

$$\lim_{n\to\infty} P\{|\theta - \xi| < \varepsilon\} = 1 \quad \text{или} \quad \theta \xrightarrow{P} \xi. \tag{9.2}$$

Статистика θ, удовлетворяющая этому условию, получила название *состоятельной*. Состоятельность — асимптотическое свойство (справедливо при $n \to \infty$), которое означает, что распределение статистики θ с ростом n концентрируется в сколь угодно малой окрестности параметра ξ (рис. 9.1) и представляет собой не что иное, как закон больших чисел применительно к статистике θ.

Например, если для оценки доли признака p выбрать статистику $\theta = \frac{m}{n}$, то свойство (9.2) будет утверждением теоремы Бернулли; при оценке генеральной средней $\bar{X}$ с помощью выборочной средней $\bar{X}^*$ свойство (9.2) представляет собой утверждение теоремы Чебышева и т. д.

Единственным условием выполнимости (9.2) является существование p или $\bar{X}$ у генерального распределения. Однако это усло-

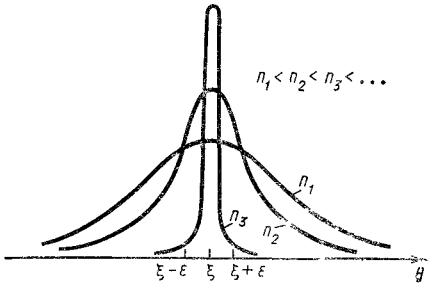


Рис. 9.1.
Иллюстрация свойства состоятельности

вие не тривиально. Так, если генеральное распределение есть распределение Коши, задаваемое плотностью вероятности $f(x) = \frac{1}{\pi(1+x^2)}$ (для $-\infty < x < +\infty$), то свойство состоятельности не имеет места, так как для этого распределения вообще не существует математического ожидания. Действительно,

$$\bar{X} = \frac{1}{\pi}\int_{-\infty}^{\infty} x\,\frac{dx}{1+x^2} = \frac{1}{2\pi}\int_{0}^{\infty}\frac{x\,dx}{1+x^2} = \frac{1}{4\pi}\lim_{b\to\infty}[\ln(1+b^2)] = \infty.$$

Иногда свойство состоятельности заменяют более жесткими, но тоже асимптотическими требованиями, предъявляемыми к оценке:

$$\lim_{n\to\infty} M(\theta) = \xi \quad \text{и} \quad \lim_{n\to\infty} D(\theta) = 0. \tag{9.3}$$

Первое из этих равенств, получившее название *асимптотической несмещенности*, есть условие того, что центр распределения θ неограниченно приближается к ξ, а второе — что распределение θ стягивается в пределе к своему центру.

Свойства (9.3) являются достаточными, но не необходимыми условиями состоятельности, так как могут быть указаны статистики, для которых, например, вообще не существует $\lim_{n\to\infty} M(\theta)$, и вместе с тем удовлетворяющие условию (9.2). Однако в дальнейшем под состоятельными будем часто понимать статистики, удовлетворяющие условиям (9.3).

Указанные асимптотические свойства, предполагающие бесконечно возрастающий объем выборки из бесконечной гипотетической совокупности, практически могут быть использованы как приближенные свойства для выборок конечного, но достаточно большого объема n.

Статистику θ, принимающую для данной выборки определенное числовое значение, будем называть *точечной оценкой* параметра ξ, и обозначать той же буквой, что и оцениваемый параметр, помечая ее звездочкой.

Условия (9.2) или (9.3) позволяют для конечного n записать лишь приближенное равенство

$$\xi \approx \xi^*. \tag{9.4}$$

Так как выборка носит случайный характер, то для различных возможных выборок случайная величина ξ^* может принимать различные значения.

Так, если в рассмотренном выше примере оценивать генеральную среднюю $\xi = \overline{X} = 5 \cdot \frac{2}{7} + 6 \cdot \frac{4}{7} + 7 \cdot \frac{1}{7} = 5{,}857$ с помощью выборочной средней $\xi^* = \overline{X}^*$, то, как видно из табл. 9.1, получим:

$$\overline{X}_1^* = 5;\ \overline{X}_2^* = \overline{X}_4^* = 5{,}5;\ \overline{X}_3^* = \overline{X}_5^* = \overline{X}_7^* = 6;\ \overline{X}_6^* = \overline{X}_8^* = 6{,}5 \text{ и } \overline{X}_9^* = 7.$$

Поэтому возникает задача дополнить точечную оценку информацией о возможной ее погрешности и надежности, т. е. оценить ошибку выборки $\delta = \xi - \xi^*$.

В рассмотренном примере эта ошибка для различных выборок составит:

$$\delta_1 = 0{,}857;\ \delta_2 = \delta_4 = 0{,}387;\ \delta_3 = \delta_5 = \delta_7 = -0{,}143;\ \delta_6 = \delta_8 = -0{,}643 \text{ и } \delta_9 = -1{,}143.$$

Величина δ, как и ξ^*, есть случайная величина, имеющая то же распределение, что и ξ^*. Пусть плотность распределения ξ^* изобра-

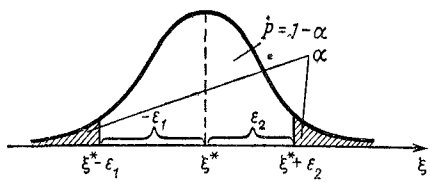


Рис. 9.2.
Доверительные границы

жена на рис. 9.2. Выберем интервал $(-\varepsilon_1, \varepsilon_2)$, в котором с достаточно близкой к 1 вероятностью будет заключаться величина $\delta = \xi^* - \xi$, т. е.

$$P(-\varepsilon_1 < \xi - \xi^* < \varepsilon_2) = 1 - \alpha, \tag{9.5}$$

где α — величина, близкая к нулю. Это означает, что в большинстве выборок (доля которых составляет $1 - \alpha$) ошибка выборок попадет в данный интервал, и лишь в относительно малом числе выборок (доля которых равна α) ошибка δ выйдет за пределы интервала $(-\varepsilon_1, \varepsilon_2)$. Поскольку произведена одна выборка, то с практической достоверностью (т. е. с вероятностью $1 - \alpha$) можно полагать, что ее ошибка попадает в данный интервал, и, наоборот, практически невозможно (т. е. с вероятностью α), чтобы она вышла за границы интервала.

Но если $-\varepsilon_1 < \xi - \xi^* < \varepsilon_2$, то $\xi^* - \varepsilon_1 < \xi < \xi^* + \varepsilon_2$ и равенство (9.5) запишется в виде

$$P(\xi^* - \varepsilon_1 < \xi < \xi^* + \varepsilon_2) = 1 - \alpha. \qquad (9.5')$$

В силу изложенного интервал $(\xi^* - \varepsilon_1;\ \xi^* + \varepsilon_2)$ называется *доверительным интервалом,* числа $\xi^* - \varepsilon_1;\ \xi^* + \varepsilon_2$ — *доверительными границами,* вероятность $P = 1 - \alpha$ — *доверительной вероятностью* и α — *уровнем значимости (существенности).* Термины эти объясняются тем, что вероятность $P = 1 - \alpha$ как бы характеризует степень доверия к данному интервалу, в котором заключается ошибка выборки, а уровень значимости α, наоборот, характеризует риск, что ошибка выйдет за пределы интервала. Чем более существенны последствия такой ошибки, тем меньшим уровнем значимости α следует задаваться. Величина α (на рис. 9.2 она численно равна площади заштрихованных фигур) характеризует вероятность события, которое полагается с таким *уровнем значимости* невозможным.

Следует заметить, что равенство (9.5′) отличается от ранее встречавшихся выражений, в которых определялась вероятность попадания случайной величины в заданный интервал. В данном случае границы интервала случайны (так как содержат случайную величину ξ^*), а оцениваемый параметр ξ, заключенный в этом интервале, не случаен. Поэтому выражение (9.5′) следовало бы трактовать как вероятность того, что интервал со случайными границами $\xi^* - \varepsilon_1$ и $\xi^* + \varepsilon_2$ включает постоянный параметр ξ. Доверительный интервал дополняет точечную оценку ξ^* оценкой ошибки выборки, или *интервальной оценкой* параметра ξ.

Важно отметить, что при определении доверительного интервала, как и при расчете точечной оценки, используется информация только одной выборки. Отличие состоит лишь в том, что если для точечной оценки необходимо знать лишь выражение для ξ^* как функцию данных выборки, то для построения доверительного интервала необходимо знать также закон распределения ξ^*, с помощью которого рассчитывается вероятность (9.5′). Здесь возникают как технические, так и принципиальные трудности. Последние вызваны тем, что для определения вероятностей выборок (см. (9.1) и (9.2)) необходимо знать вид генерального распределения и значение параметров $\xi_1, \xi_2, \ldots$, которые и подлежат оценке. Первое затруднение преодолевается тем, что иногда вид генерального распределения может постулироваться (распределение качественного признака, равномерное распределение, нормальное распределение и т. д.). В некоторых случаях при возможности неограниченного увеличения объема выборки n (а практически при достаточно больших n), опираясь на предельные теоремы, можно получить асимптотические распределения статистических оценок, практически не зависящие от вида генерального распределения. Так, в частности, из теоремы Ляпунова (см. § 8.5) следует, что распределения выборочной доли $\frac{\overline{X}_{(n)}}{n}$ или выборочной средней $\overline{X}^*$

при $n \to \infty$ будут асимптотически-нормальными при любом виде генерального распределения.

Второе затруднение приводит к двум способам построения интервальной оценки: *приближенному и точному*. Приближенный способ заключается в том, что неизвестные параметры генеральной совокупности, от которых зависит распределение ξ^*, заменяются на их точечные оценки, полученные в результате выборки. Точный способ используется при условии, что вид генерального распределения известен. Он состоит в том, что строятся специальные статистики, распределение которых не зависит от неизвестных параметров генеральной совокупности. Так, в частности, при оценке средней нормально распределенной генеральной совокупности строится статистика

$$\theta = T = \frac{\overline{X} - \overline{X}^*}{\sigma^*} \sqrt{n}, \qquad (9.6)$$

которая подчиняется распределению Стьюдента (см. § 7.8), зависящему только от n и т. д. С этим способом мы познакомимся

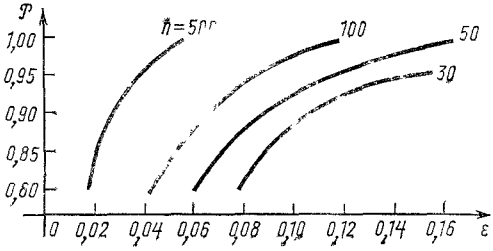


Рис. 9.3.
Зависимость доверительной вероятности P от предельной ошибки ε при оценке доли признака

ниже. Вернемся к построению доверительного интервала. Как ясно из изложенного, для этого необходимо располагать функцией распределения θ. Часто при симметричном характере этой функции относительно ξ можно и доверительный интервал рассматривать как симметричный относительно ξ. В таком случае уравнение (9.5′) можно заменить на более простое

$$P(\xi^* - \varepsilon < \xi < \xi^* + \varepsilon) = P(|\xi - \xi^*| < \varepsilon) = 1 - \alpha. \qquad (9.6')$$

Зависимость между P и ε схематически показана на рис. 9.3 Чем шире доверительный интервал, тем больше величина P, т. е тем вероятнее, что ошибка выборки будет охвачена этим интервалом. Величина ε, называемая *предельной ошибкой выборки*, характеризует ее точность. Очевидно, при данном уровне значимости α она будет зависеть от свойств генерального распределения и от объема выборки n. Зависимость ε от n и α схематически показана на рис. 9.4. Так как свойства генеральной совокупности не зависят от исследователя, то влиять на повышение точности выборки можно только * путем увеличения ее объема.

* Как мы увидим ниже, точность выборки можно увеличить также путем изменения ее организации (например, применить типическую выборку).

С интервальной оценкой связано решение трех типов задач: 1) определение доверительной вероятности по заданному доверительному интервалу и объему выборки; 2) определение доверительного интервала по заданной доверительной вероятности $P = 1 - \alpha$ и объему выборки n; 3) определение необходимого объема выборки n по заданным доверительной вероятности и доверительному интервалу.

Если решение задач первого типа практически не зависит от того, какую формулу — (9.5′) или (9.6) — использовали для расче-

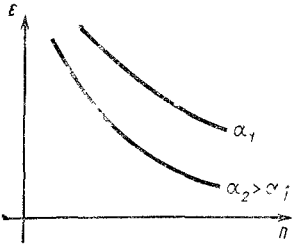


Рис. 9.4.
Зависимость предельной ошибки от объема выборки

тов, то решение задач второго и третьего типов существенно меняется. Так, в частности, решение задачи второго типа по формуле (9.5′) неопределенно, так как имеется одно уравнение для нахождения двух неизвестных ε_1 и ε_2. Поэтому приходится вводить дополнительные условия, например, заменяя уравнение (9.5′) двумя уравнениями:

$$P(\delta < \varepsilon_1) = \frac{\alpha}{2} \quad \text{и} \quad P(\delta > \varepsilon_2) = \frac{\alpha}{2}.$$

Примечание. Более точно в качестве дополнительного условия можно использовать требование минимизации ширины интервала при заданных α и n. Можно показать, что минимум достигается при равенстве плотностей вероятности на концах интервала. Добавив это условие к уравнению (9.5′), можно для конкретных законов распределения однозначно определить границы доверительного интервала в случае несимметричного закона распределения. Такие расчеты были произведены для некоторых типовых распределений (например, биномиального, Пуассона и др.), и составлены соответствующие таблицы и графики.

9.3. Требования, предъявляемые к статистическим оценкам

Мы уже говорили в предыдущем параграфе об условиях состоятельности (9.3), выражающих собой асимптотические свойства закона распределения статистики $\theta = \xi^*$. Нетрудно убедиться в том, что этим условиям может удовлетворять бесчисленное множество статистик. Так, если условиям (9.3) удовлетворяет статистика θ, то им будет удовлетворять и бесчисленное множество статистик θ', полученных из выражения $\theta' = \frac{n+a}{n+b}\theta$, где a и b — любые фиксированные числа. Действительно, используя свойства

математического ожидания и дисперсии и предполагая выполненными условия (9.3) для статистики θ, получим:

$$M(\theta') = \frac{n+a}{n+b} M(\theta);$$

$$\lim_{n\to\infty} M(\theta') = \lim_{n\to\infty} \left[\frac{n+a}{n+b} M(\theta)\right] = \lim_{n\to\infty} M(\theta) = \xi;$$

$$D(\theta') = \frac{(n+a)^2}{(n+b)^2} D(\theta);$$

$$\lim_{n\to\infty} D(\theta') = \lim_{n\to\infty} \left[\frac{(n+a)^2}{(n+b)^2} D(\theta)\right] = \lim_{n\to\infty} D(\theta) = 0.$$

Естественно, возникает проблема, как среди этого множества состоятельных статистик выбрать такую, которая оказалась бы более удобной для построения на ее основе точечной и интервальной оценок. Эта проблема тем более важна, что на практике мы имеем дело с конечным объемом выборки n и формулировка одних только асимптотических свойств оказывается недостаточной.

Оценка $\theta = \xi^*$ называется *несмещенной,* если выполняется равенство

$$M(\xi^*) = \xi \tag{9.7}$$

независимо от числа n.

Таким образом, вместо условия *асимптотической несмещенности* предполагается более жесткое требование. Очевидно, из (9.7) следует асимптотическая несмещенность, но обратное заключение неверно, так как из первого условия (9.3) при конечном n следует лишь приближенное равенство $M(\xi^*) \approx \xi$, тем более точное, чем больше величина n.

При выполнении (9.7) значения ξ^*, получаемые для различных возможных выборок конечного объема, будут группироваться вокруг величины $M(\xi^*) = \xi$. Если (9.7) не выполняется, т. е. $M(\xi^*) \neq \xi$, то оценка ξ^* называется *смещенной.*

Пусть $M(\xi^*) = \xi + a$. В этом случае значения ξ^* для разных выборок будут группироваться вокруг $\xi + a$ и, следовательно, более вероятно, что значение ξ^*, полученное в результате данной выборки, будет близко к $\xi + a$, а не к ξ. Приняв в качестве точечной оценки ξ значение ξ^*, получим *систематическую погрешность* $\xi - \xi^* = a$. Поэтому эта оценка и названа *смещенной.* Естественно, что для построения статистической оценки целесообразно выбрать несмещенную статистику.

Поясним эти понятия на примере из § 9.2, где рассматривались две оценки: $\xi^* = \bar{X}^*$ для $\xi = \bar{X}$ и $\xi^* = D^*$ для $\xi = \sigma^2$. По данным табл. 9.3 и 9.4 находим:

$$M(\bar{X}^*) = 5 \cdot {}^4/_{49} + 5{,}5 \cdot {}^{16}/_{49} + 6 \cdot {}^{20}/_{49} + 6{,}5 \cdot {}^8/_{49} + 7 \cdot {}^1/_{49} = {}^{41}/_7;$$

$$M(D^*) = 0 \cdot {}^{21}/_{49} + 0{,}25 \cdot {}^{24}/_{49} + 1 \cdot {}^4/_{49} = {}^{10}/_{49}.$$

Рассчитав по табл. 9.1 значения $\bar{X} = {}^{41}/_7$ и $\sigma^2 = {}^{20}/_{49}$, устанавливает, что $M(\bar{X}^*) = \bar{X}$, т. е. $\bar{X}^*$ есть несмещенная оценка для $\bar{X}$, и $M(D^*) = \sigma^2 - {}^{10}/_{49}$, т. е. D^* есть смещенная оценка для σ^2 на величину $a = -{}^{10}/_{49}$.

Свойство несмещенности связано с качеством точечной оценки. Однако различные несмещенные статистики, дающие одну и ту же точечную оценку, могут существенно отличаться бо́льшим или меньшим рассеянием вокруг математического ожидания. Хотя второе условие состоятельности и предполагает $D(\xi^*) \to 0$, но при конечном n дисперсия для различных оценок может оказаться различной. А от величины $D(\xi^*)$ непосредственно зависит доверительный интервал, ширина которого при прочих равных условиях пропорциональна $D(\xi^*)$.

На рис. 9.5 показаны кривые распределения трех статистик. Из них θ и θ' — несмещенные, и потому для построения оценки

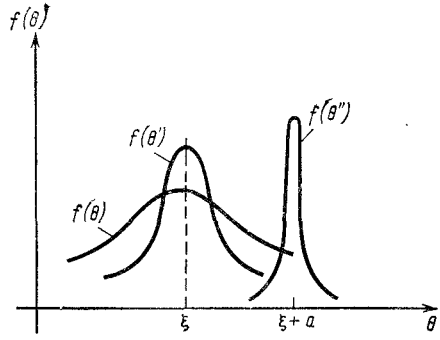


Рис. 9.5
Сравнение свойств трех статистик

предпочтение должно быть отдано статистике θ' с меньшей дисперсией. В то же время статистика θ'', обладающая еще меньшей дисперсией, менее удобна в качестве оценки, так как ее выборочные значения хотя и группируются кучно вблизи центра распределения, но последний значительно отличается от параметра ξ. Поэтому для повышения точности оценки и уменьшения ошибки выборки целесообразно выбирать для статистической оценки несмещенную статистику, обладающую наименьшей дисперсией по сравнению с любыми другими оценками того же параметра. Такое свойство статистики получило название *эффективности*. Если свойство несмещенности обеспечивает лучшую точечную оценку, то свойство эффективности дает повышение ее точности.

Так, для оценки генеральной средней нормально распределенной совокупности могут быть выбраны средняя выборочная $\bar{X}^*$, мода $\bar{X}_{Mo}$ или медиана $\bar{X}_{Me}$. Все они будут несмещенными оценками генеральной средней $\bar{X}$. Однако можно показать, что минимальной дисперсией из них обладает $\bar{X}^*$, которая и будет поэтому эффективной оценкой для $\bar{X}$.

К сожалению, практически проверить свойство эфективности оказывается сложно, так как требуется сравнить дисперсии возможных статистик, предназначенных для оценки того же параметра. Однако в случае, когда оценка ξ^* определяется по методу максимального правдоподобия, может быть указана нижняя грань для дисперсий любой оценки того же параметра. Поэтому если

дисперсия данной оценки оказывается равной этой нижней грани, то этим самым устанавливается эффективность оценки.

Различают еще одно свойство статистики, получившее название *достаточности*. Сущность его заключается в следующем. Результат выборки есть совокупность n чисел $(x_1, \ldots, x_n)$, по которым вычисляется точечная оценка ξ^*, которая иногда весьма значительно может отличаться от ξ. Поэтому возникает вопрос: как наиболее эффективно использовать результаты выборки, чтобы получить из них максимум данных о параметре ξ? Статистика, обладающая тем свойством, что при известной ξ^* выборочные данные не могут дать никакой дополнительной информации о параметре ξ, называется достаточной *. Естественно, что целесообразно пользоваться именно такими статистиками, так как они позволяют максимально использовать информацию, полученную при выборочном наблюдении. Все основные оценки, которые рассматриваются ниже, построены на основе достаточных статистик.

9.4. Методы построения статистических оценок

До сих пор говорилось о требованиях, которым должна удовлетворять статистика, предназначенная для оценки параметра генеральной совокупности. Эти требования сравнительно легко проверяются, если исходить из уже выбранной статистики, т. е. из данной функции $\theta = \theta(x_1 \ldots, x_i, \ldots, x_n)$. Однако, естественно, возникает вопрос: как выбрать эту функцию при оценке данного параметра, чтобы построенная с ее помощью оценка удовлетворяла указанным выше требованиям состоятельности, несмещенности, эффективности и т. д.? Существует два подхода к решению этой проблемы.

Первый заключается в том, что из заданных требований выводится формула для построения оценки. Этот подход не может быть реализован в общем виде, если заранее не задаться хотя бы видом этой функции. Однако если строить так называемые *линейные оценки*, т. е. статистики, выражаемые линейной функцией от выборочных данных $\theta = \Sigma a_i X_i$, то достаточно лишь предположить что существуют (конечны) генеральная средняя и дисперсия. Коэффициенты a_i определяются из условий выполнения требований, предъявляемых к статистической оценке. Однако этот путь связан, как правило, со значительными трудностями, поэтому обычно используется второй подход. Он заключается в том, что вначале статистика выбирается по некоторым дополнительным соображениям, а затем проверяется, удовлетворяет ли она необходимым требованиям. При этом в некоторых случаях, если какое-то требование не выполняется, удается внести соответствующие для его выполне-

---

* Строгое определение достаточности заключается в том, что условная вероятность выборки при известном ξ^* не зависит от оцениваемого параметра ξ. В следующем параграфе это свойство будет проиллюстрировано на примере оценки доли признака.

ния коррективы. Таким способом, например, в дальнейшем (см. § 9.6) будет определена несмещенная оценка для дисперсии.

В качестве дополнительных соображений при предварительном выборе статистики чаще всего используются метод аналогии, метод наименьших квадратов и метод максимального правдоподобия.

Метод аналогии. Этот метод заключается в том, что для оценки параметров генерального распределения выбираются аналогичные параметры (характеристики) выборочного распределения. Так, для оценки доли признака $p_k = \frac{N_k}{N}$, генеральной средней $\bar{X} = \frac{1}{N}\sum_k x_k N_k$ и генеральной дисперсии $\sigma^2 = \frac{1}{N}\sum (x_k - \bar{X})^2 N_k$ выбираются статистики: выборочная доля $\xi^* = p^*$, выборочная средняя $\bar{X}^* = \frac{1}{n}\sum_k (x_k - \bar{X}^*) m_k$ и выборочная дисперсия $D^* = \frac{1}{n}\sum (x_k - \bar{X}^*)^2 m_k$. При этом в результате дальнейшей проверки устанавливается, что первые две обладают свойством несмещенности, а последняя для этой цели должна быть умножена на корректирующий множитель $\frac{n}{n-1}$.

Метод наименьших квадратов. Этот метод является одним из наиболее простых приемов построения оценок. Суть его заключается в том, что статистика определяется из условия минимизации суммы квадратов отклонений выборочных данных от определяемой оценки. Пусть, например, необходимо выбрать оценку ξ^* для генеральной средней $\bar{X}$. Следуя методу наименьших квадратов, определим ее из условия минимизации выражения $u = \sum_{j=1} (x_j - \xi^*)^2$. Используя необходимые условия экстремума получим

$$\frac{du}{d\xi^*} = -2\sum (x_j - \xi^*) = 0,$$

откуда

$$\xi^* = \frac{1}{n}\sum_{j=1}^{n} x_j = \bar{X}^*,$$

т. е. искомая оценка есть *средняя выборочная*.

Метод максимального правдоподобия. Этот метод, указанный еще К. Гауссом и разработанный Р. Фишером, играет большую роль в теории статистического оценивания, поэтому мы остановимся на нем более подробно. Пусть генеральное распределение зависит от одного параметра и задается плотностью распределения $f(x, \xi)$. Тогда, как уже указывалось в § 9.2, плотность распределения вероятностей выборки задается функцией

$$f(x_1, \ldots, x_j, \ldots, x_n, \xi) = f(x_1, \xi) \times f(x_2, \xi) \ldots f(x_n, \xi). \qquad (9.8)$$

В этой функции $x_1, \ldots, x_n$ — переменные, а ξ — данное значение параметра. Будем рассматривать данную выборку с получен-

ными фиксированными значениями $x_1, \ldots, x_n$ и поставим вопрос: какова вероятность получить эту выборку при различных значениях параметра ξ? Плотность распределения этой вероятности будет задаваться той же функцией (9.8), в которой теперь следует $x_1, \ldots, x_n$ рассматривать как постоянные, а оцениваемый параметр ξ — как переменную θ. При таком рассмотрении (9.8) называют *функцией правдоподобия выборки.*

Пусть, например, распределение генеральной совокупности подчиняется нормальному закону с параметрами $\xi_1 = \bar{X}$ и $\xi_2 = \sigma^2$, т. е.

$$f_N(x, \bar{X}, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\bar{X})^2}{2\sigma^2}}.$$

В таком случае функция правдоподобия возвратной выборки будет иметь вид

$$f_N(x_1, \ldots, x_n, \theta_1, \theta_2) = \frac{e^{-\frac{\sum (x_j-\theta_1)^2}{2\theta_2}}}{(2\pi\theta_2)^{n/2}}. \qquad (9.9)$$

Пусть теперь генеральное распределение характеризуется некоторой долей p признака A. Тогда возвратную выборку объема n можно рассматривать как n повторных независимых испытаний и вероятность получить выборку с относительной частотой признака $p^* = \frac{m}{n}$ можно определить по биномиальной формуле

$$f_B\left(\frac{m}{n}, p\right) = C_n^m p^m (1-p)^{n-m}.$$

Если теперь в этом равенстве рассматривать $\frac{m}{n}$ как постоянную величину, а параметр p — как переменную θ, то получим функцию правдоподобия выборки для этого случая:

$$f\left(\frac{m}{n}, \theta\right) = C_n^m \theta^m (1-\theta)^{n-m}. \qquad (9.10)$$

Идея метода максимального правдоподобия заключается в том, что в качестве оценки неизвестного параметра принимается то его значение, при котором данная выборка наиболее вероятна (*правдоподобна*), т. е. при котором функция правдоподобия достигает максимума.

При практической реализации этого метода удобно от функции правдоподобия перейти к ее логарифму:

$$L(x_1, \ldots, x_n, \theta) = \ln f(x_1, \ldots, x_n, \theta) = \sum_j \ln f(x_j; \theta), \qquad (9.11)$$

условия максимизации которого остаются теми же.

Обратимся к анализу рассмотренных выше двух функций правдоподобия. Из (9.9) получаем

$$L(x_1, \ldots, x_n, \theta_1, \theta_2) = \frac{-n}{2}(\ln 2\pi + \ln \theta_2) - \frac{\sum (x_j - \theta_2)^2}{2\theta_2}.$$

Необходимые условия экстремума этой функции дают

$$\frac{\partial L}{\partial \theta_1} = \frac{1}{2\theta_2}\sum 2(x_j - \theta_1) = 0; \quad \frac{\partial L}{\partial \theta_2} = \frac{-n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum (x_j - \theta_1)^2 = 0.$$

Из первого уравнения получаем

$$\theta_1 = \frac{1}{n}\sum_j x_j,$$

т. е. для оценки параметра $\bar{X}$ должна быть выбрана статистика $\theta_1 = \bar{X}^*$.

Из второго уравнения имеем

$$\theta_2 = \frac{1}{n}\sum_j (x_j - \theta_1)^2 = \frac{1}{n}\sum_j (x_j - \bar{X}^*)^2,$$

т. е. для оценки параметра σ^2 должна быть выбрана статистика $\theta_2 = D^*$ (выборочная дисперсия).

Теперь рассмотрим функцию правдоподобия (9.10). После логарифмирования получим

$$L\left(\frac{m}{n}, \theta\right) = \ln C_m^n + m\ln\theta + (n-m)\ln(1-\theta).$$

Необходимое условие максимума дает $\frac{dL}{d\theta} = \frac{m}{\theta} - \frac{n-m}{1-\theta} = 0$, откуда следует $\theta = \frac{m}{n}$.

Таким образом, оценкой доли признака по методу максимального правдоподобия является относительная частота $p^* = \frac{m}{n}$.

9.5. Оценка доли признака

Мы уже показали в предыдущем параграфе, что при определении доли признака в генеральной совокупности $\xi = p$ оценкой его по методу максимального правдоподобия является выборочная доля $p^* = \frac{m}{n}$. Эта оценка является *состоятельной,* что следует из закона больших чисел в форме Бернулли, и *несмещенной.* Последнее вытекает из доказанного ранее соотношения

$$M(p^*) = M\left(\frac{m}{n}\right) = p.$$

Покажем, что оценка p^* является *достаточной.* Если доля признака в генеральной совокупности равна p, то вероятность любой выборки, относительная частота признака в которой есть $m^* = np$, будет равна $C_n^m p^m q^{n-m}$, а вероятность одной данной выборки — $p^m q^{n-m}$.

Следовательно, по определению условной вероятности получим,

что вероятность данной выборки при заданном значении $m = p^*n$ будет равна

$$\frac{p^m q^{n-m}}{C_n^m p^m q^{n-m}} = \frac{1}{C_n^m}.$$

Таким образом, эта вероятность оказалась не зависящей от параметра p, что является условием *достаточности* оценки.

Наконец, можно доказать, что p^* является и эффективной оценкой, т. е. обладает наименьшей дисперсией из всех возможных оценок доли признака p. Для оценки погрешности приближенного равенства $p \approx p^*$, т. е. нахождения предельной ошибки выборки, необходимо обратиться к определению интервальной оценки с помощью равенства

$$P(p^* + \varepsilon_1 < p < p^* + \varepsilon_2) = 1 - \alpha. \qquad (9.12)$$

Дальнейшие расчеты зависят от выбранной для статистики p^* функции распределения.

Расчет по биномиальному закону распределения. Если доля признака в генеральной совокупности p, то распределение возвратной выборки при любом n в точности описывается биномиальным распределением

$$P_n(m) = C_n^m p^m q^{n-m}.$$

Биномиальное распределение при $p \neq 0{,}5$ несимметричное, поэтому согласно примечанию в § 9.4 определим для заданного α интервал $(p_1; p_2)$ как наименьший, для которого вероятности попадания левее p_1 и правее p_2 будут равны $\frac{\alpha}{2}$:

$$\sum_{m=0}^{m^*-1} C_n^m p^m q^{n-m} = \frac{\alpha}{2}; \quad \sum_{m=m^*}^{n} C_n^m p^m q^{n-m} = \frac{\alpha}{2}, \qquad (9.13)$$

где $m^* = p^*n$.

Из первого уравнения найдем p_1 и из второго — p_2, в результате чего получим интервальную оценку $p_1 < p < p_2$ с доверительной вероятностью $p = 1 - \alpha$. Для удобства решения уравнений (9.13), задавшись различными значениями n и обычно используемыми на практике уровнями значимости α (0,001; 0,005; 0,01; 0,05; 0,1), строят графики зависимостей $p_1 = p_1(p^*)$ и $p_2 = p_2(p^*)$ (см. рис. 9.6). На рисунке показано определение $p_1 = 0{,}4$ и $p_2 = 0{,}77$ для $n = 30$ и $\alpha = 0{,}05$ по выборочному значению $p^* = 0{,}6$. Есть специальные таблицы для определения p_1 и p_2 по данным n, α и p^*. Заметим, что изложенный точный метод расчета следует применять лишь при малом n, так как при большом n (примерно $n \geqslant 20$) можно использовать для выборки асимптотические распределения Лапласа (если $np \geqslant 10$) или Пуассона (если p мало).

Расчет по асимптотической формуле Лапласа. Этот расчет применим при достаточно большом объеме выборки n (практически начиная с $n = 20$—30) и не очень малом значении

ожидаемой доли признака p (так что величина $np \geqslant 10$). При этих условиях биномиальное распределение хорошо аппроксимируется нормальным распределением с параметрами

$$a = M(p^*) = p; \quad b^2 = D(p^*) = \frac{pq}{n}, \quad \text{или} \quad \sigma(p^*) = \sqrt{\frac{pq}{n}}. \quad (9.14)$$

Нормальное распределение симметричное, и для него расчет доверительного интервала может быть произведен по формуле

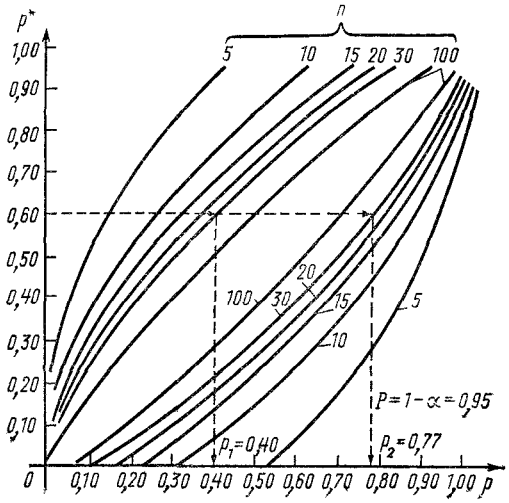


Рис. 9.6.
График определения доверительного интервала при оценке доли признака по биномиальному распределению

Лапласа

$$P(|p^* - p| < \varepsilon_\alpha) = 2\Phi(z_\alpha) = 1 - \alpha, \quad (9.15)$$

где предельная ошибка выборки ε_α определяется из равенства *

$$\varepsilon_\alpha = z_\alpha \sigma(p^*). \quad (9.16)$$

Таким образом, с вероятностью $P = 1 - \alpha$ выполняется неравенство

$$|p - p^*| < z_\alpha \sqrt{\frac{p(1-p)}{n}}.$$

Возведя обе его части в квадрат, получим равновероятное неравенство $(p - p^*)^2 < \frac{z^2 p(1-p)}{n}$, решая которое относительно p, получим границы доверительного интервала:

$$p_{1,2} = \frac{p^* + \frac{z^2}{2n} \pm z\sqrt{\frac{p^*(1-p^*)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}. \quad (9.17)$$

* Обозначением ε_α подчеркивается, что предельная ошибка выборки определяется в зависимости от дачного уровня значимости.

Однако при большом n можно сделать и второй шаг в использовании асимптотических соотношений, заменив в (9.14) неизвестный параметр p на его точечную оценку p^*, откуда получим

$$\sigma(p^*) = \sqrt{\frac{p^*(1-p^*)}{n}}, \tag{9.18}$$

называемую средней квадратической ошибкой выборки. Теперь можно из (9.16) определить предельную ошибку выборки, а значит, и границы доверительного интервала:

$$p^* - \varepsilon_\alpha < \quad < p^* + \varepsilon_\alpha. \tag{9.19}$$

Формула (9.15) связывает между собой в конечном счете три величины: *доверительную вероятность* $P = 1 - \alpha$, предельную ошибку выборки ε и объем выборки n. Соотношение между этими величинами при $p = 0{,}5$ и разных значениях n показано на рис. 9.4.

В каждой конкретной задаче две из этих величин задаются и определяется третья величина. В результате получаем следующие три типа задач:

I. Даны n и P, определяется ε.

II. Даны n и ε, определяется P.

III. Даны P и ε, определяется n.

Первые два типа связаны с анализом результатов уже произведенной выборки данного объема n, следовательно, и с найденной точечной оценкой p^*.

Целью решения задач третьего типа является расчет необходимого объема выборки n, который обеспечит заданную предельную ошибку выборки ε при выбранной величине доверительной вероятности $P = 1 - \alpha$.

Расчет производится следующим образом: по заданной доверительной вероятности $P = 1 - \alpha$ определяем из равенства $2\Phi(z) = 1 - \alpha$ величину z. Из (9.16) с учетом (9.18) получаем

$$n = \frac{z^2 p^*(1-p^*)}{\varepsilon^2}. \tag{9.20}$$

Трудности при решении задачи обусловлены тем, что в (9.20) входит величина p^*, получаемая в результате выборки. А между тем речь идет об определении n до осуществления выборки.

Поскольку p^* неизвестно, то определяем из этого равенства, при каком значении p^* величина n будет максимальной. Используя обычный метод исследования функции на максимум, получаем

$$\frac{dn}{dp^*} = \frac{z^2}{\varepsilon^2}(1 - 2p^*) = 0,$$

откуда

$$p^* = \frac{1}{2}; \quad \frac{d^2n}{dp^{*2}} = -2\frac{z^2}{\varepsilon^2} < 0.$$

Следовательно,

$$n_{\max} = \frac{z^2}{4\varepsilon^2}. \tag{9.21}$$

Если имеются некоторые предположения о доле признака $p = p_1$, то вычисляется значение n_1 по формуле (9.20). После проведения выборки находится значение $p^* = p_2$ и вновь вычисляется по (9.20) объем выборки n_2. Если окажется, что величина n_2 существенно больше, чем n_1, то производится повторная выборка объема n_2. Если же окажется, что $n_2 \leqslant n_1$ (или больше, но незначительно), то результат первой выборки окончательный и принятые доверительная вероятность и ошибка выборки обеспечены.

Расчет по асимптотической формуле Пуассона. При малом значении ожидаемой величины p и большом n (так

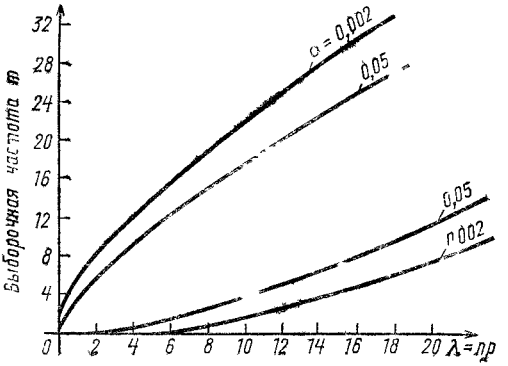


Рис. 9.7.
График определения доверительного интервала при оценке доли признака по закону распределения Пуассона

что $np < 10$) расчет по формуле Лапласа может привести к большой погрешности. В этом случае лучшее приближение дает асимптотическая формула Пуассона с использованием таблицы функции $F_{\text{П}}(m, \lambda)$, где параметр $\lambda = np \approx np^*$.

Распределение Пуассона, как и биномиальное, несимметрично. Поэтому для определения доверительного интервала необходимо найти левую и правую его границы из условия, чтобы площадь отсекаемых «хвостов» распределения в сумме составляла α (см. рис. 9.2). Как и для биномиального распределения, границы доверительного интервала можно найти из решения уравнений

$$F_{\text{П}}(m^*, \lambda_1) = \sum_{m=0}^{m^*-1} \frac{\lambda_1^m e^{-\lambda_1}}{m!} = \frac{\alpha}{2}; \quad F_{\text{П}}(m^*, \lambda_2) = \sum_{m=0}^{m^*} \frac{\lambda_2^m e^{-\lambda_2}}{m!} = 1 - \frac{\alpha}{2},$$

где $m^* = np^*$, $p_1 = \frac{\lambda_1}{n}$ и $p_2 = \frac{\lambda_2}{n}$, или с помощью графика, построенного на рис. 9.7.

В заключение отметим, что рассмотренные выше (с. 164) типы задач и приемы их решения можно распространить и на безвозвратную выборку. Следует лишь в этом случае ввести корректирующий множитель при расчете средней квадратической ошибки выборки:

$$\sigma'(p^*) = \sigma(p^*) \sqrt{\frac{N-n}{N-1}} \approx \sigma(p^*) \sqrt{1 - \frac{n}{N}}, \qquad (9.18')$$

где N — объем генеральной совокупности. При относительно небольшом объеме выборки в сравнении с объемом генеральной совокупности этот поправочный множитель близок к 1 и, следовательно, результаты расчетов для возвратной и безвозвратной выборок почти совпадают. Поэтому формулу (9.18′) следует применять лишь при большом удельном весе выборки. Существенное отличие будет лишь при решении задачи III типа — определении необходимого объема выборки n. Так, из равенства $\varepsilon = z\sigma'(p^*)$ получим

$$n = \frac{z^2 p(1-p)N}{z^2 p(1-p) + \varepsilon^2 (N-1)}. \qquad (9.20')$$

Соответственно изменится и формула для

$$n_{\max} = \frac{t^2 N}{1 + 4\varepsilon^2 (N-1)}. \qquad (9.21')$$

Можно показать, что если определен необходимый объем возвратной выборки n, то соответствующий объем безвозвратной выборки n' найдется из формулы

$$n' = \frac{1}{1/n + 1/N}. \qquad (9.22)$$

9.6. Точечные оценки для средней и дисперсии генеральной совокупности

Обозначим через $\bar{X}$ и σ^2 среднюю и дисперсию генеральной совокупности. В основе вывода последующих соотношений лежит вероятностная модель выборки как системы n случайных величин X_j, имеющих одно и то же распределение, совпадающее с генеральными, и, следовательно, для которых

$$M(X_j) = \bar{X}; \quad D(X_j) = \sigma^2.$$

Рассмотрим вначале возвратную выборку объемом n, для которой X_j — независимые случайные величины. Построим статистику $\theta = \bar{X}^* = \frac{1}{n}\sum_{j=1}^{n} X_j$, представляющую собой выборочную среднюю. Используя свойства математического ожидания и дисперсии, получим:

$$M(\bar{X}^*) = \frac{1}{n} M\left(\sum_j X_j\right) = \frac{1}{n}\sum_j M(X_j) = \frac{n\bar{X}}{n} = \bar{X}; \qquad (9.23)$$

$$D(\bar{X}^*) = \frac{1}{n^2} D\left(\sum_j X_j\right) = \frac{1}{n^2}\sum_j D(X_j) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \qquad (9.24)$$

Выберем статистику $\theta = \bar{X}^*$ для оценки генеральной средней $\bar{X}$. На основании закона больших чисел (теорема Чебышева) имеем $\bar{X}^* \xrightarrow{P} M(\bar{X}^*) = \bar{X}$, откуда следует *состоятельность* этой оценки. Из соотношения (9.23), верного при любом n, следует *несмещенность* оценки $\bar{X}^*$. Наконец, если распределение генеральной совокупности нормальное, то $\bar{X}^*$ будет также и *эффективной* оценкой.

Следовательно, $\bar{X}^*$ удовлетворяет всем требованиям, предъявляемым к оценке, и может служить точечной оценкой параметра $\bar{X}$, т. е. может быть записано приближенное равенство

$$\bar{X} \approx \bar{X}^*.$$

Оценка точности этого равенства будет произведена в следующем параграфе с помощью построения доверительного интервала. Теперь обратимся к оценке генеральной дисперсии, для чего воспользуемся статистикой

$$\theta = D^* = \frac{1}{n}\sum_i (X_i - \bar{X}^*)^2 = \frac{1}{n}\sum_i X_i^2 - \bar{X}^{*2},$$

представляющей собой выборочную дисперсию. Подставив выражение для $\bar{X}^* = \frac{1}{n}\sum_i X_i$, откуда $\bar{X}^{*2} = \frac{1}{n^2}\left(\sum_i X_i^2 + \sum_i \sum_{\substack{k \\ i \neq k}} X_i X_k\right)$, получим после приведения подобных членов

$$D^* = \frac{n-1}{n^2}\sum_i X_i^2 - \frac{1}{n^2}\sum_{i \neq k}\sum X_i X_k.$$

Вычитая из всех X_i (или X_k) постоянное число $\bar{X}$, отчего значение дисперсии не изменится, и вычисляя математическое ожидание от обеих частей равенства, придем к выражению

$$M(D^*) = \frac{n-1}{n^2}\sum_i M(X_i - \bar{X})^2 - \frac{1}{n^2}\sum_{i \neq k}\sum M[(X_i - \bar{X})(X_k - \bar{X})].$$

Но

$$M(X_i - \bar{X})^2 = \sigma^2 \text{ и } M[(X_i - \bar{X})(X_k - \bar{X})] = 0$$

из-за независимости системы X_i. Получим окончательно

$$M(D^*) = \frac{n-1}{n}\sigma^2. \qquad (9.25)$$

Таким образом, статистика $\theta = D^*$ является смещенной оценкой для генеральной дисперсии $\left(\text{смещение } a = \frac{-1}{n}\sigma^2\right)$. Поэтому для построения несмещенной точечной оценки дисперсии воспользуемся «исправленной» выборочной дисперсией

$$s^2 = \frac{n}{n-1}D^* = \frac{1}{n-1}\sum_i (X_i - \bar{X}^*)^2. \qquad (9.26)$$

Эта оценка будет несмещенной, так как

$$M(s^2) = \frac{n}{n-1} M(D^*) = \sigma^2,$$

и состоятельной в силу того же закона больших чисел. Поэтому можно записать при конечном n приближенное равенство

$$\sigma^2 \approx s^2.$$

Заметим, что при большом n отношение $\frac{n}{n-1} \approx 1$, и потому значение $s^2 \approx D^*$.

Мы рассмотрели построение точечных оценок для $\bar{X}$ и σ^2 для возвратной выборки. В случае безвозвратной выборки можно показать, что точечная оценка средней будет той же (т. е. $\bar{X}^*$), а точечная оценка дисперсии должна быть заменена на

$$s'^2 = \frac{N-1}{N} s^2. \tag{9.27}$$

Здесь множитель $\frac{N-1}{N}$, где N — объем генеральной совокупности, учитывает безвозвратность выборки. В случае безвозвратной выборки изменится и выражение для $D(\bar{X}^*)$, которое потребуется для построения доверительного интервала при оценке средней:

$$D'(\bar{X}) = \frac{N-n}{N-1} \frac{\sigma^2}{n}. \tag{9.28}$$

При относительно небольшом объеме выборки n по сравнению с объемом генеральной совокупности N (например, при 5%-ной выборке, для которой $\frac{n}{N} = 0{,}05$) корректирующий множитель

$$\frac{N-n}{N-1} \approx 1 \text{ и } D'(\bar{X}) \approx D(\bar{X}^*) = \frac{\sigma^2}{n}.$$

Наконец, при построении доверительного интервала для дисперсии нам необходимо знать $D(s^2)$. Приведем без доказательства соответствующую формулу:

$$D(s^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right), \tag{9.29}$$

где μ_4 — центральный момент 4-го порядка.

Приведенные формулы могут быть продемонстрированы на примере, рассмотренном в § 9.2.

В примере на с. 149 мы имеем $n = 2$; $\bar{X} = \frac{41}{7}$; $\sigma^2 = \frac{20}{49}$. Приведенные в явном виде все девять возвратных выборок позволили построить распределение $\bar{X}^*$ и D^* (табл. 9.2 и 9.3). Из этих распределений мы получили $M(\bar{X}^*) = {}^{41}/_{7} = \bar{X}$ и $M(D^*) = {}^{10}/_{49} = \frac{1}{2}\sigma^2 = \frac{n-1}{n}\sigma^2$, что соответствует формулам (9.23) и (9.25). Наконец, по данным табл. 9.2 можно рассчитать $D(\bar{X}^*)$:

$$D(\bar{X}^*) = {}^{10}/_{49} = \frac{\sigma^2}{2} = \frac{\sigma^2}{n},$$

что соответствует (9.24).

Теперь задавшись общей численностью генеральной совокупности $N = 7$, построим всевозможные безвозвратные выборки того же объема $n = 2$:

Таблица 9.5

	Номер выборки				
	1	2	3	4	5
Результат выборки	5; 5	5; 6	5; 7	6; 6	6; 7
Вероятность выборки	1/21	8/21	2/21	6/21	4/21
Выборочная средняя	5	5,5	6	6	6,5
Выборочная дисперсия	0	0,25	1,0	0	0,25

Аналогично предыдущему получаем $M(\overline{X}^*) = {}^{41}/_{7} = \overline{X}$. Далее, $M(D^*) = \frac{5}{21}$. Но если ввести согласно (9.26) исправленную выборочную дисперсию $s'^2 = \frac{N-1}{N}\frac{n}{n-1}D^*$, то должны получить $M(s'^2) = \sigma^2 = {}^{20}/_{49}$. Действительно,

$$M(s'^2) = \frac{N-1}{N}\frac{n}{n-1}M(D^*) = \frac{7-1}{7}\cdot\frac{2}{2-1}\cdot\frac{5}{21} = \frac{20}{49} = \sigma^2.$$

Наконец, рассчитав $D'(\overline{X}^*)$ по табл. 9.5 и сравнив с расчетом по формуле (9.28), убеждаемся в ее справедливости.

9.7. Интервальные оценки средней и дисперсии нормально распределенной генеральной совокупности

Для построения интервальных оценок необходимо знать распределения статистик, выбранных для построения оценок. Этого можно достигнуть либо прямым методом, если исходить из заданного генерального распределения, откуда как следствие получается выборочное распределение и из него — распределения статистик (см. § 9.2), либо косвенным методом, позволяющим при некоторых общих предположениях получить асимптотические (при $n \to \infty$) распределения статистик. Первый метод ведет к точному построению доверительного интервала, но, очевидно, он имеет больше теоретическое значение, потому что, как правило, при проведении выборочного наблюдения генеральное распределение неизвестно. Второй метод дает возможность лишь приближенного построения доверительного интервала для выборки конечного объема n.

Учитывая широкую распространенность нормального распределения, рассмотрим точный метод для выборки из нормально распределенной генеральной совокупности, заданной плотностью вероятности:

$$f_N(x, \overline{X}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\overline{X})^2}{2\sigma^2}}.$$

Оценка средней при известной генеральной дисперсии σ^2. Выборочная средняя $\overline{X}^*$ как средняя

арифметическая n нормально распределенных случайных величин $\overline{X}^* = \frac{1}{n}\sum_j X_j$ также имеет нормальный закон распределения с математическим ожиданием $\overline{X}$ и дисперсией $\sigma^2(\overline{X}^*) = \frac{\sigma^2}{n}$. Применяя к $\overline{X}^*$ (8.41), получим

$$P(|\overline{X}^* - X| < \varepsilon) = P(\overline{X}^* - \varepsilon < \overline{X} < \overline{X}^* + \varepsilon) = 2\Phi(z), \quad (9.30)$$

где

$$z = \frac{\varepsilon}{\sigma(\overline{X}^*)} = \frac{\varepsilon\sqrt{n}}{\sigma}.$$

Задавшись доверительной вероятностью $P = 1 - \alpha$, определяем из равенства $2\Phi(z) = 1 - \alpha$ значение z_α, далее величину

$$\varepsilon_\alpha = z_\alpha \sigma(\overline{X}^*) = z_\alpha \frac{\sigma}{\sqrt{n}} \quad (9.31)$$

и, наконец, доверительный интервал

$$(\overline{X}^* - \varepsilon_\alpha;\ \overline{X}^* + \varepsilon_\alpha).$$

Наоборот, если задана предельная ошибка ε и требуется определить доверительную вероятность P, то схема решения задачи следующая:

$$\varepsilon \longrightarrow z = \frac{\varepsilon\sqrt{n}}{\sigma} \longrightarrow \Phi(z) \longrightarrow P = 2\Phi(z). \quad (9.32)$$

Наконец, определение объема выборки n по данным P и ε проводится по следующей схеме:

$$P = 2\Phi(z) \longrightarrow z \longrightarrow n = \frac{z^2\sigma^2}{\varepsilon^2}. \quad 9.33)$$

1. Измерение длины 50 случайно отобранных после изготовления деталей дало $\overline{X}^* = 10$ см. Определить с вероятностью $P = 0{,}95$ доверительные границы для среднего размера деталей $\overline{X}$ во всей партии, если есть основания полагать, что генеральная дисперсия $\sigma^2 = 0{,}09$. На основании теоремы Ляпунова можно исходить из предположения о нормальном распределении размеров деталей. Дано: $n = 50$, $\overline{X}^* = 10$ см, $\sigma^2 = 0{,}09$ и $P = 0{,}95$. Из равенства $P = 2\Phi(z) = 0{,}95$ по таблицам функций Лапласа находим $z = 1{,}96$, откуда $\varepsilon = \frac{z\sigma}{\sqrt{n}} = 0{,}083$. Таким образом, получаем $10 - 0{,}083 \leqslant \overline{X} \leqslant 10 + 0{,}083$.

2. Определить при тех же условиях, с какой доверительной вероятностью можно гарантировать ошибку выборки,не превышающую 0,05.

По величине $\varepsilon = 0{,}05$ вычисляется $z = \frac{\varepsilon\sqrt{n}}{\sigma} = \frac{0{,}05 \cdot 50}{0{,}3} = 1{,}1785$, откуда по таблице $\Phi(z)$

$$P = 2\Phi(1{,}1785) \approx 0{,}76.$$

3. Определить по данным предыдущего примера объем выборки n, при котором указанная предельная ошибка $\varepsilon = 0{,}05$ гарантируется с вероятностью $P = 0{,}95$.

Из $P = 2\Phi(z) = 0{,}95$ находим $z = 1{,}96$, откуда

$$n = \frac{z^2\sigma^2}{\varepsilon^2} = \frac{1{,}96^2 \cdot 0{,}09}{0{,}0025} \approx 70.$$

Предположение о том, что генеральная дисперсия σ^2 известна при неизвестной генеральной средней $\bar{X}$, имеет с первого взгляда лишь теоретическое значение, так как для расчета $\sigma^2 = \frac{1}{N}\sum_j (x_j - \bar{X})^2 N_j$ необходимо знать $\bar{X}$. Однако рассмотренный случай имеет и большое практическое значение, как это будет показано в следующем параграфе.

Оценка средней при неизвестной генеральной дисперсии σ^2. Чтобы обойти затруднение, вызванное неопределенностью распределения $\bar{X}^*$, зависящего от неизвестного параметра, воспользуемся статистикой

$$T = \frac{\bar{X}^* - \bar{X}}{s}\sqrt{n}, \tag{9.34}$$

подчиняющейся распределению Стьюдента с $\nu = n - 1$ степенями свободы, не зависящему от параметра σ. Используя функцию распределения Стьюдента $S(t, \nu) = P(T < t)$, можно записать равенство, аналогичное формуле Лапласа:

$$P\left\{\frac{|\bar{X}^* - \bar{X}|}{s}\sqrt{n} < t\right\} = 2S(t, \nu) = 1 - \alpha.$$

Раскрыв неравенство, записанное с помощью абсолютной величины, получим после несложных преобразований

$$P(\bar{X}^* - \varepsilon < \bar{X} < \bar{X}^* + \varepsilon) = 2S(t, \nu), \text{ где } \varepsilon = \frac{ts}{\sqrt{n}} \tag{9.35}$$

Решение задач с помощью этого равенства аналогично решению с помощью равенства (9.30). Лишь определение n несколько осложняется из-за того, что оно входит также в параметр $\nu = n - 1$. Поэтому можно воспользоваться схемой последовательных приближений. Вначале находят n_1, приняв $\sigma^2 = s^2$. По найденному n_1 и соответственно $\nu_1 = n_1 - 1$ и заданному значению $P = 1 - \alpha$ определяют t_1 и вычисляют $n_2 = \frac{t_1^2 S^2}{\varepsilon^2}$. Теперь можно снова повторить расчет по $\nu_2 = n_2 - 1$ и т. д. Итерации заканчиваются, если окажется $n_i \approx n_{i-1}$.

Оценки дисперсии. Предположение о нормальном распределении генеральной совокупности позволяет построить интервальную оценку дисперсии. Для этого введем статистику

$$\chi^2 = \frac{s^2}{\sigma^2}(n - 1), \tag{9.36}$$

имеющую χ^2-распределение с $\nu = n - 1$ степенями свободы. Обозначив его функцию распределения через $F_{\chi^2}(u, \nu) = P(\chi^2 < u)$, получим, как обычно,

$$P(u_1 < \chi^2 < u_2) = F_{\chi^2}(u_2, \nu) - F_{\chi^2}(u_1, \nu).$$

Перейдя в неравенствах к обратным величинам, подставив выражение для χ^2 и умножив все члены неравенства на $s^2(n - 1) > 0$,

получим

$$P\left(\frac{s^2(n-1)}{u_2} < \sigma^2 < \frac{s^2(n-1)}{u_1}\right) = F_{\chi^2}(u_2, \nu) - F_{\chi^2}(u_1, \nu). \qquad (9.37)$$

Этого уравнения недостаточно, чтобы при заданном уровне значимости α найти u_1 и u_2. Поэтому, как и ранее, воспользуемся приближенным расчетом из уравнений:

$$P(\chi^2 < u_1) = F_{\chi^2}(u_1, \nu) = \frac{\alpha}{2};\ P(\chi^2 < u_2) = F_{\chi^2}(u_2, \nu) = 1 - \frac{\alpha}{2}. \qquad (9.38)$$

В результате получим доверительный интервал для дисперсии

$$\left(\frac{s^2(n-1)}{u_2};\ \frac{s^2(n-1)}{u_1}\right). \qquad (9.39)$$

Определим по данным примера на с. 170 доверительный интервал для генеральной дисперсии σ^2, если в результате выборки получено $s^2 = 0{,}09$.

Имеем $n = 50$, $\alpha = 0{,}05$ и $s^2 = 0{,}09$. При определении u_1 и u_2 по формулам (9.38) нельзя воспользоваться таблицей $F_{\chi^2}(u, \nu)$, которая содержит значения ν только от 1 до 30, а в данном примере $\nu = 49$. Поэтому воспользуемся ранее указанным асимптотическим свойством χ^2-распределения, согласно которому при $\nu \geqslant 30$ величина $z = \sqrt{2\chi^2} - \sqrt{2\nu - 1}$ подчиняется стандартному нормальному распределению. Поэтому по заданной величине α найдем z из формулы Лапласа:

$$P(|Z| < z) = P(-z < Z < z) = 2\Phi(z) = 1 - \alpha.$$

Подставим вместо z его выражение в двустороннее неравенство $-z < \sqrt{2\chi^2} - \sqrt{2\nu - 1} < z$, получим доверительный интервал для χ^2:

$$\frac{1}{2}(\sqrt{2\nu - 1} - z)^2 < \chi^2 < \frac{1}{2}(\sqrt{2\nu - 1} + z)^2.$$

По значению $P = 1 - \alpha = 0{,}95$ находим из таблицы $\Phi(z)$ $z = 1{,}96$, откуда $u_1 = \frac{1}{2}(\sqrt{2 \cdot 49 - 1} - 1{,}96)^2 = 31{,}1$: $u_2 = \frac{1}{2}(\sqrt{2 \cdot 49 - 1} + 1{,}96)^2 = 69{,}7$.

Теперь согласно (9.39) находим доверительные границы для дисперсии

$$\frac{s^2(n-1)}{u_2} = \frac{0{,}09 \cdot 49}{69{,}7} = 0{,}063 \text{ и } \frac{s^2(n-1)}{u_1} = \frac{0{,}09 \cdot 49}{31{,}1} = 0{,}142$$

9.8. Приближенный метод интервальной оценки генеральной средней

В отличие от предыдущего параграфа не будем исходить из заданного вида генерального распределения, а прибегнем к косвенному методу построения распределения для статистик. Очевидно, что в этом случае определение доверительного интервала будет приближенным и притом существенно зависящим от объема выборки n.

Оценка средней для больших выборок. Будем предполагать объем выборки n настолько большим, чтобы заключение теоремы Ляпунова о нормальном распределении выборочной средней при $n \to \infty$ было приблизительно верным при данном n. Как уже указывалось в § 8.5, приближение распределения средней арифметической к нормальному происходит достаточно

быстро с ростом n, так, что уже начиная с $n = 20$ можно полагать его практически нормальным с параметрами $a = M(\bar{X}^*) = \bar{X}$ и $b = \sigma(\bar{X}^*) = \frac{\sigma}{\sqrt{n}}$. Это позволяет при оценке средней использовать уже применявшуюся формулу (9.30).

Так как величина σ при проведении выборки неизвестна, то, заменив ее точечной оценкой (несмещенной s или смещенной σ^*), получим выражение для *средней квадратической ошибки выборки*:

$$\sigma(\bar{X}^*) = \frac{\sigma^*}{\sqrt{n}}, \text{ или } \sigma(\bar{X}^*) = \frac{s}{\sqrt{n}}. \tag{9.40}$$

Из (9.31) находим предельную ошибку выборки

$$\varepsilon_\alpha = z_\alpha \sigma(\bar{X}^*) \tag{9.41}$$

и доверительные границы для оценки средней

$$\bar{X}^* - \varepsilon_\alpha < \bar{X} < \bar{X}^* + \varepsilon_\alpha. \tag{9.42}$$

Дальнейшие схемы решения трех типов задач, рассмотренных подробно и иллюстрированных примерами в § 9.7, остаются прежними.

В случае безвозвратной выборки прежде всего изменится согласно (9.27) точечная оценка для σ. Кроме того, при расчете $D(\bar{X}^*)$ следует пользоваться вместо (9.21) формулой (9.28).

В результате получим

$$\sigma'(\bar{X}^*) = \sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{N-1}{N}} \frac{s}{\sqrt{n}} = \sqrt{\frac{N-n}{N}} \frac{s}{\sqrt{n}}.$$

Теперь формула 9.40 запишется в виде

$$\sigma'(\bar{X}^*) = \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}. \tag{9.40'}$$

Корректирующий множитель $\sqrt{1 - \frac{n}{N}} \approx 1 - \frac{n}{2N}$ при малой величине $\frac{n}{N}$ (например, для 1 или 5% выборок) близок к 1, и потому расчет может производиться как для возвратной выборки.

Оценка средней по данным малой выборки. При малых значениях n ($n < 10$—20), т. е. при проведении так называемых *малых выборок*, в случае произвольного (заранее не известного) генерального распределения, приведенный выше приближенный метод построения интервальной оценки неприменим в силу двух обстоятельств:

1. Необоснованным становится предположение о нормальном распределении $\bar{X}^*$, верное, как это вытекает из центральной предельной теоремы, при больших n.

2. Необоснованной становится замена неизвестной генеральной дисперсии σ^2 на ее точечную оценку s^2, справедливая только для большого n в силу *свойства состоятельности*.

Поэтому для малых выборок можно использовать лишь выборочную среднюю $\bar{X}^*$ как точечную оценку, без оценки точности приближенного равенства $\bar{X} \approx \bar{X}^*$, т. е. без расчета доверительного интервала. При этом выбор $\bar{X}^*$ в качестве точечной оценки не может опираться на свойство состоятельности, которое, являясь асимптотическим, не может применяться для характеристики качества оценивания при конечном и малом n.

Остается условие несмещенности $\bar{X}^*$, не зависящее от n, но этого не достаточно. Как указывалось в § 9.2, асимптотическое условие состоятельности для суждения о качестве оценки может быть заменено требованием, чтобы распределение статистики $\theta = \xi^*$, выбираемой в качестве оценки параметра ξ, концентрировалось как можно больше около значения ξ. Естественной мерой такой концентрации может служить дисперсия статистики или, другими словами, условие минимизации $D(\xi^*)$ по сравнению с другими статистиками θ, получившее название *условия эффективности.*

Остановимся на одном из наиболее простых классов оценок, получивших название *линейных оценок* и являющихся линейными функциями результатов выборки.

Пусть $\theta = a_1X_1 + \ldots + a_jX_j + \ldots + a_nX_n$, где a_j — заданные постоянные, сумма которых $a_1 + \ldots + a_j + \ldots + a_n = 1$. Очевидно, $M(\theta) = \sum_j a_j M(X_j) = \bar{X} \sum_j a_j = \bar{X}$, т. е. θ есть несмещенная оценка. Далее, $D(\theta) = \sum a_j^2 D(X_j) = \sigma^2 \sum_j a_j^2$. Теперь нетрудно показать, что минимум $D(\theta)$ достигается при условии $a_1 = \ldots = a_j = \ldots = a_n = \frac{1}{n}$, откуда следует, что минимальной дисперсией из линейных оценок параметра $\bar{X}$ обладает средняя выборочная $\bar{X}^*$.

9.9. Статистические оценки при многоступенчатом отборе

Как мы увидим ниже, организация выборки по способу многоступенчатого отбора позволяет увеличить точность выборки при сохранении ее объема. Многоступенчатый отбор заключается в следующем: на первом этапе производится выбор некоторых групп (первичные единицы), на которые разбита генеральная совокупность, затем на втором этапе из отобранных групп отбираются подгруппы (вторичные элементы), на которые разбита каждая группа, и т. д., пока на каком-то этапе не отбираются первичные элементы наблюдения. Обычно многоступенчатый отбор производится в два этапа.

Таким образом, существует много разнообразных способов организации многоступенчатой выборки, что, естественно, приводит и к наличию разных расчетных формул. Ограничимся рассмотрением трех видов выборки, организованных с помощью многоступенчатого отбора.

Пусть генеральная совокупность разбита на s непересекающихся (т. е. не имеющих общих элементов) групп, или *серий* (гнезд), численностью в $N_1, \ldots, N_i, \ldots, N_s$ элементов. Каждая группа характеризуется групповой средней $\bar{X}_i$ и групповой дисперсией σ_i^2, а вся генеральная совокупность — генеральной средней $\bar{X}$ и генеральной дисперсией σ^2. На основании свойств средней и дисперсии (см. § 2.5) можно записать:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{s} \bar{X}_i N_i; \tag{9.43}$$

$$\sigma^2 = \bar{\sigma}^2_{\text{в гр}} + \sigma^2_{\text{м гр}}, \tag{9.44}$$

где $\bar{\sigma}^2_{\text{в гр}} = \frac{1}{N} \sum_i \sigma_i^2 N_i$ — *средняя групповых дисперсий* и $\sigma^2_{\text{м гр}} = \frac{1}{N} \sum (\bar{X}_i - \bar{X})^2 N_i$ — *дисперсия групповых средних* (межгрупповая дисперсия).

При одной и той же величине генеральной дисперсии соотношение между ее слагаемыми в равенстве (9.44) может меняться в зависимости от способа разбиения генеральной совокупности на группы.

Типическая выборка. Пусть генеральная совокупность разбита на отдельные группы, более однородные по изучаемому признаку, чем генеральная совокупность. Такие группы будем называть «типами» или «стратами». Так, при изучении производительности труда в отрасли предприятия могут быть разбиты на группы, более однородные по технической оснащенности или фондовооруженности; при бюджетных обследованиях населения можно рассматривать группы населения, более однородные по уровню доходов, чем все население, и т. д.

Для выборочной оценки генеральной средней произведем случайный отбор n_1 элементов из первой группы, n_2 элементов из второй группы, n_i элементов из i-й группы и т. д. Величинами n_i можно располагать по своему усмотрению, но так, чтобы $\sum_i n_i = n$. Выборка, произведенная указанным образом, получила название *типической* (*типологической* или *стратифицированной*).

Каждые n_i элементов, отобранных из i-й группы генеральной совокупности, образуют i-ю группу в выборке, и для них могут быть определены выборочные групповые средние $\bar{X}_i^*$ и выборочные групповые дисперсии D_i^*. Соответственно для всей выборки, состоящей из s групп, можно определить выборочную среднюю $\bar{X}^*$ и выборочную дисперсию D^*. При этом аналогично (9.43) может быть записано соотношение

$$\bar{X}^* = \frac{1}{n} \sum_i \bar{X}_i^* n_i. \tag{9.43'}$$

Для проверки несмещенности этой оценки найдем

$$M(\overline{X}^*) = \frac{1}{n}\sum_i M(\overline{X}_i^*)\, n_i.$$

Но каждая групповая средняя $\overline{X}_i^*$ есть несмещенная оценка для своего прототипа $\overline{X}_i$, т. е. $M(\overline{X}_i^*) = \overline{X}_i$. Подставляя в предыдущее равенство, получим

$$M(\overline{X}^*) = \frac{1}{n}\sum \overline{X}_i\, n_i. \tag{9.45}$$

В этом равенстве можно по разному задавать n_i. Простейшим видом типической выборки является *пропорциональная типическая выборка*, при которой n_i пропорциональны соответствующим численностям групп N_i, т. е.

$$\frac{n_i}{n} = \frac{N_i}{N} \quad (i = 1, \ldots, s). \tag{9.46}$$

Подставляя эти отношения в (9.45) и используя (9.43), получим

$$M(\overline{X}^*) = \frac{1}{n}\sum_i \overline{X}_i \cdot \frac{N_i}{N} n = \frac{1}{N}\sum_i \overline{X}_i N_i = \overline{X},$$

что подтверждает несмещенность $\overline{X}^*$ для типической выборки.

Перейдем к построению интервальной оценки, для чего необходимо определить $D(\overline{X}^*)$. Будем предполагать, что отбор производится по схеме возвратной выборки. Групповые средние в этом случае будут независимыми, и на основании свойства дисперсии получим из (9.43′)

$$D(\overline{X}^*) = \sum_i \frac{n_i}{n^2} D(\overline{X}_i^*). \tag{9.47}$$

Рассматривая каждую выборочную группу как возвратную выборку объема n_i, получим

$$D(\overline{X}_i^*) = \frac{\sigma_i^2}{n_i}.$$

Подставляя это выражение в (9.47), используя (9.46) и определение средней групповых дисперсий, получим последовательно

$$D(\overline{X}^*) = \sum_i \frac{n_i^2}{n^2}\frac{\sigma_i^2}{n_i} = \frac{1}{n}\sum_i \sigma_i^2 \frac{n_i}{n} = \frac{1}{n}\sum_i \sigma_i^2 \frac{N_i}{N} = \frac{\overline{\sigma}_{\text{вгр}}}{n}. \tag{9.48}$$

Из этого равенства можно сделать важное заключение о повышении точности типической выборки по сравнению с простой случайной выборкой. Действительно, при проведении последней

точность выборки оценивалась величиной

$$D(\bar{X}^*) = \frac{\sigma^2}{n} = \frac{\bar{\sigma}^2_{\text{вгр}}}{n} + \frac{\sigma^2_{\text{мгр}}}{n},$$

т. е. больше предыдущей на $\frac{\sigma^2_{\text{м гр}}}{n}$.

Таким образом, если при организации типической выборки удалось образовать однородные группы с небольшой внутригрупповой дисперсией, то результаты будут более точными, чем при простой выборке, так как разброс групповых средних элиминируется.

Формула (9.48) может быть использована непосредственно для построения доверительного интервала, как это было показано в § 9.8.

При произвольном генеральном распределении и большом n, опираясь на асимптотическое нормальное распределение $\bar{X}^*$, получим:

$$P(|\bar{X}^* - \bar{X}| < \varepsilon| = 2\Phi(z), \text{ где } \varepsilon = \frac{z}{\sigma(\bar{X}^*)} = \frac{z\sqrt{n}}{\sigma_{\text{згр}}}.$$

Но

$$\bar{\sigma}^2_{\text{вгр}} = \frac{1}{N}\sum_i \sigma_i^2 N_i = \frac{1}{n}\sum_i \sigma_i^2 n_i \approx \frac{1}{n}\sum_i s_i^2 n_i, \tag{9.49}$$

где s_i^2 — несмещенные оценки * для групповых дисперсий.

Дальнейшие схемы расчетов аналогичны рассмотренным в § 9.8. Можно показать, что более точной будет типическая выборка, в которой размеры выборочных групп n_i выбираются не из соотношения (9.46), а пропорционально групповым дисперсиям, т. е.

$$n_i = \frac{N_i\sigma_i}{\sum \sigma_i N_i}. \tag{9.46'}$$

Такая типическая выборка, получившая название *оптимальной*, обладает минимальной дисперсией $\sigma^2(\bar{X}^*)$. Однако для ее проведения необходимо знать хотя бы ориентировочно групповые дисперсии σ_i^2 генеральной совокупности.

В заключение отметим, что типическая выборка может производиться и для оценки доли признака. Тогда точность ее характеризуется величиной

$$\sigma^2(p^*) = \frac{1}{n}\sum p_i^*(1 - p_i^*)\frac{n_i}{n}.$$

Механическая выборка. Если все группы имеют одинаковый объем, т. е. $N_1 = N_2 = \ldots = N_i = \ldots = N_s = \frac{N}{s}$, и из каждой группы отбирается по одному элементу, т. е. $n_1 = n_2 = \ldots = n_i = \ldots = n_s = 1$, откуда $n = \sum_i n_i = s$, то выборка назы-

* Мы использовали здесь приближенный метод интервального оценивания.

вается *механической*. В этом случае формула (9.48) остается в силе. Однако заменить, как это делалось ранее, неизвестные величины σ_i^2 на их групповые выборочные оценки нельзя, так как «группы» содержат по 1 элементу в каждой. Поэтому приходится при расчете доверительного интервала среднюю групповых дисперсий заменить на дисперсию, рассчитанную как для простой случайной выборки объемом $n = s$. Конечно, такой расчет может считаться обоснованным, если при выборе элемента из каждой группы и деление совокупности на группы производится по принципу, не зависящему от изучаемой вариации.

Серийная выборка. Организация такой выборки предполагает случайный отбор некоторого числа групп из генеральной совокупности, все элементы которых полностью включаются в выборку. Таким образом, здесь объектами случайного отбора оказываются лишь группы генеральной совокупности.

Изучение серийной выборки при различных по объему группах генеральной совокупности сложно, поэтому мы ограничимся рассмотрением случая, когда генеральная совокупность разбита на s равновеликих серий, т. е. $N_i = \frac{N}{s}$. В таком случае из (9.43) получаем

$$\overline{X} = \frac{1}{N}\sum_i \overline{X}_i N_i = \frac{1}{s}\sum_i \overline{X}_i. \tag{9.50}$$

Отобрав k групп, получим выборочную совокупность, состоящую из $n = k\frac{N}{s}$ элементов.

Поскольку случайный отбор производится сразу группами, то можно считать, что в результате j-го отбора (т. е. отбора j-й группы, где $j = 1, 2, \ldots, k$) получим случайную величину $\overline{X}_j$, которая с равной вероятностью $p = \frac{1}{s}$ может оказаться равной $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_i, \ldots, \overline{X}_s$. Следовательно,

$$M(\overline{X}_j') = \sum_i \overline{X}_i \frac{1}{s} = \frac{1}{s}\sum_i \overline{X}_i = \overline{X}.$$

Будем определять суммарный результат выборки $\overline{X}^*$ как среднюю арифметическую из средних в отобранных группах, т. е.

$$\overline{X}^* = \frac{1}{k}\sum_j \overline{X}_j'. \tag{9.51}$$

Определив

$$M(\overline{X}^*) = \frac{1}{k}\sum M(\overline{X}_j') = \frac{k\overline{X}}{k} = \overline{X},$$

заключаем, что (9.51) дает несмещенную оценку для генеральной средней. Далее можно показать, что

$$D(\overline{X}^*) = \frac{1}{k^2}\sum_j D(\overline{X}_j') = \frac{1}{sk}\sum_i (X_i - \overline{X})^2 = \frac{\sigma^2_{\text{мгр}}}{k}. \tag{9.52}$$

Таким образом, точность серийной выборки определяется межгрупповой дисперсией в генеральной совокупности. Она, естественно, также была бы выше точности простой случайной выборки (поскольку $\sigma^2_{\text{м гр}} < \sigma^2$), если бы число отобранных серий $k = n$. Но так как $k = \frac{ns}{N} < n$, то $D(\bar{X}^*) = \frac{\sigma^2_{\text{м гр}}}{ns}$ может оказаться большей, чем $\frac{\sigma^2}{n}$. Поэтому, как правило, серийная выборка дает бóльшую ошибку, чем простая случайная выборка при том же числе n наблюдаемых элементов и тем более бóльшую ошибку, чем типическая.

К преимуществам серийной выборки относится простота ее организации. Поэтому к ней часто прибегают, когда необходимо быстро произвести выборочную оценку ценой снижения ее точности.

При практическом построении доверительного интервала формула (9.46) непригодна, так как содержит неизвестные величины $\bar{X}$ и $\bar{X}_i$ для не попавших в выборку групп. Поэтому межгрупповую дисперсию генеральной совокупности заменяют на межгрупповую дисперсию отобранных групп, т. е. рассчитывают $D(\bar{X}^*)$ по формуле

$$D(\bar{X}^*) \approx \frac{1}{k}\left(\frac{1}{k}\sum_i (\bar{X}_i - \bar{X}^*)\right)^2. \qquad (9.52')$$

Дальнейшее построение доверительного интервала производится, как для типической выборки.

Глава 10

Статистическая проверка гипотез

10.1. Общая постановка задачи

Данные выборочных обследований часто являются основой для принятия одного из нескольких альтернативных решений. При этом в силу воздействия механизма случайного отбора любое суждение о генеральной совокупности, сделанное на основании выборки, будет сопровождаться случайной погрешностью и потому должно рассматриваться не как категорическое, а лишь как предположительное. Подобные предположения о свойствах и параметрах генерального распределения получили название *статистических гипотез.*

Сущность проверки статистической гипотезы заключается в том, чтобы установить, согласуются или нет данные наблюдения и выдвинутая гипотеза, можно ли расхождение между гипотезой и результатом выборочных наблюдений отнести за счет случайной погрешности, обусловленной механизмом случайного отбора. Эта

задача решается с помощью специальных методов математической статистики — методов статистической проверки гипотез.

Отметим общность логики методов статистической проверки гипотез и методов статистической оценки, изложенной в предыдущей главе. Как будет видно из дальнейшего, каждая задача, связанная с проверкой гипотезы, базируется на определении доверительного интервала соответствующего показателя.

Статистическая проверка гипотез имеет большое значение для практики. В частности, на ней основаны приемы статистического контроля качества произведенной партии продукции и контроля технологического процесса изготовления деталей (предупреждение появления брака). Допустим, что на предприятии о качестве продукции судят по результатам выборочного контроля. Если выборочная доля брака не превышает заранее установленной (нормативной) величины p_0, то партия принимается. Однако вывод о соответствии качества продукции установленным требованиям делается на основании выборочной проверки, поэтому на него влияют случайные обстоятельства отбора. Таким образом, суждение о качестве продукции не может рассматриваться как категорическое. По существу, речь идет о предположении (гипотезе), что доля брака во всей партии (генеральной совокупности) равна или меньше p_0. Эта гипотеза и должна быть проверена на основании результатов выборочного обследования.

Введем основные понятия статистической проверки гипотез. Подлежащая проверке гипотеза называется *основной*. Так как она часто состоит в предположении об отсутствии систематического расхождения (нулевом расхождении) между неизвестным параметром генеральной совокупности и заданной величиной, то ее называют также *нулевой гипотезой* и обозначают через H_0. Содержание гипотезы записывается после двоеточия. Так в рассмотренном примере $H_0 : p \leqslant p_0$.

Различают *простые* и *сложные* гипотезы. Простая гипотеза, относящаяся к параметру, характеризует его однозначно. Например, гипотеза утверждает равенство неизвестного параметра заданной величине — $H_0 : x = a$. В сложных гипотезах, например $H_0 : x > a$ и $H_0 : x < a$, указывается некоторая область вероятных значений. Очевидно, сложная гипотеза включает множество простых: $H_0' : x = a'$, $H_0'' : x = a''$ и т. д.

Каждой нулевой гипотезе противопоставляют альтернативную или конкурирующую гипотезу, которую обозначают символом H_a или H_1. Обычно по отношению к основной можно указать бесчисленное множество альтернативных гипотез. Так, если гипотеза $H_0 : x = a$ проверяется против альтернативы $x = b$, то $H_a : x = b$. Если же гипотеза $H_0 : x = a$ проверяется против сложной альтернативы $H_a : x \neq a$, то чаще всего ограничиваются одной из двух альтернатив: или $H_a : x < a$, или $H_a : x > a$.

Нулевая гипотеза может относиться не только к одному (как в приведенном примере), но и к нескольким параметрам или даже к закону распределения переменной.

При проверке гипотезы выборочные данные могут противоречить гипотезе H_0, тогда гипотеза *отклоняется*. В противном случае, т. е. если эти данные согласуются с ней, гипотеза H_0 *не отклоняется*. В практике в таких случаях часто говорят, что гипотеза *принимается* (такая формулировка не совсем точна, однако она широко распространена). Отсюда видно, что статистическая проверка гипотез, основанная на результатах выборки, неизбежно связана с риском (вероятностью) принять ложное решение. При этом возможны ошибки двух родов. *Ошибка первого рода* произойдет, когда будет принято решение отклонить нулевую гипотезу, хотя в действительности она оказывается верной. Принятие гипотезы также может быть ошибочным, если принята гипотеза H_0, в то время как в действительности оказывается верной альтернативная гипотеза H_a. В таком случае говорят об *ошибке второго рода*. Очевидно, что и правильные выводы также могут быть получены в двух случаях (табл. 10.1).

Таблица 10.1

Результат проверки гипотезы	Возможные состояния проверяемой гипотезы	
	верна гипотеза H_0	верна альтернативная гипотеза H_a
Гипотеза отклоняется	Ошибка первого рода	Правильное решение
Гипотеза не отклоняется	Правильное решение	Ошибка второго рода

В большинстве случаев последствия указанных ошибок неравнозначны. Одна из ошибок приводит к более осторожному, консервативному решению, другая — к неоправданным действиям. Что лучше или хуже — зависит от конкретной постановки задачи и содержания нулевой гипотезы. Так, если в рассмотренном примере допущена ошибка первого рода, то мы забракуем годную продукцию. Наоборот, допустив ошибку второго рода, мы отправим потребителю брак. Очевидно, что последствия второй альтернативы могут быть значительно более серьезными.

Поскольку исключить ошибки первого и второго рода невозможно, то естественно стремиться в каждой конкретной задаче минимизировать потери от этих ошибок. Конечно, желательно уменьшить обе ошибки одновременно, но, как мы увидим ниже, они являются конкурирующими и уменьшение вероятности допустить одну из них влечет за собой увеличение вероятности допустить другую. Поэтому в каждом конкретном случае необходимо выбирать компромиссное решение. В большинстве случаев единственный путь одновременного уменьшения риска ошибок заключается в увеличении объема выборки.

10.2. Критерий проверки. Критическая область

Суждение о совместимости выборочных данных с проверяемой гипотезой выносится на основе принятого в исследовании критерия, руководствуясь которым отклоняют или не отклоняют (при-

нимают) проверяемую гипотезу. Критерий определяет, какие данные выборки считаются не противоречащими проверяемой гипотезе и при каких она должна быть отклонена. Обычно критерий проверки гипотезы реализуют с помощью некоторой статистики θ (статистической характеристики, определяемой по выборке). Таким образом, θ есть случайная величина, закон распределения которой известен. Так, для проверки гипотезы $H_0: \bar{X} = a$ (генеральная средняя равна данному числу a) можно выбрать в качестве критерия величину выборочной средней $\theta' = \bar{X}^*$, разность $\theta'' = \bar{X}^* - a$ или нормированную разность $\theta''' = \frac{\bar{X}^* - a}{\sigma(\bar{X}^*)}$. Для всех приведенных статистик при большом объеме выборки n можно постулировать нормальный закон распределения с центром в точке a (для θ') или 0 (для θ'' и θ''').

Для проверки гипотезы о равенстве двух дисперсий можно в качестве критерия выбрать отношение дисперсий F, а для проверки

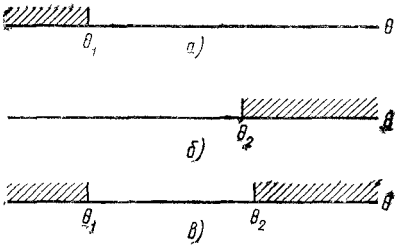


Рис. 10.1.
Односторонние и двусторонние критические области:
а — левосторонняя; б — правосторонняя; в — двусторонняя

гипотезы о равенстве средних двух совокупностей — разность выборочных средних и т. д.

Выделим в множестве возможных значений критерия θ подмножество таких его значений ω_0, при которых гипотеза H_0 отклоняется. Это подмножество ω_0 называется *критической областью*. Соответственно *областью принятия гипотезы* (допустимой областью) называют подмножество значений критерия θ, при которых гипотезу *не отклоняют*. В общем случае критерий θ может представлять собой многомерную случайную величину (когда, например, одновременно проверяется гипотеза о нескольких параметрах генеральной совокупности), однако мы ограничимся в дальнейшем рассмотрением простейшего одномерного критерия. В этом случае множество его возможных значений, а следовательно, критическая и допустимая области есть одномерные числовые множества (интервалы).

Точки, разделяющие критическую область и область принятия гипотезы называют *критическими точками*. На рис. 10.1 штриховкой показаны три вида критических областей.

Перейдем к определению критических точек, а следовательно, и критической области. В основу этого определения положен принцип практической невозможности маловероятных событий, который в данном случае используется следующим образом. Зададимся достаточно малой величиной α, называемой *уровнем зна-*

чимости критерия проверки, и определим критическую область ω_0 как множество таких значений θ, вероятность принадлежности которых к области ω_0 равнялась бы данной величине α, т. е.

$$P\{\theta \in \omega_0\} = \alpha. \tag{10.1}$$

Таким образом, событие, состоящее в том, что выборочное значение θ будет принадлежать к области ω_0, имея вероятность, равную α, может рассматриваться как маловероятное событие (с уровнем значимости α). Если по данным проведенной выборки все же получено $\theta \in \omega_0$, то это может служить основанием для отклонения гипотезы H_0. Если же θ не будет принадлежать к критической области, то можно лишь заключить, что данные опыта (выборки) не противоречат гипотезе H_0 и последняя не отклоняется.

Значения уровня значимости α задают обычно числами 0,1; 0,05; 0,01; 0,005; 0,001. В большинстве случаев берут значения 0,05 или 0,01. Как бы ни была мала величина α, попадание θ в критическую область есть только маловероятное, но не абсолютно невозможное событие. Поэтому не исключено, что при верной гипотезе H_0 значение θ может оказаться в критической области. Отклоняя в этом случае гипотезу H_0, мы, таким образом, допускаем ошибку первого рода, вероятность которой как раз и характеризуется величиной α. Чем меньше α, тем менее вероятно допустить ошибку первого рода. Однако с уменьшением α уменьшается критическая область, а следовательно, становится менее возможным попадание в нее выборочного значения θ даже когда гипотеза H_0 неверна. При $\alpha=0$ гипотеза всегда будет приниматься независимо от результатов выборки. Поэтому уменьшение α влечет за собой увеличение вероятности принять неверную гипотезу (обозначим эту вероятность β), т. е. совершить ошибку второго рода. В этом смысле ошибки первого и второго рода являются конкурирующими.

Пусть проверяется гипотеза о равенстве некоторого параметра ξ генерального распределения, например генеральной средней $\bar{X}$, данному числу a. Выберем для проверки статистику θ, распределение которой показано на рис. 10.2. При этом если верна гипотеза $H_0: \bar{X} = a$, то $M(\theta) = a$, т. е. θ есть несмещенная оценка для параметра $\bar{X}$.

Пусть гипотеза $H_0: \bar{X} = a$ проверяется против альтернативной гипотезы $H_0: \bar{X} < a$. В результате случайного разброса результатов выборки значения θ будут группироваться вокруг центра распределения (т. е. вокруг $M(\theta) = a$), однако если они окажутся значительно меньше, чем a, то предпочтительнее становится гипотеза H_a. Поэтому естественно критическую область определить неравенством $\theta < \theta_1$, т. е. в виде *левосторонней*. Задавшись значением уровня значимости α, определим критическую область (вернее, критическую точку θ_1) из уравнения (см. рис. 10.2, а)

$$P\{\theta < \theta_1\} = \alpha. \tag{10.2}$$

Соответственно относительно альтернативной гипотезы $H_a': \xi > a$ критическая область определяется из уравнения

$$P\{\theta > \theta_2\} = \alpha. \tag{10.3}$$

Наконец, если необходимо провести проверку против альтернативной гипотезы $H_a'': \xi \neq a$, то строится двусторонняя критическая область, критические точки которой определяются из уравнения

$$P\{\theta < \theta_1\} + P\{\theta > \theta_2\} = \alpha. \tag{10.4}$$

Однако найти две критические точки θ_1 и θ_2 из одного уравнения можно бесчисленным количеством способов. Поэтому оказывается возможным ввести дополнительное условие. Чаще всего

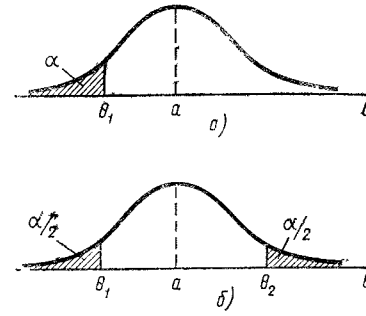


Рис. 10.2.
Левосторонняя и двусторонняя критические области

двустороннюю критическую область (см. рис. 10.2, б) строят как симметричную, распределяя величину α поровну между концами распределения, т. е. определяя θ_1 и θ_2 из уравнений

$$P\{\theta < \theta_1\} = \frac{\alpha}{2};\ P\{\theta > \theta_2\} = \frac{\alpha}{2}. \tag{10.5}$$

Использование уравнений (10.2) — (10.5) для определения критических точек требует знания распределения статистики θ, выбранной в качестве критерия проверки гипотезы.

Теперь обратимся к более подробному анализу связи между ошибками первого и второго рода. Пусть для определенности проверяется гипотеза $H_0: \bar{X} = a_0$ при альтернативной гипотезе $H_a: \bar{X} \neq a_0$ с помощью статистики $\theta = \bar{X}^*$. При достаточно большом n критерий $\bar{X}^*$ подчиняется асимптотически-нормальному закону распределения с параметрами $M(\bar{X}^*) = \bar{X}$ и $\sigma(\bar{X}^*) = \sigma/\sqrt{n}$. Таким образом, если гипотеза H_0 верна, то $M(\bar{X}^*) = a_0$. На рис. 10.3 это распределение показано в виде кривой 1. По выбранному значению уровня значимости α определяем из (10.5) критические точки θ_1 и θ_2 для двусторонней проверки гипотезы. Пусть гипотеза H_0 проверяется против альтернативной $H_a: \bar{X} = a_1$. Если верна эта простая гипотеза, то распределение $\bar{X}^*$ будет следовать нормальной кривой 2 с центром $M(\bar{X}^*) = a_1$, которая смещена вправо относительно кривой 1. Если эти кривые не пересекаются,

,о проверка гипотезы тривиальна: попадание выборочного значения $\bar{X}^*$ в область под кривой 1 свидетельствовало бы о верности гипотезы H_0, а в область под кривой 2 — о верности альтернативной гипотезы H_a. При относительно близких значениях a_0 и a_1 кривые пересекаются, как это и изображено на рис. 10.3.

Если значение $\bar{X}^*$ окажется в точке А, то гипотеза H_0 отклоняется с вероятностью ошибки первого рода, равной $\alpha/2$. Если же

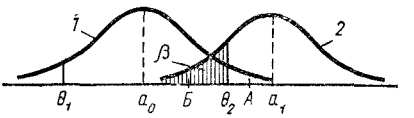


Рис. 10.3.
Вероятность ошибок первого и второго рода

оно окажется в точке Б, то гипотеза H_0 не отклоняется, хотя может быть верной гипотеза H_a. Таким образом, допускается ошибка второго рода. Вероятность этой ошибки β может быть определена по кривой 2 как вероятность того, что $\bar{X}^*$ будет лежать левее критической точки θ_2, т. е.

$$\beta = P\{\bar{X}^* \leqslant \theta_2\}. \qquad (10.6)$$

На рис. 10.3 величина этой вероятности численно равна заштрихованной площади, лежащей левее точки θ_2. С целью уменьшения риска отклонить верную гипотезу будем уменьшать величину

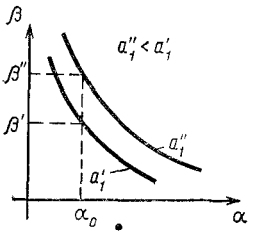


Рис. 10.4. Зависимость β от α

$\alpha/2$. При этом точка θ_2 будет сдвигаться вправо, что приведет к увеличению вероятности ошибки второго рода β. Чем ближе выбрана альтернатива к нулевой гипотезе, тем при той же величине α вероятность β будет больше, т. е. растет риск принять неверную гипотезу H_0. На рис. 10.4 показана зависимость β от α для двух значений a_1. При $a_1'' < a_1'$ для одного и того же α_0 находим, что $\beta'' > \beta'$. График зависимости β от a_1 (точнее, от расстояния между a_0 и a_1) можно построить и иначе — см. рис. 10.5. Подобного рода кривые получили название *кривых эффективности критерия.*

Разность $1-\beta$, представляющая собой вероятность обоснованного отклонения гипотезы H_0 (верной является альтернативная гипотеза H_a), называется *мощностью критерия.* Мощность, очевидно, желательно иметь наибольшей, так как это обеспечит минимум вероятности допустить ошибку второго рода.

При решении практических задач часто сталкиваются с проблемой выбора критерия из нескольких возможных. Очевидно, при

заданном α следует стремиться уменьшить β, или, что одно и то же, максимизировать мощность критерия. Чем выше значение $1-\beta$, тем меньше вероятность допустить ошибку второго рода.

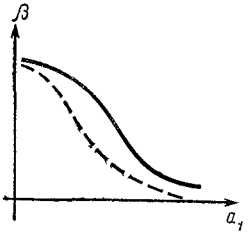


Рис. 10.5.
Кривые эффективности критерия
(пунктирная линия для выборки бо́льшего объема)

Если для данной статистики имеются кривые эффективности критерия, то они дают возможность определить такой объем выборки, который соответствует требуемой мощности критерия [12].

10.3. Общая схема проверки гипотезы

Суммируя все изложенное, можно указать следующий общий подход к статистической проверке гипотезы.

1. Формулируется проверяемая гипотеза H_0. Если она задается в виде простой гипотезы $H_0: \xi = a$, то всегда можно привести ее к нулевой гипотезе вида $H_0: (\xi - a) = 0$. Одновременно указывается, относительно каких альтернатив должна быть произведена проверка, т. е. формулируется альтернативная гипотеза H_a.

2. Выбирается статистика θ, по значениям которой можно судить о справедливости гипотезы. Статистика θ определяется по выборочным данным. Обычно для проверки одной и той же гипотезы можно предложить различные статистики при оценке параметра, если проверяется гипотеза о заранее постулированном его значении, их линейные функции и др. Для того чтобы принять статистику θ в качестве критерия проверки, необходимо, чтобы она удовлетворяла следующим требованиям: а) должен быть известен закон распределения θ в зависимости от гипотезы; б) статистика θ должна быть *несмещенной, состоятельной* и *эффективной.*

Условие несмещенности гарантирует симметричность области принятия гипотезы H_0; состоятельность обеспечивает более сжатое распределение θ вокруг центра, а следовательно, уменьшение вероятности ошибки второго рода β при том же α; эффективность дает уменьшение и α и β при прочих равных условиях.

3. Формулируется правило проверки, определяется соответствующий объем выборки по заданным α и β или из условия минимизации β при данных α и n.

4. В зависимости от проверяемой гипотезы и ее альтернатив выбирается одно- или двусторонняя проверка. Так, для сложной альтернативной гипотезы $H_a: \xi < a$ (или $H_a: \xi > a$) соответствующие критические области будут задаваться неравенством $\theta < \theta_1$ (или $\theta > \theta_2$). Для альтернативной гипотезы $H_a: \xi \neq a$ строится

двусторонняя критическая область $\theta < \theta_1$ и $\theta > \theta_2$. Выбор альтернативной гипотезы диктуется существом проверки. Так, если проверяется гипотеза, что процент брака составляет 5% (т. е. $H_0 : \xi = 0{,}05$), то альтернативной, очевидно, будет $H_a : \xi > 0{,}05$, ибо при меньшем проценте брака партия тем более будет принята. Если же в проверяемой гипотезе $H_0 : \xi = a$ ξ обозначает предел прочности материала на разрыв, то альтернативной, очевидно, будет $H_a : \xi < a$.

5. Критические точки определяются на основании известного распределения критерия по уравнениям (10.2)—(10.5).

6. Производится выборка намеченного объема n, по данным которой вычисляется выборочное значение критерия θ^*. Если это значение попадет в критическую область, то гипотеза H_0 признается не соответствующей данным наблюдения и потому отклоняется. Если значение θ^* попадет в допустимую область, то гипотеза H_0 признается не противоречащей выборочным данным и может быть признана правдоподобной.

Для каждого вида проверяемых гипотез разработаны соответствующие критерии. Для построения критериев наиболее часто используются z-статистика, t-статистика Стьюдента, F-статистика Фишера — Снедекора, χ^2-статистика Пирсона. Соответствующие законы распределения рассмотрены в гл. 7.

В описанной схеме проверка гипотез основывалась на предположении об известном законе распределения генеральной совокупности, из которого следовало определенное распределение критерия. Назовем данные методы проверки *параметрическими*. На практике часто приходится производить статистические проверки гипотез и при неизвестном генеральном распределении. Соответствующие методы получили название *непараметрических*. Естественно, что непараметрические критерии обладают значительно меньшей мощностью, чем параметрические. Это означает, что для сохранения той же величины β нужно иметь значительно больше опытных данных. В этом смысле непараметрические критерии «консервативны», т. е. с их помощью труднее отклонить неверную гипотезу H_0. Зато они могут применяться при любом законе распределения генеральной совокупности и применимы как к количественным, так и к качественным (ранговым) признакам. Единственным условием их применения остается взаимная независимость данных наблюдения, т. е. требование, чтобы все данные, по которым проводится статистическая проверка гипотез, представляли собой результат случайного независимого отбора. Непараметрические критерии обладают еще одним достоинством — относительной простотой вычислений. В настоящее время они все больше привлекают внимание исследователей. Параметрические методы проверки гипотез рассматриваются в § 10.4—10.7, непараметрическим методом посвящен § 10.8.

Укажем теперь на основные типы задач, решаемых с помощью проверки гипотез. Это прежде всего задачи сравнения: сравнения выборочных характеристик (средней, доли, дисперсии) с со-

ответствующими нормативами, сравнения характеристик двух выборок между собой (проверка гипотезы о принадлежности этих выборок одной совокупности). Более общими являются задачи, в которых проверяются гипотеза о принадлежности данного выборочного распределения к определенному виду или гипотеза о значимости расхождения двух эмпирических (выборочных) распределений. Каждый из указанных типов задач подразделяется на различные виды в зависимости от того, какими данными располагает исследователь при проведении проверки.

Выбор конкретного метода проверки зависит от объема выборки (большие и малые выборки), наличия генерального среднего квадратического отклонения изучаемой характеристики и т. д.

Идеи проверки гипотез используются в дисперсионном анализе (гл. 11) и теории корреляции (гл. 12, 13), где в конечном счете также проверяются гипотезы о значимости соответствующих параметров.

10.4. Проверка гипотез относительно доли признака

Рассмотрим два основных вида задач: сравнение доли признака с нормативом (стандартом) и сравнение долей признака в двух совокупностях.

Сравнение доли признака с нормативом. Пусть подлежит проверке гипотеза о том, что доля p некоторого признака в генеральной совокупности равна нормативной величине a, т. е. $H_0: p = a$. Рассмотрим вначале проверку против альтернативы $H_a: p \neq a$, т. е. по двустороннему критерию.

В качестве критерия примем статистику $\theta = \frac{m}{n}$ — частость признака в выборке. Эта статистика при любом n распределена по биномиальному (для возвратной выборки) или гипергеометрическому (для безвозвратной выборки) закону распределения. Однако при достаточно большом n можно воспользоваться при расчетах асимптотическими распределениями (Пуассона или нормальным). Исходя из нормального распределения и задавшись уровнем значимости α, найдем $z_{\alpha/2}$ из равенства

$$P\left\{\left|\frac{m}{n} - a\right| \leqslant z_{\alpha/2}\,\sigma\left(\frac{m}{n}\right)\right\} = 2\Phi(z) = 1 - \alpha, \qquad (10.7)$$

где $\Phi(z)$ — функция Лапласа.

Взяв выражение для $\sigma\left(\frac{m}{n}\right) = \sqrt{\frac{a(1-a)}{n}}$, получим доверительные границы для частости $\frac{m}{n}$, откуда находим критические точки:

$$\theta_1 = a - z_{\alpha/2}\sqrt{\frac{a(1-a)}{n}} \text{ и } \theta_2 = a + z_{\alpha/2}\sqrt{\frac{a(1-a)}{n}}. \qquad (10.8)$$

Отсюда следует правило проверки: если выборочная частость $\frac{m}{n}$ окажется в допустимой области (θ_1, θ_2), то гипотеза H_0 не от-

клоняется, если же $\frac{m}{n} < \theta_1$ или $\frac{m}{n} > \theta_2$, то гипотеза H_0 отклоняется.

Пусть теперь гипотеза H_0 проверяется против альтернативы $H_a: p > a$. В этом случае используется односторонняя проверка с помощью z_α, определяемой из уравнения

$$P\left\{\frac{m}{n} > \theta_2\right\} = 0{,}5 - \Phi(z) = \alpha, \tag{10.9}$$

где

$$\theta_2 = a + z_\alpha \sqrt{\frac{a(1-a)}{n}}. \tag{10.10}$$

Гипотеза H_0 отклоняется, если $\frac{m}{n} > \theta_2$, и принимается, если $\frac{m}{n} \leqslant \theta_2$.

Пусть производится проверка соответствия содержания активного вещества в продукции стандарту, который равен 10%, т. е. проверяется гипотеза H_0: $p = 0{,}1$, где p — доля активного вещества в продукции. Для контроля произведена выборка из 100 проб, которая дала $\frac{m}{n} = 0{,}152$. Считать ли гипотезу верной или продукцию следует забраковать, как не соответствующую нормативам?

Сначала рассмотрим случай, когда отклонения от норматива в обе стороны нежелательны, т. е. когда проверка производится по двустороннему критерию. Задавшись уровнем значимости $\alpha = 0{,}05$, находим по таблице функции Лапласа $z_{\alpha/2} = 1{,}96$, откуда из (10.8)

$$\theta_1 = 0{,}1 - 1{,}96\sqrt{\frac{0{,}1\,(1-0{,}1)}{100}} = 0{,}041;$$

$$\theta_2 = 0{,}1 + 1{,}96\sqrt{\frac{0{,}1 \cdot 0{,}9}{100}} = 0{,}159.$$

Так как $\frac{m}{n} = 0{,}152$ оказывается в допустимой области, то гипотеза H_0 не отклоняется.

Пусть теперь недопустимым является лишь превышение нормативного содержания активного вещества, т. е. проверка должна быть произведена против альтернативы H_a: $p > a$. Задавшись тем же уровне значимости $\alpha = 0{,}05$, имеем $z_\alpha = 1{,}65$, откуда по (10.10)

$$\theta_2 = 0{,}1 + 1{,}65\sqrt{\frac{0{,}1 \cdot 0{,}9}{100}} = 0{,}149.$$

Теперь получили $\frac{m}{n} = 0{,}152 > \theta_2$, т. е. лежит в критической области. Это дает основание с уровнем значимости $\alpha = 0{,}05$ отклонить гипотезу H_0 и признать партию продукции не соответствующей стандарту.

Задавшись конкретной альтернативной гипотезой $H_a: p = a_1$, можно определить мощность критерия $1 - \beta$, определив β. Так, для случая двусторонней проверки (см. (8.40)) получим

$$\beta = P_a\left\{\theta_1 \leqslant \frac{m}{n} \leqslant \theta_2\right\} = \Phi\left(\frac{\theta_2 - a_1}{\sigma_1}\right) - \Phi\left(\frac{\theta_1 - a_1}{\sigma_1}\right),$$

где θ_1 и θ_2 определяются из (10.8), а индекс «а» указывает, что вероятность должна вычисляться в предположении справедливо-

сти гипотезы H_a, т. е. при условиях

$$M\left(\frac{m}{n}\right)=a_1 \text{ и } \sigma_1=\sqrt{\frac{a_1(1-a_1)}{n}}.$$

Если $a_1 > a$, то можно показать, что минимизация β при заданном α приводит к *односторонней критической области* вида $\frac{m}{n} > \theta_2$. Поэтому критическую точку θ_2 следует определить из (10.9). В этом случае для β получим выражение

$$\beta = P_a\left\{\frac{m}{n} \leqslant \theta_2\right\} = 0{,}5 + \Phi\left(\frac{\theta_2 - a_1}{\sigma_1}\right). \qquad (10.11)$$

Пользуясь этим выражением, можно определить β при заданном n, или найти n при заданном β.

В предыдущем примере при односторонней проверке гипотеза H_0 о соответствии продукции стандарту была отклонена Найдем теперь значение $1-\beta$, т. е. вероятность правомерного отклонения этой гипотезы Выше для этого случая было получено $\theta_2 = 0{,}149 \approx 0{,}15$. Теперь можно рассчитать β по формуле (10.11):

$$\beta = P_a\left\{\frac{m}{n} \leqslant \theta_2\right\} = 0{,}5 + \Phi\left(\frac{0{,}15-0{,}20}{0{,}04}\right) = 0{,}5 + \Phi(-1{,}25).$$

Здесь $\Phi(-1{,}25) = -0{,}3944$, следовательно, $\beta = 0{,}11$ и $1-\beta = 0{,}89$.

Продолжим анализ и решим обратную задачу: каков должен быть объем выборки n, чтобы вероятность ошибки второго рода при односторонней проверке и прочих равных условиях составляла 0,05?

Используя (10.11) и подставляя $\beta = 0{,}05$, получим $\Phi\left(\frac{\theta_2 - a_1}{\sigma_1}\right) = -0{,}45$.

По таблице функции Лапласа находим $\frac{\theta_2 - a_1}{\sigma_1} = -1{,}65$. Подставляя значения $\theta_2 = 0{,}15$, $a_1 = 0{,}20$ и $\sigma_1 = \sqrt{\frac{0{,}2 \cdot 0{,}8}{n}}$, получим $\sigma = \frac{0{,}15-0{,}2}{-1{,}65} = 0{,}03$, $n = 174$.

Сравнение долей признака в двух совокупностях. Пусть $\frac{m_1}{n_1}$ и $\frac{m_2}{n_2}$ — частости одного и того же признака в в двух совокупностях из n_1 и n_2 единиц. Гипотезой H_0 является предположение, что обе совокупности представляют собой две выборки из одной генеральной совокупности с некоторой долей признака p, а отмеченное расхождение выборочных частостей есть результат случайностей, сопровождающих отбор. Поэтому проверку подобной гипотезы называют также *сравнением* долей признака или еще проверкой *значимости (существенности)* расхождения между ними.

В дальнейшем, приступая к построению критерия проверки, будем различать большие и малые выборки.

1. *Большие выборки.* Если n_1 и n_2 — большие числа (примерно более 30), то распределение выборочных частостей будет близко к нормальному с параметрами $M\left(\frac{m_1}{n_1}\right) = M\left(\frac{m_2}{n_2}\right) = p$ и дисперсиями $\sigma^2\left(\frac{m_1}{n_1}\right) = \frac{p(1-p)}{n_1}$, $\sigma^2\left(\frac{m_2}{n_2}\right) = \frac{p(1-p)}{n_2}$.

Введем для проверки гипотезы статистику $\theta = \frac{m_1}{n_1} - \frac{m_2}{n_2}$. При справедливости гипотезы H_0 значение θ может лишь случайно отличаться от нуля. Статистика θ также подчиняется нормальному распределению с параметрами:

$$M(\theta) = M\left(\frac{m_1}{n_1} - \frac{m_2}{n_2}\right) = M\left(\frac{m_1}{n_1}\right) - M\left(\frac{m_2}{n_2}\right) = p - p = 0;$$

$$\sigma^2(\theta) = \sigma^2\left(\frac{m_1}{n_1} - \frac{m_2}{n_2}\right) = \sigma^2\left(\frac{m_1}{n_1}\right) + \sigma^2\left(\frac{m_2}{n_2}\right) = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Теперь нетрудно перейти к проверке нулевой гипотезы с помощью двустороннего критерия, который в данном случае диктуется смыслом проверки. Задавшись уровнем значимости α, найдем $z_{\alpha/2}$ из уравнения

$$P\{|\theta| < z_{\alpha/2}\sigma(\theta)\} = 2\Phi(t) = 1 - \alpha, \qquad (10.12)$$

откуда получим критические точки:

$$\theta_1 = -z_{\alpha/2}\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$
$$\theta_2 = z_{\alpha/2}\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \qquad (10.13)$$

где величину p заменяем ее оценкой на основании данных двух выборок:

$$p = \frac{m_1 + m_2}{n_1 + n_2}.$$

Правило проверки: если выборочное значение θ^* лежит в интервале (θ_1, θ_2), то гипотеза H_0 не отклоняется, т. е. расхождение между $\frac{m_1}{n_1}$ и $\frac{m_2}{n_2}$ несущественно; если θ^* окажется вне этого интервала, то H_0 отклоняется, т. е. расхождение между $\frac{m_1}{n_1}$ и $\frac{m_2}{n_2}$ можно рассматривать как значимое.

Пусть число бракованных изделий в экспериментальной партии составило 4 из 100, в то время как в контрольной — 12 из 500. Оценить с уровнем значимости $\alpha = 0{,}01$ существенность расхождений долей брака в этих двух партиях.

По величине $\alpha = 0{,}01$ находим табличное значение $z_{\alpha/2} = 2{,}58$. Далее имеем

$$p = \frac{4+12}{100+500} = 0{,}027, \text{ откуда } \sigma = \sqrt{0{,}027\,(1-0{,}027)\left(\frac{1}{100} + \frac{1}{500}\right)} = 0{,}0177$$

Следовательно, $\theta_1 = -2{,}58 \cdot 0{,}0177 = -0{,}0458$ и $\theta_2 = 0{,}0458$.

Опытное расхождение составило $\theta^* = \frac{4}{100} - \frac{12}{500} = 0{,}016$, т. е. лежит в допустимой области. Таким образом, данные опыта не противоречат гипотезе H_0 и полученное расхождение между долями брака в двух партиях можно считать несущественным *.

* Расчет следовало бы провести, используя распределение Пуассона, поскольку p_1 и p_2 малы. Однако практически мы придем к тем же результатам, что и при предположении о нормальном распределении.

2. *Малые выборки.* Если n_1 и n_2 — малые числа, то использование нормального распределения для статистики $\theta = \frac{m_1}{n_1} - \frac{m_2}{n_2}$ становится неверным. В этом случае пользуются критерием χ^2, с помощью которого в общем случае оценивается расхождение между теоретическими и выборочными частотами (см. § 10.8). Распределение χ^2 рассмотрено подробно в § 7.8.Соответствующие таблицы приводятся в [12], [14], [15], [21], [23].

В данном случае критерий χ^2 вычисляется следующим образом. Пусть в табл. 10.2 указаны результаты группировки в каж-

Таблица 10.2

Совокупность	Фактические частоты			Теоретические частоты	
	A	$\bar{A}$	всего	A	$\bar{A}$
Выборка 1	m_1	$\bar{m}_1$	n_1	pn_1	$(1-p)\,n_1$
Выборка 2	m_2	$\bar{m}_2$	n_2	pn_2	$(1-p)\,n_2$
Всего	m_1+m_2	$\bar{m}_1+\bar{m}_2$	n_1+n_2	—	—

дой выборочной совокупности по признаку A; соответственно через $\bar{m}_1$ и $\bar{m}_2$ обозначено число элементов в каждой выборке, не обладающих признаком A. Если это выборки из одной и той же генеральной совокупности с долей признака p, то можно определить теоретические частоты pn_1, $(1-p)n_1$ и т. д., которые указаны в последних двух столбцах табл. 10.2. При этом для p принимается оценка $p = \frac{m_1+m_2}{n_1+n_2}$.

Теперь можно вычислить χ^2 по формуле

$$\chi^2 = \frac{(m_1 - pn_1)^2}{pn_1} + \frac{[\bar{m}_1 - (1-p)\,n_1]^2}{(1-p)\,n_1} + \frac{(m_2 - n_2 p)^2}{pn_2} + \frac{[\bar{m}_2 - (1-p)\,n_2]^2}{(1-p)\,n_2} \qquad (10.14)$$

При этом так как между четырьмя теоретическими частотами существуют три независимых соотношения, то независимой является только одна величина, т. е. в распределении χ^2 следует учесть одну степень свободы ($\nu = 1$).

Из (10.14) видно, что если нулевая гипотеза (согласно которой обе совокупности есть выборки из одной генеральной совокупности) верна, то расхождение между теоретическими и опытными частотами можно отнести только за счет случайностей отбора. Поэтому, определив для данного уровня значимости α значение χ^2_0, примем решение об отклонении гипотезы H_0, если $\chi^2 > \chi^2_0$, и о незначимости расхождений, если $\chi^2 \leqslant \chi^2_0$.

При испытании нового препарата животные были распределены на две группы. В экспериментальной группе из 50 животных к концу лечения осталось

9 больных, в контрольной (где лечение осуществлялось без данного препарата) из 30 осталось 7 больных животных. Необходимо проверить существенность воздействия препарата. Расчет теоретических частот производится по оценке p: $p = \frac{9+7}{50+30} = 0{,}2$. По данным задачи заполняем табл. 10.3.

Таблица 10.3

Обследуемые группы	A	$\overline{A}$	Всего	A	$\overline{A}$
Экспериментальная	9	41	50	10	40
Контрольная	7	23	30	6	24
Всего	16	64	80		

Теперь можно вычислить значение критерия:

$$\chi^2 = \frac{(9-10)^2}{10} + \frac{(41-40)^2}{40} + \frac{(7-6)^2}{6} + \frac{(23-24)^2}{24} = 0{,}333.$$

В то же время при $\alpha = 0{,}05$ и $\nu = 1$ получаем $\chi_0^2 = 3{,}8$. Таким образом, получили $\chi^2 = 0{,}333 < \chi_0^2 = 3{,}8$, т. е. отмеченное расхождение долей незначимо и не может быть отнесено за счет влияния препарата.

10.5. Проверка гипотез относительно средней

Рассмотрим, как и в предыдущем параграфе, два основных типа задач: сравнение средней со стандартом и сравнение средних двух совокупностей.

Сравнение средней с нормативом. Подлежит проверке гипотеза $H_0 : \overline{X} = a$, т. е. генеральная средняя $\overline{X}$ равна данному числу a. С подобного рода задачами встречаются при проверке качества продукции, характеризуемого некоторым средним показателем: средняя продолжительность службы радиодеталей, средняя зольность топлива, средняя эффективность препарата, средняя влажность вещества и т. д. Аналогичная задача возникает при проверке гипотезы о принадлежности данной выборки к генеральной совокупности, обладающей определенными свойствами.

При выборе статистики для построения критерия проверки гипотезы H_0 необходимо различать случаи больших и малых выборок. Если имеется большая выборка, то на основании центральной предельной теоремы выборочная средняя будет асимптотически подчиняться нормальному закону распределения. В качестве критерия проверки целесообразно выбрать не среднюю выборочную, а нормированное ее отклонение, т. е.

$$z = \frac{\overline{X}^* - a}{\sigma(\overline{X}^*)}. \qquad (10.15)$$

Так как $M(z) = 0$ и $\sigma(z) = 1$, то распределение z есть нормированное нормальное распределение, которое удобно использовать в расчетах.

При справедливости гипотезы H_0 $M(\bar{X}^*) = \bar{X} = a$ и потому отклонение z от нуля есть результат случайных погрешностей выборки. Задавшись уровнем значимости α, найдем из уравнения

$$P\{|z| < z_{\alpha/2}\} = 2\Phi(z_{\alpha/2}) = 1 - \alpha \qquad (10.16)$$

значение $z_{\alpha/2}$ и критические точки для двусторонней проверки: $z_1 = -z_{\alpha/2}$ и $z_2 = z_{\alpha/2}$. Если найденное на основании выборки значение z^* будет удовлетворять неравенству $|z^*| < z_{\alpha/2}$, то гипотеза H_0 не отклоняется, если же $|z^*| > z_{\alpha/2}$, то H_0 отклоняется.

Проверяется гипотеза о среднем размере детали $H_0 : \bar{X} = 5{,}8$ мм относительно альтернативной $H_a : \bar{X} \neq 5{,}8$ мм Из предыдущих исследований известно, что $\sigma = 0{,}4$ мм. По результатам выборки объемом $n = 100$ получено $\bar{X}^* = 4{,}8$ мм.

Задавшись уровнем значимости $\alpha = 0{,}05$, получим $z_{\alpha/2} = 1{,}96$. Итак, допустимая область значений параметра z будет $(-1{,}96;\ 1{,}96)$. Так как выборочное значение $z^* = \dfrac{4{,}8 - 5{,}8}{0{,}4} = -2{,}5 < -1{,}96$, т. е. попало в критическую область, то гипотеза H_0 отклоняется.

Если альтернативная гипотеза имеет вид $H_a : \bar{X} < a$ или $H_a : \bar{X} > a$, то производится односторонняя проверка при помощи значения z_α, определяемого из уравнения

$$\alpha = P\{|z| < z_\alpha\} = 0{,}5 + \Phi(z_\alpha).$$

В этом случае критическая область задается одним из неравенств $z < z_\alpha$ или $z > -z_\alpha$.

Если указана конкретная альтернативная гипотеза $H_a : \bar{X} = a_1$, то относительно нее можно определить мощность критерия $1 - \beta$. Пусть $a_1 > a$. Тогда величина β может быть определена из уравнения $\beta = P\{|z| < z_\alpha\}$.

Переходя от величины z к нормированному отклонению от нового центра a_1, получим

$$z' = \frac{\bar{X}^* - a_1}{\sigma(\bar{X}^*)} = \frac{\bar{X}^* - a}{\sigma(\bar{X}^*)} + \frac{a - a_1}{\sigma(\bar{X}^*)} = z + \lambda.$$

Следовательно,

$$\beta = P\{|z| < z_{\alpha/2}\} = \Phi\{\lambda - z_{\alpha/2} < z < \lambda + z_{\alpha/2}\} = \\ = \Phi(\lambda + z_{\alpha/2}) - \Phi(\lambda - z_{\alpha/2}). \qquad (10.17)$$

Определить вероятность β ошибочного принятия гипотезы $H_0 : \bar{X} = 5{,}8$ мм, если верна гипотеза $H_a : \bar{X} = a_1 = 4{,}5$ мм Как и в предыдущем примере, будем исходить из $\sigma(\bar{X}^*) = 0{,}4$ мм, $n = 100$ и $\alpha = 0{,}05$. Вычисляем $\lambda = \dfrac{a - a_1}{\sigma(\bar{X}^*)} = \dfrac{5{,}8 - 4{,}5}{0{,}4} = 3{,}25$. Ранее имели $z_{\alpha/2} = 1{,}96$. Следовательно,

$$\beta = \Phi(3{,}25 + 1{,}96) - \Phi(3{,}25 - 1{,}96) = \Phi(5{,}21) - \Phi(1{,}29) = 0{,}0985.$$

Для применения изложенных правил проверки, построенных на статистике (10.15), необходимо знать дисперсию генеральной совокупности (величину σ). В случае, когда она неизвестна (что

соответствует более реальной ситуации), можно при большом n (т. е. $n > 30$) заменить σ на ее выборочную несмещенную оценку

$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \overline{X}^*)^2}.$$

Однако в этом случае, а также при малых выборках (см. § 9.7) удобнее использовать статистику

$$t = \frac{\overline{X}^* - a}{s}\sqrt{n}, \tag{10.18}$$

которая имеет t-распределение Стьюдента с $\nu = n - 1$ степенями свободы.

Дальнейшее построение критической области для двух- и односторонней проверок проводится в точности, как изложено выше, с той лишь разницей, что критические точки (здесь вместо $-z_\alpha$ и z_α они будут обозначаться через $-t_\alpha$ и t_α) определяются по таблице распределения Стьюдента, а не Лапласа. Распределение Стьюдента симметрично, как и нормальное, но имеет меньший (отрицательный) эксцесс, поэтому в «хвостах» распределения заключена бо́льшая площадь. Следовательно, при том же уровне значимости α значение t_α будет больше, чем z_α, т. е. доверительный интервал шире, чем построенный на базе нормального распределения.

Для проверки гипотезы о среднем размере детали $H_0: \overline{X} = 5{,}8$ мм взята выборка объемом $n = 15$, в результате которой найдены среднее значение $\overline{X}^* = 4{,}8$ мм и средняя квадратическая ошибка $s = 0{,}4$ мм.

По выбранному уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $\nu = n - 1 = 15 - 1 = 14$ находим табличное значение $t_\alpha = 2{,}145$. Следовательно, если $|t^*| < 2{,}145$, то H_0 принимается; если $|t^*| > 2{,}145$, то H_0 отклоняется. В данном случае имеем $t^* = \dfrac{4{,}8 - 5{,}8}{0{,}4/\sqrt{14}} = -0{,}935$, которое лежит в допустимой области. Следовательно, H_0 *не отклоняется.*

Сравнение средних двух совокупностей. Пусть имеются две совокупности, характеризуемые средними $\overline{X}$, $\overline{Y}$ и дисперсиями σ_x^2, σ_y^2. Выдвигается гипотеза, что эти средние равны, т. е. $H_0: \overline{X} = \overline{Y}$. Для ее проверки из каждой совокупности производится выборка: из первой — объемом n_1, в результате которой получаются $\overline{X}^*$ и s_x^2, из второй — объемом n_2, в результате которой получаются $\overline{Y}^*$ и s_y^2. Здесь s_x^2 и s_y^2 — дисперсии в двух выборках.

По этим данным необходимо проверить основную гипотезу. Для ее проверки используется статистика

$$\theta = \frac{\overline{X}^* - \overline{Y}^*}{\sigma(\overline{X}^* - \overline{Y}^*)}. \tag{10.19}$$

Так как $M(\overline{X}^*) = \overline{X}$ и $M(\overline{Y}^*) = \overline{Y}$, то при справедливости гипотезы H_0 будем иметь $M(\theta) = 0$. Используя свойство дисперсии и полагая выборки (а следовательно, и выборочные средние) неза-

висимыми, получим

$$\sigma^2(\bar{X}^* - \bar{Y}^*) = \sigma^2(\bar{X}^*) + \sigma^2(\bar{Y}^*) = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}.$$

Теперь сделаем дополнительное предположение, что дисперсии обеих совокупностей равны т. е. $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Это предположение нуждается в специальной проверке, о чем речь будет идти в следующем параграфе. Если принять это предположение, то

$$\sigma^2(\bar{X}^* - \bar{Y}^*) = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Подставляем это выражение в (10.19), получаем

$$\theta = \frac{\bar{X}^* - \bar{Y}^*}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Если обе выборки достаточно большого объема, то $\bar{X}^*$ и $\bar{Y}^*$ распределены нормально, поэтому нормально будет распределена и статистика θ. Кроме того, заменив неизвестную дисперсию генеральной совокупности σ^2 ее несмещенной выборочной оценкой *

$$s^2 \approx \frac{\sum(x_i - \bar{X}^*)^2 + \sum(y_i - \bar{Y}^*)^2}{n_1 + n_2 - 2} = \frac{n_1 s_x^2 + n_2 s_y^2}{n_1 + n_2 - 2},$$

придем к нормально распределенному критерию

$$z = \frac{\bar{X}^* - \bar{Y}^*}{\sqrt{n_1 s_x^2 + n_2 s_y^2}} \cdot \frac{\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \qquad (10.20)$$

Дальнейшая проверка ведется обычным образом с использованием таблиц функций распределения Лапласа. Если выборки малого объема и применение нормального распределения может привести к ошибкам, то для того же критерия (10.20) используется распределение Стьюдента.

Для проверки эффективности новой технологии отбираются две группы рабочих: в первой группе численностью $n_1 = 40$ чел., где применяется новая технология, получены следующие данные: средняя выработка в штуках $\bar{X}^* = 84$, при этом $s_x = 10{,}1$, во второй группе численностью $n_2 = 54$ $\bar{Y}^* = 77{,}5$; $s_y = 8{,}4$ Определим с уровнем значимости $\alpha = 0{,}05$, действительно ли новая технология оказала влияние на среднюю производительность.

Вычисляем

$$z^* = \frac{84 - 77{,}5}{\sqrt{40 \cdot 10{,}1^2 + 54 \cdot 8{,}4^2}} \cdot \frac{\sqrt{40 + 54 - 2}}{\sqrt{\frac{1}{40} + \frac{1}{54}}} = 3{,}364.$$

Критическое табличное значение критерия при $\alpha = 0{,}05$ составляет 1,96. Так как $z^* > 1{,}96$, то заключаем, что расхождение о средних нельзя приписать

* При этом число степеней свободы будет равно $n_1 + n_2 - 2$, так как при расчете s_x^2 и s_y^2 используются два соотношения для средних $\bar{X}^*$ и $\bar{Y}^*$.

случайностям отбора и, следовательно, нулевая гипотеза H_0 должна быть отклонена. Иначе говоря, новая технология оказала положительное воздействие.

К такому же результату можно прийти, если, считая n_1 и n_2 небольшими числами, воспользоваться распределением Стьюдента.

10.6. Сравнение дисперсий двух совокупностей

Задача проверки гипотезы о равенстве двух дисперсий возникает достаточно часто; например, при анализе стабильности производственного процесса до и после введения технического новшества (колеблемость в выпуске продукции измеряется с помощью квадратического отклонения), при изучении качества измерительных приборов (сопоставление дисперсий показателей отдельных приборов), при изучении степени однородности двух совокупностей в отношении какого-либо признака (квалификации рабочих, стажа персонала и т. д.). Необходимость в проверке равенства дисперсий возникает, как мы убедились выше, и при сравнении средних величин совокупностей, поскольку при этом в большинстве случаев предполагается, что генеральные дисперсии одинаковы.

Итак, пусть имеются две нормально распределенные совокупности, дисперсии которых равны σ_1^2 и σ_2^2, в этом случае естественно проверить гипотезу H_0: $\sigma_1^2 = \sigma_2^2$. Поскольку дисперсии генеральных совокупностей являются неизвестными, то проверку выдвинутой гипотезы осуществляют на основе сопоставления выборочных дисперсий s_1^2 и s_2^2. Если отношение $s_1^2 : s_2^2$ близко к 1, то, очевидно, нет оснований для отклонения гипотезы. Если же это отношение значительно отличается от 1, то данный факт может служить указанием на необходимость ее отклонения. Возникает вопрос: при каком значении отношения выборочных дисперсий такой вывод будет достаточно обоснованным? Для решения этого вопроса обратимся к распределению отношения

$$F = \frac{s_1^2}{s_2^2}, \quad \text{где} \quad s_1^2 > s_2^2, \tag{10.21}$$

получаемому при многократном повторении выборок из двух нормально распределенных совокупностей с равными согласно гипотезе дисперсиями. Указанное распределение, получившее наименование F-распределения Фишера — Снедекора, подробно рассмотрено в § 7.8. Оно зависит от двух параметров — числа степеней свободы числителя и знаменателя: $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$, где n_1 и n_2 — объемы выборок. Числа ν_1 и ν_2 указываются в фигурных скобках рядом с вычисленным значением F:

$$F = \frac{s_1^2}{s_2^2}; \quad \left\{ \frac{\nu_1}{\nu_2} \right\}.$$

Обозначим критическое значение F, определенное по таблице распределения Фишера — Снедекора при заданном α, через F_α. Проверка гипотезы H_0 сводится к следующему правилу: если отно-

шение выборочных дисперсий F^* окажется больше критического (т. е. $F^* > F_\alpha$), то гипотеза H_0 отклоняется; если же $F^* \leqslant F_\alpha$, то H_0 не отклоняется. Очевидно, если альтернативная гипотеза формулируется как $H_a: \sigma_1^2 \neq \sigma_2^2$, то используется двусторонний критерий. Если же известно, что одна из дисперсий предположительно больше другой, то альтернативной является гипотеза $H_a: \sigma_1^2 > \sigma_2^2$ и используется односторонний критерий.

Один и тот же химический продукт получают на двух производственных линиях. Так как качество продукта во многом зависит от продолжительности процесса обработки сырья, то на второй линии в порядке эксперимента введены некоторые усовершенствования, сократившие вариацию времени обработки, в связи с чем продукт стал более однородным.

Выборочные измерения вариации времени обработки на первой и второй линиях дали $s_1^2 = 1{,}05$ мин^2 при $n = 25$ наблюдениям и $s_2^2 = 0{,}5$ мин^2 при том же числе наблюдений. Можно ли считать существенным расхождение между вариациями продолжительности процесса обработки сырья на первой и второй линиях?

Нулевую гипотезу сформулируем в виде H_0: $\sigma_1^2 = \sigma_2^2$. Альтернативной будет H_a: $\sigma_1^2 > \sigma_2^2$, т. е предположение, что колебания в продолжительности процесса обработки на второй линии после введения усовершенствований сократились. Из вида альтернативной гипотезы следует необходимость проведения только односторонней проверки с помощью критической точки F_α, определяемой из условия $P\{F > F_\alpha\} = \alpha$. Имеем

$$F^* = \frac{1{,}05}{0{,}5}; \left\{\frac{24}{24}\right\}, \text{ или } F^* = 2{,}1; \left\{\frac{24}{24}\right\}.$$

По значениям $\alpha = 0{,}05$, $\nu_1 = \nu_2 = 24$ находим по таблице $F_\alpha \approx 2$. Так как $F^* > F_\alpha$, то заключаем, что F^* попадает в критическую область и гипотеза H_0 должна быть отклонена. Это означает, что расхождения между дисперсиями следует считать существенными, свидетельствующими о положительном влиянии введенных усовершенствований на повышение однородности качества продукции.

В заключение заметим, что проверка гипотезы относительно равенства дисперсий производилась без каких-либо предположений относительно средних и соотношений между ними.

Проведенное выше сравнение дисперсий двух выборок может быть обобщено на случай любого числа выборок (более двух). Для проверки гипотезы о равенстве дисперсий предложены различные критерии: Неймана — Пирсона, Бартлетта, Хартли и др. (см. [12], [15]). В основе построения этих критериев лежит некоторая величина, определяемая по выборочным дисперсиям и распределение которой известно.

10.7. Сравнение двух зависимых выборок (парные сопоставления)

В практике статистической проверки гипотез часто сталкиваются со случаями, когда две сопоставляемые выборки не могут рассматриваться как независимые. Например, эффективность новой технологии может проверяться по результатам работы одной и той же бригады до и после внедрения новой технологии, спрос на два вида товаров изучается по одной и той же группе магазинов и т. д.

Регистрируя по каждому объекту наблюдения два показателя x и y (например, производительность каждого рабочего до и после внедрения новой технологии), получим два ряда наблюдений:

$$x\colon\ x_1,\ x_2,\ \ldots,\ x_i,\ \ldots,\ x_n;$$

$$y\colon\ y_1,\ y_2,\ \ldots,\ y_i,\ \ldots,\ y_n.$$

Таким образом, в данном случае речь идет о парных наблюдениях, т. е. об n связанных парах $(x_i,\ y_i)$. Если исследуемый фактор оказывает влияние только на признак x или y, то между спаренными наблюдениями будет отмечаться существенное различие. Задача заключается в том, чтобы установить, когда расхождение спаренных наблюдений можно отнести за счет случайных отклонений, а когда оно существенно и должно быть связано с влиянием изучаемого фактора.

Пусть разности между наблюдениями в каждой паре составляют величину $d_i = x_i - y_i$. Тогда обобщенной величиной расхождения спаренных наблюдений может служить средняя разность $\bar{d}$:

$$\bar{d} = \frac{\sum d_i}{n},$$

где n — число пар наблюдений. Чем меньше величина $\bar{d}$, тем правдоподобнее предположение о несущественности расхождения между рядами наблюдений. Таким образом, проверке подлежит гипотеза $H_0 : \bar{d} = 0$; альтернативной к ней является одна из следующих гипотез: для односторонней проверки — $H_a : \bar{d} < 0$ или $\bar{d} > 0$ и для двусторонней проверки — $H_a : \bar{d} \neq 0$. Критерием для проверки может служить статистика

$$t = \frac{\bar{d}}{s(\bar{d})}, \qquad (10.22)$$

где $s(\bar{d}) = \sqrt{\dfrac{\sum (d_i - \bar{d})^2}{n-1}}$. При нормальном распределении разностей $x_i - y_i$ (строго говоря, само это предположение нуждается в проверке) t-статистика имеет распределение Стьюдента с числом степеней свободы $\nu = n - 1$. Дальнейший механизм проверки не отличается от проверки расхождения средних (см. § 10.5).

Для проверки двух типов автомобильных покрышек (А и Б) выделено 12 автомобилей, каждый из которых прошел с покрышками А и Б по равному числу километров. В первых двух строках табл. 10.4 указано число покрышек на N километров пути. Оценить, существенно ли различие между двумя типами покрышек

Нулевая гипотеза состоит в предположении, что износоустойчивость обоих типов покрышек примерно одинакова и расхождение между результатами спаренных наблюдений является несущественным

Производим проверку гипотезы H_0 с помощью t-критерия. Для этого, используя результаты расчетов в 3-й и 4-й строках табл 10 4, вычисляем $\bar{d} = 0{,}183$, $s(\bar{d}) = 0{,}046$ По формуле (10 22) находим $t^* = 3{,}99$ Задавшись уровнем значимости $\alpha = 0{,}05$ при числе степеней свободы $\nu = 12 - 1 = 11$, определяем по таблице $t_\alpha = 2{,}2$. Так как $t^* > t_\alpha = 2{,}2$, то гипотеза H_0 о несущественном расхождении данных о расходе покрышек типа А и Б отклоняется.

Таблица 10.4

Число покрышек														Итого
	А (x_i)	9,7	8,8	9,5	7,9	8,1	7,4	8,8	9,9	7,8	7,4	8,7	9,8	
	Б (y_i)	9,5	8,9	9,1	7,1	8,7	7,8	8,1	9,2	7,6	7,4	8,9	9,3	
$d_i \cdot 10$		2	—1	4	8	—6	—4	7	7	2	0	—2	5	22
$d_i^2 \cdot 10^2$		4	1	16	64	36	16	49	49	4	0	4	25	268
Ранг		4	—1	6,5	12	—9	—6,5	10,5	10,5	4	0	—4	8	—

Следует заметить, что парные сравнения — более тонкий аппарат проверки, чем обычное сравнение двух средних. Дело в том, что разность средних сопоставляется с суммарной дисперсией выборочных совокупностей. Эта дисперсия связана с влиянием различных факторов, воздействующих на средние (в нашем примере — умением водителей, отрегулированностью машин и т. д.). При парном сравнении часть подобного рода факторов перестает оказывать влияние на величину дисперсии. Отсюда для одних и тех же данных $s(\bar{d}) < s(\bar{X} - \bar{Y})$. В связи с этим может оказаться, что там, где при проверке разности средних нулевая гипотеза не отклоняется, парные сравнения дадут основание для ее отклонения.

10.8. Непараметрические сравнения двух выборок

Выше были рассмотрены методы проверки гипотез, которые предполагали, что закон генерального распределения изучаемого признака известен. Перейдем к непараметрическим методам, которые не требуют этой предпосылки.

Сравнение двух независимых выборок. Даны результаты двух выборок:

$$1)\ x_1,\ x_2,\ x_3,\ \ldots,\ x_i,\ \ldots,\ x_{n_1};$$
$$2)\ x'_1,\ x'_2,\ x'_3,\ \ldots,\ x'_j,\ \ldots,\ x'_{n_2}.$$

Проверяется гипотеза H_0, состоящая в том, что эти совокупности имеют одинаковые распределения, т. е. выборки могут предполагаться взятыми из одной совокупности. Проверка основывается на том интуитивно-ясном положении, что если H_0 верна, то закономерности расположения значений признаков в первой и второй выборках должны быть однообразными. Для проверки гипотезы H_0 применим несколько критериев:

а) *Критерии положения.* Гипотеза H_0 отклоняется с вероятностью ошибки первого рода $\alpha = 0{,}05$, если при равных объемах выборок ($n_1 = n_2 = n$) будет наблюдаться одно из следующих положений: 1) по меньшей мере 5 наибольших или наименьших выборочных значений при $n \leqslant 25$ или 6 значений при $n > 25$ содержит

одна и та же выборка; 2) по меньшей мере 7 значений одной выборки с бóльшим размахом лежат вне размаха другой выборки.

б) *Критерий медианы.* Расположим данные двух выборок в один упорядоченный ряд (в порядке возрастания или убывания) и найдем медиану для объединенного ряда $\bar{X}_{\text{Ме}}$. Определим число наблюдений в каждой выборке со значениями признака меньшими и бóльшими, чем $\bar{X}_{\text{Ме}}$. Обозначим эти частоты как m_{11}, m_{12} для первой выборки и m_{21}, m_{22} — для второй. В результате получим четырехклеточную таблицу частот. Соответствующие теоретические частоты m'_{ij} могут быть определены из условия совпадения медиан первой и второй выборок:

$$m'_{11} = m'_{12} = 0{,}5n_1 \quad \text{и} \quad m'_{21} = m'_{22} = 0{,}5n_2.$$

Теперь проверка гипотезы может быть сведена к оценке расхождения теоретических и опытных частот с помощью критерия χ^2, как это было показано в § 10.4 (сравнение долей двух малых выборок).

в) *Ранговые критерии.* В основе этих критериев лежит замена наблюденных значений признака на их порядковые номера или ранги. Остановимся подробнее на одном из ранговых критериев — *критерии Вилкоксона.* Ранжируем результаты обеих выборок сразу, приписав каждому ранг R, соответствующий его положению в объединенном ряду. Теперь найдем среднее значение ранга для наблюдений одной (допустим, первой) выборки, $\bar{R}$. Понятно, что если выборки взяты из одной совокупности, то средние ранги для обеих совокупностей будут близки друг к другу. Отсюда критерием сравнения средних может служить нормированное отклонение среднего ранга от его математического ожидания:

$$u = \frac{\bar{R} - M(\bar{R})}{\sigma(\bar{R})}, \tag{10.23}$$

где

$$M(\bar{R}) = \frac{n_1 + n_2 + 1}{2}; \qquad \sigma(\bar{R}) = \sqrt{\frac{(n_1 + n_2 + 1)\, n_2}{12 n_1}} \tag{10.24}$$

(n_1 — объем первой выборки; n_2 — объем второй выборки).

Приведенная формула для средней квадратической ошибки имеет смысл при условии, что ранги имеют значения 1, 2, ..., $(n_1 + n_2)$.

Иногда некоторые значения признака в выборке оказываются одинаковыми; следовательно, им надо приписать одинаковые (связанные) ранги. В качестве связанного ранга обычно берется среднее для этой группы рангов значение (например, третье и четвертое места в объединенном ранжированном ряду занимают одинаковые выборочные значения; соответственно им приписываются связанные ранги, равные 3,5). Введение средних рангов, естественно, не оказывает влияния на $M(\bar{R})$, однако при расчете $\sigma(\bar{R})$ следует учесть некоторое влияние «связок». Расчет $\sigma(\bar{R})$ в этом слу-

чае ведется по формуле *

$$\sigma(\bar{R}) = \sqrt{\frac{(n^2-1)\,n - \sum T_i}{12 n_1 n} \cdot \frac{n_2}{n-1}}, \qquad (10.25)$$

где $n = n_1 + n_2$; $T_i = (t_i - 1)\ t_i(t_i + 1)$. Здесь t_i — число равных рангов в одной «связке».

Нормированное отклонение u при больших n_1 и $n_2 (\geqslant 20)$ приближенно подчиняется нормальному распределению с параметрами $M(u) = 0$ и $D(u) = 1$. На этом и основано применение (10.23) в качестве критерия для проверки гипотезы H_0. Задавшись уровнем значимости α, определяем по таблице функции Лапласа критическое значение z_α и затем устанавливаем, принадлежит ли допустимому интервалу $(-z_\alpha, z_\alpha)$ вычисленное по (10.23) значение u *.

Ранговый критерий парных сопоставлений. В § 10.7 парные сопоставления были осуществлены на основе параметрического критерия. Решим теперь эти же задачи на основе рангового критерия, который не требует нормального распределения разностей. Для его построения значения разности d_i располагают в неубывающем порядке по абсолютной величине, нулевые разности отбрасывают и приписывают единицу наименьшему значению разности: равным величинам присваивают средний (связанный) ранг. Каждому рангу приписывают знак соответствующей разности. Проверка гипотезы производится на основе *критерия Вилкоксона*

$$u = \frac{R^* - M(R)}{\sigma(R)}, \qquad (10.26)$$

где R^* — сумма положительных рангов;

$$M(R) = \frac{n(n+1)}{4};$$

$$\sigma^2(R) = \frac{n(n+1)(2n+1)}{24};$$

$$\sigma^2(R) = \frac{n(n+1)(2n+1)}{24} - \frac{\sum T_i}{48}.$$

Второе выражение для $\sigma^2(R)$ применяется в тех случаях, когда имеются связанные ранги. Здесь, как и выше, $T_i = (t_i - 1)\ t_i(t_i + 1)$; t_i — число равных рангов в каждом случае их связи (заметим, что учет связок при условии, что $n > 10$, и небольших значениях t_i дает незначительное уточнение в величине $\sigma(R)$). Критерий u при $n > 25$ подчиняется нормированному нормальному распределению. При $n \leqslant 25$ критические значения u_α можно получить по специальной таблице [7], [23].

Для иллюстрации метода проверки обратимся к примеру, приведенному в § 10.7. Произведем проверку гипотезы об одинаковой износоустойчивости покрышек с помощью рангового критерия. Так как число пар наблюдений ($n = 12$)

* Доказательство выражений для $M(\bar{R})$ и $\sigma(\bar{R})$ можно найти в работе [12].

недостаточно для использования асимптотического нормального распределения, то необходимы специальные таблицы В последней строке табл 10 4 показаны ранги, приписанные разностям. Заметим, что абсолютные значения разностей совпадают в трех случаях, поэтому в таблице проставлены соответствующие средние ранги (4; 6,5; 10,5). По этим данным находим $R^* = 55{,}5$; $M(R) = 33{,}0$; $\sigma(R) = 10{,}5$ откуда

$$u^* = \frac{55{,}5 - 33}{10{,}5} = 2{,}14.$$

Табличное значение u_α для $\alpha = 0{,}05$ и $n = 11$ равно 17 Поскольку $u^* < u_\alpha$ то гипотеза H_0 не отклоняется.

Полученный результат оказался более консервативным, чем результат параметрической проверки в § 10 7.

10.9. Критерии согласия

В предыдущих параграфах рассматривались методы проверки гипотез относительно отдельных параметров генерального распределения. Особую группу составляют так называемые *критерии согласия,* предназначенные для проверки гипотезы о согласованности выборочного распределения с теоретическим (генеральным) распределением. Критерии согласия позволяют ответить на вопрос о том, является ли расхождение между опытным и теоретическими распределениями столь незначительным, что оно может быть приписано влиянию случайности, или нет. Так как предположение о нормальном распределении лежит в основе многих критериев, особую роль играют критерии согласия, проверяющие справедливость гипотезы о наличии нормального распределения в совокупности.

Основанием для выдвижения гипотезы о генеральном распределении могут быть теоретические предпосылки о характере изменения признака. К ним, в частности, можно отнести выполнение условий, составляющих посылки теоремы Ляпунова, обосновывающие асимптотическое нормальное распределение, условий, приводящих к закону распределения Пуассона, и т. д. В некоторых случаях основанием для выдвижения гипотезы о теоретическом распределении могут служить некоторые формальные свойства наблюденного распределения (например, равенство нулю асимметрии и эксцесса может служить признаком нормального распределения, равенство средней и дисперсии — признаком распределения Пуассона и т. д.). Иногда предположения о характере гипотетического распределения могут основываться на графическом анализе наблюденного ряда (так, по форме полигона или гистограммы можно судить о законе распределения; изображение ряда накопленных частостей на вероятностной бумаге в виде прямой линии служит признаком близости к нормальному распределению и т д.) Так как все эти предположения носят характер гипотез, а не категорических утверждений, то они, естественно, должны быть подвергнуты статистической проверке с помощью критериев согласия.

Критерий χ^2 Пирсона. Одним из наиболее широко упот-

ребимых критериев согласия является критерий χ^2, определяемый по формуле

$$\chi^2 = \sum_{i=1}^{q} \frac{(m_i^* - m_i)^2}{m_i}, \qquad (10.27)$$

где q — число классов (групп), на которые разбито опытное распределение; m_i^* — наблюденная частота признака в i-й группе; m_i — теоретическая частота, рассчитанная по предполагаемому распределению.

Эта статистика подчиняется распределению χ^2 с числом степеней свободы $\nu = q - 1 - k$, где k — число параметров генерального распределения, оцениваемых на основании наблюденных данных. Так, например, если проверяется согласие опытного распределения с распределением Пуассона, единственный параметр λ которого оценивается по выборочным данным, то $\nu = q - 2$; если проверяется согласие с нормальным распределением, для которого по выборочным данным оцениваются два параметра $\bar{X}$ и σ, то $\nu = q - 3$ и т. д.

В некоторых случаях сравнение опытного распределения может производиться с заранее данным распределением или с распределением, у которого часть параметров указана (а не рассчитана по выборочным данным). В этом случае число k используемых связей уменьшается.

При полном совпадении теоретического и опытного распределений $\chi^2 = 0$, в противном случае $\chi^2 > 0$. Определив по заданному уровню значимости α табличное критическое значение χ^2_α, отклоняем гипотезу H_0 о несущественности расхождения между теоретическим и опытным рядом (т. е. о согласии опытных данных с гипотетическим распределением), если $\chi^2_{\text{наб}}$, вычисленное по (10.27), окажется больше или равно χ^2_α, и принимаем гипотезу H_0, если $\chi^2_{\text{наб}} < \chi^2_\alpha$. Применение критерия χ^2 требует соблюдения следующих условий: 1) экспериментальные данные должны быть независимыми, т. е. представлять собой результат случайного отбора; 2) объем выборки должен быть достаточно большим (практически не менее 50 единиц), при этом численность (частота) каждой группы должна быть не менее 5. Если последнее условие не выполняется, то производится предварительное объединение малочисленных интервалов.

Для опытного распределения, приведенного в первых двух столбцах табл 10 5, подобрано выравнивающее его нормальное распределение (о методе подбора нормального распределения см § 7 7). Соответствующие теоретические частоты показаны в столбце 3. Необходимо проверить согласие опытного и теоретического распределений.

В качестве критерия согласия воспользуемся χ^2. Данные для расчета этой характеристики приведены в последних трех столбцах табл 10 5 В результате получили $\chi^2_{\text{наб}} = 7{,}06$. Задавшись уровнем значимости $\alpha = 0{,}05$ и учитывая число степеней свободы $\nu = 11 - 1 - 2 = 8$, находим табличное критическое значение $\chi^2_\alpha = 15{,}5$.

Так как $\chi^2_{\text{наб}} = 7{,}06 < \chi^2_{\alpha} = 15{,}5$, то заключаем, что опытный ряд хорошо согласуется с гипотезой о нормальном распределении данного признака.

Таблица 10.5

x_i	m_i^*	$m_i = n\,\Delta\Phi$	$m_i^* - m_i$	$(m_i^* - m_i)^2$	$\frac{(m_i^* - m_i)^2}{m_i}$
1	2	3	4	5	6
9,0—9,1	2	1,4	—0,8	0,64	0,08
9,1—9,2	5	6,4			
9,2—9,3	27	23,2	3,8	14,4	0,62
9,3—9,4	52	62,0	—10	100	1,61
9,4—9,5	117	126	—9	81	0,64
9,5—9,6	203	189	4	16	0,08
9,6—9,7	228	214,2	13,8	190,4	0,89
9,7—9,8	180	181,3	1,3	1,7	0,01
9,8—9,9	105	115,7	10,7	174,5	0,99
9,9—10,0	60	54,7	5,3	28,1	0,51
10,0—10,1	14	19,5	5,5	30,3	1,55
10,1—10,2	4	5,2			
10,2—10,3	2	1,1	0,7	0,49	0,08
10,3—10,4	1	0			
Итого	1000	999,7	—	—	7,06

В заключение отметим, что критерий применяется не только для проверки согласованности опытного и теоретического распределений, но и для проверки независимости принципов классификации.

Критерий Колмогорова. Величина χ^2 зависит от группировки совокупности по интервалам, вносящей некоторый элемент случайности. Поэтому в ряде случаев целесообразнее воспользоваться критерием Колмогорова, основанным на сопоставлении функций распределения накопленных частот. Для этого вычисляется значение критерия

$$D = \max_i |F^*(x) - F(x)|,$$

характеризующего максимум абсолютной величины разности между опытной $F^*(x)$ и теоретической $F(x)$ функциями распределения накопленных относительных частот.

Для малых n ($n < 35$) при проверке гипотезы используется критическое значение D (см. таблицы в [17], [23]). При больших n используется предельное распределение критерия, предложенное Колмогоровым [17].

Результаты наблюдения над 11 случаями опоздания вылета самолета по различным причинам приведены в первых двух столбцах табл. 10.6. Допустим, что проверяется гипотеза о том, что время опоздания x_i подчиняется нормальному распределению. Проверим эту гипотезу с помощью критерия D, задавшись уровнем значимости $\alpha = 0{,}05$.

Таблица 10.6

Время опоздания x_i	Частота m_i^*	Накопленная частость F^*	По нормальному закону F	F^*-F
1	2	3	4	5
0,8	1	0,0909	0,0179	0,0730
1,0	1	0,1818	0,0228	0,1590
1,9	1	0,2727	0,1757	0,1370
2,1	1	0,3636	0,1841	0,1795
2,7	1	0,4545	0,3821	0,0724
2,8	1	0,5454	0,4207	0,1247
3,2	1	0,6363	0,5793	0,0570
3,6	1	0,7272	0,7257	0,0015
3,9	1	0,8181	0,8159	0,0022
4,2	1	0,9090	0,8849	0,0241
5,0	1	0,9999	0,9821	0,0178

В столбце 3 табл. 10.6 показаны накопленные частости. Для определения теоретической функции распределения $F(x)$ вычисляем среднее время опоздания $\overline{X} = 2{,}84$ ч и среднюю квадратическую ошибку $s = 1{,}32$ ч. Теперь с помощью таблиц функции Лапласа можно определить значения теоретической функции распределения (см. столбец 4).

$$F(x) = 0{,}5 + \Phi\left(\frac{x_i - \overline{X}}{s}\right).$$

Максимальная разность составляет $D^* = 0{,}1795$. При $\alpha = 0{,}05$ табличное значение $D_\alpha = 0{,}39$. Так как $D^* < D_\alpha$, то гипотеза о нормальном распределении не отклоняется, хотя зарегистрировано равномерное распределение.

Глава 11

Планирование эксперимента и дисперсионный анализ

11.1. Модели эксперимента

Методы математической статистики находят применение в различного рода экспериментальных исследованиях. Причем их задача не ограничивается пассивной обработкой опытных результатов, они призваны играть активную роль при планировании эксперимента и анализе полученных при его проведении результатов. Цель *планирования экспериментов* заключается в получении бо́льшего объема информации, чем это можно сделать при тех же затратах на эксперимент, применяя обычные методы. Собранные в ходе контролируемого эксперимента данные подвергаются специальному статистическому анализу, получившему название *дисперсионный анализ*. Кратко суть этого анализа сводится к расчленению общей дисперсии признака на компоненты, обусловленные влиянием конкретных факторов, и проверке гипотез о значимо-

сти их влияния. В основе дисперсионного анализа лежит предположение, согласно которому значение результата эксперимента можно представить в виде суммы ряда компонент. Например, если исследуется влияние одного фактора, то модель, описывающая структуру результата эксперимента, есть

$$x_{ij} = \bar{X} + \alpha_j + \varepsilon_{ij}, \tag{11.1}$$

где x_{ij} — значение признака, полученное на i-м уровне фактора (под уровнем фактора понимают некоторую его меру, например, если фактором является удобрение, то уровень фактора характеризуется количеством вносимого удобрения); $\bar{X}$ — общая средняя; α_j — эффект фактора на j-м уровне; ε_{ij} — случайная компонента, вызванная влиянием всех прочих факторов. Принимается предположение, что ε_{ij} имеет нормальное распределение со средней, равной нулю. В модели (11.1) выделено влияние одного фактора и случайной погрешности на результат эксперимента. Это наиболее простая модель. Более сложные модели эксперимента предусматривают влияния нескольких факторов и их взаимодействий. В частности, при анализе влияния двух факторов (A и B) и их взаимодействий структура результативного признака описывается моделью

$$x_{ijk} = \bar{X} + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{11.2}$$

где x_{ijk} — результат эксперимента в k-м наблюдении на i-м уровне фактора A и j-м уровне фактора B; $\bar{X}$ — общая средняя; α_i — эффект фактора A; β_j — эффект фактора B; γ_{ij} — эффект, вызванный совместным влиянием обоих факторов; ε_{ijk} — случайная компонента.

Как следует из предыдущего, при применении дисперсионного анализа исследуемая совокупность данных расчленяется на группы, отличающиеся по уровню факторов. Предполагается, что распределения признаков являются нормальными, а дисперсии в каждой отдельной группе данных одинаковы. Оба предположения нуждаются в проверке. Небольшое различие дисперсий и умеренное отклонение от нормальности распределения существенно не отразятся на конечных результатах анализа.

11.2. Однофакторный анализ при полностью случайном плане эксперимента

Рассмотрим наиболее простые планы экспериментов и соответствующие им схемы дисперсионного анализа *. Обратимся вначале к очень важной в теоретическом и практическом отношении и наиболее простой схеме опыта — полностью случайному (рандомизированному) плану эксперимента. Пусть подлежит изучению влияние на значение признака x только одного фактора. Разобьем ре-

* Более сложные планы экспериментов рассмотрены в работах [13], [19], [21]. В [19] уделено внимание и экономической стороне выбора плана эксперимента.

зультаты наблюдения на p групп (выборок), различающихся между собой по уровню фактора. Численность этих групп может быть различной, поэтому обозначим число наблюдений в j-й группе ($j = 1, 2, \ldots, p$) через n_j. Значения признака в j-й группе обозначим через x_{ij}, где $i = 1, 2, \ldots, np$ — порядковый номер наблюдения в j-й группе.

Результаты наблюдений удобно представить в упорядоченном виде, например так, как это сделано в табл. 11.1.

Таблица 11.1

Номер выборки	Наблюденные значения признака	Объем выборки	Сумма	Групповая средняя
1	$x_{11}, x_{21}, \ldots, x_{i1}, \ldots, x_{n_1 1}$	n_1	$\sum_i x_{i1} = T_1$	$\bar{X}_1 = \frac{1}{n_1}\sum_i x_{i1}$
...	. . . . .	...	. . . . .	. . . . .
j	$x_{1j}, x_{2j}, \ldots, x_{ij}, \ldots, x_{n_j j}$	n_j	$\sum_i x_{ij} = T_j$	$\bar{X}_j = \frac{1}{n_j}\sum_i x_{ij}$
...	. . . . .	...	. . . . .	. . . . .
p	$x_{1p}, x_{2p}, \ldots, x_{ip}, \ldots, x_{n_p p}$	n_p	$\sum_i x_{ip} = T_p$	$\bar{X}_p = \frac{1}{n_p}\sum x_{ip}$
	Итого	$N = \sum_j n_j$	$\sum_i \sum_j x_{ij} = G$	$\bar{X} = \frac{1}{N}\sum_i \sum_j x_{ij}$

В итоговой строке таблицы показаны общий объем совокупности N, сумма всех наблюденных значений признака $\sum_i \sum_j x_{ij}$ и общая средняя $\bar{X}$.

В соответствии с моделью эксперимента (11.1) для выполнения дисперсионного анализа необходимо определить две дисперсии: межгрупповую (дисперсию групповых средних), обусловленную влиянием изучаемого фактора, и внутригрупповую, величина которой рассматривается как случайная. Необходимые для расчета дисперсии суммы квадратов отклонений получим, воспользовавшись разложением общей суммы квадратов отклонений:

$$\sum_i \sum_j (x_{ij} - \bar{X})^2 = \sum_i \sum_j (x_{ij} - \bar{X}_j)^2 + \sum_j n_j (\bar{X}_j - \bar{X})^2.$$

Обозначив

$$\sum_i \sum_j (x_{ij} - \bar{X})^2 = Q_0, \quad \sum_i \sum_j (x_{ij} - \bar{X}_j)^2 = Q_1 \text{ и } \sum_j n_i (\bar{X}_j - \bar{X})^2 = Q_2,$$

перепишем разложение в виде

$$Q_0 = Q_1 + Q_2,$$

где Q_0 — общая сумма квадратов отклонений; Q_1 — сумма квадратов отклонений от групповых средних (сумма квадратов остаточных отклонений) и Q_2 — взвешенная сумма квадратов отклонений групповых средних от общей средней.

Для получения несмещенных оценок дисперсий необходимо каждую сумму квадратов разделить на число степеней свободы ν. Обозначим ν_0 — число степеней свободы, учитываемое при расчете общей дисперсии, ν_1 — при расчете внутригрупповой дисперсии и ν_2 — при расчете межгрупповой дисперсии. Напомним, что при расчете несмещенной оценки дисперсии число степеней свободы равно $N-1$, так как одна степень свободы теряется при определении средней. Таким образом, $\nu_0 = N-1$. Аналогично при оценке внутригрупповых дисперсий $\nu_1 = N-p$, так как используются p соотношений при вычислении p групповых средних $\bar{X}_j$. Наконец, при оценке межгрупповой дисперсии $\nu_2 = p-1$, так как групповые средние варьируют вокруг одной общей средней. Суммируя ν_1 и ν_2, получим

$$\nu_1 + \nu_2 = N - p + p - 1 = N - 1 = \nu_0. \tag{11.3}$$

Используя полученные значения сумм квадратов и чисел степеней свободы, можно вычислить несмещенные оценки трех дисперсий:

$$s_0^2 = \frac{Q_0}{N-1}; \quad s_1^2 = \frac{Q_1}{N-p}; \quad s_2^2 = \frac{Q_2}{p-1}. \tag{11.4}$$

Если p групп, на которые разбита вся совокупность результатов наблюдений, соответствуют p уровням воздействующего фактора, то s_1^2 характеризует рассеяние внутри групп (случайная вариация признака; s_1^2 называют также остаточной дисперсией), а s_2^2, характеризует рассеяние групповых средних (*систематическая* вариация).

Задачу проверки существенности различия между выборками можно конкретизировать и представить как задачу о различии дисперсий. В самом деле, если влияние фактора отсутствует, то s_1^2 и s_2^2 можно рассматривать как независимые оценки дисперсии совокупности σ^2. Наоборот, если фактор оказывает заметное влияние, то отношение $s_2^2 : s_1^2$ превзойдет критический предел и выборки следует считать взятыми из разных совокупностей (совокупностей с разным уровнем воздействия фактора).

Сравнение дисперсий двух выборок рассматривалось в § 10.6. Напомним порядок проверки. Выдвигается гипотеза $H_0 : \sigma_1^2 = \sigma_2^2$ о равенстве дисперсий в двух выборках. По выборочным данным вычисляются оценки дисперсий s_1^2 и s_2^2 и их отношение

$$F = \frac{s_2^2}{s_1^2}; \quad \left\{ \frac{p-1}{N-p} \right\} \tag{11.5}$$

Здесь и далее в фигурных скобках после формулы будем показывать числа степеней свободы, учтенных при расчете s_2^2 и s_1^2.

Задавшись уровнем значимости α, определяем по таблице критическое значение F_α, затем сравниваем вычисленное по формуле (11.5) значение F^* с критическим. Если $F^* \leqslant F_\alpha$, то нулевая гипотеза об отсутствии влияния фактора не отклоняется, а если $F^* > F_\alpha$, то H_0 отклоняется и, следовательно, статистически подтверждается влияние фактора.

Расчет оценок дисперсий удобно располагать в специальной таблице, получившей название таблицы дисперсионного анализа. В табл. 11.2 приводится схема однофакторного дисперсионного анализа (полностью случайный отбор).

Таблица 11.2

Характер вариации	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Систематическая (межгрупповая)	$Q_2 = \sum_j n_j (\overline{X}_j - \overline{X})^2$	$p-1$	$s_2^2 = \frac{Q_2}{p-1}$
Остаточная (внутригрупповая)	$Q_1 = \sum_i \sum_j (x_{ij} - \overline{X}_j)^2$	$N-p$	$s_1^2 = \frac{Q_1}{N-p}$
Итого	$Q_0 = Q_1 + Q_2 = \sum_i \sum_j (x_{ij} - \overline{X})^2$	$N-1$	—

Данные табл. 11.2 позволяют найти искомое отношение дисперсии

$$F^* = \frac{s_2^2}{s_1^2}; \quad \left\{ \frac{p-1}{N-p} \right\}.$$

В практике дисперсионного анализа наиболее трудоемкие расчеты связаны с вычислением сумм Q_1 и Q_2. Поэтому напомним формулы, облегчающие их вычисления:

$$Q_0 = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}; \quad Q_2 = \sum_j \frac{T_j^2}{n_j} - \frac{G^2}{N}, \tag{11.6}$$

откуда

$$Q_1 = Q_0 - Q_2. \tag{11.7}$$

Напомним также, что T_j — сумма значений признака в выборке j, а G — общая сумма значений признака.

Для проверки влияния методики обучения производственным навыкам на качество подготовки отбираются случайным образом из выпускников ПТУ четыре группы учеников, которые после окончания обучения (по разным методикам) показали следующие производственные результаты (табл. 11.3).

Дисперсии в группах примерно одинаковы. По приведенным в таблице данным имеем $G = 1782$, $N = 23$. Далее вычисляем:

$$\sum_i \sum_j x_{ij}^2 = 140481; \quad \sum_j \frac{T_j^2}{n_j} = 138984; \quad \frac{G^2}{N} = 138066{,}3.$$

Таблица 11.3

Группа (методика)	Выработка, шт./день	Число учеников	Суммарная выработка, шт.	Групповая средняя
1	60, 80, 75, 80, 85, 70	6	450	75
2	75, 66, 85, 80, 70, 80, 90	7	546	78
3	60, 80, 65, 60, 86, 75	6	426	71
4	95, 85, 100, 80	4	360	90
	Итого	23	1 782	77,48

Таким образом, $Q_0 = 140\,481 - 138066{,}3 = 2414{,}7$; $Q_2 = 138\,984 - 138066{,}3 = 917{,}7$. Откуда $Q_1 = 2414{,}7 - 917{,}7 = 1497$.

Теперь заполним таблицу дисперсионного анализа.

Таблица 11.4

Характер вариации	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Систематическая	$Q_2 = 917{,}7$	3	$\frac{917{,}7}{3} = 305{,}9$
Остаточная	$Q_1 = 1497{,}0$	19	$\frac{1497{,}0}{19} = 78{,}8$
Итого	$Q_0 = 2414{,}7$	22	

Таким образом, $F^* = 305{,}9 : 78{,}8 = 3{,}88$. При $\alpha = 0{,}05$ табличное значение $F_\alpha = 3{,}13$. Поскольку $3{,}88 > 3{,}13$, то с вероятностью 0,95 можно отклонить нулевую гипотезу H_0: $\sigma_2^2 = \sigma_1^2$, т. е. выбор методик обучения оказывает влияние на производственные навыки (в гл. 10 при сравнении дисперсий применялся двусторонний критерий; здесь же проверка односторонняя, поэтому берется табличное значение F_α, при заданном значении α).

11.3. Однофакторный анализ при группировке по случайным блокам

Пусть проверяется различие в урожайности нескольких сортов сельскохозяйственной культуры. Если все участки земли по плодородию примерно одинаковые, то лучше всего прибегнуть к полностью случайному плану размещения сортов по участкам. Однако участки чаще всего различаются между собой (иногда существенно) и это будет вызывать дополнительный разброс в экспериментальных данных. *Для устранения влияния неоднородности* выделенную для эксперимента площадь делят на участки, которые назовем блоками, с примерно одинаковым качеством земли в пределах каждого блока (между блоками могут существовать большие различия в отношении качества земли). Затем каждый блок делят

на столько делянок, сколько испытывается сортов культуры. Распределение сортов по делянкам производится в случайном порядке. Такой метод планирования эксперимента получил название *метод случайных блоков*. В отличие от полностью случайного плана число единиц наблюдения для каждого уровня фактора должно быть одинаковым, т. е. $n_1 = n_2 = \ldots = n_j = \ldots = n_p = n$. В нашем примере с урожайностью для каждого сорта выделяется одинаковое число участков. Модель экспериментального результата в этом случае можно записать в виде

$$x_{ij} = \bar{X} + \alpha_i + \beta_j + \varepsilon_{ij},$$

где α_i — эффект блоков; β_j — эффект уровня фактора; ε_{ij} — случайная компонента.

Как следует из сказанного, преимущество метода случайных блоков заключается в том, что с его помощью уменьшается разброс данных наблюдения. При применении метода случайных блоков данные эксперимента сводятся в таблицу (табл. 11.5):

Таблица 11.5

Уровень фактора	Результат наблюдения по блокам						Сумма по строкам	Средняя по уровням фактора
	1	2	...	i	...	n		
1	x_{11}	x_{21}	...	x_{i1}	...	x_{n1}	T_1	$\bar{X}_1$
2	x_{12}	x_{22}	...	x_{i2}	...	x_{n2}	T_2	$\bar{X}_2$
...	...	...	...	...	...	...	...	...
j	x_{1j}	x_{2j}	...	x_{ij}	...	x_{nj}	T_j	$\bar{X}_j$
...	...	...	...	...	...	...	...	...
p	x_{1p}	x_{2p}	...	x_{ip}	...	x_{np}	T_p	$\bar{X}_p$
Сумма по вертикали	B_1	B_2	...	B_i	...	B_n	G	$\bar{X}$
Средняя по блокам	$\bar{B}_1$	$\bar{B}_2$	...	$\bar{B}_i$	...	$\bar{B}_n$	—	—

Общую сумму квадратов отклонений $\sum_{ij} (x_{ij} - \bar{X})^2$ здесь нужно расчленить на три составляющие:

$$\sum_i \sum_j (x_{ij} - \bar{X})^2 = \sum_i \sum_j (x_{ij} - \bar{B}_i + \bar{B}_i - \bar{X}_j + \bar{X}_j - X + X - \bar{X})^2 =$$
$$= \sum_i \sum_j [(x_{ij} - \bar{B}_i - \bar{X}_j + \bar{X}) + (\bar{B}_i - \bar{X}) + (\bar{X}_j - \bar{X})]^2. \quad (11.8)$$

Раскрывая квадрат суммы трех слагаемых и учитывая, что суммы удвоенных перекрестных произведений равны нулю *, получим окончательно

$$\sum_i \sum_j (x_{ij} - \bar{X})^2 = \sum_i \sum_j (x_{ij} - \bar{B}_i - \bar{X}_j + \bar{X})^2 + \sum_i \sum_j (\bar{B}_i - \bar{X})^2 +$$
$$+ \sum_i \sum_j (\bar{X}_j - \bar{X})^2 = \sum_i \sum_j (x_{ij} - \bar{B}_i - \bar{X}_j + \bar{X})^2 +$$
$$+ p \sum_i (\bar{B}_i - \bar{X})^2 + n \sum_j (\bar{X}_j - \bar{X})^2. \qquad (11.9)$$

В последнем выражении первая сумма

$$Q_1 = \sum_i \sum_j (x_{ij} - \bar{B}_i - \bar{X}_j + \bar{X})^2 \qquad (11.10)$$

характеризует вариацию между блоками, а третий член

$$Q_3 = n \sum_j (\bar{X}_j - \bar{X})^2 \qquad (11.11)$$

— межгрупповую вариацию, т. е. вариацию, обусловленную изменением уровня фактора.

Наиболее трудоемким здесь является вычисление Q_1, поэтому, обозначив

$$Q_0 = \sum_i \sum_j (x_{ij} - \bar{X})^2, \qquad (11.12)$$

находим Q_1 из соотношения

$$Q_1 = Q_0 - (Q_2 + Q_3). \qquad (11.13)$$

При этом расчет Q_0, Q_2 и Q_3 удобно вести по формулам:

$$Q_0 = \sum\sum x_{ij}^2 - \frac{G^2}{N}; \quad Q_2 = \frac{\sum B_i^2}{p} - \frac{G^2}{N}; \quad Q_3 = \frac{\sum T_j^2}{n} - \frac{G^2}{N}. \qquad (11.14)$$

Как и в предыдущем случае, расчет оценок дисперсии удобно представлять в специальной таблице (табл. 11.6).

Числа степеней свободы очевидны для Q_2, Q_3 и Q_0. Число степеней свободы для Q_1 находим из соотношения $\nu_1 = \nu_0 - (\nu_2 + \nu_3)$.

Для проверки значимости влияния фактора, как и в предыдущем случае, используем критерий отношения дисперсий

$$F^* = \frac{s_2^2}{s_1^2}; \quad \left\{ \frac{p-1}{(n-1)(p-1)} \right\}. \qquad (11.15)$$

Дальнейшие выводы строятся как обычно: при $F^* \leqslant F_\alpha$ гипотеза о влиянии фактора не отклоняется, при $F^* > F_\alpha$ гипотеза отклоняется.

---

* Например, $\sum_i \sum_j 2(\bar{B}_i - \bar{X})(\bar{X}_j - \bar{X}) = 2 \sum_j (\bar{X}_j - \bar{X}) \sum_i (\bar{B}_i - \bar{X}) = 0$,

что следует из очевидного равенства $\sum_i (\bar{B}_i - \bar{X}) = \sum_i \bar{B}_i - n\bar{X} = 0$.

При рассмотрении табл. 11.6 становится очевидным, что введение блоков сокращает долю остаточной вариации, т. е. часть остаточной вариации теперь объясняется различием в блоках.

Таблица 11.6

Характер вариации	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Систематическая (межгрупповая)	$Q_3 = n \sum_j \left(\overline{X}_j - \overline{X}\right)^2$	$p - 1$	$s_3^2 = \frac{Q_3}{p-1}$
Между блоками	$Q_2 = p \sum_i \left(\overline{B}_i - \overline{X}\right)^2$	$n - 1$	$s_2^2 = \frac{Q_2}{n-1}$
Остаточная	$Q_1 = Q_0 - \left(Q_2 + Q_3\right)$	$(n-1)(p-1)$	$s_1^2 = \frac{Q_1}{(n-1)(p-1)}$
Итого	$Q_0 = \sum_i \sum_j \left(x_{ij} - \overline{X}\right)^2$	$N - 1$	

Рассмотрим классическую ситуацию, для которой применяется метод случайных блоков. Проверяется урожайность четырех сортов пшеницы. Для экспериментов выделено пять участков земли (пять блоков), на каждом из которых

Таблица 11.7

Сорт	Урожайность (ц/га) по блокам					Сумма по блокам T_j	Средняя по факторам $\overline{X}_j$
	1	2	3	4	5		
1	28,7	26,7	21,6	25,0	28,2	130,2	26,04
2	24,5	28,5	27,7	28,7	32,5	141,8	28,38
3	23,2	24,7	20,0	24,0	24,0	115,9	23,18
4	29,0	28,7	22,5	28,0	27,0	135,2	27,04
Суммы по сортам (B_i)	105,4	108,6	91,8	105,7	111,7	523,2	26,16
Средняя по блокам $(\overline{B}_i)$	26,35	27,15	22,95	26,42	27,92	—	—

высеяно все четыре сорта пшеницы (делянки по 1 га). Результаты эксперимента приведены в табл. 11.7

По данным табл. 11.7 получим $G = 523{,}2$; $G^2 : N = 13692{,}14$; $\sum_i \sum_j x_{ij}^2 = 13\,867{,}38$; $\sum_i B_i^2 = 54979{,}74$; $\sum_j T_j^2 = 67771{,}1$. Теперь находим значения сумм

квадратов отклонений:

$$Q_0 = 13867{,}38 - 13692{,}14 = 175{,}24;$$

$$Q_3 = \frac{67771{,}1}{5} - 13692{,}14 = 62{,}08;$$

$$Q_2 = \frac{54979{,}74}{4} - 13692{,}14 = 52{,}80;$$

$$Q_1 = 175{,}24 - (62{,}08 + 52{,}8) = 60{,}36.$$

Соответствующие данные представлены в табл. 11.8.

Таблица 11.8

Источник вариации	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Между сортами	$Q_3 = 62{,}08$	3	$s_3^2 = 20{,}69$
Между блоками	$Q_2 = 52{,}80$	4	$s_2^2 = 13{,}20$
Остаточная	$Q_1 = 60{,}36$	12	$s_1^2 = 5{,}03$
Итого	$Q_0 = 175{,}24$	19	

Следовательно, $F^* = 20{,}69 : 5{,}03 = 4{,}11$.

Табличное значение F_α при $\alpha = 5\%$ и числе степеней свободы 3 и 12 составит 3,49. Следовательно, $F^* > F_\alpha$ и нулевая гипотеза (об отсутствии различий в урожайности испытываемых сортов) отклоняется.

Определение суммы квадратов отклонений, связанной с различием между блоками, позволило в данном примере выделить часть систематических изменений, что в свою очередь привело к заметному уменьшению остаточной дисперсии. Следует заметить, что если бы данный эксперимент производился по схеме случайного плана, то изменения урожайности, определяемые различиями в плодородии отдельных блоков, относились бы к случайным. В этом случае $Q_1 = 175{,}24 - 62{,}08 = 112{,}6$ Остаточная дисперсия (уже при 16 степенях свободы) составила бы $112{,}6 : 16 = 7{,}04$ и $F^* = 20{,}69 : 7{,}04 = 2{,}94$, следовательно, нулевая гипотеза не отклонялась бы.

11.4. Двухфакторный анализ при полностью случайном плане эксперимента

Полностью случайный план и метод случайных блоков при одном факторе являются наиболее простыми представителями планов экспериментов. Существуют и другие, более сложные планы, позволяющие не только проверить влияние ряда факторов порознь, но и их взаимодействия. Рассмотрим один из них. Пусть необходимо выявить влияние двух факторов (A и B) и их взаимодействий (обозначим $A \times B$) на некоторый результативный признак. Модель эксперимента имеет вид (11.2). Опыт проводится при фиксированных уровнях факторов A и B. Поскольку для каждого сочетания факторов опыт повторяется n раз, то соответственно получим n значений признака. Данные эксперимента удобно представить в виде таблицы (см. табл. 11.9). В таблице значение каж-

дого результата (отклика) обозначено через x_{ijk}, где $i = 1, \ldots, p$ (p — число уровней фактора A); $j = 1, \ldots, q$ (q — число уровней фактора B); $k = 1, \ldots, n$ (k — порядковый номер эксперимента для каждого сочетания уровней).

Таблица 11.9

Уровень фактора B	Уровень фактора A				Сумма
	A_1	A_2	...	A_p	
B_1	$x_{111}, x_{112}, \ldots, x_{11n}$	$x_{211}, x_{212}, \ldots, x_{21n}$	...	$x_{p11}, x_{p12}, \ldots, x_{p1n}$	$\sum_{ik} x_{i1k}$
B_2	$x_{121}, x_{122}, \ldots, x_{12n}$	$x_{221}, x_{222}, \ldots, x_{22n}$	...	$x_{p21}, x_{p22}, \ldots, x_{p2n}$	$\sum_{ik} x_{i2k}$
...	...	...	...	...	...
B_q	$x_{1q1}, x_{1q2}, \ldots, x_{1qn}$	$x_{2q1}, x_{2q2}, \ldots, x_{2qn}$	...	$x_{pq1}, x_{pq2}, \ldots, x_{pqn}$	$\sum_{ik} x_{iqk}$
Сумма	$\sum_{jk} x_{1jk}$	$\sum_{jk} x_{2jk}$	...	$\sum_{jk} x_{pjk}$	$\sum_{ijk} x_{ijk}$

По данным таблицы получим следующие средние:
общая средняя

$$\bar{X} = \frac{\sum_{ijk} x_{ijk}}{pqn}; \tag{11.16}$$

средние по строкам

$$\bar{T}_j = \frac{\sum_{ik} x_{ijk}}{pn}; \tag{11.17}$$

средние по столбцам

$$\bar{T}_i = \frac{\sum_{jk} x_{ijk}}{qn}; \tag{11.18}$$

средние для каждого сочетания уровней факторов (для каждой отдельной клетки таблицы)

$$\bar{X}_{ij} = \frac{\sum_{k} x_{ijk}}{n}. \tag{11.19}$$

Суммы квадратов отклонений от общей средней $\sum_{ijk} (x_{ijk} - \bar{X})^2$ разложим на составляющие. Поскольку все перекрестные произве-

дения равны нулю, то в итоге получим

$$\sum_{ijk}(x_{ijk}-\bar{X})^2=\sum_{ijk}(\bar{T}_i-\bar{X})^2+\sum_{ijk}(\bar{T}_j-\bar{X})^2+$$
$$+\sum_{ijk}(\bar{X}_{ij}-\bar{T}_i-\bar{T}_j+\bar{X})^2+\sum_{ijk}(x_{ijk}-\bar{X}_{ij})^2. \qquad (11.20)$$

Здесь четыре составляющие: сумма квадратов, связанная с влиянием фактора A, фактора B и их взаимодействия, и составляющая, которая характеризует остаточную сумму квадратов (сумму квадратов внутри каждой ячейки таблицы).

Общее число степеней свободы, очевидно, равно $N-1$, где N — общее число наблюдений ($N=pqn$). Число степеней свободы между столбцами равно $p-1$, между строками $q-1$, для взаимодействия $(p-1)(q-1)$, внутри ячеек $(n-1)pq=N-pq$. Для проверки найдем сумму $N-1+p-1+q-1+(p-1)(q-1)+$ $+(n-1)pq=N-1$.

Схема дисперсионного анализа, основанная на полученных выше данных, представлена в табл. 11.10.

Таблица 11.10

Источник вариации	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Фактор A	$Q_4=nq\sum_{i}^{p}(\bar{T}_i-\bar{X})^2$	$p-1$	$s_4^2=\frac{Q_4}{p-1}$
Фактор B	$Q_3=np\sum_{j}^{q}(\bar{T}_j-\bar{X})^2$	$q-1$	$s_3^2=\frac{Q_3}{q-1}$
Взаимодействие $A\times B$	$Q_2=n\sum_{ij}^{pq}(\bar{X}_{ij}-\bar{T}_i-\bar{T}_j+\bar{X})^2$	$(p-1)(q-1)$	$s_2^2=\frac{Q_2}{(p-1)(q-1)}$
Остаточная вариация	$Q_1=\sum_{ijk}^{pqn}(x_{ijk}-\bar{X}_{ij})^2$	$N-pq$	$s_1^2=\frac{Q_1}{N-pq}$
Итого	$Q_0=\sum_{ijk}(x_{ijk}-\bar{X})^2$	$N-1$	—

Поскольку при проведении анализа интерес представляет влияние каждого фактора порознь и влияние их взаимодействия, то находим соответственно три значения F^*:

$$F_A^*=\frac{s_4^2}{s_1^2};\ \left\{\frac{p-1}{N-pq}\right\};\ F_B^*=\frac{s_3^2}{s_1^2};\ \left\{\frac{q-1}{N-pq}\right\};$$

$$F_{AB}^*=\frac{s_2^2}{s_1^2};\ \left\{\frac{(p-1)(q-1)}{N-pq}\right\}.$$

Необходимые для анализа суммы квадратов отклонений можем получить по следующим формулам:

$$Q_4 = \sum_i^p \frac{T_i^2}{nq} - \frac{G^2}{N}; \qquad (11.21)$$

$$Q_3 = \sum_j^q \frac{T_j^2}{np} - \frac{G^2}{N}; \qquad (11.22)$$

$$Q_2 = \sum_{ij}^{pq} \frac{T_{ij}^2}{np} - \sum_i^p \frac{T_i^2}{nq} - \sum_j^q \frac{T_j^2}{np} + \frac{G^2}{N}; \qquad (11.23)$$

$$Q_1 = \sum_{ij}^{pq} \left(\sum_k^n x_{ijk}^2 - \frac{T_{ij}^2}{n} \right), \qquad (11.24)$$

где

$$G = \sum_{ijk}^{pqn} x_{ijk}; \qquad T_{ij} = \sum_k^n x_{ijk}, \qquad (11.25), \quad (11.26)$$

T_i, T_j — суммы по i и j.

Пусть экспериментально проверяется влияние на износоустойчивость детали таких факторов, как материал (два вида) и технология изготовления (три метода). Данные эксперимента (число месяцев работы детали) приведены в табл. 11.11.

Таблица 11.11

Материал (фактор B)	Технология (фактор A)			Сумма по строке	Средняя
	1	2	3		
1	10; 8; 7; 10	8; 12; 14; **12**	15; 8; 10; **10**	124	10,33
2	12; 8, 8; 7	12; 13; 11; **14**	13; 15; 12; **10**	135	11,25
Сумма по столбцу	70	96	93	259	—
Средняя	8,75	12,00	11,63	—	—

На основе формул (11.21) — (11.26) получены следующие суммы квадратов отклонений: $Q_0 = 143{,}8$; $Q_1 = 85{,}4$; $Q_2 = 2{,}8$; $Q_3 = 5{,}0$ и $Q_4 = 50{,}6$. Данные дисперсионного анализа приведены в табл. 11.12.

Определим значения критерия:

$$F_A^* = \frac{25{,}3}{4{,}8} = 5{,}3, \left\{\frac{2}{18}\right\}; \quad F_B^* = \frac{5{,}0}{4{,}8} = 1{,}1, \left\{\frac{1}{18}\right\};$$

$$F_{AB}^* = \frac{1{,}4}{4{,}8} = 0{,}28, \left\{\frac{2}{18}\right\},$$

и сравним их с соответствующими критическими значениями F_α при $\alpha = 0{,}05$. Так как $F_A^* = 5{,}3 > F_\alpha^* = 3{,}55$, то нулевая гипотеза об отсутствии влияния фактора A отклоняется и следует сделать вывод о значимом влиянии этого фактора.

Таблица 11.12

Источник вариации	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Фактор A	50,6	2	$s_4^2 = 25{,}3$
Фактор B	5,0	1	$s_3^2 = 5{,}0$
Взаимодействие $A \times B$	2,8	2	$s_2^2 = 1{,}4$
Остаточная вариация	85,4	18	$s_1^2 = 4{,}8$
Итого	143,8	23	—

Что касается влияния фактора B и взаимодействия факторов, то, так как $F_B^* < F_\alpha$ и $F_{AB}^* < F_\alpha$, нет статистического основания для отклонения соответствующих нулевых гипотез.

Глава 12

Основы теории корреляции и регрессии. Парная корреляция и регрессия

12.1. Корреляционный и регрессионный анализ

Учение о корреляции и регрессии * широко используется при анализе связей различных явлений. Назовем лишь несколько областей приложения соответствующих методов анализа. Прежде всего это экономика — исследования зависимости объемов производства от целого ряда факторов (размера основных фондов, их возраста, обеспеченности предприятий производственным персоналом и т. д.), зависимости производительности труда на предприятиях от уровня механизации и электрификации производства, стажа и квалификации рабочих и т. д., зависимости спроса или потребления населения от уровня дохода, цен на товары и т. д. Интенсивно применяется регрессионный анализ при изучении зависимости в агротехнике и экономике сельского хозяйства. Классическим примером здесь может служить исследование зависимости урожайности от факторов, характеризующих применяемую технологию (сроки посева, количество удобрений и т. д.), и факторов погодно-климатического характера. Широко применяются методы корреляционного и регрессионного анализа в психологии, социологии, педагогике и в ряде других областей науки и практики.

* Первый термин произошел от латинского correlatio — соотношение, взаимосвязь. Второй (*латинское regressio — движение назад*) введен *английским статистиком* Ф. Гальтоном, который, изучая зависимость между ростом родителей и их детей, обнаружил явление «регрессии к среднему» — у детей, родившихся у очень высоких родителей, рост имел тенденцию быть ближе к средней его величине.

Интересной областью приложения регрессионного анализа является обобщение экспертных оценок. Если один или несколько экспертов оценивают некоторый объект (обозначим соответствующую оценку через y_i) в зависимости от его свойств (каждое свойство измеряется неслучайной характеристикой x_{ij}), то зависимость экспертной оценки от свойств объекта можно охарактеризовать соответствующей регрессией $\hat{y} = f(x_j)$, с помощью которой обобщают вклад каждого свойства в общую оценку объекта.

Следует отметить, что статистический анализ зависимостей сам по себе не вскрывает существо причинных связей между явлениями. Иначе говоря, статистический анализ не решает вопрос, в силу каких причин одно явление влияет на другое. Решение этой задачи лежит в другой плоскости и является результатом качественного (содержательного) изучения связей, которое обязательно должно либо предшествовать статистическому анализу, либо сопровождать его.

В естественных науках большей частью сталкиваются с функциональными зависимостями, которые отличаются тем, что каждому возможному значению аргумента поставлено в однозначное соответствие одно строго определенное значение функции. Функциональная зависимость может существовать и между случайными переменными. В самом деле, пусть $y = x^2$, где x — число очков, выпадающих на игральной кости. Здесь x — случайная переменная, которая принимает шесть различных значений, каждое с вероятностью $\frac{1}{6}$. Однако если известно значение x, то y однозначно определяется указанной выше функцией.

В большинстве случаев между переменными, характеризующими экономические величины, существуют зависимости, отличающиеся от функциональных. Такого рода связи характеризуются нежесткими соотношениями между переменными, их множественностью. Например, уровень производительности труда на предприятиях в общем тем выше, чем больше его электровооруженность. Вместе с тем нет никаких оснований утверждать об однозначности этой зависимости.

Множественность результатов при анализе связи x и y объясняется прежде всего тем, что зависимая переменная y испытывает влияние не только фактора x, но и целого ряда других факторов, которые не учитываются. Кроме того, влияние выделенного фактора может быть не прямым, а проявляться через цепочку других факторов. Поэтому в рассматриваемых зависимостях каждому данному значению независимой переменной соответствует не одно, а ряд значений y. Сказанное становится более понятным, если обратиться к специальному виду представления числовых данных, характеризующих взаимосвязь двух признаков в виде *корреляционной таблицы* (см. табл. 12.1, в которой приведены данные о себестоимости и производительности труда сорока предприятий).

Таблица 12.1

Себестоимость единицы продукции, руб. (y)	Месячная производительность труда, тыс шт. (x)					Итого
	10–12	12–14	14–16	16–18	18–20	
6—8	—	—	1	1	2	4
8—10	—	—	3	4	1	8
10—12	—	3	7	4	—	14
12—14	2	4	5	—	—	11
14—16	2	1	—	—	—	3
Итого	4	8	16	9	3	40

В § 6.6 было введено понятие стохастической (вероятностной) зависимости. Данные табл. 12.1 иллюстрируют такого рода зависимость. Распределение показателей себестоимости изменяется здесь вместе с ростом производительности труда. При отсутствии

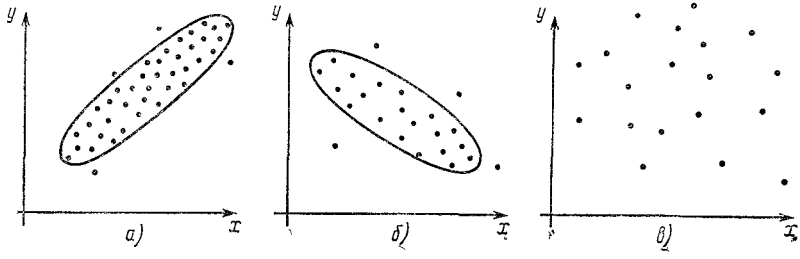


Рис. 12.1.
Корреляционное поле

связи между признаками условные распределения в каждом столбце (строке) были бы примерно одинаковыми.

Частным случаем стохастической зависимости является *корреляционная* зависимость, которая характеризуется тем, что изменение x сопровождается изменением только условной средней распределения y. Связь или корреляция двух переменных называется *парной*. Если с увеличением x переменная y в среднем растет, то такая парная корреляция будет *положительной*, если же y имеет тенденцию к уменьшению с ростом x, то говорят о наличии *отрицательной корреляции*. Нулевая корреляция наблюдается при отсутствии связи между x и y.

Диаграмма, на которой изображается совокупность значений двух признаков, называется *корреляционным полем*. Каждая точка этой диаграммы имеет координаты x_i, y_i, соответствующие размерам признаков в i-м наблюдении. Три варианта распределения точек на корреляционном поле показаны на рис. 12.1. На первом из них основная масса точек укладывается в эллипс, главная диагональ которого образует положительный угол с осью x. Это

график положительной корреляции. Второй вариант распределения соответствует отрицательной корреляции. Равномерное распределение точек в пространстве xy свидетельствует об отсутствии корреляционной зависимости (рис. 12.1, *в*).

При изучении зависимости явлений сталкиваются с двумя различными типами систем предположений. В первом случае экспериментатор задает определенные значения зависимой переменной

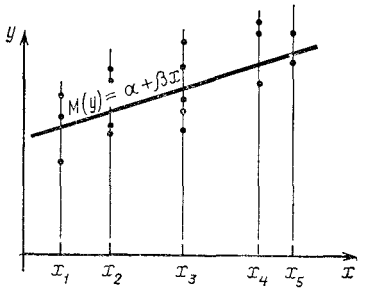


Рис. 12.2.
Регрессионная зависимость

x, для которых и наблюдаются соответствующие значения переменной y. Таким образом, величины x здесь неслучайные (фиксированные), каждому значению x соответствует некоторое генеральное распределение y с дисперсией σ^2 (см. рис. 12.2). Наблю-

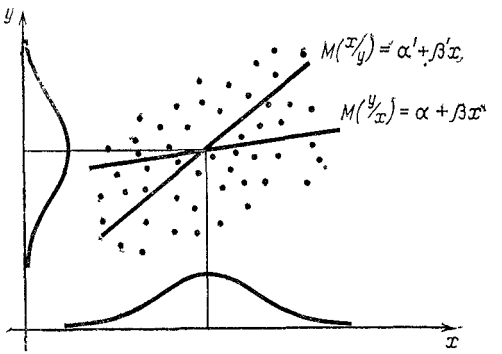


Рис. 12.3.
Корреляционная зависимость

даемые значения y рассматриваются как выборочные значения из этих распределений. В этом случае связь зависимой переменной y от независимой переменной x может быть представлена в виде уравнения регрессии $M(y) = \alpha + \beta x$ (если предполагается, что связь линейная), а саму модель называют *регрессионной*.

Исходные данные, однако, могут иметь и иной характер, а именно: наблюдаемые значения y и x могут представлять собой выборки из двумерного нормального распределения. Следовательно, в отличие от предыдущего здесь x уже не являются фиксированными или контролируемыми переменными. Модель такого рода называют *корреляционной*. В этом случае можно определить две регрессии. математические ожидания y лежат на линии регрессии, описываемой уравнением $M(y/x) = \alpha + \beta x$, а математи-

ческие ожидания x — на линии регрессии $M(x/y) = \alpha' + \beta' y$ (см. рис. 12.3). Различие между двумя моделями имеет принципиальное значение. Однако применяемый для анализа статистический аппарат и в том и другом случае в основном одинаков. Различным будет содержание некоторых получаемых результатов. Следует также упомянуть, что на практике далеко не всегда можно провести четкое разграничение между двумя моделями, особенно при работе с экономическими данными.

12.2. Уравнение парной регрессии

Связь зависимой переменной с одной или несколькими независимыми переменными представляют в виде уравнения регрессии $\hat{y} = f(x_1 \ldots x_n)$. Уравнение регрессии — наиболее часто встречающийся в практике вид статистической модели. Подобные модели применяются, во-первых, непосредственно для экономического и технико-экономического анализа, где с помощью уравнений регрессии измеряют влияние отдельных факторов на зависимую переменную Тем самым анализ становится более конкретным, познавательная его ценность существенно увеличивается. Во-вторых, уравнения регрессии интенсивно применяются в прогностических работах.

Построение уравнения регрессии предполагает решение двух основных задач. Первая заключается в выборе независимых переменных, существенно влияющих на зависимую величину, и в определении вида уравнения регрессии. Этот этап в разработке регрессии называют *спецификацией*. Данная задача решается с помощью качественного анализа изучаемой взаимосвязи. Например, известно, что затраты на изготовление одного изделия определяются целым рядом факторов, причем одни из них связаны с масштабами производства, другие нет. Соответственно статистическую зависимость себестоимости единицы изделия от масштабов производства можно, очевидно, представить в виде уравнения гиперболы с постоянным членом, т. е. $y = a_0 + \frac{a_1}{x}$, где x — объем производства. Процессы, где влияние фактора-аргумента происходит с постоянным ускорением или замедлением, хорошо описываются параболическими кривыми. Наконец, в ряде случаев для описания зависимостей экономических переменных пригодны и более сложные типы функций, например S-образные кривые. В частности, хорошую экономическую интерпретацию могут получать так называемые логистические функции типа $y = \frac{k}{1 + ab^{f(x)}}$. Эти функции оказываются полезными там, где процесс сначала ускоренно развивается, а затем, после достижения некоторого уровня, затухает и приближается к некоторому пределу. Наиболее простой является гипотеза о линейной зависимости переменных.

Значительную помощь при выборе вида уравнения могут оказать графики. Рассмотрим простой случай. Пусть необходимо

описать в виде некоторой функции зависимость двух переменных величин x и y. Если зависимость y от x существует, то эмпирические данные покажут, что $\bar{Y}_x$ (напомним, что этим символом мы обозначали условную среднюю переменной y для фиксированных значений x) изменяется с изменением значения x. Соединим на корреляционном поле значения $\bar{Y}_x$, получим ломаную линию. Поскольку $\bar{Y}_x$ являются выборочными оценками, то с увеличением объема наблюдений ошибки выборки этих средних будут уменьшаться, а ломаная линия приближаться к некоторой сглаженной кривой — линии регрессии. В пределе, при неограниченно большом числе наблюдений, все значения $\bar{Y}_x$ должны лежать на кривой, если, повторяем, между переменными действительно существует зависимость соответствующего вида. Практически, однако, число наблюдений всегда ограничено и эмпирическая линия, характеризующая зависимость Y от x, будет в большей или меньшей мере отличаться от плавной кривой.

Вторая задача — оценивание параметров (коэффициентов) уравнения — решается с помощью того или иного математико-статистического метода обработки данных. Поскольку параметры уравнений представляют собой выборочные характеристики, то процесс оценивания должен сопровождаться статистической проверкой существенности полученных параметров.

Линейное уравнение парной регрессии. Пусть между переменными x и y теоретически существует некоторая линейная зависимость. Однако в действительности между x и y наблюдается не столь жесткая связь. Отдельные наблюдения y будут отклоняться от линейной зависимости в силу воздействия различных причин. Обычно зависимая переменная находится под влиянием целого ряда факторов, в том числе и не известных исследователю, а также случайных причин (возмущения и помехи); существенным источником отклонений в ряде случаев являются ошибки измерения. Отклонения от предполагаемой формы связи, естественно, могут возникнуть и в силу неправильной спецификации уравнения, т. е. неправильного выбора вида самого уравнения, описывающего эту зависимость. В дальнейшем будем полагать, что спецификация выполнена правильно. Учитывая возможные отклонения, линейное уравнение связи двух переменных (парную регрессию) представим в виде

$$y = \alpha + \beta x + \varepsilon, \tag{12.1}$$

где α и β — неизвестные параметры (коэффициенты) регрессии; ε — случайная переменная, характеризующая отклонение от теоретически предполагаемой регрессии — возмущение. Таким образом, в уравнении (12.1) значение y представлено как сумма двух частей — систематической $(\alpha + \beta x)$ и случайной (ε). В свою очередь систематическую часть можно представить в виде уравнения

$$\hat{y} = \alpha + \beta x, \tag{12.2}$$

где $\hat{y}$ характеризует некоторое среднее значение y для данного значения x. Соответственно уравнение (12.1) показывает значения y с учетом возможных отклонений от средних значений.

Относительно возмущения ε сделаем следующие предположения 1) возмущение ε является нормально распределенной случайной переменной; 2) математическое ожидание ε равно нулю: $M(\varepsilon)=0$; 3) дисперсия возмущений постоянна: $\sigma_\varepsilon^2=\text{const}$; 4) последовательные значения ε не зависят друг от друга.

Итак, при построении регрессии (в данном случае линейной парной регрессии) принимается гипотеза о том, что для каждого наблюдения i справедлива следующая зависимость:

$$y_i=\alpha+\beta x_i+\varepsilon_i.$$

В результате статистического наблюдения исследователь имеет ряд характеристик независимой переменной x и соответствующие значения зависимой переменной y. Задача заключается в определении параметров α и β. Однако истинные значения этих параметров получить нельзя, так как мы опираемся на выборку ограниченного объема. Поэтому найденные значения параметров являются статистическими оценками истинных (неизвестных нам) параметров. Обозначим соответствующие оценки как a и b. Таким образом, уравнение парной регрессии $\hat{y}=a+bx$ есть оценка модели $y=\alpha+\beta x+\varepsilon$.

Метод наименьших квадратов для парной регрессии. Приняв некоторую гипотезу о форме кривой, описывающей взаимосвязь переменных x и y (пусть для определенности это будет простая линейная взаимосвязь), мы тем не менее не можем однозначно подобрать параметры уравнения, если не будут выдвинуты условия, которым должно отвечать уравнение с численно оцененными параметрами В самом деле, через область корреляционного поля, в которой расположены точки, соответствующие отдельным наблюдениям, можно провести множество прямых (например, соединить первую и последнюю точку или первую и предпоследнюю и т. д.). Необходимо принять некоторый критерий, в соответствии с которым будет осуществляться подбор параметров. Различные методы оценивания параметров опираются на различные критерии подбора и, разумеется, дают неодинаковые значения оценок параметров для одной и той же совокупности наблюдений. При этом оказывается, что получаемые оценки обладают разными статистическими свойствами.

Наиболее часто оценивание параметров регрессий осуществляют на основе *метода наименьших квадратов* (МНК), разработка которого восходит к К. Гауссу и П Лапласу. МНК первоначально имел довольно узкую сферу применения — главным образом при обработке результатов наблюдений в астрономических и геодезических расчетах. Этот метод получил новую и широкую область приложения в экономико-статистических расчетах после создания теории регрессии. Согласно МНК параметры уравнения

регрессии подбираются так, чтобы сумма квадратов отклонений наблюдений от линии регрессии была минимальной. Рассмотрим график (см. рис. 12.4), на котором показаны результаты наблюдений значений переменных x и y. Через область, занимаемую точками, проведена прямая $\hat{y} = a + bx$. Отклонение (возмущение) какой-либо точки с координатами x_i, y_i составит величину e_i:

$$e_i = y_i - \hat{y}_i = y_i - (a + bx_i). \tag{12.3}$$

Здесь y_i — фактическое, а $\hat{y}_i$ — расчетное значение зависимой переменной.

Как видно из (12.3), величина e_i (ее часто называют остаточным членом) есть функция параметров a и b. Точно так же функ-

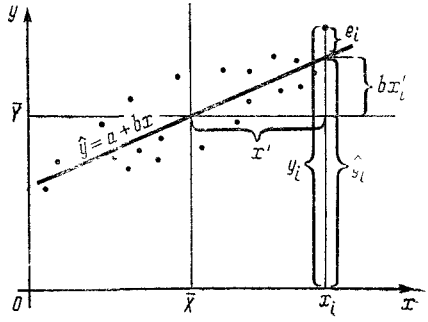


Рис. 12.4.
Парная регрессия

цией этих параметров является обобщенный показатель рассеяния точек вокруг прямой, а именно $\sum e_i^2$. Отсюда логично принять критерий, согласно которому коэффициенты регрессии a и b должны быть подобраны так, чтобы сумма квадратов величин e_i была минимальна, т. е.

$$\sum e_i^2 = \min.$$

Необходимым условием существования минимума функции является равенство нулю частных производных по неизвестным параметрам a и b. Итак, найдем для функции

$$Q = \sum e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - a - bx_i)^2$$

частные производные по a и b и приравняем их нулю. Получим

$$\left.\begin{aligned} \frac{\partial Q}{\partial a} &= -2 \sum_i (y_i - a - bx_i) = 0; \\ \frac{\partial Q}{\partial b} &= -2 \sum_i (y_i - a - bx_i)\, x_i = 0. \end{aligned}\right\} \tag{12.4}$$

Преобразовав систему (12.4), получим стандартную форму *нормальных уравнений* *:

$$\left.\begin{aligned} \sum y_i &= na + b\sum x_i; \\ \sum x_i y_i &= a\sum x_i + b\sum x_i^2. \end{aligned}\right\} \qquad (12.5)$$

Разделим первое уравнение системы (12.5) на n, получим

$$\bar{Y} = a + b\bar{X}. \qquad (12.6)$$

Таким образом, МНК дает такие оценки a и b, что найденная прямая проходит через точку с координатами $\bar{X}$, $\bar{Y}$.

Значения переменных x_i и y_i могут быть измерены в отклонениях от средней, т. е. как $x_i - \bar{X}$; $y_i - \bar{Y}$. Обозначим эти разности как x'_i и y'_i соответственно. Начало координат при этом переместится в точку $\bar{X}$, $\bar{Y}$, а система нормальных уравнений упростится до

$$\sum x'_i y'_i = b\sum (x')^2,$$

так как $\sum y'_i$ и $\sum x'_i$ равны нулю. Решение этого уравнения относительно b дает

$$b = \frac{\sum x'_i y'_i}{\sum (x'_i)^2}. \qquad (12.7)$$

Из (12.6) получим

$$a = \bar{Y} - b\bar{X}. \qquad (12.8)$$

Необходимые для расчета параметра b суммы могут быть получены по следующим формулам (индексы i опущены):

$$\sum (x')^2 = \sum x^2 - \frac{(\sum x)^2}{n} = \sum x^2 - n\bar{X}^2; \qquad (12.9)$$

$$\sum x'y' = \sum xy - \frac{\sum x \sum y}{n} = \sum xy - n\bar{X}\bar{Y}. \qquad (12.10)$$

Как следует из изложенного выше, метод оценки параметров a и b не зависит от того, являются ли x фиксированными или выборочными значениями.

Пусть годовая производительность труда в расчете на одного рабочего и энерговооруженность труда на предприятиях одной отрасли характеризуются следующими данными (табл. 12.2):

Таблица 12.2

Производительность, тыс. руб./чел.	6,7	6,9	7,2	7,3	8,4	8,8	9,1	9,8	10,6	10,7	11,1	11,8	12,1	12,4
Энерговооруженность, кВт/чел.	2,8	2,2	3,0	3,5	3,2	3,7	4,0	4,8	6,0	5,4	5,2	5,4	6,0	9,0

* Здесь и далее предполагается, что суммирование производится с $i = 1$ по $i = n$.

Требуется оценить линейное уравнение регрессии. Здесь, очевидно, x_i — энерговооруженность, y_i — производительность труда, $i = 1, \ldots 14$, расчет параметров начнем с определения сумм квадратов и средних:

$$\sum y = 132{,}9; \quad \sum x = 64{,}2,$$

$$\sum (x')^2 = 335{,}26 - \frac{64{,}2^2}{14} = 40{,}857;$$

$$\sum x'y' = 650{,}99 - \frac{64{,}2 \cdot 132{,}9}{14} = 41{,}549;$$

$$\overline{Y} = \frac{132{,}9}{14} = 9{,}493; \quad \overline{X} = \frac{64{,}2}{14} = 4{,}586.$$

Откуда

$$b = \frac{\sum x'y'}{\sum (x')^2} = \frac{41{,}549}{40{,}857} = 1{,}017.$$

По формуле (12 8) определяем $a = \overline{Y} - b\overline{X} = 9{,}493 - 1{,}017 \cdot 4{,}586 = 4{,}829$. Уравнение регрессии с оцененными параметрами имеет вид:

$$\hat{y}_i = 4{,}829 + 1{,}017\, x_i.$$

Таким образом, рост энерговооруженности труда на 1 кВт/чел. на данных предприятиях приводит к увеличению производительности труда в среднем на 1,017 тыс. руб/чел. в год.

Свойства оценок параметров, получаемых МНК. Оценки МНК при условии, что сделанные выше предположения относительно ε справедливы, обладают рядом ценных для последующего применения регрессий свойств *, а именно:

а) оценки параметров являются *несмещенными*, т. е. математическое ожидание параметра равно истинному его значению; в частности, для парной регрессии $M(a) = \alpha$ и $M(b) = \beta$. Несмещенность означает, что выборочные оценки параметров концентрируются вокруг истинных неизвестных значений параметров;

б) оценки *состоятельны*, иначе говоря, дисперсия оценки параметра стремится к нулю с возрастанием n. Для парной регрессии это свойство можно отразить как

$$\lim_{n \to \infty} \sigma_a^2 = 0 \quad \text{и} \quad \lim_{n \to \infty} \sigma_b^2 = 0.$$

Свойство состоятельности означает, что при увеличении объема наблюдения оценки параметров становятся более надежными в вероятностном смысле, т. е. с ростом n оценки все плотнее концентрируются вокруг истинных неизвестных значений параметров;

в) оценки являются *эффективными* в том смысле, что они имеют минимальную дисперсию по сравнению с любыми другими линейными и несмещенными оценками этого параметра.

Если предположение 3 или 4 относительно ε нарушено, то свойство несмещенности оценок сохраняется, однако оценки оказываются менее эффективными, чем в случае, когда эти допущения соблюдаются.

* Доказательство свойств оценок параметров уравнений регрессии см., например, в [4], [5], [6].

Ошибки коэффициентов регрессии. Итак, мы оценили параметры a и b и получили регрессию, на основе которой можно оценивать значения y для заданных значений x. Естественно полагать, что действительные значения зависимой переменной не будут совпадать с расчетными, так как сама линия регрессии описывает взаимосвязь лишь в среднем, в общем. Отдельные наблюдения рассеяны вокруг нее. Таким образом, надежность получаемых по уравнению регрессии расчетных значений во многом определяется рассеянием наблюдений вокруг линии регрессии. В качестве меры рассеяния примем такую общую характеристику, как дисперсия относительно регрессии. Для ее определения найдем сумму квадратов отклонений фактических наблюдений от линии регрессии, т. е. $\sum e_i^2$. Искомая дисперсия, таким образом, равна:

$$s^2 = \frac{\sum e_i^2}{n-2}. \tag{12.11}$$

Число степеней свободы равно $n-2$, так как две степени свободы теряются при определении двух параметров уравнения прямой.

Величина s^2 является выборочной оценкой дисперсии возмущений ε_i, содержащихся в теоретической модели (12.1). Величину $\sum e_i^2$ можно определить, не рассчитывая e_i. В самом деле, из рис. 12.2 легко установить, что

$$e_i = y_i' - bx_i'. \tag{12.12}$$

Возведем это выражение в квадрат и просуммируем. Преобразуем результат, используя соотношения (12.7). Получим

$$\sum e_i^2 = \sum (y_i')^2 - b \sum x_i' y_i'. \tag{12.13}$$

Определим теперь дисперсию коэффициента регрессии b. По определению $\mathrm{D}(b) = \mathrm{M}(b-\beta)^2$. Можно доказать, что

$$\mathrm{D}(b) = s_b^2 = \frac{s^2}{\sum (x_i')^2}. \tag{12.14}$$

Дисперсию коэффициента a парной линейной регрессии находим следующим образом:

$$\mathrm{D}(a) = s_a^2 = \frac{\sum x_i^2}{n \sum (x_i')^2} s^2. \tag{12.15}$$

В случае, если регрессия имеет вид $\hat{y}_i = \bar{Y} + bx_i'$, выражение для оценки дисперсии свободного члена ($a = \bar{Y}$) упрощается до

$$s_a^2 = \frac{s^2}{n}. \tag{12.16}$$

Опираясь на значения дисперсии коэффициентов регрессии, можно проверить значимость этих коэффициентов, применив обычную схему проверки статистических гипотез. При этом

предполагаем, что отклонения от регрессии следуют нормальному распределению.

Обычно в парной регрессии проверяется только значимость коэффициента b. Нулевая гипотеза имеет вид $H_0 : \beta = 0$. Проверка осуществляется с помощью нормированного отклонения $t = \frac{b}{s_b}$, следующего t-распределению Стьюдента. Если $t^* > t_\alpha$ с числом степеней свободы $\nu = n - 2$, то нулевая гипотеза отклоняется. Последнее является статистическим подтверждением существования линейной зависимости соответствующих переменных.

В предыдущем примере получена регрессия $\hat{y}_i = 4{,}829 + 1{,}017x_i$. Найдем ошибку коэффициента регрессии и проверим его существенность. По данным того же примера определим следующие суммы: $\sum (x')^2 = 40{,}822$; $\sum (y')^2 = 52{,}35$ $\sum x'y' = 41{,}502$. Теперь по формуле (12.13) найдем сумму квадратов отклонений от регрессии $\sum e_i^2 = 52{,}352 - 1{,}017 \cdot 41{,}502 = 10{,}142$. Дисперсия относительно регрессии равна:

$$s^2 = \frac{10{,}142}{14 - 2} = 0{,}845.$$

Квадратическая ошибка параметров b равна в соответствии с (12.14):

$$s_b = \frac{0{,}845}{\sqrt{40{,}822}} = 0{,}132.$$

Следовательно,

$$t^* = \frac{1{,}017 - 0}{0{,}132} = 7{,}7.$$

Критическое значение t_α значительно меньше, чем полученное нами значение t^* (при числе степеней свободы равном 12 $t_{0\,05} = 2{,}179$). Следовательно, гипотеза H_0 отклоняется.

Доверительные интервалы линии регрессии Рассмотрим теперь метод определения доверительных границ для значения $\hat{y}$, т. е. тех границ, в пределах которых с заданной доверительной вероятностью будет находиться это значение. Итак, в силу того, что оценивание параметров осуществляется по выборочным данным, оценки a и b содержат некоторую погрешность. Причем погрешность в значении a приводит к вертикальному сдвигу линии регрессии, а колеблемость оценки b, связанная с ее выборочным происхождением,— к «покачиванию» линии регрессии. Таким образом, дисперсия значения зависимой переменной $\hat{y}_i$, обозначим ее $s_{\hat{y}}^2$, будет складываться из двух компонент — дисперсии параметра a и дисперсии параметра b. Выражение для $s_{\hat{y}}^2$ удобнее определить исходя из соотношения $\hat{y}_i = \bar{Y} + bx_i'$. В этом случае непосредственно получаем

$$s_{\hat{y}}^2 = \frac{s^2}{n} + \frac{s^2}{\sum (x_i')^2}(x_p')^2 = s^2\left(\frac{1}{n} + \frac{(x_p')^2}{\sum (x_i')^2}\right), \qquad (12.17)$$

где x_p' — значение переменной x (выраженное в виде отклонения от средней), для которого определяется $\hat{y}$. Из (12.17) видно, что

$s^2_{\hat{y}}$ имеет минимальное значение в точке $x'_p = 0$. В этом случае $s^2_{\hat{y}} = \frac{s^2}{n}$.

Зная дисперсию $\hat{y}_i$, легко определить доверительные границы

$$\hat{y}_i \pm t_\alpha s_{\hat{y}} \tag{12.18}$$

Если нанести доверительные границы на график, то они расположатся выше и ниже линии регрессии.

Доверительная зона (12.18) определяет местоположение линии регрессии, но не отдельных возможных значений зависимой переменной, которые отклоняются от нее. Следовательно, если мы хотим определить доверительные интервалы для отдельных значений зависимой переменной, то при определении дисперсии необходимо учитывать еще один источник вариации — рассеяние вокруг линии регрессии, иначе говоря, в суммарную дисперсию следует еще включить величину s^2. Таким образом, дисперсия прогнозных значений y_p равна:

$$s_p^2 = s^2 + \frac{s^2}{n} + \frac{s^2}{\sum (x')^2}(x'_p)^2 = s^2\left(1 + \frac{1}{n} + \frac{(x'_p)^2}{\sum (x')^2}\right). \tag{12.19}$$

Доверительные интервалы для прогнозов индивидуальных значений будут, следовательно, равны

$$\hat{y}_i \pm t_\alpha s_p.$$

Пусть по регрессии $\hat{y}_i = 4{,}829 + 1{,}017x_i$ (см. пример на с. 227) необходимо оценить среднее значение производительности труда, соответствующее энерговооруженности, равной 5,5 кВт/чел. Тогда

$$\hat{y} = 4{,}829 + 1{,}017 \cdot 5{,}5 = 10{,}42;$$

$$s_p^2 = 0{,}845\left(1 + \frac{1}{14} + \frac{(5{,}5 - 4{,}586)^2}{40{,}822}\right) = 0{,}919.$$

Откуда при $t_{0,05} = 2{,}179$ получим

$$y_p = 10{,}42 \pm 2{,}179\sqrt{0{,}919} = 10{,}42 \pm 2{,}09.$$

Если взять значение независимой переменной, которое в большей мере удалено от $\bar{X}$, чем $x_p = 5{,}5$, то доверительный интервал, естественно, увеличится при той же доверительной вероятности.

12.3. Коэффициент корреляции

При изучении двумерного нормального распределения случайных величин X и Y в гл. 6 была введена такая характеристика тесноты связи, как коэффициент корреляции

$$\rho = \frac{M(x - \bar{X})(y - \bar{Y})}{\sqrt{M(x - \bar{X})^2 M(y - \bar{Y})^2}}.$$

С помощью этого показателя, как говорилось выше, измеряется степень коррелированности (линейной зависимости) признаков в генеральной совокупности. Для выборочных данных

эмпирическая мера связи x и y определяется как

$$r = \frac{\sum x'_i y'_i}{n s_x s_y}, \tag{12.20}$$

где

$$s_x = \sqrt{\frac{\sum (x'_i)^2}{n}}; \quad s_y = \sqrt{\frac{\sum (y'_i)^2}{n}}$$

— средние квадратические ошибки x и y, которые определены без учета числа степеней свободы.

Формулу (12.20) обычно применяют в несколько преобразованном виде. Так, подставив в нее развернутые выражения s_x и s_y, легко получить

$$r = \frac{\sum\limits_i x'_i y'_i}{\sqrt{\sum\limits_i (x'_i)^2 \sum\limits_i (y'_i)^2}}, \tag{12.21}$$

а использовав (12.9) и (12.10), найдем еще одну рабочую формулу:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}. \tag{12.22}$$

Коэффициент корреляции (12.22) находится непосредственно через данные наблюдений, следовательно, округления, связанные с расчетом средних и отклонений от них, не повлияют на значение r.

Напомним следующие свойства коэффициента корреляции:

а) коэффициент не имеет размерности, следовательно, он сопоставим для разных статистических рядов;

б) величина r лежит в пределах от -1 до 1. Значение $r = +1$ свидетельствует о том, что между переменными существует полная положительная корреляция, т. е. функциональная зависимость — все данные наблюдения лежат на прямой с положительным углом наклона в плоскости xy, иначе говоря, с увеличением x растет y; $r = -1$ указывает на полную обратную линейную связь; $r = 0$ (а это может быть тогда, когда $\Sigma x'y' = 0$) не означает, что x и y статистически независимы, а лишь указывает на отсутствие линейной связи между ними, что, естественно, не отрицает возможность существования иной формы зависимости между переменными.

Итак, при наличии двумерного нормального распределения коэффициент корреляции является мерой линейной согласованности между переменными, их взаимного варьирования. Высокий коэффициент корреляции подтверждает наличие линейной связи между переменными. Последняя может быть, если x есть причина (или следствие) y, если x и y являются совместно зависимыми переменными — x зависит от y, а y от x (например, цены на один вид продукции на разных колхозных рынках), наконец, если x и

y являются следствиями некоторой общей для них причины (например, существенную корреляцию имеют такие признаки, как рост и вес человека; однако ни один из них не может рассматриваться как непосредственная причина другого).

Если анализ связи переменных x и y выполнялся на основе модели, предполагающей, что x — фиксированные значения, устанавливаемые экспериментатором, то коэффициент корреляции может быть вычислен и в этом случае, однако его не следует рассматривать как строгую меру взаимосвязи явлений. Для модели с фиксированными x это просто мера близости эмпирических точек к линии регрессии. Заметим, что в этом случае величина коэффициента корреляции будет заметно зависеть от того, какие значения x выбраны экспериментатором. Если же совокупность значений x представляет собой выборку, то предполагается, что выборка отражает соответствующее генеральное распределение и отклонения от него есть следствие только случайности.

В практике статистического анализа не являются исключением случаи, когда с помощью корреляционного анализа обнаруживают существование достаточно сильной «зависимости» признаков, в действительности не имеющих причинной связи между собой. Такие корреляции принято называть ложными или бессмысленными. Как правило, бессмысленные корреляции получают при коррелировании временных рядов двух признаков, не связанных причинной зависимостью. В дальнейшем будем полагать, что между рассматриваемыми переменными существует причинная зависимость и, следовательно, применение теории корреляции имеет логическое основание.

Определим коэффициент корреляции по данным примера на с 227. Для этого предварительно рассчитаем необходимые суммы:

$$\Sigma x = 64{,}2; \quad \Sigma y = 132{,}9; \quad \Sigma xy = 650{,}99;$$

$$\Sigma x^2 = 335{,}26; \quad \Sigma y^2 = 1313{,}95;$$

откуда

$$r = \frac{14 \cdot 650{,}99 - 64{,}2 \cdot 132{,}9}{\sqrt{(14 \cdot 335{,}26 - 64{,}2^2)\,(14 \cdot 1313{,}95 - 132{,}9^2)}} = 0{,}9.$$

Полученная величина коэффициента корреляции подтверждает наличие линейной зависимости между переменными.

Из (12.7) и (12.21) следует, что между коэффициентом корреляции и параметром парной линейной регрессии существует зависимость, которая применительно к выборочным оценкам может быть представлена следующим образом:

$$b = r \frac{s_{\overline{Y}}}{s_{\overline{X}}}, \tag{12.23}$$

где $s_{\overline{Y}}$ и $s_{\overline{X}}$ — средние квадратические ошибки. Приведенное выражение позволяет оценить параметр регрессии без решения системы нормальных уравнений при условии, что коэффициент корреляции уже определен. На основе последней формулы легко

показать, что коэффициент корреляции есть средняя геометрическая двух сопряженных коэффициентов регрессии В самом деле, обозначим как b_{yx} коэффициент регрессии y по x. Допустим, что нам необходимо найти «обратную» регрессию — x по y. Тогда очевидно, что $b_{xy} = \frac{\sum x'y'}{\sum (y')^2}$, следовательно,

$$r = \sqrt{b_{yx} \frac{\sum x'y'}{\sum (y')^2}} = \sqrt{b_{yx} b_{xy}}. \qquad (12.24)$$

Если зависимость между признаками функциональная, то $b_{yx} = \frac{1}{b_{xy}}$ и, следовательно, $r = 1$. Наоборот, при полном отсутствии зависимости между признаками $b_{yx} = 0$, $b_{xy} = 0$ и $r = 0$.

Проверка значимости r. Применяя коэффициент корреляции в качестве меры связи, нужно иметь в виду, что он получен на основе данных выборки и, следовательно, подвержен влиянию случайности. Если объем выборки небольшой, то найти выборочную ошибку этой величины достаточно сложно, поэтому в практике обычно вместо определения ошибки коэффициента корреляции проверяют гипотезу о его значимости (существенности), т. е. существенно ли r отличается от нуля или это отличие можно приписать влиянию случайности, связанной с выборкой. Иначе говоря, проверяется гипотеза H_0: $\rho = 0$. Проверка значимости осуществляется путем сопоставления табличного и расчетного значений t-статистики. Последняя определяется по формуле

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (12.25)$$

Величина t здесь следует t-распределению Стьюдента. Найденное по данной формуле значение t^* сопоставляют с табличным значением t_α при $n - 2$ степенях свободы.

При изучении взаимосвязи роста и веса человека (выборка охватила 42 мужчин) коэффициент корреляции оказался равным 0,65. Необходимо проверить гипотезу о существенности отличия ρ от нуля Применяя формулу (12.25), получим

$$t^* = \frac{0{,}65\sqrt{42-2}}{\sqrt{1-0{,}65^2}} = 5{,}27.$$

При $\alpha = 0{,}05$ и числе степеней свободы, равном 40, $t_\alpha = 2{,}02$ Следовательно, коэффициент корреляции существенно отличается от нуля ($\rho \neq 0$), что, впрочем, и следовало ожидать

Как следует из формулы (12.25), значение t здесь полностью определяется числом наблюдений (n) и величиной выборочного коэффициента корреляции. Поэтому нетрудно для заданного числа степеней свободы найти наименьшее значение выборочного коэффициента корреляции, при котором гипотеза H_0: $\rho = 0$ может быть отклонена с заданной вероятностью. Таблица таких значений приведена в [15].

Для данных предыдущего примера минимальное значение коэффициента корреляции равно 0,30 (при числе степени свободы $42 - 2 = 40$), что значительно меньше 0,65

Коэффициент детерминации как мера качества подбора линии регрессии. Коэффициент корреляции служит основой для другой статистической характеристики — *коэффициента детерминации*. Последний определяют как r^2. Этот коэффициент позволяет ответить на вопрос о том, каково качество описания зависимости с помощью уравнения регрессии. Очевидно, чем теснее наблюдения примыкают к линии регрессии, тем лучше регрессия описывает соответствующую зависимость переменных и с большей надежностью может быть применена для практических расчетов — оценивания значения y для заданных значений x.

Для того чтобы выяснить смысл коэффициента детерминации, значение зависимой переменной расчленим на составляющие:

$$y_i = \bar{Y} + k_i + e_i. \tag{12.26}$$

Здесь $k_i = \hat{y}_i - \bar{Y} = bx'_i$ и $y'_i = y_i - \bar{Y} = k_i + e_i$.

Определим теперь сумму квадратов отклонений:

$$\sum (y'_i)^2 = \sum (k_i + e_i)^2 = \sum k_i^2 + 2 \sum k_i e_i + \sum e_i^2.$$

Однако $2\Sigma k_i e_i = 0$. Это следует из того, что

$$e_i = y'_i - k_i = y'_i - bx'_i \tag{12.27}$$

и

$$\sum k_i e_i = \sum bx'_i (y'_i - bx'_i) = b\left[\sum x_i y'_i - b \sum (x'_i)^2\right].$$

Но

$$b = \frac{\sum x'_i y'_i}{\sum (x'_i)^2}$$

и, следовательно

$$\sum k_i e_i = b\left[\sum x'_i y'_i - \frac{\sum x'_i y'_i}{\sum (x'_i)^2} \sum (x'_i)^2\right] = b \cdot 0 = 0. \tag{12.28}$$

Таким образом,

$$\sum (y_i)^2 = \sum k_i^2 + \sum e_i^2. \tag{12.29}$$

Итак, сумма квадратов отклонений от средней расчленена на две составляющие. Первая из них характеризует *систематическую вариацию*, т е изменение $\hat{y}_i$ в соответствии с уравнением регрессии (такую вариацию иногда называют *объясненной*, так как она объясняется уравнением регрессии), вторая — *случайную вариацию* (отклонение от линии регрессии).

Найдем теперь отношение суммы квадратов отклонений, обусловленных линией регрессии, к общей сумме квадратов:

$$\frac{\sum k_i^2}{\sum (y_i')^2} = \frac{\sum (bx_i')^2}{\sum (y_i')^2} = b^2 \frac{\sum (x_i')^2}{\sum (y_i')^2} =$$

$$= \left[\frac{\sum x_i' y_i'}{\sum (x_i')^2}\right]^2 \frac{\sum (x_i')^2}{\sum (y_i')^2} = \frac{(\sum x_i' y_i')^2}{\sum (x_i')^2 \sum (y_i')^2}.$$

Напомним, что

$$r = \frac{\sum x_i' y_i'}{\sqrt{\sum (x_i')^2 \sum (y_i')^2}}.$$

Следовательно,

$$\frac{\sum k_i^2}{\sum (y_i')^2} = \frac{\sum (\hat{y} - \bar{Y})^2}{\sum (y_i - \bar{Y})^2} = r^2. \qquad (12.30)$$

Величина r^2 получила название выборочного *коэффициента детерминации*. Заменим теперь суммы квадратов отклонений в формуле (12.30) на соответствующие дисперсии. После такой замены получим

$$r^2 = \frac{s_{\hat{y}}^2}{s_{\bar{Y}}^2} \quad \text{или} \quad r^2 = 1 - \frac{s^2}{s_{\bar{Y}}^2}, \qquad (12.31)$$

где $s_{\hat{y}}^2$ — дисперсия линии регрессии относительно средней; s^2 — дисперсия остаточных членов относительно линии регрессии; $s_{\bar{Y}}^2$ — общая дисперсия.

Из (12.31) следует, что коэффициент детерминации характеризует долю объясненной регрессией дисперсии в общей величине дисперсии зависимой переменной.

Так, в нашем примере с производительностью труда $r = 0{,}9$ Это значит, что $r_2 = 0{,}81$ и 81% общей дисперсии производительности труда (относительно средней) связан с изменением его энерговооруженности.

В силу сказанного r^2 может рассматриваться как мера качества описания зависимости признаков x и y с помощью уравнения регрессии. Чем ближе $s_{\hat{y}}$ к значению $s_{\bar{Y}}$, тем выше r^2 и тем теснее примыкают отдельные наблюдения к линии регрессии. Если $s_{\hat{y}} = s_{\bar{Y}}$, то $s = 0$, а $r^2 = 1$. В этом случае все эмпирические точки лежат на линии регрессии. Если $s_{\hat{y}} = 0$ (а это может быть в том случае, когда все $\hat{y} = \bar{Y}$ и изменения y не связаны с изменениями x), то $r^2 = 0$.

Из (12.30) и (12.31) следует, что коэффициент корреляции может быть определен и как

$$r = \sqrt{\frac{\sum (\hat{y} - \bar{Y})^2}{\sum (y_i - \bar{Y})^2}} = \sqrt{1 - \frac{\sum e_i^2}{\sum (y_i - \bar{Y})^2}} \qquad (12.32)$$

и, наконец,

$$r = \sqrt{1 - \frac{s^2}{s_Y^2}}. \qquad (12.33)$$

В таком виде коэффициенг принято называть *индексом корреляции* или *корреляционным отношением.*

12.4. Ранговая корреляция

Коэффициент ранговой корреляции. Выше была рассмотрена мера степени (тесноты) взаимосвязи признаков, каждый из которых можно было измерить. Иногда сталкиваются со случаями, когда признак не имеет абсолютной шкалы измерения (профессии, различного рода предпочтения, качественные особенности и т. д.) или, хотя теоретически признак и можно измерить, практически это неосуществимо. Оказывается, что и в этих условиях проблема измерения тесноты связи между признаками разрешима. Для того чтобы определить соответствующий измеритель, следует прежде всего упорядочить или *ранжировать* данные. Для этого объекты анализа располагаются по порядку в соответствии с некоторым признаком (качественным или количественным). При этом каждому объекту присваивается порядковый номер, который называется рангом. Ранги представляют собой члены натурального ряда чисел от 1 до n. Ранг 1 приписывается наиболее важному или крупному объекту, ранг 2 — следующему и т. д. Ничего практически не изменится, если за начало будет взят наименее важный или наименьший по величине признака объект. Если объекты ранжированы по двум признакам, то имеется возможность измерить силу связи между признаками, основываясь на значениях рангов Для этого обычно применяют коэффициент ранговой корреляции Спирмэна *

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}, \qquad (12.34)$$

где d — разность значений рангов, расположенных в двух рядах у одного и того же объекта.

Коэффициент r_s является частным случаем коэффициента парной корреляции. В самом деле, его можно получить, преобразовав выражение (12.22) путем введения в него вместо значений x и y соответствующих рангов.

Величина r_s для двух рядов, состоящих из n рангов, зависит только от $\sum d^2$. Посмотрим, какие значения эта сумма может принимать и как она влияет на r_s Если два ряда рангов полностью совпадают, то $\sum d^2 = 0$ и, следовательно, $r_s = 1$. Таким образом, при полной прямой связи $r_s = 1$ Можно доказать, что при полной

* Детальное обсуждение различных методов ранговой корреляции приводится в [9].

обратной связи, т. е. когда ранги двух рядов расположены в обратном порядке, $\sum d^2 = \frac{n(n^2-1)}{3}$. Откуда $r_s = -1$. Как видим, крайние значения r_s, равные 1 и -1, симметрично расположены относительно нуля, поэтому значение суммы квадратов отклонений, соответствующее $r_s = 0$, находим как среднюю из двух крайних значений Σd^2, т. е.

$$\left[0 + \frac{n(n^2-1)}{3}\right] : 2 = \frac{n(n^2-1)}{6}.$$

Подставив найденное значение в (12.37), получим $r_s = 0$.

Допустим, что при ранжировании отметок на вступительных экзаменах и средних баллов за первую экзаменационную сессию одних и тех же лиц получены следующие ранги:

Таблица 12.3

	Студент	А	Б	В	Г	Д	Е	Ж	З	И	К
Ранг	Вступительные экзамены	2	5	6	1	4	10	7	8	3	9
	Экзаменационная сессия	3	6	4	1	2	7	8	10	5	9
d		-1	-1	2	0	2	3	-1	-2	-2	0
d^2		1	1	4	0	4	9	1	4	4	0

В результате

$$\sum d^2 = 28; \quad r_s = 1 - \frac{6 \cdot 28}{10(10^2-1)} = 0{,}83.$$

Коэффициент ранговой корреляции подтвердил наличие достаточно высокой связи между результатами двух видов экзаменов.

При ранжировании иногда сталкиваются со случаями, когда трудно выделить один из двух или более объектов, настолько они оказываются подобными в отношении какого-нибудь признака. Такие случаи возникают тогда, когда или числовые характеристики признака одинаковы, или ранжирование осуществляется на основе субъективных оценок, а исследователь не в состоянии найти существенные различия между объектами. Объекты, как говорят, оказываются *связанными*. Связанным объектам приписываются одинаковые ранги. Для того чтобы сумма всех рангов оставалась такой же, как и в случае, когда нет связанных рангов, необходимо подобным объектам присвоить их средний ранг. Допустим, что два объекта получили ранги 1 и 2. Следующие четыре объекта оказались равноценными в отношении соответствующего признака. Средний ранг, приписываемый каждому из них, равен $\frac{1}{4}(3+4+5+6) = 4{,}5$.

Проверка значимости r_s. Поскольку r_s определяется на основе выборки, возникает необходимость в проверке гипотезы $H_0: \rho_s = 0$, где ρ_s — генеральный коэффициент ранговой корреляции. При проверке исходят из того, что распределение ρ_s стремится к нормальному с увеличением n.

Среднюю квадратическую ошибку находят по следующей простой формуле (при отсутствии связей между рангами):

$$s_r = \frac{1}{\sqrt{n-1}}. \tag{12.35}$$

Пусть уровень существенности при проверке гипотезы H_0 равен 5% ($\alpha = 0{,}05$), тогда

а) гипотеза отклоняется, если

$$r_s < \frac{-1{,}96}{\sqrt{n-1}} \quad \text{или} \quad r_s > \frac{1{,}96}{\sqrt{n-1}};$$

б) гипотеза не отклоняется, если

$$\frac{-1{,}96}{\sqrt{n-1}} \leqslant r_s \leqslant \frac{1{,}96}{n-1}.$$

Выше мы нашли, что $r_s = 0{,}95$ Средняя квадратическая ошибка в нашем примере равна: $1 : \sqrt{14-1} = 0{,}2773$. Таким образом, нулевая гипотеза отклоняется, так как r_s чуть ли не вдвое превышает значение $\frac{1{,}96}{\sqrt{n-1}}$.

Коэффициент согласованности. В практике иногда сталкиваются со случаями, когда совокупность объектов характеризуется не двумя, а несколькими рядами рангов. Например, оценка предпочтительности (приоритета) каких-либо объектов может быть поручена группе экспертов, каждый из которых соответствующим образом ранжирует объекты. Если экспертов двое, то степень согласованности их суждений можно измерить с помощью коэффициента ранговой корреляции. Однако группа экспертов часто состоит из большего числа людей. Таким образом, возникает задача определения общей меры согласованности экспертных оценок. В качестве такого измерителя применяют *коэффициент согласованности* (*конкордации*). Поскольку каждый объект характеризуется совокупностью рангов, полученных от m экспертов, то в основу статистической меры согласованности, очевидно, может быть положена средняя сумма рангов одного объекта и отклонения от нее.

Если имеется n объектов и m экспертов, то сумма рангов у одного эксперта равна $\frac{n(n+1)}{2}$ (как сумма n членов натурального ряда), а общая сумма рангов составит $\frac{mn(n+1)}{2}$. Следовательно, сумма рангов, приписываемых одному из n объектов, в среднем равна $\frac{m(n+1)}{2}$. Возьмем экстремальный случай. Пусть все ряды рангов совпадают (полная согласованность мнений

экспертов). Отклонения сумм оценок от их средней (обозначим их через D) составят для объекта, получившего ранги, равные 1, величину $m - \frac{m(n+1)}{2} = -\frac{1}{2}m(n-1)$, ранги, равные 2,— величину $2m - \frac{m(n+1)}{2} = -\frac{1}{2}m(n-3)$ и т. д. В рассмотренном случае сумма квадратов отклонений будет максимальной: $\frac{1}{12}m^2 \times \times (n^3 - n)$.

Эта величина и взята за основу формулы коэффициента согласованности, которая имеет следующий вид:

$$W = \frac{12\sum D^2}{m^2(n^3 - n)}, \qquad (12.36)$$

где n — число объектов; m — число рядов рангов (если каждый ряд соответствует оценкам одного эксперта, то m — число экспертов); D — отклонение суммы рангов объекта от средней их суммы для всех объектов.

Если оценки всех экспертов совпадают (полная согласованность), то $W = 1$. Если же оценки экспертов полностью не совпадают, то сумма квадратов отклонений меньше, чем $\frac{1}{12}m^2(n^3 - n)$ и коэффициент $W < 1$. Наименьшее возможное его значение равно нулю.

Что же касается проверки значимости коэффициента согласованности, то при ее осуществлении прибегают к различным «обходным приемам», поскольку распределение W исследовано лишь для нескольких значений m и n. Таблица критических значений ΣD^2, при которых величина W является значимой, приведена в [9].

Пусть 5 дегустаторов следующим образом выразили предпочтения вкусовым качествам продукта, выпускаемого пятью заводами:

Таблица 124

Дегустатор	Завод					Итого
	I	II	III	IV	V	
А	1	3	4	2	5	
Б	1	2	5	3	4	
В	2	2	5	3	4	
Г	1	2	4	5	3	
Д	3	1	4	2	5	
Сумма рангов	8	9	22	15	21	75
D	−7	−6	7	0	6	—
D^2	49	36	49	0	36	170

Величины D здесь получены как отклонения суммы рангов от средней, равной 15. Найдем коэффициент согласованности:

$$W = \frac{12 \cdot 170}{5^2 \cdot (5^3 - 5)} = 0{,}68.$$

Заметим, что $\Sigma D^2 = 170$ здесь превышает критическое значение этой суммы при 5%-ном уровне значимости (112,3), так что наблюдаемая согласованность в оценках не может быть приписана игре случая.

Глава 13

Множественная регрессия

13.1. Линейная регрессия с двумя независимыми переменными

Аппарат парной регрессии, рассмотренный в предыдущей главе, позволяет изучить взаимосвязь лишь двух переменных. Если выбранная в качестве независимой переменной величина представляет соответственно доминирующий фактор, то соответствующая парная регрессия достаточно полно описывает механизм причинно-следственной связи. Чаще изменение y связано с влиянием не одного, а нескольких факторов (иногда действующих в противоположных направлениях). В этих случаях естественно стремление исследователя ввести в уравнение регрессии несколько объясняющих переменных. Такая регрессия называется *множественной*. Уравнение множественной регрессии позволяет лучше, полнее объяснить поведение зависимой переменной, чем парная регрессия, кроме того, оно дает возможность сопоставить эффективность влияния различных факторов.

Прежде чем перейти к уравнению регрессии с m независимыми переменными, рассмотрим простой случай — линейную регрессию с двумя независимыми переменными x_1 и x_2:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \varepsilon_1. \qquad (13.1)$$

Оценка параметров этого уравнения производится на основе выборочных данных; в каждом наблюдении (обозначим номер наблюдения через i) регистрируются значения трех признаков: y_i, x_{i1} и x_{i2}. Член ε_i представляет собой случайное возмущение. Предполагаем, что оно имеет нулевую среднюю и дисперсию, равную σ^2. Уравнение регрессии с оцененными параметрами запишем в виде

$$\hat{y}_i = a_0 + a_1 x_{i1} + a_2 x_{i2}.$$

Значения a_0, a_1 и a_2, очевидно, должны быть подобраны так, чтобы минимизировать сумму квадратов отклонений $\Sigma (y_i - \hat{y}_i)^2$, где $i = 1, \ldots, n$; n — число наблюдений.

Взяв частные производные от $\Sigma (y_i - \hat{y}_i)^2$ по a_0, a_1 и a_2 и приравняв их нулю, получим после несложных преобразований следующую систему нормальных уравнений.

$$\begin{aligned} \sum y_i &= n a_0 + a_1 \sum x_{i1} + a_2 \sum x_{i2}; \\ \sum y_i x_{i1} &= a_0 \sum x_{i1} + a_1 \sum x_{i1}^2 + a_2 \sum x_{i1} x_{i2}; \\ \sum y_i x_{i2} &= a_0 \sum x_{i2} + a_1 \sum x_{i1} x_{i2} + a_2 \sum x_{i2}^2. \end{aligned} \qquad (13.2)$$

Суммирование везде производится от $i = 1$ до $i = n$.

Оценку искомых параметров можно осуществлять, применяя систему (13 2). Удобнее, однако, найти готовые формулы для расчета параметров, а не решать систему каждый раз. Для этого решим первое уравнение относительно a_0:

$$a_0 = \frac{\sum y_i}{n} - a_1 \frac{\sum x_{i1}}{n} - a_2 \frac{\sum x_{i2}}{n} \tag{13.3}$$

или

$$a_0 = \bar{Y} - a_1 \bar{X}_1 - a_2 \bar{X}_2.$$

Во втором и третьем уравнениях системы заменим исходные переменные, т. е. y_i, x_{i1} и x_{i2}, на соответствующие отклонения от средних y'_i, x'_{i1} и x'_{i2}:

$$\begin{aligned} \sum y'_i x'_{i1} &= a_1 \sum x'^2_{i1} + a_2 \sum x'_{i1} x'_{i2}; \\ \sum y'_i x'_{i2} &= a_1 \sum x'_{i1} x'_{i2} + a_2 \sum x'^2_{i2}. \end{aligned} \tag{13.4}$$

Решим систему (13.4) относительно a_1 и a_2:

$$\begin{aligned} a_1 &= \frac{\left(\sum x'^2_{i2}\right)\left(\sum y'_i x'_{i1}\right) - \left(\sum x'_{i1} x'_{i2}\right)\left(\sum y'_i x'_{i2}\right)}{D}; \\ a_2 &= \frac{\left(\sum x'^2_{i1}\right)\left(\sum y'_i x'_{i2}\right) - \left(\sum x'_{i1} x'_{i2}\right)\left(\sum y'_i x'_{i1}\right)}{D}, \end{aligned} \tag{13.5}$$

где

$$D = \left(\sum x'^2_{i1}\right)\left(\sum x'^2_{i2}\right) - \left(\sum x'_{i1} x'_{i2}\right)^2. \tag{13.6}$$

Пусть переменная y находится в линейной зависимости от переменных x_1 и x_2. Для оценок параметров уравнения регрессии собраны следующие данные:

Таблица 13.1

i	y_i	x_{i1}	x_{i2}	i	y_i	x_{i1}	x_{i2}
1	10	2	1	6	10	3	4
2	12	2	2	7	14	5	7
3	17	8	10	8	12	3	3
4	13	2	4	9	16	9	10
5	15	6	8	10	18	10	11
				Итого	137	50	60

По приведенным данным получены также следующие суммы: $\sum y_i^2 = 1947$ (эта сумма для расчетов параметров не нужна, однако она понадобится в дальнейшем); $\sum x_{i1}^2 = 336$; $\sum x_{i2}^2 = 480$; $\sum y_i x_{i1} = 756$; $\sum y_i x_{i2} = 908$; $\sum x_{i1} x_{i2} = 398$.

Построим теперь систему нормальных уравнений

$$\begin{aligned} 137 &= 10a_0 + 50a_1 + 60a_2; \\ 756 &= 50a_0 + 336a_1 + 398a_2; \\ 908 &= 60a_0 + 398a_1 + 480a_2. \end{aligned}$$

Решение системы относительно неизвестных параметров дает искомые оценки: $a_0 = 9{,}3873$, $a_1 = 0{,}1285$, $a_2 = 0{,}6117$. Уравнение регрессии имеет вид:

$$\vartheta = 9{,}3873 + 0{,}1285x_1 + 0{,}6117x_2$$

Таким образом, прирост значения x_1 на одну единицу в среднем увеличивает y на 0,1285 единицы, а аналогичный прирост x_2 — на 0,6117 единицы при всех прочих равных условиях.

Для решения этого же примера с помощью выражений (13.4) и (13.5) найдем

$$\overline{Y} = \frac{137}{10} = 13{,}7; \quad \overline{X}_1 = \frac{50}{10} = 5; \quad \overline{X}_2 = \frac{60}{10} = 6;$$

$$\sum \left(x'_{i1}\right)^2 = \sum x_{i1}^2 - n\overline{X}_1^2 = 336 - 10 \cdot 5^2 = 86:$$

$$\sum \left(x'_{i2}\right)^2 = 480 - 10 \cdot 6^2 = 120;$$

$$\sum y'_i x'_{i1} = \sum y_i x_{i1} - n\overline{Y}\,\overline{X}_1 = 756 - 10 \cdot 13{,}7 \cdot 5 = 71;$$

$$\sum y'_i x'_{i2} = 908 - 10 \cdot 13{,}7 \cdot 6 = 86;$$

$$\sum x'_{i1} x'_{i2} = \sum x_{i1} x_{i2} - n\overline{X}_1 \overline{X}_2 = 398 - 10 \cdot 5 \cdot 6 = 98.$$

Откуда

$$D = 86 \cdot 120 - 98^2 = 716;$$

$$a_1 = \frac{120 \cdot 71 - 98 \cdot 86}{716} = 0{,}1285;$$

$$a_2 = \frac{86 \cdot 86 - 98 \cdot 71}{716} = 0{,}6117.$$

Наконец,

$$a_0 = 13{,}7 - 0{,}1285 \cdot 5 - 0{,}6117 \cdot 6 = 9{,}3873.$$

О степени близости расчетных и исходных значений зависимой переменной можно судить по данным, приведенным в табл. 13.2.

Таблица 13.2

y_i	ϑ_i	$e_i = y_i - \vartheta_i$	y_i	ϑ_i	$e_i = y_i - \vartheta_i$
10	10,2559	—0,2559	10	12,2195	—2,2195
12	10,8676	1,1324	14	14,3117	—0,3117
17	16,5324	0,4676	12	11,6078	0,3922
13	12,0910	0,9090	16	16,6609	—0,6609
15	15,0520	—0,0520	18	17,4012	0,5988
			Итого 137	137,0	0

13.2. Множественная линейная регрессия (общий случай)

Линейная модель с m независимыми переменными имеет вид:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_m x_{im} + \varepsilon_i. \qquad (13.7)$$

В принципе метод оценивания параметров уравнений регрессии с числом независимых переменных больше двух не отличается от метода оценивания уравнений с одной или двумя переменными.

Так, нетрудно вывести нормальные уравнения для $m=3$. Соответствующая система будет содержать уже четыре уравнения и суммы, состоящие из различных комбинаций произведений зависимых и независимых переменных. Система для $m=4$ будет включать пять уравнений и т. д. Таким образом, с увеличением числа независимых переменных становится более громоздкой запись уравнений и возрастает сложность обработки данных. Поскольку наблюдение за каждой переменной дает не одно, а ряд ее значений, то удобно записывать каждый такой ряд с помощью вектора, а совокупность значений всех независимых переменных — с помощью матрицы.

МНК в матричной записи. Введем соответствующие векторные и матричные обозначения. Пусть

$\boldsymbol{\alpha}=(\alpha_j)$, $j=0,1,\ldots,m$ — вектор неизвестных параметров;

$\mathbf{a}=(a_j)$ — вектор оценок параметров;

$\mathbf{y}=(y_i)$, $i=1,\ldots,n$ — вектор значений зависимой переменной;

$\mathbf{X}=(x_{ij})$ — матрица значений независимых переменных размерностью $n \times m$;

$\boldsymbol{\varepsilon}=(\varepsilon_i)$ — вектор ошибок;

$\mathbf{e}=(e_i)$ — вектор ошибок.

Перепишем линейную модель (13.7), используя принятые матричные обозначения:

$$\mathbf{y}=\mathbf{X}\boldsymbol{\alpha}+\boldsymbol{\varepsilon}. \tag{13.8}$$

Уравнение множественной регрессии с оцененными параметрами

$$y_i=a_0+a_1x_{i1}+a_2x_{i2}+\ldots+a_mx_{im}+e_i \tag{13.9}$$

запишем теперь в виде

$$\mathbf{y}=\mathbf{X}\mathbf{a}+\mathbf{e}.$$

Сумму квадратов отклонений получим следующим образом:

$$Q=\sum e_i^2=\mathbf{e}^T\mathbf{e}=(\mathbf{y}-\mathbf{X}\mathbf{a})^T(\mathbf{y}-\mathbf{X}\mathbf{a})=$$
$$=\mathbf{y}^T\mathbf{y}-\mathbf{a}^T\mathbf{X}^T\mathbf{y}^T-\mathbf{y}^T\mathbf{X}\mathbf{a}+\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a}.$$

Здесь и далее через T обозначена операция транспонирования вектора или матрицы. Так как $\mathbf{a}^T\mathbf{X}^T\mathbf{y}=\mathbf{y}^T\mathbf{X}\mathbf{a}$, то

$$Q=\mathbf{y}^T\mathbf{y}-2\mathbf{a}^T\mathbf{X}^T\mathbf{y}+\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a}.$$

Продифференцируем Q по a и получим

$$\frac{\partial Q}{\partial a}=-2\mathbf{X}^T\mathbf{y}+2(\mathbf{X}^T\mathbf{X})\,\mathbf{a}.$$

Приравняем данный результат нулю. После этого легко находим систему нормальных уравнений, которая в матричной форме записывается как

$$\mathbf{X}^T\mathbf{y}=\mathbf{X}^T\mathbf{X}\mathbf{a},$$

отсюда

$$\mathbf{a}=(\mathbf{X}^T\mathbf{X})^{-1}\cdot\mathbf{X}^T\mathbf{y}. \tag{13.10}$$

Оценку **a**, найденную по формуле (13.10), будем называть *оценкой метода наименьших квадратов* или *оценкой МНК*. Для определения вектора **a** необходимо по данным наблюдения найти матрицу, обратную матрице $\mathbf{X}^T\mathbf{X}$, и вектор $\mathbf{X}^T\mathbf{y}$. Матрицу $\mathbf{X}$ в развернутом виде можно записать как

$$\mathbf{X}=\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix},$$

откуда

$$\mathbf{X}^T\mathbf{X}=\begin{bmatrix} n & \sum x_{i1} & \dots & \sum x_{im} \\ \sum x_{i1} & \sum x_{i1}^2 & \dots & \sum x_{i1}x_{im} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \sum x_{im} & \sum x_{i1}x_{im} & \dots & \sum x_{im}^2 \end{bmatrix};$$

$$\mathbf{X}^T\mathbf{y}=\begin{bmatrix} \sum y_i \\ \sum y_i x_{i1} \\ \cdot \\ \cdot \\ \cdot \\ \sum y_i x_{im} \end{bmatrix}.$$

Суммирование здесь производится от $i=1$ до $i=n$, т. е. по всем наблюдениям.

Проиллюстрируем сказанное на частном примере. Пусть число независимых переменных равно 2. В этом случае

$$\mathbf{X}=\begin{bmatrix} 1 & x_{11} & x_{12} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & x_{n1} & x_{n2} \end{bmatrix}; \qquad \mathbf{y}=\begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix};$$

$$\mathbf{X}^T\mathbf{X}=\begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 \end{bmatrix}; \qquad \mathbf{X}^T\mathbf{y}=\begin{bmatrix} \sum y_i \\ \sum y_i x_{i1} \\ \sum y_i x_{i2} \end{bmatrix}.$$

Подставив элементы матриц в уравнение $\mathbf{X}^T\mathbf{X}\mathbf{a}=\mathbf{X}^T\mathbf{y}$, придем уже к знакомой нам системе нормальных уравнений (13.3).

По данным примера на с. 242 имеем

$$\mathbf{y}=\begin{bmatrix} 10 \\ 12 \\ 17 \\ \cdot \\ \cdot \\ \cdot \\ 18 \end{bmatrix}; \qquad \mathbf{X}=\begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 1 & 8 & 10 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 10 & 11 \end{bmatrix}.$$

Откуда

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 2 & 2 & 8 & \dots & 10 \\ 1 & 2 & 10 & \dots & 11 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 1 & 8 & 10 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 10 & 11 \end{bmatrix} = \begin{bmatrix} 10 & 50 & 60 \\ 50 & 336 & 398 \\ 60 & 398 & 480 \end{bmatrix}$$

Матрица, обратная к $\mathbf{X}^T\mathbf{X}$, имеет следующий вид:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0{,}40168 & -0{,}01676 & -0{,}03631 \\ -0{,}01676 & 0{,}16760 & -0{,}13687 \\ -0{,}03631 & -0{,}13687 & 0{,}12011 \end{bmatrix}$$

Наконец,

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 2 & 2 & 8 & \dots & 10 \\ 1 & 2 & 10 & \dots & 11 \end{bmatrix} \begin{bmatrix} 10 \\ 12 \\ 17 \\ \cdot \\ \cdot \\ \cdot \\ 18 \end{bmatrix} = \begin{bmatrix} 137 \\ 756 \\ 908 \end{bmatrix}.$$

Теперь находим

$$\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1} \cdot \mathbf{X}^T\mathbf{y} = \begin{bmatrix} 0{,}40168 & -0{,}01676 & -0{,}03631 \\ -0{,}01676 & 0{,}16760 & -0{,}13687 \\ -0{,}03631 & -0{,}13687 & 0{,}12011 \end{bmatrix} \begin{bmatrix} 137 \\ 756 \\ 908 \end{bmatrix} = \begin{bmatrix} 9{,}3872 \\ 0{,}1285 \\ 0{,}6117 \end{bmatrix}.$$

Сравнение коэффициентов регрессии. В § 13.1 отмечалось, что множественная регрессия позволяет соизмерить влияние различных воздействующих на явление факторов. В самом деле, возьмем регрессию $\hat{y} = 9{,}3873 + 0{,}1285x_1 + 0{,}6117x_2$. Если бы переменные x_1 и x_2 измерялись в одних и тех же единицах, то, непосредственно сопоставляя a_1 и a_2, можно было бы сразу сказать, что x_2 сильнее воздействует на y, чем x_1 (примерно в 4,5 раза). В общем случае, когда x_1 и x_2 представляют собой факторы, имеющие различные единицы измерения, скажем, рубли, тонны или метры, такое простое сопоставление лишено смысла. Поэтому если необходимо сравнить влияние факторов и установить относительную важность каждого из них (в самом простом случае — ранжировать их), то прибегают к нормированию коэффициентов регрессии. Нормирование осуществляют следующим образом:

$$\beta_j = a_j\left(\frac{s_{x_j}}{s_y}\right), \tag{13.11}$$

где β_j — коэффициент регрессии после нормирования (бэта-коэффициент); s_{x_j} — средняя квадратическая ошибка переменной x_j; s_y — средняя квадратическая ошибка y.

Из сказанного ясно, что нормирование коэффициентов регрессии возможно лишь при условии, что x_j — случайные переменные. В самом деле, если бы x были бы не случайными, а детерминированными переменными, то величина s_{x_j} во многом определялась бы экспериментатором.

По данным нашего примера получим $s_{x_1} = 3{,}09$, $s_{x_2} = 3{,}65$ и $s_y = 2{,}79$. Откуда $\beta_1 = 0{,}1285 \cdot \frac{3{,}09}{2{,}79} = 0{,}1423$ и $\beta_2 = 0{,}6117 \cdot \frac{3{,}65}{2{,}79} = 0{,}8002$. Таким образом, воздействие переменной x_2 более чем в 5,5 раз выше, чем x_1.

Коэффициент множественной корреляции. Теснота линейной взаимосвязи переменной y с рядом переменных x_i, рассматриваемых в целом, измеряется с помощью коэффициента множественной корреляции R, который представляет собой обобщение парного коэффициента корреляции. Расчет обычно ведется по формуле

$$R = \sqrt{1 - \frac{\sum e_i^2}{\sum (y_i - \bar{Y})^2}}. \qquad (13.12)$$

Чем теснее данные примыкают к линии регрессии, тем выше значение этого показателя. Таким образом, коэффициент множественной корреляции, как и коэффициент парной корреляции, может служить измерителем качества подбора уравнения регрессии. Чем ближе R к ± 1, тем лучше подобрана регрессия.

Квадрат величины R представляет собой *коэффициент множественной детерминации*. Этот показатель также является мерой качества подбора уравнения. Он характеризует долю «объясненной» с помощью регрессии дисперсии в общей дисперсии зависимой переменной.

По данным сквозного примера необходимые для расчета R суммы квадратов отклонений составят:

$$\sum (y_i - \bar{Y})^2 = \sum y_i^2 - n\bar{Y}^2 = 1947 - 10 \cdot 13{,}7^2 = 70{,}1; \quad \sum e_i^2 = 8{,}4,$$

откуда

$$R = \sqrt{1 - \frac{8{,}4}{70{,}1}} = \sqrt{0{,}8806} = 0{,}938.$$

Следовательно, регрессия y на x_1 и x_2 объясняет почти 88% ($R^2 = 0{,}938^2 = 0{,}88$) колебаний значений y

При большом числе наблюдений и переменных можно воспользоваться другой формулой для определения R, соответственно которой нет необходимости подсчитывать e_i и Σe^2. В самом деле, поскольку

$$\sum e_i^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{Xa})^T(\mathbf{y} - \mathbf{Xa}) = \mathbf{y}^T\mathbf{y} - 2\mathbf{a}^T\mathbf{X}\mathbf{y} + \mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{y}^T\mathbf{y} - \mathbf{a}^T\mathbf{X}^T\mathbf{y}$$

и

$$\sum (y_i - \bar{Y})^2 - \sum e_i^2 = \sum y_i^2 - n\bar{Y}^2 - \sum e_i^2 = \mathbf{y}^T\mathbf{y} - (\mathbf{y}^T\mathbf{y} - \mathbf{a}^T\mathbf{X}^T\mathbf{y}) - n\bar{Y}^2 = \mathbf{a}^T\mathbf{X}^T\mathbf{y} - n\bar{Y}^2,$$

$$R = \sqrt{\frac{\mathbf{a}^T\mathbf{X}^T\mathbf{y} - n\bar{Y}^2}{\sum y_i^2 - n\bar{Y}^2}} = \sqrt{\frac{\mathbf{a}^T\mathbf{X}^T\mathbf{y} - n\bar{Y}^2}{\mathbf{y}^T\mathbf{y} - n\bar{Y}^2}}. \qquad (13.13)$$

Коэффициент частной корреляции. Если исследуемая модель является регрессионной (напомним, что в этом случае

x — неслучайная переменная), то, как было показано выше, коэффициент корреляции характеризует качество подбора уравнения регрессии. Если же x — случайная переменная, то коэффициент корреляции является мерой тесноты линейной взаимосвязи между y и набором переменных x. Иногда представляет интерес измерение частных зависимостей (между y и x_j) при условии, что воздействие остальных охваченных исследованием факторов элиминировано. В качестве соответствующих измерителей приняты *коэффициенты частной корреляции.*

Рассмотрим коэффициенты частной корреляции для случая, когда во взаимосвязи находятся три случайные переменные, например y, x и z (переход к большему их числу принципиально ничего не изменяет). Для этих переменных можно получить простые коэффициенты парной корреляции r_{yx}, r_{yz} и r_{xz} (индексы указывают на коррелируемые переменные). Остановимся на одном из этих коэффициентов, например r_{yx}. Он характеризует степень линейной взаимосвязи между y и x. Однако большая величина этого коэффициента может возникнуть не только потому, что y и x действительно связаны между собой, но и в силу того, что оба эти признака испытывают влияние третьего фактора — z. Коэффициент частной корреляции отличается от простого коэффициента парной корреляции тем, что он измеряет парную корреляцию соответствующих признаков (допустим, y и x) при условии, что влияние на них третьего фактора (z) устранено.

Вывод формул для коэффициентов частной корреляции основывается на соответствующем расчленении общей дисперсии. В итоге получены следующие выражения в случае, когда имеются три переменные:

$$r_{yx\cdot z} = \frac{r_{yx} - r_{yz}\cdot r_{xz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}; \qquad (13.14)$$

$$r_{yz\cdot x} = \frac{r_{yz} - r_{yx}\cdot r_{xz}}{\sqrt{(1-r_{xz}^2)(1-r_{yx}^2)}}. \qquad (13.15)$$

Здесь $r_{yx\cdot z}$ — коэффициент частной корреляции y и x; влияние z элиминировано, $r_{yz\cdot x}$ — коэффициент частной корреляции y и z при устранении влияния x. Поскольку все три переменные равноправны, то можно найти и третий коэффициент

$$r_{xz\cdot y} = \frac{r_{xz} - r_{xy}\cdot r_{yz}}{\sqrt{(1-r_{yz}^2)(1-r_{xy}^2)}}. \qquad (13.16)$$

Из (13.14) видно, что если $r_{yz} = 0$, то $|r_{yx\cdot z}| > |r_{yx}|$, так как в этом случае $r_{yx\cdot z} = r_{yx} : \sqrt{(1-r_{xz}^2)}$. То же можно сказать и о случае, когда $r_{xz} = 0$. Если же, наоборот, один из коэффициентов r_{xz} или r_{yz} приближается к единице, то соответственно частный коэффициент корреляции стремится к нулю.

По данным сквозного примера получены следующие коэффициенты парной корреляции (пусть $x = x_1$ и $z = x_2$): $r_{yx} = 0{,}9144$; $r_{yz} = 0{,}9377$ и $r_{xz} = 0{,}9647$.

Отсюда коэффициенты частной корреляции составят:

$$r_{yx\cdot z}=\frac{0{,}9144-0{,}9377\cdot 0{,}9647}{\sqrt{(1-0{,}9647^2)\,(1-0{,}9377^2)}}=0{,}107;$$

$$r_{yz\cdot x}=\frac{0{,}9377-0{,}9144\cdot 0{,}9647}{\sqrt{(1-0{,}9144^2)\,(1-0{,}9647^2)}}=0{,}520;$$

$$r_{xz\cdot y}=\frac{0{,}9647-0{,}9144\cdot 0{,}9377}{\sqrt{(1-0{,}9377^2)\,(1-0{,}9144^2)}}=0{,}763.$$

13.3. Доверительные интервалы множественной регрессии

Определение доверительных интервалов регрессий и получаемых по ним прогностических оценок значений зависимой переменной базируется на целом ряде предположений, лежащих в основе МНК. Кратко охарактеризуем их.

Основные предположения МНК и свойства получаемых оценок параметров. При применении МНК относительно ошибок ε_i в модели (13.1) принимают следующие предположения, которые являются аналогами предположений, сделанных выше для МНК, применяемого при оценивании параметров парной регрессии.

1. *Для каждого наблюдения* i ε_i — *случайная нормально распределенная величина.* Предположение о том, что ошибки ε_i нормально распределены, не является необходимым для оценки коэффициентов регрессии, однако оно необходимо как для получения доверительных интервалов отдельных параметров коэффициентов регрессии и уравнения регрессии в целом, так и при проверке гипотез о значимости (существенности) коэффициентов регрессии. Сделанное замечание относится не только к множественной, но и к парной регрессии.

2. $M(\varepsilon_i)=0$ — математическое ожидание ошибки равно нулю.

3. $\sigma_\varepsilon^2=\text{const}$ — ошибки ε_i имеют одинаковую дисперсию.

4. ε_{i1} и ε_{i2} не коррелируют для $i_1\neq i_2$. Используем здесь простейшую характеристику связи между случайными величинами — ковариацию (напомним, под ковариацией x и y понимается математическое ожидание произведений отклонений x и y от их средних). Для случайных ошибок ε_{i1} и ε_{i2} имеем

$$\operatorname{cov}(\varepsilon_{i1},\varepsilon_{i2})=M(\varepsilon_{i1}\cdot\varepsilon_{i2})=0.$$

Кроме того, примем предположение о свойстве независимых переменных. Матрица $\mathbf{X}$ состоит из линейно-независимых векторов-столбцов, т. е. между векторами $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_m$ нет линейных зависимостей. Последнее обстоятельство эквивалентно тому, что ранг матрицы $\mathbf{X}$ равен m, а это в свою очередь означает, что $|\mathbf{X}^T\mathbf{X}|\neq 0$, т. е. матрица $\mathbf{X}^T\mathbf{X}$ обратима. Матрица $\mathbf{X}$ не содержит ошибок.

Если сделанные предположения отвечают действительности, то получаемые оценки МНК являются *несмещенными, состоятельными и эффективными* *.

* Доказательства свойств оценок МНК приводятся в работах [4], [5], [6].

Дисперсия оценок параметров линейной множественной регрессии. Оценки параметров множественной регрессии, являясь выборочными характеристиками, естественно, будут отклоняться от истинных их значений. Для того чтобы измерить дисперсии оценок параметров, найдем матрицу ковариаций для **a**. По определению имеем

$$\operatorname{cov}(\mathbf{a}) = M[(\mathbf{a}-\alpha)(\mathbf{a}-\alpha)^T].$$

Уже из этого выражения следует, что элементами главной диагонали матрицы ковариаций являются дисперсии параметров a_j. Поэтому рассмотрим ее более подробно. Прежде всего определим, чему равно выражение $(\mathbf{a}-\alpha)\cdot(\mathbf{a}-\alpha)^T$. Из (13.8) и (13.10) следует, что

$$\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\alpha+\varepsilon) = \alpha + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon,$$

откуда

$$\mathbf{a}-\alpha = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon.$$

Использовав найденное выражение для $(\mathbf{a}-\alpha)$, получим

$$\operatorname{cov}(\mathbf{a}) = M\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\cdot\varepsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] = M(\varepsilon\varepsilon^T)\cdot(\mathbf{X}^T\mathbf{X})^{-1}. \quad (13.17)$$

Полученное выражение содержит матрицу $M(\varepsilon\varepsilon^T)$. В ней все элементы, не лежащие на главной диагонали, равны нулю (в силу того, что ошибки не коррелируют между собой; см. предположение 4). Поскольку все ошибки имеют одинаковую дисперсию (см. предположение 3), то

$$M(\varepsilon\varepsilon^T) = \sigma^2\cdot\mathbf{I}, \quad (13.18)$$

где **I** — единичная матрица.

Подставив найденное значение $M(\varepsilon\varepsilon^T)$ в (13.17), получим

$$\operatorname{cov}(\mathbf{a}) = \sigma^2\cdot(\mathbf{X}^T\mathbf{X})^{-1}. \quad (13.19)$$

Матрица ковариаций (13.19) не может быть точно определена, поскольку в нее помимо обратной матрицы $(\mathbf{X}^T\mathbf{X})^{-1}$ входит в качестве множителя дисперсия ошибок σ^2, значение которой нам неизвестно. В качестве статистической оценки σ^2 можно воспользоваться наблюденной дисперсией ошибок Примем, что $\mathbf{e} = (e_i) = \mathbf{y} - \hat{\mathbf{y}}$, тогда в качестве оценки σ^2 найдем

$$s^2 = \frac{Q}{n-m-1} = \frac{\mathbf{e}^T\mathbf{e}}{n-m-1} = \frac{\sum e_i^2}{n-m-1}. \quad (13.20)$$

Знаменатель этой формулы представляет собой число степеней свободы Последнее равно числу наблюдений за вычетом числа оцениваемых параметров. Как известно, полученная таким образом оценка σ_2 будет несмещенной и состоятельной Теперь на основе (13.20) определим значения дисперсий оценок a_j, взяв вместо σ^2 оценку этой величины s^2. Получим

$$s_a^2 = s^2 b_{jj}, \quad (13.21)$$

где b_{jj} — диагональный элемент матрицы $(\mathbf{X}^T\mathbf{X})^{-1}$.

В нашем примере диагональные элементы матрицы $(X'X)^{-1}$ равны 0,4017; 0,1676 и 0,1201. По данным этого же примера находим $\sum e_i^2 = 8{,}3678$, откуда

$$\frac{\sum e_i^2}{n-m-1} = \frac{8{,}3678}{10-2-1} = 1{,}1954;$$

$$s_{a_0}^2 = 1{,}1954 \cdot 0{,}4017 = 0{,}4802; \quad s_a = 0{,}6930;$$

$$s_{a_1}^2 = 1{,}1954 \cdot 0{,}1676 = 0{,}2003; \quad s_{a_1} = 0{,}4476;$$

$$s_{a_2}^2 = 1{,}1954 \cdot 0{,}1201 = 0{,}1436, \quad s_{a_2} = 0{,}3789.$$

Полученные квадратические ошибки могут быть использованы для определения доверительных интервалов параметров и для проверки значимости отличия a_j от нуля.

В нашем примере уже простое сопоставление оценки a_1 с ее квадратической ошибкой ($a_1 = 0{,}1285$ и $s_{a_1} = 0{,}4476$) говорит о том, что этот параметр оказывается незначимым в статистическом смысле Проверка параметра a_2 дает

$$t^* = \frac{a_2 - 0}{s_{a_2}} = \frac{0{,}6174}{0{,}3789} = 1{,}442.$$

Табличное значение t при 95% ной доверительной вероятности и двусторонней проверке равно 2,365 следовательно, нет оснований отклонять нулевую гипотезу Значимость данного коэффициента регрессии здесь можно признать, лишь приняв 80%-ную доверительную вероятность В этом случае критическое значение t-статистики будет равно 1,415 Подобный результат обусловлен тем, что выборка охватила небольшое число наблюдений В практических разработках при определении регрессии на два независимых признака обычно требуется не менее $15 \div 17$ наблюдений (при умеренной колеблемости значений переменных).

Обычно если проверка приводит к тому, что параметр оказывается незначимым (несущественно отличается от нуля), то он исключается из регрессии и оценивание параметров повторяется уже для нового набора независимых переменных.

Доверительные интервалы регрессии и прогноза. Пусть значение y определяется по уравнению регрессии с оцененными параметрами:

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m.$$

В силу того, что a_j — несмещенные оценки некоторых неизвестных параметров, то $\hat{y}$ — лишь одно из возможных значений прогнозируемой величины для заданных значений x, точнее, это оценка среднего значения $\hat{y}$ в этих условиях. Поскольку a_j — случайная величина, то и оценка $\hat{y}$ также случайна и имеет дисперсию. Определим ее значение. Обозначим операцию определения дисперсии символом D, тогда

$$\mathrm{D}(\hat{y}) = \mathrm{D}(a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m).$$

Использовав теорему о дисперсии суммы зависимых величин и воспользовавшись матричной записью, перепишем полученное выражение следующим компактным способом:

$$\mathrm{D}(\hat{y}) = \mathbf{x}_p^T \operatorname{cov}(\mathbf{a})\, \mathbf{x}_p,$$

где $\mathbf{x}_p = (1 x_{p1}\ x_{p2}\ \ldots\ x_{pm})$ — вектор заданных значений независимых переменных; $\mathrm{cov}(\mathbf{a})$ — матрица ковариаций оценок a. Откуда, использовав выражение (13.19), получим

$$\mathrm{D}(\hat{y}) = \sigma^2 \mathbf{x}_p^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_p.$$

Поскольку значение σ нам неизвестно, то воспользуемся оценкой s^2, тогда

$$\mathrm{D}(\hat{y}) = s_{\hat{y}}^2 = s^2 \mathbf{x}_p^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_p. \qquad (13.22)$$

Таким образом, «истинное» среднее значение $\hat{y}$ лежит в пределах

$$\hat{y} \pm t_\alpha s \sqrt{\mathbf{x}_p^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_p}. \qquad (13.23)$$

Допустим, что прогноз определяется по уравнению $\hat{y} = 9{,}3873 + 0{,}1285x_1 + +0{,}6117x_2$, в котором, несмотря на то, что параметр a_1 оказался несущественным, мы сохранили его (так обычно поступают, когда важно сохранить соответствующий член регрессии исходя из каких-либо содержательных предпосылок). Требуется найти доверительные границы для значения $\hat{y}$, соответствующего $x_{p1} = 10$ и $x_{p2} = 10$. Тогда $\mathbf{x}_p^T = [1\ 10\ 10]$ и $\hat{y} = 16{,}789$ Для определения $s_{\hat{y}}^2$ нам прежде всего необходима матрица $(\mathbf{X}^T\mathbf{X})^{-1}$. Эта матрица была уже определена при оценке параметров регрессии (см. с 246). В соответствии с (13.22) имеем

$$s_{\hat{y}}^2 = 1{,}1954\,[1\ \ 10\ \ 10] \begin{bmatrix} 0{,}4017 & -0{,}0168 & -0{,}0363 \\ -0{,}0168 & 0{,}1676 & -0{,}1368 \\ -0{,}0363 & -0{,}1368 & 0{,}1201 \end{bmatrix} \begin{bmatrix} 1 \\ 10 \\ 10 \end{bmatrix} \approx 0{,}9201.$$

Найдем теперь 95%-ный доверительный интервал. Число степеней свободы здесь равно: $n - m - 1 = 10 - 2 - 1 = 7$, $t_\alpha = 2{,}365$. Откуда

$$\hat{y} \pm t_\alpha s_{\hat{y}} = 16{,}789 \pm 2{,}365 \sqrt{0{,}9201} \approx 16{,}79 \pm 2{,}27.$$

Под прогнозным значением y можно понимать его математическое ожидание, т. е. $\mathbf{Xa}$. Доверительный интервал в этом случае соответствует доверительному интервалу регрессии (13.23). Однако более естественно в прогнозное значение y включить и возможное отклонение от регрессии. В этом случае к дисперсии $\hat{y}$ необходимо добавить и дисперсию e, т. е. s^2. Таким образом, дисперсия прогнозного значения y составит:

$$s_p^2 = s^2 \left[1 + \mathbf{x}_p^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_p\right]. \qquad (13.24)$$

Доверительные интервалы для индивидуальных прогнозных значений y равны $\hat{y} \pm t_\alpha s_p$.

В нашем примере $s_p^2 = 1{,}1954 + 0{,}9201 = 2{,}1155$ и интервал прогноза равен: $16{,}789 \pm 2{,}365 \sqrt{2{,}1155} \approx 16{,}79 \pm 3{,}44$.

13.4. Нелинейная регрессия

Все, что выше было сказано о регрессиях и МНК, относилось к линейным моделям, т. е. к таким уравнениям, в которых переменные имели первую степень (модель, линейная по переменным), а параметры выступали в виде коэффициентов при этих

переменных (модель, линейная по параметрам). Именно такие регрессии используются наиболее широко в силу своей простоты, относительно малой трудоемкости их получения и, наконец, в силу достаточно глубокой их изученности. В практике, однако, нельзя ограничиться моделями, линейными по переменным и параметрам. Сравнительно часто линейная модель неадекватно отображает исследуемое явление и наилучшее приближение дают уравнения, нелинейные как по переменным, так и по параметрам.

Оценивание параметров регрессий, нелинейных по переменным. Проблема оценивания параметров решается в этом случае достаточно просто. Независимые переменные, имеющие степень, отличающуюся от первой, заменяются другими независимыми переменными в первой степени, и к новой системе переменных применяется обычный МНК. После того как получено уравнение с оцененными параметрами, введенные в него новые независимые переменные заменяются на первоначальные. Так, пусть необходимо рассчитать параметры уравнения

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_2^2 + e.$$

Введем переменные $z = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_2^2$, после чего запишем

$$y = a_0 + a_1z_1 + a_2z_2 + a_3z_3 + a_4z_4 + e.$$

В уравнении (13.10) заменим матрицу $(\mathbf{X}^T\mathbf{X})^{-1}$ на $(\mathbf{Z}^T\mathbf{Z})^{-1}$ и вектор $\mathbf{X}^T\mathbf{y}$ на $\mathbf{Z}^T\mathbf{y}$, получим

$$\mathbf{a} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}.$$

Для полученной регрессии можно найти ошибки параметров и доверительные интервалы.

Пусть зарегистрированы следующие данные, характеризующие интенсивность орошения (x) и урожайность (y) некоторой зерновой культуры:

Таблица 13.3

y, ц/га	6	7	13	16	20	24	22	20
x, дм	0,9	1,0	1,8	2,4	4,0	5,8	7,6	8,5

Подберем к этим данным параболу $\hat{y} = a_0 + a_1x + a_2x^2 = a_0 + a_1z + a_2z_2$. Для оценки параметров воспользуемся формулами (13.4) и (13.6). По приведенным данным получим:

$$\bar{Y} = 16;\ \bar{Z}_1 = 4{,}0;\ \bar{Z}_2 = 23{,}81;\ \sum z_1'^2 = 62{,}46;\ \sum z_2'^2 = 5453{,}9;$$
$$\sum y'z_1' = 118{,}6;\ \sum y'z_2' = 943;\ \sum z_1'z_2' = 572.$$

Откуда

$$D = 62{,}46 \cdot 5453{,}9 - 572^2 = 13466{,}6;$$
$$a_1 = \frac{5453{,}9 \cdot 118{,}6 - 572 \cdot 943}{13466{,}6} = 7{,}7231;$$
$$a_2 = \frac{62{,}46 \cdot 943 - 572 \cdot 118{,}6}{13466{,}6} = -0{,}6638;$$
$$a_0 = 16 - 7{,}7231 \cdot 4 + 0{,}6638 \cdot 23{,}81 = 0{,}9127.$$

Таким образом

$$\hat{y} = 0{,}9127 + 7{,}7231x - 0{,}6638x^2.$$

Соответствующая кривая показана на рис. 13.1 (кривая 1)

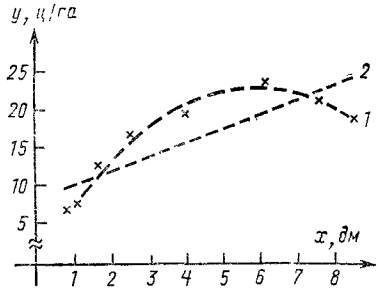


Рис. 13.1.
Линейная и параболическая регрессия

Индекс корреляции. Измерение тесноты связи в случае нелинейной регрессии осуществляется с помощью индекса корреляции или корреляционного отношения, смысл которого был раскрыт в гл. 12 на примере линейной регрессии. В принципе ничего не изменится, если регрессия будет нелинейной. В этом случае в формуле

$$I = \sqrt{1 - \frac{\sum e_i^2}{\sum (y_i - \bar{Y})^2}}, \qquad (13.25)$$

где I — индекс корреляции, величины e_i будут означать отклонение фактических значений переменной y_i от расчетных, полученных по уравнению нелинейной регрессии.

По данным предыдущего примера $\sum e_i^2 = 7{,}25$, $\sum (y_i - \bar{Y})^2 = 322$, отсюда

$$I = \sqrt{1 - \frac{7{,}25}{322}} = 0{,}989.$$

Если бы, например, в качестве первого и самого грубого приближения была применена линейная парная регрессия, то

$$\hat{y}_i = 8{,}405 + 1{,}899x_i; \quad \sum e_i^2 = 96{,}7 \text{ и } r = 0{,}83.$$

Введение дополнительного члена в регрессию, т. е. переход от линейной к параболической регрессии $\hat{y}_i = 0{,}9127 + 7{,}7231x_i - 0{,}6638x_i^2$, повышает качество описания исследуемой зависимости, как это и видно на рис 13.1. При этом резко снижается значение «необъясненной» суммы квадратов отклонений и вырастает измеритель тесноты связи.

Дальнейшее усложнение формы кривой, например введение членов в третьей и более степени, практически не увеличит меру тесноты связи, так как «объясняющая способность» уравнения регрессии, по существу, не изменится.

Оценивание параметров регрессий, нелинейных по параметрам. Возьмем, например, такие функции, как

$$f(a_1, a_2) = a_1 x^{a_2};$$

$$f(a_1, a_2) = a_1 K^{a_2} L^{1-a_2};$$

$$f(a_1, a_2, a_3) = a_1 + a_2 e^{a_3 t};$$

$$f(a_1, a_2, a_3) = \frac{a_1}{1 + a_2 e^{a_3 t}} \text{ и т. д.}$$

Все эти функции нелинейны по параметрам. Непосредственное применение МНК для их оценивания невозможно. Первое и второе уравнения можно оценить, предварительно приведя их с помощью логарифмирования к линейному виду. Однако такое преобразование приводит к тому, что оценка параметров базируется не на минимизации суммы квадратов отклонений, а на минимизации суммы квадратов отклонений в логарифмах. Причем

$$Q = \sum (\log y_i - \log \hat{y}_i)^2 = \sum \log^2 \frac{y_i}{\hat{y}_i} \neq \sum (y - \hat{y}_i)^2 = \sum e_i^2.$$

Следствием этого является некоторое смещение оценок параметров, получаемых обычным (линейным) МНК. В общем случае оценивание нелинейных по параметрам уравнений достигается с помощью так называемого *нелинейного метода наименьших квадратов (НМНК)*. Соответственно НМНК параметры уравнений подбираются так, чтобы максимально приблизить кривую $f(a)$ к полученным из наблюдения значениям зависимой переменной y. Таким образом, здесь, как и в линейном МНК, минимизируется сумма квадратов отклонений:

$$Q = \sum e_i^2 = \sum [y_i - f(a)]^2.$$

В результате дифференцирования Q по параметрам a получают систему уравнений, которые, однако, являются нелинейными. Решение этой системы в общем случае достаточно сложно и обычно не проще непосредственной минимизации Q. Последняя достигается с помощью ряда итеративных процедур, применяемых при минимизации функции многих переменных. Основными методами оценивания параметров, базирующихся на таких процедурах, являются метод *линеаризации* (разложение функции в ряд Тейлора и использование МНК) и метод наискорейшего спуска (использование градиентов для итеративной минимизации функции). Поскольку каждый из этих методов обладает рядом недостатков, был создан компромиссный метод (метод Марквардта), сочетающий в себе положительные черты метода линеаризации и градиентного метода [4], [5].

Мультипликативная регрессия. В последнее десятилетие в экономическом анализе получила применение нелинейная регрессия, которая в самом простом варианте имеет вид

$$y = a_1 x_1^{a_2} x_2^{a_3}. \qquad (13.26)$$

Подобного рода функции обычно применяются для измерения влияния на какой-либо результативный признак, например объем производства, таких факторов, как численность занятых (x_1) и размер основных фондов (x_2). Существует множество модификаций функции (13.26). Одна из них заключается во введении в нее члена, характеризующего влияние времени, а тем самым и всех в явном виде не учтенных факторов, которые можно связать с временем. В этом случае

$$y = a_1 x_1^{a_2} x_2^{a} e^{a_4 t}, \tag{13.27}$$

где e — основание натурального логарифма. Мультипликативные функции с численно оцененными параметрами позволяют найти взаимосвязь между темпами прироста результативного признака и темпами приростов факторов. В самом деле, допустим, что функция (13.26) характеризует взаимосвязь показателей y, x_1 и x_2 во времени, т. е.

$$y_t = a_1 x_{1t}^{a_2} x_{2t}^{a_3}, \tag{13.28}$$

Приросты переменных можно, очевидно, охарактеризовать с помощью соответствующих производных, т. е.

$$\frac{\partial y_t}{\partial t}; \quad \frac{\partial x_{1t}}{\partial t}; \quad \frac{\partial x_{2t}}{\partial t},$$

а темпы приростов как

$$\tau_y = \frac{\partial y_t}{\partial t} \cdot \frac{1}{y_t}; \quad \tau_{x_1} = \frac{\partial x_{1t}}{\partial t} \cdot \frac{1}{x_{1t}}; \quad \tau_{x_2} = \frac{\partial x_{2t}}{\partial t} \cdot \frac{1}{x_{2t}}.$$

Продифференцируем (13.28) по времени:

$$\frac{\partial y_t}{\partial t} = \frac{\partial y_t}{\partial x_{1t}} \cdot \frac{\partial x_{1t}}{\partial t} + \frac{\partial y_t}{\partial x_{2t}} \cdot \frac{\partial x_{2t}}{\partial t}. \tag{13.29}$$

Частные производные, необходимые для определения $\frac{\partial y_t}{\partial t}$, равны:

$$\frac{\partial y_t}{\partial x_{1t}} = a_1 a_2 x_{1t}^{a_2 - 1} x_{2t}^{a_3};$$

$$\frac{\partial y_t}{\partial x_{2t}} = a_1 a_3 x_{1t}^{a_2} x_{2t}^{a_3 - 1}.$$

Подставив полученные выражения в (13.29) и разделив результат на y_t, находим после сокращений

$$\frac{\partial y_t}{\partial t} \cdot \frac{1}{y_t} = a_2 \frac{\partial x_{1t}}{\partial t} \cdot \frac{1}{x_{1t}} + a_3 \frac{\partial x_{2t}}{\partial t} \cdot \frac{1}{x_{2t}}$$

или

$$\tau_y = a_2 \tau_{x_1} + a_3 \tau_{x_2}.$$

Таким образом, темп прироста результативного признака равен взвешенной сумме темпов прироста определяющих его факторов. В качестве весов здесь выступают параметры соответствующих независимых переменных. Если повторить только что проде-

ланную процедуру для функции (13.27), то в результате получим

$$\tau_y = a_2\tau_{x_1} + a_3\tau_{x_2} + a_4 t.$$

Это выражение характеризует и вклад всех прочих (помимо x_1 и x_2) факторов в темп прироста y_t.

Что касается оценивания параметров рассмотренных мультипликативных регрессий, то необходимо заметить, что часто их оценивают, приводя функции к линейному виду с помощью логарифмирования. Как уже было отмечено выше, такое преобразование иногда приводит к существенному смещению оценок.

13.5. Автокорреляция остатков

При рассмотрении основных предположений, положенных в основу МНК, мы приняли, что ошибки ε_t представляют собой случайные независимые (некоррелируемые) величины со средней, равной нулю. При работе с фактическими данными, особенно с динамическими рядами, такое допущение далеко не всегда имеет под собой реальную почву. В самом деле, если вид функции (закон развития) выбран неудачно, то вряд ли можно говорить о том, что отклонения от регрессии являются независимыми. В этом случае обычно наблюдается заметная концентрация положительных и отрицательных отклонений от регрессии и можно сомневаться в их случайном характере. Если последовательные значения e_t коррелируют между собой, то говорят, что имеет место *автокорреляция ошибок*. Метод наименьших квадратов и в случае автокорреляции дает несмещенные и состоятельные оценки. Однако получаемая при наличии высокой автокорреляции стандартная ошибка и соответственно доверительный интервал имеют мало смысла в силу своей ненадежности. Во всяком случае значительная автокорреляция говорит о том, что спецификация регрессии неправильна. При анализе соответствующей регрессии необходимо определить, присутствует или нет автокорреляция в остаточных членах e_t, особенно если исследуется временно́й ряд.

Существует ряд приемов обнаружения автокорреляции. Наиболее простым и достаточно обоснованным из них является метод Дарбина — Уотсона. Соответствующий критерий связан с гипотезой о существовании автокорреляции первого порядка, т. е. автокорреляции между соседними остаточными членами ряда. Соответствующий критерий имеет вид:

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}.$$

Здесь для того чтобы подчеркнуть, что мы имеем дело с временными рядами, вместо i введен индекс t, означающий номер

члена в хронологическом порядке. Предполагая, что $\sum_{t=2}^{n} e_t^2 \approx$ $\approx \sum_{t=2}^{n} e_{t-1}^2$, так как при высоком значении n эти две суммы мало отличаются друг от друга, получим

$$d \approx \frac{2\sum_{t=2}^{n} e_t^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} = 2\left(1 - \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}\right).$$

Вычитаемая из единицы дробь равна нулю, если автокорреляция отсутствует. Если же наблюдается полная автокорреляция, то эта дробь равна 1 или —1. Отсюда следует, что при отсутствии автокорреляции (при случайных отклонениях от регрессии) $d \approx 2$, а при полной положительной автокорреляции $d \approx 0$, при отрицательной $d \approx 4$.

Для d-статистики найдены критические границы (d_u — верхняя и d_l — нижняя), позволяющие принять или отклонить гипотезу об отсутствии автокорреляции при 1; 2,5 и 5%-ном уровне существенности. Сокращенная таблица значений этого критерия приведена в [22] для n от 15 до 100. Если эмпирическое значение d находится в пределах от d_u до $(4 - d_u)$, то гипотеза об отсутствии автокорреляции не отклоняется, если же эта статистика находится в пределах от d_l до d_u или между $(4 - d_u)$ и $(4 - d_l)$, то нет статистических оснований ни принять, ни отклонить эту гипотезу (области неопределенности). Наконец, если вычисленное значение $d < d_l$, то это указывает на положительную автокорреляцию, а если $d > (4 - d_l)$, то наблюдается отрицательная автокорреляция.

Вероятно, в настоящее время следует признать обязательным требование, согласно которому уравнение регрессии с численно оцененными параметрами должно сопровождаться расчетным значением d-статистики. Если с помощью критерия Дарбина — Уотсона обнаружена существенная автокорреляция отклонений от регрессии, то логично признать наличие ошибки в спецификации уравнения. Следовательно, надо вернуться к этой проблеме, пересмотреть набор включаемых в уравнение переменных и уточнить форму уравнения.

Определим значение d по данным примера на с 242. Выше было найдено, что $\sum e_t^2 = 8{,}3678$. Сумма $\sum (e_t - e_{t-1})^2$ составит для этих же данных величину 25,0169, откуда

$$d = \frac{25{,}0169}{8{,}3678} = 1{,}79.$$

Ближайшее табличное значение d определено для $n = 15$ ($d_l = 0{,}95$, $d_u = 1{,}54$), поэтому прямое сопоставление с табличными данными невозможно. Однако, учитывая верхнюю критическую границу (для $n = 15$) и тот факт, что расчетное значение оказалось близким к 2, можно полагать, что наличие автокорреляции остатков не подтверждается

13.6. Проверка уравнения регрессии (дисперсионный анализ)

В гл. 11 дисперсионный анализ рассматривался как самостоятельный инструмент статистического анализа. Вместе с тем он применим и как вспомогательное средство в процессе корреляционного и регрессионного анализа. Прежде всего отметим, что с помощью дисперсионного анализа можно проверить гипотезу о наличии связи между двумя переменными (без постановки вопроса о том, какую форму имеет эта связь). Далее последовательно можно испытать пригодность линейной регрессии и регрессии более высоких степеней. Если испытывается регрессия на несколько независимых переменных, то дисперсионный анализ позволяет проверить существенность вклада дополнительных переменных.

Наличие зависимости между переменными проверяется с помощью простой схемы дисперсионного анализа (полностью случайный отбор) при условии, что для каждого уровня фактора имеется несколько наблюдений зависимой переменной. В качестве фактора здесь выступает независимая переменная. Если систематическая дисперсия существенно выше случайной, то гипотеза об отсутствии связи отклоняется.

Проверка гипотезы линейной связи состоит в расчленении суммы квадратов отклонений на две составляющие:

$$\sum (y_i - \bar{Y})^2 = \sum (\hat{y} - \bar{Y})^2 + \sum e_i^2,$$

где $e_i = y_i - \hat{y}$. Первое слагаемое в этой сумме характеризует сумму квадратов, определяемую изменением по линейной регрессии (кратко — сумма квадратов регрессии), второе — случайную составляющую (отклонения от регрессии). Схема дисперсионного анализа имеет следующий вид (табл. 13.4):

Таблица 13.4

Источник вариации	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Линейная регрессия	$Q_2 = \sum (\hat{y} - \bar{Y})^2$	1	$s_2^2 = Q_2$
Остаточная вариация	$Q_1 = \sum (y_i - \hat{y})^2$	$n-2$	$s_1^2 = \frac{Q_1}{n-2}$
Итого	$Q_0 = \sum (y_i - \bar{Y})^2$	$n-1$	

Соответственно находится отношение дисперсий

$$F^* = \frac{s_2^2}{s_1^2}; \quad \left\{\frac{1}{n-2}\right\}.$$

В гл. 12 было показано, что

$$\sum (\hat{y} - \bar{Y})^2 = b^2 \sum (x_i - \bar{X})^2;$$

$$\sum e_i^2 = \sum (y_i - \bar{Y})^2 - b \sum (x_i - \bar{X})(y_i - \bar{Y}).$$

Этими соотношениями можно воспользоваться при определении Q_2 и Q_1.

Общее число степеней свободы равно $n-1$. Число степеней свободы для оценки остаточной дисперсии равно $n-2$. Таким образом, на долю линейной регрессии остается $n-1-(n-2)=1$ степень свободы. Отношение дисперсий F позволяет проверить нулевую гипотезу о том, что переменные не имеют линейной связи. Аналогичный подход сохраняется и при проверке гипотез о наличии регрессий более высоких степеней, т. е. дисперсия, обусловленная регрессией, сравнивается с остаточной дисперсией.

Обратимся теперь к проверке гипотезы о существовании нелинейной регрессии. Эту проблему можно решить, прибегая к дисперсионному анализу, следующим образом. Сначала определяются линейная регрессия и сумма квадратов отклонений от нее $\left(\sum e_{i\,(л)}^2\right)$. Затем находятся нелинейная регрессия, допустим, параболическая, и соответствующая остаточная сумма квадратов $\left(\sum e_{i\,(нл)}^2\right)$. Разность $\left(\sum e_{i\,(л)}^2 - \sum e_{i\,(нл)}^2\right)$ показывает, как уменьшается (или увеличивается) случайная составляющая суммы квадратов отклонений. Можно проверить, значимо или нет это уменьшение. Приводим схему дисперсионного анализа для этого случая (табл. 13.5).

Таблица 13.5

Источник вариации	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Линейная регрессия	$Q_0 = \sum e_{i\,(л)}^2$	$n-2$	—
Параболическая регрессия	$Q_1 = \sum e_{i\,(нл)}^2$	$n-3$	$s_1^2 = \frac{Q_1}{n-3}$
Криволинейность	$Q_2 = Q_0 - Q_1$	1	$s_2^2 = Q_2$

Для проверки нулевой гипотезы (криволинейность и соответственно сокращение остаточной суммы квадратов несущественны) находим отношение дисперсий

$$F^* = \frac{s_2^2}{s_1^2}; \quad \left\{\frac{1}{n-3}\right\}.$$

В приведенном на с. 254 примере оценены регрессии:

$$\hat{y} = 8{,}405 + 1{,}899x; \quad \hat{y} = 0{,}9127 + 7{,}7231x - 0{,}6638x^2$$

Существен ли переход от линейной к квадратической регрессии? Для ответа на этот вопрос составим таблицу дисперсионного анализа (табл. 13.6).

Таблица 13.6

Источник вариации	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Линейная регрессия	96,7	8—2	—
Параболическая регрессия	7,5	8—3	1,5
Криволинейность	89,2	1	89,2

По полученным данным находим $F^* = 59{,}5$. Табличное значение $F_{0,05} = 6{,}61$ при числе степеней свободы 1 и 5. Следовательно, $F^* > F_\alpha$ и нулевая гипотеза отклоняется.

13.7. Структура уравнения регрессии. Проблема мультиколлинеарности

При решении вопроса о наборе независимых переменных, включаемых в уравнение регрессии, стремятся, с одной стороны, охватить как можно больше число факторов и тем самым не упустить из виду важнейшие из них. Увеличение числа переменных в общем повышает точность аналитического описания взаимосвязи и тем самым «прогностическую способность» соответствующего уравнения. С другой стороны, всегда приходится учитывать существенное увеличение объема работ, связанное с ростом числа переменных в уравнении. Поэтому обычно стремятся ограничить число учитываемых факторов и включать только те из переменных, которые вносят ощутимый вклад в объяснение изменения зависимой переменной. Существует несколько статистических приемов отбора наиболее важных переменных, которые можно разделить на две группы.

Приемы первой группы кратко сводятся к следующему. Разрабатывается развернутое уравнение регрессии, включающее все основные переменные, имеющие отношение к зависимой переменной. Затем оцениваются параметры регрессионного уравнения и с помощью тех или иных процедур определяется вклад каждой переменной в сумму квадратов, обусловленную регрессией. Переменные, чей вклад оказывается меньшим, чем некоторый выбранный минимальный уровень, исключаются из регрессии. Затем оценивание параметров регрессии повторяется, но уже для нового набора переменных.

Согласно второму пути структура уравнения определяется в ходе постепенного его «наращивания» переменными. В соответствии с одним из методов процедура состоит в следующем. Сначала определяются парные коэффициенты корреляции между зависимой и всеми независимыми переменными. Переменная, имеющая наибольшее значение этого коэффициента, первой включается в уравнение. Затем подсчитываются коэффициенты частной корреляции между зависимой переменной и всеми независимыми пере-

менными, не включенными в регрессию, и выбирается переменная с наибольшим значением этого параметра. Далее оценивается регрессия на две включенные в регрессию переменные и с помощью дисперсионного анализа проверяется существенность первой из включенных переменных. Если она существенна, то переходят к выбору следующей переменной. Для чего вновь подсчитываются частные коэффициенты корреляции для оставшихся переменных и вся процедура повторяется. Заметим, что каждый раз после введения новой переменной должны проверяться на значимость все включенные ранее переменные. В ходе такой проверки переменные, введенные в уравнения на предыдущих шагах процедуры, мо-

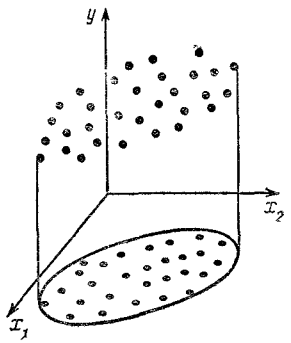


Рис. 13.2.
Регрессионная зависимость y от x_1 и x_2

гут оказаться незначимыми и должны быть исключены из регрессии.

При разработке структуры уравнения регрессии сталкиваются с явлением *мультиколлинеарности*. Под мультиколлинеарностью понимают взаимосвязь независимых переменных уравнения регрессии. Так, в уравнении $y = a_0 + a_1x_1 + a_2x_2 + u$ переменные x_1 и x_2 могут находиться в некоторой линейной зависимости между собой. Эта зависимость может быть функциональной, т. е. $x_1 = ax_2$, тогда имеет место *строгая мультиколлинеарность* переменных. Чаще, однако, взаимосвязь между переменными не столь жестка и проявляется лишь приблизительно, в этом случае мультиколлинеарность называется *нестрогой*.

Одно из основных предположений МНК заключается в том, что между независимыми переменными нет линейной связи. Следовательно, наличие мультиколлинеарности нарушает одно из предположений МНК. При этом уравнение регрессии формально может быть получено, однако оно будет крайне ненадежно в том смысле, что незначительное изменение исходных выборочных данных приводит в этом случае к резкому изменению оценок параметров. Средние квадратические ошибки параметров резко возрастают при наличии мультиколлинеарности.

Раскроем смысл мультиколлинеарности с помощью трехмерного графика. Пусть определяется регрессия

$$y = a_0 + a_1x_1 + a_2x_2 + u.$$

Если x_1 и x_2 линейно независимы, то наблюдения расположатся в трехмерном пространстве и для них можно будет подобрать плоскость $\hat{y} = a_0 + a_1x_1 + a_2x_2$, причем сумма квадратов отклонений наблюдений от этой плоскости будет минимальной и для данной их совокупности плоскость будет единственной. Если спроецировать точки, соответствующие наблюдениям, на плоскость x_1x_2, то они займут на ней некоторую область (рис. 13.2).

Допустим теперь, что $x_1 = \alpha x_2$. Тогда все наблюдения расположатся на плоскости yz, а проекции точек окажутся на линии $x_1 = \alpha x_2$ (рис. 13,3 *а*). На плоскости yz для наблюдений можно

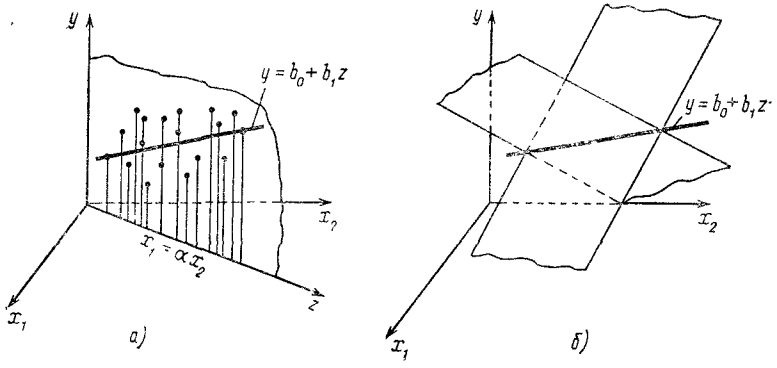


Рис. 13.3.
Мультиколлинеарность переменных x_1 и x_2

подобрать прямую $\hat{y} = b_0 + b_1z$, через которую можно провести сколько угодно плоскостей $\hat{y} = a_0 + a_1x_1 + a_2x_2$ (рис. 13.3, *б*).

Объясним явление мультиколлинеарности и с другой стороны. В § 13.2 говорилось о том, что матрица $\mathbf{X}^T\mathbf{X}$ (где $\mathbf{X}$ — матрица, характеризующая значения независимых переменных) должна быть обратима, а это означает, что $|\mathbf{X}^T\mathbf{X}| \neq 0$. Последнее, как известно из матричной алгебры, имеет место только тогда, когда векторы, составляющие матрицу, линейно независимы. Если хотя бы два вектора матрицы $\mathbf{X}$ строго линейно зависимы, то наблюдается строгая мультиколлинеарность, $|\mathbf{X}^T\mathbf{X}| = 0$ и оценки по формуле $\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ найти нельзя. В случае нестрогой мультиколлинеарности, когда $|\mathbf{X}^T\mathbf{X}| \approx 0$, формально можно определить $\mathbf{a}$ (если вычисления $|\mathbf{X}^T\mathbf{X}|$ вести с большим числом знаков), однако, как было показано выше, эти оценки мало надежны.

Общепринятый метод, применяемый для обнаружения мультиколлинеарности, заключается в вычислении матрицы парных коэффициентов корреляции, охватывающей все сочетания переменных. Коэффициенты, близкие по своему значению к ± 1, свидетельствуют о наличии мультиколлинеарности между соответствующими переменными. Более надежный подход при наличии многих переменных заключается в нахождении определителя матрицы $\mathbf{X}^T\mathbf{X}$. Если определитель близок к нулю, то имеется мультиколлинеар-

ность. Устранение мультиколлинеарности достигается путем пересмотра структуры уравнения регрессии. Самый простой способ заключается в исключении одной из двух, находящихся во взаимосвязи, переменных. Второй прием состоит в трансформации соответствующих переменных. Например, вместо абсолютных значений переменных берут их абсолютные или относительные приросты и т. д. Естественно, что при этом изменяется содержание уравнения регрессии.

13.8. Система регрессионных уравнений

При изучении сложных явлений часто сталкиваются с необходимостью статистического описания не отдельных зависимостей, а их комплекса. В этих случаях прибегают к разработке *системы уравнений регрессии*, каждое из которых описывает одну статистическую закономерность. Как правило, развернутая статистическая модель содержит помимо регрессионных уравнений математические выражения, отображающие балансовые увязки между переменными, характеристики тенденций развития отдельных переменных и т. д.

Взаимозависимая линейная модель. Система регрессионных уравнений, как и любая другая экономико-математическая модель, содержит так называемые *эндогенные* и *экзогенные* переменные. Эндогенными являются те переменные, которые в силу принятых концепций определяются внутренней структурой изучаемого явления, иначе говоря, их значения выясняются на основе модели. Экзогенные переменные по определению независимы от структуры описываемого моделью явления и их значения устанавливаются вне модели. Модель содержит также различного рода параметры (коэффициенты), которые определяются в ходе статистического оценивания путем обработки имеющейся информации (данных наблюдения). Ниже рассматриваются системы, линейные как относительно переменных, так и содержащихся в ней коэффициентов. Нелинейные системы изучены недостаточно полно и пока не нашли широкого практического применения. Разумеется, линейная зависимость далеко не всегда адекватно описывает действительную связь. Однако там, где измерения не отличаются большой точностью, линейное приближение может оказаться вполне достаточным. В ряде случаев несложными преобразованиями нелинейные соотношения могут быть сведены к линейным.

При разработке системы регрессионных уравнений часто оказывается, что зависимые переменные одних уравнений выступают в качестве независимых переменных других. Подобного рода модели (или отдельные уравнения) называют *взаимозависимыми* (или одновременными). Взаимозависимые модели имеют две формы: структурную и приведенную. *Структурная форма* модели создается в процессе формирования самой модели при стремлении отразить причинно-следственный механизм, существующий в реальности. Она позволяет проследить влияние изменений экзоген-

ных переменных модели на значения эндогенных переменных. По предположению система уравнений структурной формы модели должна быть совместной. Отсюда эта система может быть решена относительно эндогенных переменных. Результат решения и представляет собой *приведенную форму* модели. Сказанное поясним простым примером. Пусть имеется линейная взаимозависимая модель (для простоты положим, что она содержит только регрессионные уравнения), содержащая n эндогенных и m экзогенных переменных:

$$\begin{aligned} y_1 &= \gamma_{12}y_2 + \ldots + \gamma_{1n}y_n + \alpha_{11}x_1 + \ldots + \alpha_{1m}x_m + \varepsilon_1; \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= \gamma_{n1}y_1 + \ldots + \gamma_{n,n-1}y_{n-1} + \alpha_{n1}x_1 + \ldots + \alpha_{nm}x_m + \varepsilon_n. \end{aligned} \tag{13.30}$$

Эндогенные переменные обозначены здесь как y_i, экзогенные — как x_j. Система (13.30) может быть решена относительно $y_1, \ldots, y_n$, т. е. представлена как система линейных функций только от $x_1, \ldots, x_m$:

$$\begin{aligned} y_1 &= \delta_{11}x_1 + \ldots + \delta_{1m}x_m + \eta_1; \\ &\ldots\ldots\ldots\ldots\ldots \\ y_n &= \delta_{n1}x_1 + \ldots + \delta_{nm}x_m + \eta_n. \end{aligned} \tag{13.31}$$

Каждая эндогенная переменная содержится в приведенной форме только в одном уравнении и зависит, не считая возмущения η, только от значений параметров и экзогенных переменных. Ошибки η_i здесь являются линейными функциями возмущений ε_i, а коэффициенты δ_{ij} — нелинейными функциями коэффициентов структурной формы.

В общем виде переход от структурной к приведенной форме модели можно представить, используя матричную запись. Так, структурную форму модели можно записать как

$$\mathbf{\Gamma y} + \mathbf{Ax} = \boldsymbol{\varepsilon},$$

где $\mathbf{\Gamma}$ и $\mathbf{A}$ — матрицы неизвестных параметров размерностей $n \times n$ и $n \times m$ соответственно; матрица $\mathbf{\Gamma}$ — невырождена; $\mathbf{y}$ и $\mathbf{x}$ — векторы эндогенных и экзогенных переменных; $\boldsymbol{\varepsilon}$ — вектор случайных, нормально распределенных величин с нулевыми математическими ожиданиями; n — число эндогенных переменных; m — число экзогенных переменных.

Приведенную форму теперь можно получить в виде

$$\mathbf{y} = -\mathbf{\Gamma}^{-1}\mathbf{Ax} + \boldsymbol{\eta} = \mathbf{Bx} + \boldsymbol{\eta}, \tag{13.32}$$

где

$$\mathbf{B} = -\mathbf{\Gamma}^{-1}\mathbf{A}. \tag{13.33}$$

Уравнения приведенной формы с численными параметрами позволяют получить значения эндогенных переменных, однако они объясняют их только через экзогенные переменные, т. е. с существенно меньшей детализацией, чем структурная форма, здесь отсутствуют взаимосвязи зависимых переменных между собой,

Частным случаем системы (13.30) является рекурсивная модель, состоящая из уравнений регрессии, имеющих одностороннюю зависимость (без обратных связей). Приведем пример рекурсивной системы:

$$\begin{aligned} y_1 &= \alpha_{11}x_1 + \ldots + \alpha_{1m}x_m + \varepsilon_1; \\ y_2 &= \gamma_{21}y_1 + \alpha_{21}x_1 + \ldots + \alpha_{2m}x_m + \varepsilon_2; \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n &= \gamma_{n1}y_1 + \ldots + \gamma_{n,n-1}y_{n-1} + \alpha_{n1}x_1 + \ldots + \alpha_{nm}x_m + \varepsilon_n. \end{aligned} \tag{13.34}$$

Оценивание параметров взаимозависимой системы. Непосредственное оценивание параметров структурной формы с взаимозависимыми эндогенными переменными обычно не применяется, поскольку при таком подходе получают смещенные оценки (см., например, [5]). Поэтому сначала получают параметры приведенной формы, оценивая каждое уравнение в отдельности, а затем тем или иным путем переходят к структурной форме (оцениванию ее параметров). Однако этот переход далеко не всегда можно реализовать однозначным образом. Дело в том, что если приведенная форма взаимозависимой модели содержит полный набор переменных в каждом уравнении (соответственно в системе nm параметров), то в структурной форме их число равно $n(n-1+m)$. Ясно, что последние нельзя получить единственным образом на основе nm параметров приведенной формы. Иначе говоря, из (13.33) следует

$$\mathbf{\Gamma B} + \mathbf{A} = \mathbf{O}. \tag{13.35}$$

Для однозначного определения коэффициентов матриц $\mathbf{\Gamma}$ и $\mathbf{A}$ на основе (13.35) в общем случае недостаточно располагать матрицей коэффициентов приведенной формы $\mathbf{B}$. Действительно, для одной и той же матрицы $\mathbf{B}$ можно подобрать множество матриц структурной формы $\mathbf{\Gamma}$ и $\mathbf{A}$ так, чтобы удовлетворялось это соотношение. Положение, однако, исправляется тем, что часто можно принять некоторые ограничения на коэффициенты матриц $\mathbf{\Gamma}$ и $\mathbf{A}$, приравняв некоторые из них нулю (что означает исключение переменной) или положив иные ограничения, если, разумеется это не противоречит экономической теории. Если после введения ограничений матрицы $\mathbf{\Gamma}$ и $\mathbf{A}$ можно в силу их внутренних особенностей однозначно определить по матрице $\mathbf{B}$, то модель называют точно идентифицируемой. Если же таких ограничений больше, чем это нужно для точной идентификации, то мы имеем дело со случаем сверхидентификации системы — с наиболее часто встречающейся в практике ситуацией.

Таким образом, проблема идентификации модели сводится к конструированию ее формы. Формулировка полного решения проблемы довольно громоздка (она приводится, например, в [5]), поэтому ограничимся лишь рассмотрением необходимого (но не достаточного) критерия идентифицируемости, который обычно и используется в практике моделирования. Пусть n — полное число

уравнений, или, что одно и то же, число эндогенных переменных, m — число экзогенных переменных всей системы, $n_i + m_i$ — число экзогенных и эндогенных величин, присутствующих в i-м уравнении. При исключении переменных из уравнений необходимое условие идентифицируемости каждого уравнения системы имеет вид:

$$(n + m) - (n_i + m_i) \geqslant n - 1. \qquad (13.36)$$

Таким образом, число оставленных в полном наборе переменных не должно быть меньше числа уравнений системы без единицы. Если каждое уравнение структурной формы идентифицируемо, то идентифицируемой оказывается вся система уравнений.

Приведение системы к идентифицируемому виду может (хотя и не обязательно) привести к существенному обеднению модели в содержательном смысле в связи с исключением ряда переменных. Для того чтобы сказанное стало более ясным, приведем пример. Пусть структурная форма модели имеет вид:

$$y_1 = \gamma_{12} y_2 + \alpha_{11} x_1 + \alpha_{12} x_2 + \varepsilon_1;$$
$$y_2 = \gamma_{21} y_1 + \alpha_{21} x_1 + \alpha_{22} x_2 + \varepsilon_2.$$

Оба уравнения системы неидентифицируемы. Как первое, так и второе уравнения становятся идентифицируемы при исключении из правых сторон каждого по одному (разному) элементу. Например, пусть $\gamma_{12} = 0$ и $\alpha_{21} = 0$, тогда

$$y_1 = \alpha_{11} x_1 + \alpha_{12} x_2 + \varepsilon_1;$$
$$y_2 = \gamma_{21} y_1 + \alpha_{22} x_2 + \varepsilon_2.$$

Оба уравнения идентифицируемы. Если переменные y_2 в первом уравнении и x_1 во втором не представляли большой важности при объяснении соответственно y_1 и y_2, то их исключение не приведет к искажениям в описании взаимосвязи. Однако не исключено, что выбор переменных для исключения из уравнений не является очевидным. Ошибки в этом случае приведут к объединению модели.

В случае когда система точно идентифицируема, параметры структурной формы определяются с помощью так называемого *косвенного метода наименьших квадратов* (КМНК). Суть его состоит в следующем. Сначала с помощью МНК оцениваются параметры каждого уравнения приведенной формы модели в отдельности. Затем эти параметры трансформируются в параметры структурной формы модели. Например, пусть имеется структурная форма

$$\begin{aligned} y_1 &= c_{12} y_2 + a_{11} x_1 + e_1; \\ y_2 &= c_{21} y_1 + a_{22} x_2 + e_2. \end{aligned} \qquad (13.37)$$

Оба уравнения точно идентифицируемы. Решение системы относительно y_1 и y_2 дает приведенную форму:

$$\begin{aligned} y_1 &= d_{11} x_1 + d_{12} x_2 + u_{11}; \\ y_2 &= d_{21} x_1 + d_{22} x_2 + u_{12}, \end{aligned} \qquad (13.38)$$

где

$$d_{11} = \frac{a_{11}}{1 - c_{12}c_{21}}; \quad d_{12} = \frac{c_{12}a_{22}}{1 - c_{12}c_{21}};$$
$$d_{21} = \frac{c_{21}a_{11}}{1 - c_{12}c_{21}}; \quad d_{22} = \frac{a_{22}}{1 - c_{12}c_{21}}. \qquad (13.39)$$

Параметры d_{ij} легко получить при оценивании регрессий (13.38) с помощью МНК. Что же касается параметров структурной формы, то они находятся путем решения (13.39) относительно четырех неизвестных коэффициентов. В этом случае

$$c_{12} = d_{12} : d_{22}; \; c_{21} = d_{21} : d_{11};$$
$$a_{11} = \frac{d_{22}d_{11} - d_{12}d_{21}}{d_{22}}; \quad a_{22} = \frac{d_{22}d_{11} - d_{12}d_{21}}{d_{11}}.$$

Оценить параметры сверхидентифицированной системы, используя приведенную форму, таким же путем, как при точно идентифицируемой системе, не удается. Например, пусть имеется структурная форма (13.37), оба уравнения которой точно идентифицируемы. Однако если на параметры любого из них наложить хоть одно ограничение, то это уравнение станет сверхидентифицируемым. Так, если в первом уравнении $c_{12} = a_{11}$, то в структурной форме

$$y_1 = c_{12}(y_2 + x_1) + e_1;$$
$$y_2 = c_{21}y_1 + a_{22}x_2 + e_2 \qquad (13.40)$$

параметры a_{22} и c_{21} находятся однозначно. Однако параметр c_{12} не может быть однозначно определен на основе коэффициентов приведенной формы.

Если система сверхидентифицируема, то прибегают к ряду специальных методов оценивания параметров, среди которых наиболее простым и в то же время надежным является *двушаговый метод наименьших квадратов* (ДМНК). При применении ДМНК оценивание параметров производится дважды. На первом шаге оцениваются параметры приведенной формы. Это дает возможность получить оценки систематической и случайной составляющей y, т. е. предполагается, что

$$y_i = y_i^* + v_i,$$

де y_i^* — оценки значений зависимой переменной, получаемые по приведенной форме:

$$y_i^* = d_{i1}x_1 + d_{i2}x_2 + \ldots + d_{ij}x_j.$$

На втором шаге эндогенные переменные, находящиеся в правой части структурных уравнений, заменяются их оценками y^*. К преобразованному таким путем структурному уравнению применяется обычный МНК.

Что же касается оценивания параметров структурной формы рекурсивной модели, то оно возможно без преобразования последней в приведенную форму. Для этого последовательно оцени-

вают уравнения структурной формы (13.34). Сначала оценивается первое уравнение, затем второе, причем в этом случае в качестве y_1, содержащихся в правой части уравнения, берут расчетные значения (т. е. значения этой переменной, полученные по предыдущей регрессии). Аналогичным способом последовательно оцениваются остальные уравнения. Заметим, что такой подход к оцениванию рекурсивных систем возможен только в том случае, когда каждое уравнение структурной формы идентифицируемо.

Сложность получения параметров структурной формы привела к тому, что разработчики модели часто стремятся модель в целом или хотя бы отдельные ее блоки представить в виде независимых или рекурсивных уравнений. Естественно, что за удобства оценивания в этом случае приходится расплачиваться некоторым сокращением учитываемых экономических связей. Вопрос о том, в каких пределах эта жертва оправдана, не прост и требует специального анализа.

Глава 14

Введение в анализ временных рядов

14.1. Задачи анализа временных рядов

В практике статистического анализа экономист весьма часто сталкивается с тем, что исходные данные, которыми он располагает для выявления той или иной закономерности, представлены в виде временных (динамических) рядов. Такие ряды описывают изменение некоторой характеристики во времени. Каждый член (уровень) такого ряда связан с соответствующим моментом времени или временным интервалом. Разумеется, уровни ряда должны быть сопоставимыми по своему содержанию. Показатели временных рядов формируются под совокупным влиянием множества длительно и кратковременно действующих факторов и в том числе различного рода случайностей. Изменение условий развития явления приводит к более или менее интенсивной смене самих факторов, к изменению силы и результативности их воздействия и в конечном счете к вариации уровня изучаемого явления во времени. Лишь в очень редких случаях в экономике встречаются чисто *стационарные ряды*, т. е. ряды, в которых не наблюдаются систематические изменения в средних значениях уровней, их дисперсиях, и эти характеристики не зависят от начала отсчета времени. В таких случаях вариацию уровней можно изучать с помощью специального раздела математической статистики — теории стационарных процессов. В основном временные ряды, с которыми имеют дело в экономике, не являются стационарными. Последовательность расположения исследуемых данных во времени в таких рядах имеет существенное значение для анализа, т. е. время

здесь выступает как один из определяющих для изучаемого явления факторов. Естественно, что задачи и средства анализа подобных рядов будут отличаться от того, с чем имеет дело математическая статистика при изучении рядов распределения. Вместе с тем применяемые при обработке временных рядов методы во многом опираются на методы и характеристики, разработанные математической статистикой для рядов распределения. Следует сразу подчеркнуть, что в некоторых случаях выводы, полученные при математико-статистической обработке конкретных данных, имеют условный характер, так как методы математической статистики базируются на довольно жестких требованиях к качеству обрабатываемых данных (например, к их однородности) и строгих гипотезах о характере поведения анализируемых величин (их распределениях). На практике же экономист зачастую, особенно если исследуются временные ряды, имеет дело с информацией, качество которой в отношении выдвинутых требований оставляет желать лучшего или просто неизвестно. Обычно неизвестен и тип распределения переменных. Таким образом, для практика остаются две альтернативы: или вообще отказаться от применения большинства методов математико-статистического анализа и довольствоваться достаточно скудным инструментарием, или применять разнообразные статистические методы обработки данных, при этом не забывая о соответствующих этим методам требованиях. Очевидно, что если существуют сомнения в «чистоте эксперимента» в только что указанном смысле, то не следует рассматривать получаемые статистические выводы чрезмерно строго. В то же время эти выводы, как правило, оказываются полезными для практической деятельности и, в частности, для прогнозирования, которому последнее время уделяется много внимания.

К настоящему времени статистика располагает разнообразными методами анализа временных рядов — от самых элементарных до весьма изощренных и сложных. Все эти методы так или иначе призваны охарактеризовать в каком-либо аспекте развитие изучаемого явления во времени. Можно выделить три основные задачи исследования временных рядов. Первая из них заключается в *описании* изменения соответствующего показателя во времени и выявлении тех или иных свойств исследуемого ряда. Для этого прибегают к разнообразным способам: расчету обобщающего показателя изменения уровней во времени — среднего темпа роста; применению различных сглаживающих фильтров, уменьшающих колебания уровней во времени и позволяющих более четко представить тенденции развития; подбору кривых, характеризующих эту тенденцию; выделению сезонных и иных периодических и случайных колебаний; измерению зависимости между членами ряда (автокорреляции). К методам описания какого-либо свойства динамики можно с некоторым основанием отнести и методы проверки наличия или отсутствия долговременных тенденций в ряду.

Второй важной задачей анализа является *объяснение механизма* изменения уровней ряда. Для ее решения обычно прибегают к регрессионному анализу. Наконец, описание изменения временного ряда и объяснение механизма формирования ряда часто используются для *статистического прогнозирования*, которое в большинстве случаев сводится к экстраполяции обнаруженных тенденций развития.

Перечисленные задачи решаются с помощью различных методов. В рамках отдельной главы нет возможности даже кратко охарактеризовать каждый из них. Поэтому остановимся лишь на некоторых, причем внимание будет сконцентрировано в основном на методе решения классической задачи — выявление основной тенденции развития и измерение отклонений от нее. Следует также добавить, что в главе обсуждаются методы анализа временных рядов, которые так или иначе связаны с аппаратом математической статистики. Методы, которые основаны на иных подходах и принципах, здесь не рассматриваются.

Понятие тенденция развития не имеет достаточно четкого определения. В статистической литературе под тенденцией развития понимают общее направление развития, долговременную эволюцию. Обычно тенденцию стремятся представить в виде более или менее гладкой кривой, которой соответствует некоторая функция времени. Эта кривая, назовем ее *трендом*, характеризует основную закономерность движения во времени и в известной мере (но не полностью) свободна от случайных воздействий. Тренд описывает некоторую усредненную для достаточно протяженного периода наблюдения тенденцию развития во времени. В большинстве случаев полученная траектория связывается исключительно с ходом времени. Предполагается, что с помощью переменной время можно выразить влияние всех основных факторов. Механизм их влияния в явном виде не учитывается. В связи со сказанным под трендом часто понимают регрессию на время. Иногда тренд представляют как детерминированную компоненту переменной (причем не обязательно изменение этой компоненты связывают со временем). При этом предполагают, что отклонение от тренда есть некоторая случайная составляющая ε_t. Такой подход действительно удобен на практике; однако вряд ли он имеет под собой надежную теоретическую базу при описании экономических характеристик. Здесь речь идет скорее о соглашении, в силу которого обычно считают, что тренд определяется влиянием постоянно действующих факторов, а отклонения от него — влиянием случайных факторов. Однако сравнительно часто к одним и тем же рядам наблюдений можно хорошо подобрать функции различных видов. В этом случае весьма спорным или во всяком случае весьма относительным является практическое разделение уровня на детерминированную и случайную составляющие. В силу сказанного мы будем придерживаться более общего понимания термина тренд, не связывая его с конкретными формальными определениями.

14.2. Элементарные приемы описания временных рядов

Средний темп роста. Одним из наиболее часто применяемых показателей, с помощью которого в обобщенном виде характеризуют изменение ряда, является средний темп роста. Этот показатель обычно получают как геометрическую среднюю из ряда последовательных (цепных) темпов роста (см. § 2.2). Цепной темп роста характеризует отношение какого-либо уровня ряда к предыдущему. Он выражается в процентах или в долях единицы. В последнем случае его часто называют коэффициентом роста (мы будем применять только термин темп роста вне зависимости от того, выражено ли это отношение в долях единицы или в процентах, поскольку это несущественно для анализа).

Если ряд состоит из уровней $y_1, y_2, \ldots, y_t, \ldots, y_n$, то цепные темпы роста (τ_t) равны отношениям

$$\tau_1 = \frac{y_2}{y_1}, \quad \tau_2 = \frac{y_3}{y_2}, \quad \ldots, \quad \tau_{n-1} = \frac{y_n}{y_{n-1}},$$

а средняя геометрическая из них ($\bar{\tau}$) будет равна:

$$\bar{\tau} = \sqrt[n-1]{\tau_1 \cdot \tau_2 \ldots \tau_{n-1}} = \sqrt[n-1]{\frac{y_n}{y_1}}, \quad t = 1, 2, \ldots, n. \qquad (14.1)$$

Откуда

$$y_n = y_1 \bar{\tau}^{n-1}. \qquad (14.2)$$

В статистической литературе неоднократно высказывались сомнения относительно ценности среднего темпа для экономического анализа [20]. К недостаткам среднего темпа как обобщающего показателя, исчисляемого в виде (14.1), следует отнести следующее:

1. Средний темп полностью определяется двумя крайними уровнями ряда, при этом выбор периода для расчета среднего темпа существенным образом определяет его значение. Сдвиг периода даже на один шаг может привести к значительному изменению величины темпа. К сказанному следует добавить, что при достаточной протяженности анализируемого ряда можно подсчитать много значений среднего темпа (на все вкусы!) путем подбора различных временных интервалов в пределах ряда. Таким образом, значение среднего темпа будет в известном смысле случайным. Неустойчивость показателя среднего темпа особенно проявляется в том случае, когда ряд имеет чередующиеся повышения и понижения. Возьмем в качестве иллюстрации ряд, который характеризуется некоторой тенденцией к росту и умеренной колеблемостью (см. табл. 14.1).

В целом для ряда $\bar{\tau} = 1{,}0457$; для периода $1 \div 8$ $\bar{\tau} = 1{,}0729$ (максимальное значение); для периода $3 \div 11$ $\bar{\tau} = 1{,}0269$ (минимальное значение).

Очевидно также, что надежность среднего темпа как обобщающей характеристики не увеличивается с ростом числа наблюдений.

Таблица 14.1

t	1	2	3	4	5	6
y_t	87,7	97,7	108,5	107,9	111,0	122,7
t	7	8	9	10	11	12
y_t	135,4	143,5	132,5	133,7	134,2	143,4

2. При расчете среднего темпа не принимаются во внимание промежуточные члены ряда, отсюда теряется существенная для анализа информация. Именно эти члены определяют форму тенденции развития. Отсюда чем продолжительнее период, для которого исчисляется средний темп, тем больше теряется информации, тем меньше он играет роль обобщающего показателя.

3. Применение среднего темпа предполагает, что развитие явления следует геометрической прогрессии и соответственно

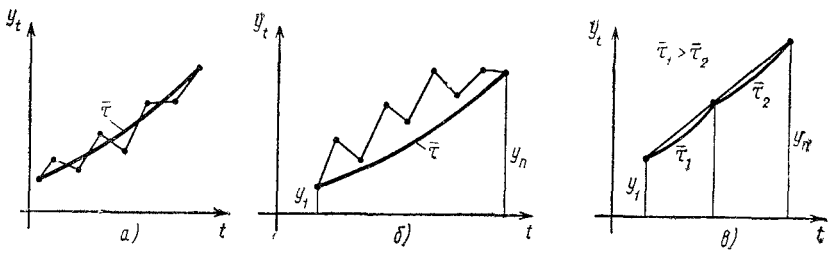


Рис. 14.1.
Экспоненциальный тренд

траектория изменения соответствующего показателя в общем приближается к экспоненциальной кривой. В этом случае средний темп действительно представляет собой обобщающий показатель динамики, так как он усредняет колебания вокруг экспоненциальной траектории развития (см. рис. 14.1, *а*). Если же траектория имеет другую форму, то, строго говоря, средний темп теряет свое содержание (см. рис. 14.1, *б*). Допустим, что ряд строго следует линейному тренду с положительным углом наклона. В этом случае вряд ли есть основание утверждать, что средний темп является обобщенным свойством ряда, так как по мере увеличения интервала наблюдения средний темп будет уменьшаться. Для первой половины ряда средний темп ($\bar{\tau}_1$) окажется меньше аналогичной характеристики второй половины ряда ($\bar{\tau}_2$) (см. рис. 14.1, *в*). Однако характер развития ряда не претерпел каких-либо изменений.

Из сказанного выше следует, что средний темп как обобщающий показатель динамики имеет весьма ограниченную ценность. Однако при всех своих недостатках средний темп по традиции,

в связи с легкостью его получения, наконец, отсутствием другой простой обобщающей характеристики развития, широко применяется в практическом анализе динамики. Особенно он удобен при сопоставлении рядов с разной закономерностью развития. К тому же сами средние темпы в основном используются как конечные характеристики, которые не участвуют в дальнейших расчетах, а выступают лишь как соответствующие индикаторы динамики.

Недостатки среднего темпа в общем-то давно осознаны экономистами. Поэтому время от времени предпринимаются попытки «модернизации» приемов оценки этого показателя. В частности, в практике иногда пытаются уменьшить погрешность, связанную с расчетом темпа по двум точкам. Средний темп исчисляют, например, беря за основу две средние величины, одна из которых исчислена для отрезка, взятого в начале ряда, другая — для отрезка в конце его. Действительно, если ряд следует экспоненциальной тенденции развития и имеет высокую колеблемость уровней, то такой прием может несколько снизить опасность выбора нерепрезентативных уровней, и, как следствие, средний темп будет более надежным (например, расчет среднего темпа роста урожайности). Однако при этом не устраняются все остальные недостатки расчета среднего темпа по формуле (14.1). Имеются и иные попытки устранения отмеченных выше недостатков.

Рассмотрим метод определения среднего темпа, соответственно которому в расчете участвуют все члены ряда и, таким образом, используется вся информация, содержащаяся во временном ряду. Расчет среднего темпа по этому методу [8] базируется на предпосылке о том, что сумма фактических уровней ряда должна быть равна сумме уровней, полученных расчетным путем на основе начального уровня ряда и среднего темпа роста. Назовем средний темп, рассчитанный таким способом, *средним кумулятивным темпом роста* и обозначим его как $\tau_к$. Для его определения найдем сперва сумму уровней ряда, начиная со второго, использовав $\tau_к$:

$$\sum_{t=2}^{n} y_t = y_1\tau_к + y_1\tau_к^2 + \ \ldots\ + y_1\tau_к^{n-1} = y_1 \sum_{t=2}^{n} \tau_к^{t-1}. \qquad (14.3)$$

На основе (14.3) легко получить отношение

$$R = \frac{\sum_{2}^{n} y_t}{y_1},$$

которое примем в качестве основы для расчета $\tau_к$. Теперь воспользуемся тем, что R есть сумма членов геометрической прогрессии. Откуда

$$R = \tau_к + \tau_к^2 + \ \ldots\ + \tau_к^{n-1} = \frac{\tau_к^n - \tau_к}{\tau_к - 1}. \qquad (14.4)$$

Таким образом, $\tau_к = f(R)$. Определение $\tau_к$ на основе (14.4) достаточно трудоемко, однако существуют таблицы значений $\tau_к$ в за-

висимости от R для заданного числа лет *. Для ряда, приведенного в табл. 14.1, средний кумулятивный темп роста равен 1,0576.

Следует принять во внимание, что $\tau_к$ — устойчивый показатель, он не меняет свое значение при изменении значений уровней, если при этом не меняется отношение R. Однако эта устойчивость нарушается при изменении значения y_1. Отсюда случайные колебания этого уровня существенным образом скажутся на $\tau_к$.

Средний темп роста можно получить также, прибегнув к аналитическому выравниванию соответствующего ряда по экспоненте. В § 14.4 будут обсуждены методы аналитического выравнивания (подбора кривой по эмпирическим данным в соответствии с некоторым критерием). Пусть в качестве такой аппроксимирующей кривой принята экспонента

$$\hat{y}_t = ab^t, \tag{14.5}$$

где $\hat{y}_t$ — ордината экспоненты (выравненный или расчетный уровень ряда) в момент t; a, b — параметры, значения которых получают по данным наблюдения. Параметры имеют следующее содержание: a — расчетное (выравненное) значение первого уровня ряда, т. е. $a = \hat{y}_0$, b — средний темп роста. Из (14.5) при условии, что $t = n$, следует

$$b = \sqrt[n]{\frac{\hat{y}_n}{a}} = \sqrt[n]{\frac{\hat{y}_n}{\hat{y}_0}}. \tag{14.6}$$

Если время характеризуется последовательностью $t = 1, 2, \ldots, \ldots, n$, то вместо (14.5) и (14.6) имеем

$$\hat{y}_t = ab^{t-1} \quad \text{и} \quad b = \sqrt[n-1]{\frac{\hat{y}_n}{\hat{y}_1}}$$

Для приведенного выше ряда получим $b = 1{,}042$ **, для этого же ряда $\bar{\tau} = 1{,}046$ и $\tau_к = 1{,}058$.

Сравнивая параметр b со средним темпом $\bar{\tau}$, легко понять, чем один средний темп отличается от другого: параметр b основывается на отношении не фактических уровней ряда, а расчетных, выравненных по экспоненте. На рис. 14.2 показаны две экспоненциальные кривые (параметр кривой 1 равен b, кривой 2 — $\bar{\tau}$).

Заканчивая обсуждение среднего темпа роста и методов его измерения, приведем иллюстрацию, характеризующую устойчивость $\bar{\tau}$, $\tau_к$ и b при увеличении периода наблюдения (см. табл. 14.2), средние темпы подсчитаны для ряда, приведенного в табл. 14.1.

Сглаживание временных рядов. Наиболее распространенным и простым путем выявления тенденции развития является *сглаживание* временного ряда. Суть различных приемов, с помощью которых осуществляется сглаживание, сводится к замене фактических уровней ряда расчетными, имеющими значительно меньшую колеблемость, чем исходные данные. Уменьшение

---

* См. [22], где приведены значения $\tau_к$ для рядов с числом членов от 5 до 15.

** Выравнивание данного ряда по экспоненте показано в табл. 14.9.

колеблемости позволяет тенденции развития проявить себя более наглядно. В последнее время подобного рода операции стали называть фильтрованием, а оператор, с помощью которого осуществля-

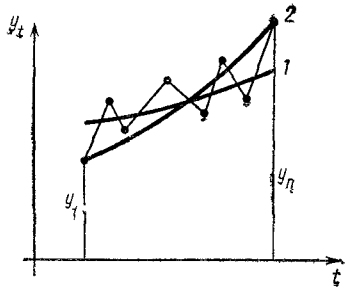


Рис. 14.2.
Тренды, полученные выравниванием по экспоненте (кривая 1) и по среднему темпу роста (кривая 2)

ют ее,— *фильтром.* В ряде случаев сглаживание ряда может рассматриваться как важное вспомогательное средство, облегчающее

Таблица 14.2

Средний темп роста	Период наблюдения, лет							Амплитуда колебаний темпа
	1÷6	1÷7	1÷8	1÷9	1÷10	1÷11	1÷12	
$\bar{\tau}$	1,0694	1,0751	1,0728	1,0529	1,0480	1,0434	1,0457	0,0317
$\bar{\tau}_к$	1,0751	1,0751	1,0744	1,0692	1,0645	1,0616	1,0576	0,0175
b	1,0605	1,0657	1,0674	1,0579	1,0503	1,0441	1,0415	0,0259

применение других методов статистического анализа временных рядов.

Наиболее употребительными в практике являются *линейные фильтры.* Общая формула линейного фильтра имеет вид:

$$\bar{y}_t = \sum_{r=-q}^{s} a_r y_{t+r}, \qquad (14.7)$$

где $\bar{y}_t$ — сглаженное («отфильтрованное») значение уровня на момент t; a_r — вес, приписываемый уровню ряда, находящемуся на расстоянии r от момента t. Фильтр (14.7) охватывает s уровней после момента t и q уровней до него.

Скользящие средние. Если принято, что $\sum a_r = 1$ и $a_r = \text{const}$, то фильтр (14.7) будет соответствовать средней (арифметической), которую в этом случае называют *скользящей средней.* Положим, что динамический ряд состоит из уровней y_t, $t = 1, \ldots, n$. Для каждых m последовательных уровней этого ряда ($m < n$) можно подсчитать среднюю величину. Вычислив значение средней для первых m уровней, переходят к расчету средней для уровней $y_2, \ldots, y_{m+1}$, затем $y_3, \ldots, y_{m+2}$ и т. д. Таким образом, интервал сглаживания, т. е. интервал, для которого подсчитывается средняя, как бы скользит по динамическому ряду с шагом, рав-

ным единице. Пусть m — нечетное число, а предпочтительнее брать нечетное число уровней, поскольку в этом случае расчетное значение уровня приходится на момент времени, для которого имеется фактически наблюдаемый уровень, так что замена y_t на $\bar{y}_t$ происходит без каких-либо дополнительных затруднений. Для определения $\bar{y}_t$ на основе (14.7) легко записать формулу, приняв $a_r = \frac{1}{2p+1}$, $r = -p, \ldots, p$:

$$\bar{y}_t = \frac{1}{2p+1} \sum_{i=t-p}^{t+p} y_i, \qquad (14.8)$$

где $\bar{y}_t$ — скользящая средняя для момента t; y_i — фактическое значение уровня в момент i; i — порядковый номер уровня в интервале сглаживания; m — интервал сглаживания, $m = 2p + 1$, откуда $p = \frac{m-1}{2}$.

Для иллюстрации воспользуемся данными об урожайности озимой пшеницы (ц/га) за 10 лет. Сглаживание осуществим с помощью трехлетней скользящей средней. Исходные данные и результаты сглаживания приведены в табл. 14.3.

Таблица 14.3

t	1	2	3	4	5	6	7	8	9	10
y_t	16,3	20,2	17,1	7,7	15,3	16,3	19,9	14,4	18,7	20,7
$\bar{y}_t$	—	17,53	15,00	13,37	13,10	17,17	16,87	17,67	17,94	—

Чаще всего сглаживание производят по трех-, пяти- и семилетней скользящей средней: чем выше колеблемость ряда, тем шире берется интервал сглаживания. Расчет скользящей средней при $m > 3$ и большой протяженности ряда можно несколько упростить, применяя рекуррентную формулу

$$\bar{y}_t = \bar{y}_{t-1} + \frac{y_{t+p} - y_{t-(p+1)}}{2p+1}. \qquad (14.9)$$

Чем больше интервал сглаживания, тем значительнее усреднение данных и в большей мере погашаются колебания, поэтому выделяемая тенденция развития получается более плавной. При этом дисперсия $\bar{y}_t$ равна $\sigma^2 : m$ (где σ^2 — дисперсия исходных показателей в интервале сглаживания). Если ряд имеет периодические колебания с жесткой продолжительностью цикла, то они полностью устраняются при сглаживании с помощью скользящей средней при интервале сглаживания, равном или кратном циклу.

Графическая иллюстрация результатов сглаживания по трех- и семилетним скользящим средним ряда, характеризующего урожайность, представлена на рис. 14.3.

В ряде случаев сглаживание с помощью простой скользящей средней оказывается настолько сильным, что тенденция развития

проявляется лишь в самом общем виде, а отдельные важные для экономического анализа детали (волны или изгибы) исчезают. Кроме того, часто после сглаживания мелкие волны меняют свой знак, т. е. на кривой получают вогнутый участок вместо выпуклого, и наоборот. В силу сказанного применение простой скользящей средней в общем случае является нежелательным.

Более тонкий прием заключается в применении *взвешенных скользящих средних*. В этом случае каждому уровню в пределах

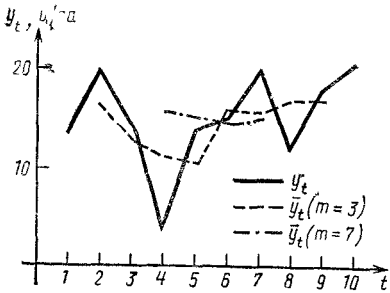


Рис. 14.3.
Сглаживание ряда с помощью трех- и семилетней скользящей средней

интервала сглаживания приписывается вес, зависящий от расстояния, измеряемого от данного уровня до середины интервала сглаживания. Для определения системы весов прибегают к различным способам. Рассмотрим два из них. Согласно первому веса соответствуют членам разложения бинома $\left(\frac{1}{2}+\frac{1}{2}\right)^{2p}$, т. е. имеют следующие значения: при $m=3$ (в этом случае $p=1$) $a_r=\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$; при $m=5$ $a_r=\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16}$; при $m=7$ $a_r=\frac{1}{64}, \frac{3}{32}, \frac{15}{64}, \frac{5}{16}, \frac{15}{64}, \frac{3}{32}, \frac{1}{64}$.

Напомним, что при росте p распределение, характеризуемое данной системой весов, стремится к нормальному.

Второй подход заключается в подборе полинома к данным, содержащимся в интервале сглаживания, и определении расчетного значения полинома для центральной точки интервала. Заметим также, что если отсчет времени в пределах интервала сглаживания ведется от его середины, то искомое расчетное значение полинома, необходимое для сглаживания, равно свободному члену полинома. Ниже мы более подробно остановимся на подборе кривых для аналитического выравнивания ряда. Сейчас же только отметим, что параметры таких кривых обычно подбирают с помощью метода наименьших квадратов (см. гл. 12 и 13). Например, пусть веса определяются на основе парабол второй степени. Тогда для каждых m последовательных членов ряда подбирается уравнение

$$\bar{y}_t = a + bi + ci^2, \quad i = -p, \ldots, -2, -1, 0, 1, 2, \ldots, p,$$

где i — номер уровня в интервале сглаживания. Система нормальных уравнения имеет вид:

$$\sum y = am + c \sum i^2;$$
$$\sum yi = b \sum i^2;$$
$$\sum yi^2 = a \sum i^2 + c \sum i^4.$$

На основе приведенных уравнений необходимо определить голько параметр a, поскольку $\bar{y}_t = a$ для $i = 0$. Пусть сглаживание производится по пяти точкам, $m = 5$. Отсюда $\sum i^2 = 10$; $\sum i^4 = 34$ и

$$\sum_{-p}^{p} y_i = 5a + 10c; \quad \sum_{-p}^{p} y_i i^2 = 10a + 34c.$$

Решим эти уравнения относительно a:

$$a = \frac{34 \sum y - 10 \sum yi^2}{5 \cdot 34 - 10^2} = \frac{17 \sum y - 5 \sum yi^2}{35} =$$

$$= \frac{1}{35}\left(-3y_{t-2} + 12y_{t-1} + 17y_t + 12y_{t+1} - 3y_{t+2}\right).$$

Таким образом, найдена система весов для расчета взвешенной скользящей средней $\bar{y}_t$ при $m = 5$. Аналогичным образом получим значения a параболы второй (и третьей) степени и для других интервалов сглаживания. Ниже приводятся соответствующие системы весов для некоторых значений m:

$$m = 5, \frac{1}{35}(-3,\ 12,\ 17,\ 12,\ -3);$$

$$m = 7, \frac{1}{21}(-2,\ 3,\ 6,\ 7,\ 6,\ 3,\ -2);$$

$$m = 9, \frac{1}{231}(-21,\ 14,\ 39,\ 54,\ 59,\ 54,\ 39,\ 14,\ -21).$$

Нетрудно убедиться в том, что сглаживание по простой скользящей средней является частным и самым простым случаем сглаживания с помощью подбора многочленов: в качестве сглаживающего многочлена берется уравнение прямой.

Как видно из приведенных формул, веса фильтра симметричны относительно центрального уровня (y_t) и их сумма с учетом общего множителя, вынесенного за скобки, равна единице. Заметим также, что в системе весов кроме положительных величин содержатся и отрицательные. Это обстоятельство приводит к тому, что сглаженная кривая в большей мере улавливает характер действительного развития, чем кривая, полученная на основе простой скользящей средней.

В табл. 14 4 приводятся результаты сглаживания по пятилетним взвешенным скользящим средним данных об урожайности пшеницы за 16 лет. В качестве весов приняты биномиальные коэффициенты и коэффициенты, полученные с помощью полинома Для сравнения в этой же таблице приводятся результаты сглаживания с помощью простых скользящих средних.

Таблица 144

Год	Урожайность, ц/га	Простая скользящая средняя	Взвешенные скользящие средние	
			биномиальные коэффициенты	веса, полученные по полиному
1	10,3	—	—	—
2	14,3	—	—	—
3	7,7	12,5	12,0	11,9
4	15,8	13,8	13,4	12,6
5	14,4	14,0	15,0	16,2
6	16,7	16,5	15,9	15,2
7	15,3	16,7	16,9	17,4
8	20,2	15,4	17,2	18,8
9	17,1	15,1	15,2	15,2
10	7,1	15,3	13,0	11,7
11	15,3	15,3	13,9	12,5
12	16,3	14,7	16,2	18,1
13	19,9	16,9	17,3	17,3
14	14,4	18,0	17,4	17,1
15	18,7	—	—	—
16	20,7	—	—	—

Как видно из таблицы, взвешивание по биномиальным коэффициентам дало почти во всех случаях промежуточные результаты между сглаживанием по простой скользящей средней и сглаживанием по скользящей средней с весами, полученными по полиному.

Дисперсия сглаженной по квадратичному полиному величины $\bar{y}_t$ (обозначим ее как σ_1^2) меньше дисперсии исходных наблюдений в интервале сглаживания (σ^2):

$$\sigma_1^2 = \frac{3(3p^2+3p-1)}{(2p-1)(2p+1)(2p+3)}\sigma^2,$$

откуда при $p=2$ $\sigma_1^2 = \frac{17}{35}\sigma^2$, при $p=3$ $\sigma_1^2 = \frac{1}{3}\sigma^2$.

Поскольку скользящая средняя заменяет центральный член каждых m уровней, то для p начальных и конечных уровней она не может быть подсчитана, например при сглаживании по трехлетней скользящей средней теряется первый и последний уровень. В большинстве случаев (особенно при анализе достаточно протяженного ряда) с потерей нескольких уровней легко примириться. Если же сглаженные концевые уровни представляют непосредственный интерес для исследователя, то их значения можно определить по полиному, на котором базировалась скользящая средняя. Для этого, естественно, надо оценить значения всех параметров этого полинома.

В предыдущем примере ряд сглаживался по взвешенной скользящей средней, равной свободному члену квадратичной параболы Найдем теперь сглаженные значения двух последних уровней ряда — $\bar{y}_{15}$ и $\bar{y}_{16}$ Для этого определим параметры b и c параболы, аппроксимирующей пять последних уровней ряда· 16,3, 19,9, 14,4, 18,7, 20,7 На основе нормальных уравнений, приведенных на с 279, и значения свободного члена $a = 17{,}057$ (в табл. 14.4 он дан с округлением до

десятичного знака) находим

$$b = \frac{\sum yi}{\sum i^2} = \frac{7{,}6}{10} = 0{,}076; \quad c = \frac{\sum y - am}{\sum i^2} = \frac{4{,}72}{10} = 0{,}472,$$

откуда по уравнению $\bar{y}_t = 17{,}057 + 0{,}076i + 0{,}472i^2$ находим $\bar{y}_{15} = 17{,}6$, $\bar{y}_{16} = 20{,}0$

Расчет скользящей средней представляется простой операцией, имеющей вполне ясное содержание. Однако эта операция трансформирует (фильтрует) временной ряд в большей мере, чем это кажется на первый взгляд. Если до сглаживания уровни ряда были независимыми, то после этого преобразования последовательные расчетные уровни находятся в некоторой зависимости между собой.

Экспоненциальные средние. Линейные фильтры (простые и взвешенные скользящие средние) симметричные. В связи с развитием методов прогнозирования в статистике были разработаны и *асимметричные фильтры.* Простейшим из них может служить скользящая средняя, которая заменяет не центральный, а последний уровень ряда в интервале сглаживания:

$$\bar{y}_t = \frac{1}{m} \sum_{r=0}^{m} y_{t-r}. \tag{14.10}$$

Преобразовав (14.10), получим рекуррентную формулу

$$\bar{y}_t = \bar{y}_{t-1} + \frac{y_t - y_{t-m}}{m}. \tag{14.11}$$

Если первый член суммы в формуле (14.11) несет в себе «груз прошлого» и характеризует инерцию развития, то второй член играет роль элемента, адаптирующего среднюю к новым условиям: средняя с каждым шагом во времени как бы обновляется, впитывая в себя новую информацию о фактически реализуемом процессе. Степень обновления определяется постоянным весом, равным $1/m$.

К определению асимметричных скользящих средних можно подойти, учитывая степень «устаревания» данных. Данная задача может быть решена с помощью применения системы весов, в которой информации приписывают вес, соответствующий степени ее новизны. Один из приемов сглаживания ряда с учетом «устаревания» данных заключается в расчете *экспоненциальных средних* по формуле, которую легко получить на основе общей записи асимметричного линейного фильтра (14.10), где $m = t$, а $a_r = \alpha(1-\alpha)^r$. Отсюда

$$\bar{y}_t = Q_t = \sum_{r=0}^{t} \alpha(1-\alpha)^r y_{t-r}, \tag{14.12}$$

где Q_t — экспоненциальная средняя (сглаженное значение уровня ряда) на момент t; α — коэффициент, характеризующий вес текущего наблюдения (параметр сглаживания), $0 < \alpha < 1$. Распишем

(14.13) по элементам суммы:

$$Q_t = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \ldots + \alpha(1-\alpha)^t y_0. \tag{14.13}$$

Суммируя все члены, начиная со второго, находим

$$Q_t = \alpha y_t + (1-\alpha)[\alpha y_{t-1} + \alpha(1-\alpha) y_{t-2} + \ldots + \alpha(1-\alpha)^{t-1} y_0]. \tag{14.14}$$

Второе слагаемое согласно (14.12) равно экспоненциальной средней для момента $t-1$, следовательно,

$$Q_t = \alpha y_t + (1-\alpha) Q_{t-1}. \tag{14.15}$$

Иначе говоря, средний уровень ряда на момент t равен линейной комбинации фактического уровня для этого же момента и среднего уровня предыдущего периода. Рассмотренная средняя имеет экспоненциально распределенные веса. Так, если $\alpha = 0{,}1$, то веса, придаваемые данным, составят: 0,1 — для текущего наблюдения; $0{,}1(1-0{,}1) = 0{,}09$ — для $t-1$; для члена $t-2$ получим вес $0{,}1(1-0{,}1)^2 = 0{,}081$ и т. д. При достаточно большом удалении в прошлое от момента t вес соответствующего уровня ряда будет настолько мал, что он практически не окажет сколь-нибудь заметного влияния на значение Q_t.

Экспоненциальная средняя обладает одним ценным свойством. Она постоянно адаптируется к новым условиям (при движении во времени). Это видно из следующей формы ее записи, которую можно получить из (14.15), перегруппировав ее члены:

$$Q_t = Q_{t-1} + \alpha(y_t - Q_{t-1}). \tag{14.16}$$

В этом виде экспоненциальная средняя выступает как обобщение скользящей средней (14.11).

Выражение (14.16) часто используется в прогностических целях (для краткосрочного прогнозирования). Пусть Q_{t-1} рассматривается как прогноз на один шаг вперед, тогда разность $y_t - Q_{t-1}$ есть погрешность этого прогноза. Следовательно, прогноз для времени $t+1$ (величина Q_t) учитывает наблюдавшуюся в момент t ошибку прогноза.

Для расчета Q_1 по формуле (14.16) требуется величина Q_0, относящаяся к моменту времени предшествующему имеющемуся ряду. Иначе говоря, возникает проблема определения начальных условий. Если есть соответствующие данные, то в качестве Q_0 возьмем некоторое среднее значение уровней, относящихся к прошлому. Если же этой величины нет, то в качестве начального можно практически принять первый уровень ряда, т. е. $Q_0 = y_1$. Вес, приписываемый этому уровню, уменьшается по мере отдаления от первого члена ряда, вместе с этим уменьшается его влияние на размер экспоненциальной средней. При высоких значениях α процесс уменьшения влияния происходит весьма быстро.

Для демонстрации расчета экспоненциальных средних обратимся к ряду показателей урожайности В качестве начального значения экспоненциальной средней возьмем величину y_1 Приняв, что коэффициент α равен, допустим, 0,1, получим $Q_1 = y_1 = 10{,}3$; $Q_2 = 0{,}1 \cdot 14{,}3 + (1 - 0{,}1)\ 10{,}3 \approx 10{,}7$ и т. д. Результаты расчета для трех вариантов коэффициента α представлены в табл. 14.5.

Таблица 14.5

Год	Урожайность, ц/га	Экспоненциальные средние		
		$\alpha=0{,}1$	$\alpha=0{,}2$	$\alpha=0{,}3$
1	10,3	10,3	10,3	10,3
2	14,3	10,7	11,1	11,5
3	7,7	10,4	10,4	10,4
4	15,8	10,9	11,5	12,0
5	14,4	11,3	12,1	12,7
6	16,7	11,8	13 0	13 9
7	15,3	12,2	13,5	14,3
8	20,2	13,0	14,8	16,8
9	17,1	13,4	15.3	16,4
10	7,7	12,8	13,8	13,8
11	15,3	13,1	14,1	14,2
12	16,3	13,4	14,5	14,9
13	19,9	14,0	15,6	16,4
14	14,4	14,1	15,4	15,8
15	18,7	14,5	16,0	16,7
16	20,7	15,2	17,0	17,9

Можно показать, что математическое ожидание $M(Q_t) = M(y_t)$. Кроме того, дисперсия экспоненциальных средних (сглаженных уровней) меньше, чем дисперсия исходных наблюдений:

$$\sigma_Q^2 = \frac{\alpha}{2 - \alpha}\,\sigma^2, \qquad (14.17)$$

где σ_Q^2 — дисперсия сглаженных уровней; σ^2 — дисперсия значений y_t. Как видно из (14.17), при большом значении α дисперсия экспоненциальных средних незначительно отличается от дисперсии наблюдений. Чем меньше α, тем больше этот коэффициент играет роль фильтра, поглощающего колебания (см. табл. 14.3).

Выше уже отмечалось, что α может находиться в пределах (0; 1). Однако на практике α чаще всего берут в диапазоне (0,1; 0,3). При выборе значения сглаживающего параметра можно поступать следующим образом: определить такое значение α, при котором сумма весовых коэффициентов (число которых равно d), ближайших к члену ряда y_t, была бы равной или близкой некоторой величине A. Иначе говоря, возникает задача определения α по заданной длине интервала d. Последняя же определяется с учетом колеблемости ряда и периода упреждения, принятого при прогнозировании. Найдем зависимость α от d и A. Поскольку веса составляют ряд α, $\alpha(1-\alpha)$, $\alpha(1-\alpha)^2$, ..., $\alpha(1-\alpha)^d$, то их сумма представляет собой сумму членов геометрической прогрессии с

первым членом α и знаменателем β, где $\beta = 1 - \alpha$. В силу этого

$$\sum_{t=0}^{d} \alpha\beta^t = \alpha \frac{\beta^d - 1}{\beta - 1}.$$

Приравняем значение суммы прогрессии заданной величине A и решим полученное уравнение относительно α:

$$\alpha = 1 - \sqrt[d]{1 - A}.$$

Пусть $A = 0{,}9$ (т. е. интервал, равный d, должен охватывать 90% суммы значений весовых коэффициентов). В этом случае получим следующие значения α (табл. 14.6):

Таблица 14.6

d	α	d	α	d	α
3	0,536	7	0,280	12	0,174
4	0,438	8	0,250	14	0,152
5	0,369	9	0,226	16	0,134
6	0,319	10	0,206	20	0,109

Из приведенных в таблице данных следует, что при $\alpha \approx 0{,}2$ десять членов ряда практически полностью определяют значение экспоненциальной средней.

Метод последовательных разностей. Особым видом сглаживания, которое позволяет выделить случайные колебания, если тренд следует полиномиальной тенденции, является вычисление последовательных разностей. Метод основывается на предположении о том, что уровень ряда может быть представлен как сумма двух компонент:

$$y_t = \hat{y}_t + \varepsilon_t, \tag{14.18}$$

где $\hat{y}_t$ — систематическая, а ε_t — случайная компонента. Систематическая компонента определяется полиномом $\hat{y} = f(t)$. Случайные составляющие ε_t не коррелированы, имеют нулевые средние и дисперсию σ^2. Пусть тренд строго следует полному степени λ. Разности ординат порядка λ тогда равны друг другу, а разности ординат порядка $\lambda + 1$ равны нулю. Например, если $\lambda = 1$, то

$$\hat{y}_2 - \hat{y}_1 = a + 2b - (a + b) = b; \quad \hat{y}_3 - \hat{y}_2 = a + 3b - (a + 2b) = b$$

и т. д.

В свою очередь последовательные разности наблюдаемых уровней при линейном тренде равны:

$$y_2 - y_1 = b + (\varepsilon_2 - \varepsilon_1); \quad y_3 - y_2 = b + (\varepsilon_3 - \varepsilon_2)$$

и т. д.

Аналогичные по смыслу выражения можно записать для случаев, когда $f(t)$ представляют собой полиномы более высоких сте-

пеней. Из сказанного следует, что примерное равенство последовательных разностей уровней ряда рассматривается как *симптом* того, что y_t следует в своем развитии полиному соответствующей степени. При применении этого метода исчисляют первые, вторые и т. д. разности уровней ряда, т. е. $u_t = y_t - y_{t-1}$, $u_t^{(2)} = u_t - u_{t-1}$, $u_t^{(3)} = u_t^{(2)} - u_{t-1}^{(2)}$ и т. д. Расчет ведется до тех пор, пока разности не будут примерно равными друг другу. Порядок разностей принимается за степень выравнивающего полинома. Такой подход далеко не универсален, он возможен при подборе только кривых, описываемых многочленами. К тому же его предпосылки могут и не быть адекватными рассматриваемому реальному процессу.

14.3. Автокорреляция

Среди методов, характеризующих свойства временного ряда, особое место отводится вычислению *коэффициентов автокорреляции* (r_τ). Последние представляют собой коэффициенты корреляции, измеряющие взаимосвязь исходного ряда и этого же ряда, но сдвинутого на τ шагов во времени (с лагом равным τ). Последовательность коэффициентов автокорреляции r_τ, $\tau = 1, 2, \ldots, n$, дает достаточно глубокое представление о внутренней структуре изучаемого процесса.

Для начала найдем коэффициент автокорреляции при условии, что $\tau = 1$, ряд сдвинут на один шаг во времени. Таким образом, получим два ряда переменных: $y_1, y_2, \ldots, y_{n-1}$ и $y_2, y_3, \ldots, y_n$, т. е. для расчета коэффициента корреляции имеются следующие пары наблюдений: (y_1, y_2), (y_2, y_3), $\ldots$, (y_{n-1}, y_n), всего $n-1$ пар. Заменив в формуле (12.23) обозначения, получим

$$r_1 = \frac{\sum_{t=1}^{n-1} (y_t - \bar{y}_1)(y_{t+1} - \bar{y}_2)}{\sqrt{\sum_{1}^{n-1} (y_t - \bar{y}_1) \sum_{2}^{n} (y_{t+1} - \bar{y}_2)^2}}. \qquad (14.19)$$

Здесь

$\bar{y}_1 = \frac{1}{n-1} \sum_{1}^{n-1} y_t$ — средний уровень первого ряда;

$\bar{y}_2 = \frac{1}{n-1} \sum_{2}^{n} y_t$ — средний уровень второго ряда.

Воспользовавшись выражениями, выведенными в гл. 12 для определения суммы квадратов отклонений и ковариаций, получим следующую рабочую формулу:

$$r_1 = \frac{\sum_{1}^{n-1} y_t y_{t+1} - (n-1)\bar{y}_1 \bar{y}_2}{\sqrt{\left[\sum_{1}^{n-1} y_t - (n-1)\bar{y}^2\right]\left[\sum_{2}^{n} y_t^2 - (n-1)\bar{y}_2^2\right]}}. \qquad (14.20)$$

Для запаздывания на τ шагов во времени имеем

$$r_\tau = \frac{\sum_{t=1}^{n-\tau} (y_t - \bar{y}_1)(y_{t+\tau} - \bar{y}_2)}{\sqrt{\sum_{1}^{n-\tau} (y_t - \bar{y}_1)^2 \sum_{\tau}^{n} (y_t - \bar{y}_2)^2}}, \qquad (14.21)$$

где

$$\bar{y}_1 = \frac{1}{n-\tau} \sum_{1}^{n-\tau} y_t; \quad \bar{y}_2 = \frac{1}{n-\tau} \sum_{\tau}^{n} y_t.$$

Рабочую формулу запишем следующим образом:

$$r_\tau = \frac{\sum_{1}^{n-\tau} y_t y_{t+\tau} - (n-\tau)\bar{y}_1\bar{y}_2}{\sqrt{\left[\sum_{1}^{n-\tau} y_t^2 - (n-\tau)\bar{y}_1^2\right]\left[\sum_{\tau}^{n} y_t^2 - (n-\tau)\bar{y}_2^2\right]}}. \qquad (14.22)$$

При расчете r_τ следует помнить, что с увеличением τ число пар наблюдений соответственно уменьшается. Как было показано в

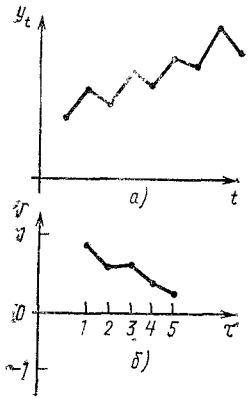


Рис. 14.4.
Коррелограмма ряда y_t

гл. 12, при небольшом числе наблюдений значимыми оказываются лишь высокие значения коэффициента корреляции (например, при $n = 12$ и уровне значимости 0,05 только $r > 0{,}576$ оказываются значимыми). Отсюда следует, что наибольшее значение τ должно быть таким, чтобы число пар наблюдений оказалось достаточным для вычисления r_τ. В практике ориентируются на правило, согласно которому $\tau \leqslant \frac{n}{4}$. При большой протяженности исследуемого ряда расчет r_τ можно упростить. Для этого находят отклонения не от средних коррелируемых рядов, а от общей средней всего ряда,

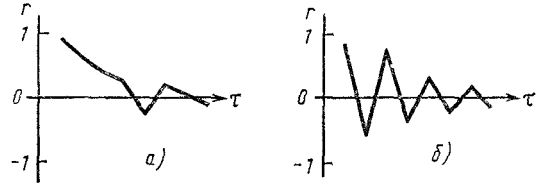
В этом случае

$$r_\tau \approx \frac{\sum_{1}^{n-\tau} (y_t - \bar{y})(y_{t+\tau} - \bar{y})}{\sum_{1}^{n} (y_t - \bar{y})^2}, \qquad (14.23)$$

где $\bar{y} = \sum y_t : n$.

График, на котором показаны значения r_τ для каждой данной величины τ, называется *коррелограммой*. Анализ коррелограмм оказывается весьма полезным при вскрытии закономерностей развития, описываемых временным рядом. Интерпретация коррело-

Рис. 14.5. Коррелограммы



грамм требует определенного навыка и не всегда легко осуществима. Рассмотрим несколько типов развития и соответствующие им коррелограммы:

а. *Нестационарный ряд.* В случае, когда ряд имеет тренд и относительно небольшие колебания вокруг него (рис. 14.4, *а*), коррелограмма при тенденции ряда к росту показывает систематическое понижение r_τ по мере увеличения τ. В этом случае все значимые r_τ больше нуля (см. рис. 14.4, *б*).

б. *Полностью случайный ряд.* Если уровни ряда суть случайные числа, то для больших n и любых положительных τ, как правило, $r \approx 0$.

в. *Краткосрочные корреляции.* Коррелограмма стационарных рядов зачастую показывает высокие значения r_1, причем следующие значения r_τ имеют тенденцию к уменьшению по мере роста τ и близки к нулю (затухают) при больших τ. Иначе говоря, наблюдается корреляция смежных уровней ряда (рис. 14.5, *а*). В этом случае хорошее описание ряда дает так называемая *авторегрессионная модель.*

г. *Ряд с чередованием роста и падения.* Если уровни ряда последовательно характеризуются положительными и отрицательными отклонениями от общей средней, то коррелограмма показывает чередование положительных и отрицательных последовательных значений r_τ. Последние стремятся к нулю по мере роста τ, если, разумеется, ряд стационарен (рис. 14.5, *б*).

В чисто иллюстративных целях рассчитаем коэффициенты автокорреляции для ряда, представленного в табл. 14.5. Поскольку ряд этот весьма короток и для расчета коэффициентов автокорреляции мало пригоден, несколько увеличим его, добавив следующие уровни: 21,2; 17,5; 22,5; 20,5. Таким образом, общее число членов ряда теперь равно 20. Расчет коэффициентов r_τ для $\tau = 1 \div 6$ выполним по формуле (14.22). Все необходимые для расчета данные приведены в табл. 14.7, в которой также показаны полученные значения r_τ.

Таблица 14.7

τ	$n-\tau$	$\bar{y}_1$	$\bar{y}_2$	$\sum_1^{n-\tau} y_t^2$	$\sum_1^{n-\tau} y_{t+\tau}^2$	$\sum_1^{n-\tau} y_t y_{t+\tau}$	r_τ
1	19	16,0789	16,6158	5226,41	5540,57	5173,21	0,308
2	18	15,7222	16,7444	4720,16	5336,08	4827,80	0,318
3	17	15,6176	17,2770	4413,91	5276,79	4631,80	0,192
4	16	15,2687	17,3688	3964,47	5027,15	4344,88	0,469
5	15	14,9067	17,5667	3535,98	4819,79	3929,58	—0,106
6	14	14,6357	17,6643	3186,29	4557,35	3611,00	—0,045

Интерпретация полученных коэффициентов затруднительна, так как почти все коэффициенты (кроме $r_4 = 0{,}469$) оказались статистически незначимыми даже при уровне значимости, равном 10%. В основном это связано с тем, что период наблюдения оказался недостаточной протяженности. Однако коэффициент r_4 (оказавшийся значимым при уровне 10% и незначимым при уровне 5%) заставляет обратить внимание на возможность существования четырехлетнего повторения направлений в колебании урожайности.

14.4. Аналитическое выравнивание временных рядов

Как было показано выше, сглаживание ряда с помощью того или иного фильтра дает возможность с известным приближением выявить тренд развития. При этом исследователь получает ряд чисел — сглаженных уровней. Вместе с тем часто желательно обобщить тенденцию развития в виде некоторой функции времени и соответствующей ей кривой. Кривые, описывающие закономерности развития явлений во времени, получают путем *аналитического выравнивания* временных рядов. Выравнивание ряда с помощью тех или иных функций (т. е. их подгонка к данным) в большинстве случаев оказывается удобным средством описания эмпирических данных, характеризующих развитие во времени исследуемого явления. Найденная функция позволяет получить выравненные (расчетные), или, как их иногда не вполне правомерно называют, теоретические значения уровней динамического ряда. Процесс выравнивания состоит из двух основных этапов: выбора типа кривой, форма которой соответствует характеру изменения временного ряда, и определения числовых значений (статистическое оценивание) параметров кривой.

Вопрос о выборе типа кривой является основным при выравнивании ряда. При всех прочих равных условиях ошибка в решении этого вопроса оказывается более серьезной по своим последствиям (особенно для прогнозирования), чем ошибка, связанная со статистическим оцениванием параметров. Самым обоснованным подходом к решению проблемы был бы подход, при котором форма кривой определялась бы путем анализа, охватывающего внутреннюю логику изучаемого процесса, специфику и взаимозависимость с другими процессами и окружающими условиями. Однако, как правило, такой анализ в лучшем случае может вскрыть характер динамики лишь в самых общих чертах (например, дви-

жение равномерное, равноускоренное, с затуханием роста и т. д.). Хотя содержательный анализ и не является практическим инструментом при выборе формы кривой, он тем не менее обязательно предшествует и сопутствует эмпирическому подходу.

Существует несколько практических приемов, которые позволяют более или менее удовлетворительно выбрать адекватную действительному развитию форму кривой. Наиболее простой путь — визуальный — выбор формы на основе графического изображения временного ряда. Риск субъективного и произвольного выбора здесь очень велик. Результат выбора в значительной мере зависит от принятого масштаба графического изображения. Вместе с тем при относительно простой конфигурации тенденций развития и незначительных колебаниях визуальный подход дает вполне приемлемые результаты. Второй путь заключается в применении метода последовательных разностей, который рассмотрен выше (см. § 14.2). Порядок разностей, при котором они будут примерно одинаковыми, принимается за степень выравнивающего полинома. Так, если примерно близкими оказываются первые разности, то для выравнивания берется полином первой степени и т. д. Однако такой подход далеко не универсален, он возможен при подборе только кривых, описываемых многочленами.

К выбору формы кривой можно подойти и иначе, например исходя из значения принятого критерия. Обычно в качестве критерия принимают сумму квадратов отклонений фактических значений уровня от расчетных, полученных выравниванием. Выбирается кривая, которой соответствует минимальное значение критерия, поскольку чем меньше значение критерия, тем ближе примыкают к кривой данные наблюдений и кривая по предположению лучше описывает тенденцию развития. Однако, строго говоря, минимум суммы квадратов отклонений еще не означает, что тенденция развития описана наилучшим образом. В самом деле, к ряду, состоящему из m точек, можно так подобрать по крайней мере один многочлен степени $m-1$, что соответствующая кривая будет проходить через все m точек. Кроме того, существует множество многочленов с более высокими степенями, кривые которых также проходят через эти точки. Сумма квадратов отклонений будет, естественно, равна нулю. Таким образом, формально это «наилучшая» кривая, и она действительно наиболее точно описывает фактическую динамику развития явления в прошлом. Однако вряд ли правомерно говорить в этом случае о выделении тенденции развития.

Применение критерия для выбора формы кривой, по-видимому, даст практически пригодные результаты в том случае, если отбор будет проходить в два этапа. На первом этапе отбираются зависимости, пригодные с позиции содержательного подхода к задаче, в результате чего происходит ограничение круга потенциально приемлемых функций. На втором этапе для этих функций подсчитываются значения критерия и выбирается та из кривых, которой соответствует минимальное его значение.

Метод наименьших квадратов при оценивании параметров полиномов. Допустим, что уровень ряда может быть представлен в виде суммы двух слагаемых.

$$y_t = \hat{y}_t + \varepsilon_t, \tag{14.24}$$

где $\hat{y}_t$ — систематическая (структурная) составляющая, характеризующая основную тенденцию в изменении явления во времени. Предположим далее, что это изменение может быть охарактеризовано уравнением тренда. Если составляющая $\hat{y}_t$ формируется под воздействием основных для изучаемого явления факторов, то вторая составляющая отражает влияние второстепенных, временных факторов. Указанное воздействие приводит к отклонениям от основной тенденции. В ряде случаев эти отклонения можно рассматривать как чисто случайные (возмущения). Итак, если одно из слагаемых y_t случайная величина ε_t и есть основания утверждать, что выполняются все предпосылки регрессионного анализа, то модель (14.24) не будет отличаться от обычной регрессионной модели и основная тенденция может быть изучена методами регрессионного анализа. Практически, однако, предположения регрессионного анализа относительно ε_t (т. е. $M(\varepsilon_t) = 0$ и $\sigma^2(\varepsilon_t) = \text{const}$) не могут быть проверены, так как в каждый момент времени ε_t встречается один раз. В этом случае уравнение тренда можно оценить путем аппроксимации с помощью МНК. Дальнейшее обсуждение особенностей модели (14.24) отложим до § 14.5.

Пусть развитие явления необходимо аппроксимировать с помощью полиномиального тренда, иначе говоря, представить зависимую переменную y_t как функцию времени в виде многочлена

$$\hat{y}_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_\lambda t^\lambda. \tag{14.25}$$

Оценки параметров $a_0, a_1, \ldots, a_\lambda$ можно получить с помощью рассмотренного в гл. 12 и 13 метода наименьших квадратов. Развернутая запись системы нормальных уравнений, в которой переменные $x_0, x_1, \ldots, x_m$ заменены на $t, t^2, \ldots, t^\lambda$, имеет вид:

$$\begin{aligned}
\sum y_t &= a_0 n + a_1 \sum t + a_2 \sum t^2 + \ldots + a_\lambda \sum t^\lambda; \\
\sum y_t t &= a_0 \sum t + a_1 \sum t^2 + a_2 \sum t^3 + \ldots + a_\lambda \sum t^{\lambda+1}; \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\sum y_t t^\lambda &= a_0 \sum t^\lambda + a_1 \sum t^{\lambda+1} + a_2 \sum t^{\lambda+2} + \ldots + a_\lambda \sum t^{\lambda+\lambda},
\end{aligned} \tag{14.26}$$

где n — число членов в динамическом ряду. Суммирование здесь производится от $t = 1$ до $t = n$. Система (14.26), состоящая из $\lambda + 1$ уравнений, содержит в качестве известных величин Σy_t, $\Sigma y_t t$, ..., $\Sigma y_t t^\lambda$ и $\lambda + 1$ неизвестных величин a_j. Решение этой системы относительно $a_0, a_1, \ldots, a_\lambda$ и дает искомые значения параметров.

Обозначим последовательные параметры полиномов невысоких степеней как a, b, c, ... Тогда нормальные уравнения для оценивания параметров прямой примут вид:

$$\begin{aligned} \sum y_t &= an + b\sum t; \\ \sum y_t t &= a\sum t + b\sum t^2; \end{aligned} \tag{14.27}$$

для параболы второй степени:

$$\begin{aligned} \sum y_t &= an + b\sum t + c\sum t^2; \\ \sum y_t t &= a\sum t + b\sum t^2 + c\sum t^3; \\ \sum y_t t^2 &= a\sum t^2 + b\sum t^3 + c\sum t^4; \end{aligned} \tag{14.28}$$

наконец, для параболы третьей степени получим:

$$\begin{aligned} \sum y_t &= an + b\sum t + c\sum t^2 + d\sum t^3; \\ \sum y_t t &= a\sum t + b\sum t^2 + c\sum t^3 + d\sum t^4; \\ \sum y_t t^2 &= a\sum t^2 + b\sum t^3 + c\sum t^4 + d\sum t^5; \\ \sum y_t t^3 &= a\sum t^3 + b\sum t^4 + c\sum t^5 + d\sum t^6. \end{aligned} \tag{14.29}$$

Поскольку полиномы степени выше третьей при обработке временных рядов встречаются крайне редко, то нормальные уравнения для них приводить не будем. Величины Σt, Σt^2, ... не зависят от конкретных уровней ряда. Если ряд состоит из уровней, равно отстоящих друг от друга (а именно с такими рядами чаще всего имеет дело экономист), то суммы Σt, Σt^2 и т. д. являются функциями только числа членов ряда. Для них легко получить расчетные формулы. В частности,

$$\sum t = \frac{n(n+1)}{2};$$

$$\sum t^2 = \sum t\,\frac{2n+1}{3} = \frac{n(n+1)(2n+1)}{6};$$

$$\sum t^3 = \left(\sum t\right)^2 = \frac{n^2(n+1)^2}{4};$$

$$\sum t^4 = \sum t^2\,\frac{3n^2+3n-1}{5} = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30};$$

$$\sum t^5 = \sum t^3\,\frac{2n^2+2n-1}{3} = \frac{n^2(n+1)^2(2n^2+2n-1)}{12};$$

$$\sum t^6 = \sum t^2\,\frac{3n^4+6n^3-3n+1}{7}.$$

Во всех приведенных формулах суммирование производится от $t = 1$ до $t = n$.

Рассмотрим систему нормальных уравнений (14.27). Решение относительно искомых параметров a и b дает

$$a = \frac{\sum y_t \sum t^2 - \sum t \sum y_t t}{n \sum t^2 - (\sum t)^2}; \tag{14.30}$$

$$b = \frac{n \sum y_t t - \sum y_t \sum t}{n \sum t^2 - (\sum t)^2}. \tag{14.31}$$

Преобразовав (14.30) и (14.31), получим следующие удобные для расчетов выражения:

$$b = \frac{\sum y_t t - \frac{\sum y_t}{n} \sum t}{\sum t^2 - \frac{(\sum t)^2}{n}}; \tag{14.32}$$

$$a = \frac{\sum y_t}{n} - b \frac{\sum t}{n}. \tag{14.33}$$

Системы нормальных уравнений существенно упрощаются, если начало отсчета времени перенести в середину ряда (разумеется, к этим упрощениям нет необходимости прибегать, когда расчет выполняется на ЭВМ). В этом случае упрощаются не только сами уравнения, но и уменьшаются абсолютные значения величин, участвующих в расчете. После переноса начала отсчета времени параметры времени имеют следующие значения: $t = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$, если число членов ряда нечетное (если же число ряда четное, то удобно приписать t следующие значения: $\ldots -5, -3, -1, 1, 3, 5, \ldots$). Следовательно, Σt и все Σt^{λ}, у которых λ нечетное число, равны нулю. Системы нормальных уравнений теперь имеют вид:
для прямой:

$$\begin{aligned} \sum y_t &= an; \\ \sum y_t t &= b \sum t^2; \end{aligned} \tag{14.34}$$

для параболы второй степени:

$$\begin{aligned} \sum y_t &= an + c \sum t^2; \\ \sum y_t t &= b \sum t^2; \\ \sum y_t t^2 &= a \sum t^2 + c \sum t^4; \end{aligned} \tag{14.35}$$

для параболы третьей степени:

$$\begin{aligned} \sum y_t &= an + c \sum t^2; \\ \sum y_t t &= b \sum t^2 + d \sum t^4; \\ \sum y_t t^2 &= a \sum t^2 + c \sum t^4; \\ \sum y_t t^3 &= b \sum t^4 + d \sum t^6. \end{aligned} \tag{14.36}$$

Решив системы относительно неизвестных параметров, получим оценки параметров соответствующих полиномов:
прямой:

$$a = \frac{\sum y_t}{n};\quad b = \frac{\sum y_t t}{\sum t^2};$$

параболы второй степени:

$$a = \frac{\sum y_t}{n} - \frac{\sum t^2}{n}\left[\frac{n\sum y_t t^2 - \sum t^2 \sum y_t}{n\sum t^4 - \left(\sum t^2\right)^2}\right];$$

$$b = \frac{\sum y_t t}{\sum t^2};$$

$$c = \frac{n\sum y_t t^2 - \sum t^2 \sum y_t}{n\sum t^4 - \left(\sum t^2\right)^2};$$

параболы третьей степени:

$$a = \frac{\sum y_t}{n} - \frac{\sum t^2}{n}\left[\frac{n\sum y_t t^2 - \sum t^2 \sum y_t}{n\sum t^4 - \left(\sum t^2\right)^2}\right];$$

$$b = \frac{\sum y_t t}{\sum t^2} - \frac{\sum t^4}{\sum t^2}\left[\frac{\sum t^2 \sum y_t t^3 - \sum t^4 \sum y_t t}{\sum t^2 \sum t^6 - \left(\sum t^4\right)^2}\right];$$

$$c = \frac{n\sum y_t t^2 - \sum t^2 \sum y_t}{n\sum t^4 - \left(\sum t^2\right)^2};$$

$$d = \frac{\sum t^2 \sum y_t t^3 - \sum t^4 \sum y_t t}{\sum t^2 \sum t^6 - \left(\sum t^4\right)^2}.$$

Значение Σt^2, Σt^4, ... можно получить по следующим формулам:
для нечетного n:

$$\sum t^2 = \frac{(n-1)\,n\,(n+1)}{12};$$

$$\sum t^4 = \sum t^2 \frac{3n^2 - 7}{20};$$

$$\sum t^6 = \sum t^2 \frac{3n^4 - 18n^2 + 31}{112}$$

(суммирование производится от $t = -p$ до $t = p$, где $p = (n - 1)/2$);
для четного n:

$$\sum t^2 = \frac{(n-1)\,n\,(n+1)}{3};$$

$$\sum t^4 = \sum t^2 \frac{3n^2 - 7}{5};$$

$$\sum t^6 = \sum t^2 \frac{3n^4 - 18n^2 + 31}{7}$$

(суммирование производится от $t = -p$ до $t = p$, но $p = n - 1$).

Обсудим теперь вопрос о влиянии переноса начала отсчета времени на значения параметров кривых. Рассмотрим полиномы первой и второй степени. Поскольку параметры b и c характеризуют, так сказать, скорость роста и ускорение — они определяют форму кривой,— то перенос начала времени на их значении не отразится (если не изменены единицы измерения времени). Что же касается параметра a, то он характеризует значение $\hat{y}_t$ при $t=0$. Соответственно перенос этой точки во времени скажется на значении этого параметра (сдвиг кривой по вертикали).

Таблица 14.8

t	y_t	$y_t t$	$\hat{y}_t$
−8	54,1	−432,8	45,7
−7	35,4	−247,8	47,0
−6	56,6	−339,6	48,3
−5	46,6	−233,0	49,6
−4	46,7	−186,8	50,8
−3	52,1	−156,3	52,1
−2	56,6	−113,2	53,4
−1	44,8	−44,8	54,7
0	68,3	0	56,0
1	36,3	36,3	57,2
2	75,0	150,0	58,5
3	57,2	171,6	59,8
4	69,0	276,0	61,1
5	55,5	277,5	62,4
6	73,3	439,8	63,7
7	64,1	448,7	65,0
8	60,0	480,0	66,3
Итого	951,6	525,6	951,6

Запишем по два уравнения прямой и параболы:

$$\hat{y}_t = a + bt;\ \hat{y}_t = a^* + bt^*;$$

$$\hat{y}_t = a + bt + ct^2;$$

$$\hat{y}_t = a^* + bt^* + ct^{*2},$$

где a^* — значение параметра a при переносе начала отсчета времени; t^* — время при переносе начала его отсчета. Отсюда

$$a = a^* - b(t - t^*);$$

$$a = a^* - b(t - t^*) - c(t^2 - t^{*2}).$$

В частности, при переносе начала отсчета времени в середину ряда с нечетным числом членов получим:

$$a = a^* - b\frac{n+1}{2};$$

$$a = a^* - b\frac{n+1}{2} - c\left(\frac{n+1}{2}\right)^2. \qquad (14.37)$$

Ежегодные затраты предприятия на первичную обработку сырья характеризуются следующими данными (тыс. руб.): 54,1; 35,4; 56,6; 46,6; 46,7; 52,1; 56,6; 44,8; 68,3; 36,3; 75,0; 57,2; 69,0; 55,5; 73,3; 64,1; 60,0. Как видно из приведенных данных, характеристики ряда существенно варьируют во времени. Без выравнивания ряда тенденция здесь трудно определима.

Подберем для этих данных уравнение прямой и параболы. Расчет соответствующих сумм для оценки параметров прямой представлен в табл. 14.8 (начало отсчета времени перенесено в середину ряда). Найдем теперь Σt^2 и $n\Sigma t^4 - (\Sigma t^2)^2$. Для $n = 17$ эти величины составят 408 и 131780. Окончательно находим

$$a = \frac{951,6}{17} = 55,976, \qquad b = \frac{525,6}{408} = 1,288.$$

Уравнение прямой с оцененными параметрами имеет вид:

$$\hat{y}_t = 55{,}976 + 1{,}288t, \quad t = -8, \ -7, \ \ldots, \ 7, \ 8.$$

Таким образом, обнаружена тенденция к росту затрат — в среднем около 1,3 тыс. руб. в год. Выравненные по прямой значения показаны в табл. 14.8 и на рис. 14.6. Если необходимо получить уравнение прямой для $t = 1, 2, \ldots, 17$,

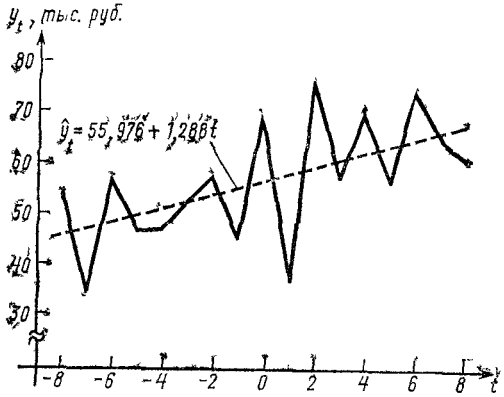


Рис. 14.6.
Выравнивание ряда по прямой

то воспользовавшись (14 37), находим $a = 55{,}976 - 1{,}288 \cdot 9 = 44{,}384$; $\hat{y}_t = 44{,}384 + 1{,}288t$.

Для выравнивания по параболе находим

$$\sum y_t t^2 = 22849{,}2; \quad \sum t^4 = 17\,544.$$

Откуда

$$a = \frac{951{,}6}{17} - \frac{408}{17}\left(\frac{17 \cdot 22849{,}2 - 408 \cdot 951{,}6}{131\,780}\right) = 55{,}943;$$

$$b = \frac{525{,}6}{408} = 1{,}288; \quad c = \frac{17 \cdot 22849{,}2 - 408 \cdot 951{,}6}{131\,780} = 0{,}00139$$

Таким образом,

$$\hat{y}_t = 55{,}943 + 1{,}288t + 0{,}00139t^2.$$

Введение квадратичного члена в данном случае оказалось практически бесполезным, так как точность описания явления при этом не увеличилась. Это легко понять даже при простом сопоставлении коэффициентов. Коэффициент квадратичного члена на несколько порядков меньше, чем коэффициент линейного члена. Заранее можно сказать, что применение дисперсионного анализа для проверки криволинейности (см. § 13 6) покажет несущественность введения в уравнение квадратичного члена Достаточно сказать, что сумма квадратов отклонений от линейного тренда равна 1520, а сумма квадратов отклонений от параболической кривой отличается от нее лишь на несколько единиц.

Оценивание параметров экспоненты и гиперболы. Выше было показано, как с помощью МНК оцениваются параметры многочленов. Однако далеко не всегда многочлены адекватно описывают фактически сложившуюся тенденцию. Например, процессы, имеющие, так сказать, «лавинообразный» характер, хорошо описываются *экспоненциальными* кривыми. Под лавинообразными процессами будем понимать развитие, при котором прирост в основном зависит от уже достигнутого уровня и при этом ограничения (например, фактор насыщения) не играют заметной

роли. Итак, пусть уровень описывается моделью

$$y_t = ab^t \varepsilon_t.$$

Тогда выравнивание ряда производится по *экспоненте*

$$\hat{y}_t = ab^t.$$

Логарифмируя это выражение, получим

$$\log \hat{y}_t = \log a + t \log b.$$

Приняв $y_t^* = \log \hat{y}_t$, $\alpha = \log a$ и $\beta = \log b$, запишем уравнение прямой

$$y_t^* = \alpha + \beta t.$$

Нормальное уравнение определим на основе минимизации $Q = \Sigma (\log y_t - \log \hat{y}_t)^2$. Откуда

$$\sum \log y_t = \alpha n + \beta \sum t;$$
$$\sum (\log y_t) t = \alpha \sum t + \beta \sum t^2.$$

Получив α и β, нетрудно перейти к оценкам параметров a и b. Рассмотренный метод является общераспространенным. Заметим, однако, что логарифмическое преобразование исходного уравнения не является столь безобидным, как это может показаться на первый взгляд. В самом деле, можно показать, что в этом случае* $Q \approx \sum \frac{\varepsilon_t^2}{y_t}$. Следовательно, с увеличением t ошибка ε_t играет меньшую роль в формировании Q, если y_t имеют тенденцию к росту. Наоборот, если уровни в общем уменьшаются во времени, то ε_t увеличивают свой вклад в Q по мере роста t. Таким образом, оценки параметров оказываются смещенными. При относительно небольшой скорости изменения y_t указанное смещение будет незначительно.

Найдем параметры экспоненты, выравнивающей ряд, приведенный в табл. 14 1, по формулам (14 32) и (14 33), заменив в них y_t на $\ln y_t$, а также a и b на α и β.

$$\beta = \frac{\sum (\ln y_t) t - \frac{\sum \ln y_t}{n} \sum t}{\sum t^2 - \frac{(\sum t^2)^2}{n}};$$

$$\alpha = \frac{\sum \ln y_t}{n} - \beta \frac{\sum t}{n}.$$

Поскольку $n = 12$, то $\Sigma t = 78$, $\Sigma t^2 = 650$.

Остальные необходимые данные получим из табл. 14.9.

* Подробнее об этом см [4].

Таблица 149

t	y_t	$\ln y_t$	$(\ln y_t)\, t$	$\hat{y}_t$
1	87,7	4,4739	4,47	96,08
2	97,7	4,5819	9,16	100,06
3	108,5	4,6868	14,06	104,21
4	107,9	4,6812	18,72	108,53
5	111,0	4,7095	23,55	113,03
6	122,7	4,8097	28,86	117,72
7	135,4	4,9082	34,36	122,60
8	143,5	4,9663	39,73	127,67
9	132,5	4,8866	43,98	132,99
10	133,7	4,8956	48,96	138,50
11	134,2	4,8993	53,89	144,24
12	143,4	4,9656	59,59	150,23
Итого	1458,2	57,46	379,33	1455,9

Таким образом,

$$\beta = \frac{379{,}33 - \frac{57{,}46}{12} \cdot 78}{650 - \frac{(78)^2}{12}} = 0{,}04065;$$

$$\alpha = \frac{57{,}46}{12} - 0{,}04065 \cdot \frac{78}{12} = 4{,}52447.$$

Потенцируя эти результаты находим $a = 92{,}247$ и $b = 1{,}04148$. Уравнение экспоненциального тренда имеет вид:

$$\hat{y}_t = 92{,}247 \cdot 1{,}0415^t, \quad t = 1, 2, \ldots$$

Оценка параметров *гиперболы*

$$\hat{y}_t = a + b\frac{1}{t}$$

осуществляется достаточно просто. Заменив в этом выражении величину $\frac{1}{t}$ на z, получим уравнение прямой, параметры которой легко оценить с помощью МНК. При этом, однако, нельзя пользоваться упрощающими приемами (переносом начала отсчета времени), так как они предполагают, что значения независимой переменной t характеризуются членами натурального ряда чисел, что нельзя сказать о z. Указанное линейное преобразование не приводит к смещению оценок.

Оценивание параметров модифицированной экспоненты и логистической кривой. Выше говорилось о том, что развитие «лавинообразного» процесса хорошо описывается экспоненциальной кривой, если нет ограничивающего фактора. Если же ограничивающий фактор все время воздействует, причем его влияние усиливается вместе с ростом достигнутого уровня, то хорошее описание этого процесса можно получить с

помощью *модифицированной экспоненты* (рис. 14.7).

$$\hat{y}_t = k + ab^t. \tag{14.38}$$

В случаях, когда ограничивающий фактор начинает влиять только после некоторого момента (точка перегиба), до которого процесс развивался, следуя экспоненциальному закону, наилучшее приближение дают *S*-образные кривые. В сущности, *S*-образные кривые описывают два последовательных лавинообразных процес-

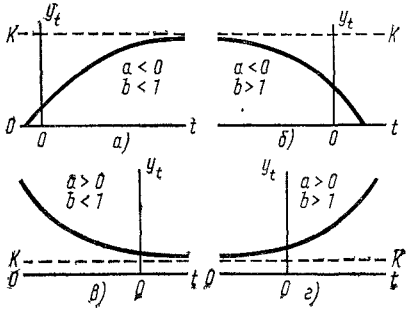


Рис. 14.7.
Графики модифицированной экспоненты

са: один — с ускорением развития, другой — с замедлением. К *S*-образным кривым относятся кривая Гомперца

$$\hat{y}_t = a^{b^t} \tag{14.39}$$

и логистическая кривая. Наиболее часто встречающаяся запись последней имеет вид:

$$\hat{y}_t = \frac{k}{1 + be^{at}}. \tag{14.40}$$

Модифицированная экспонента и *S*-образные кривые при определенных значениях своих параметров имеют асимптоты. Отсюда такие типы кривых пригодны для описания различного рода процесса «с насыщением». В частности, *S*-образные кривые находят применение в различного рода демографических и страховых расчетах, при обработке данных наблюдения в биологии, наконец, они оказались весьма полезными при решении отдельных задач прогнозирования научно-технического прогресса, так как хорошо описывают развитие во времени функциональных характеристик различных технических устройств (скорости самолетов, эффективность потребления угля на электростанциях и т. д.).

Попытаемся применить МНК для оценивания параметров приведенных кривых. Минимизируем сумму квадратов отклонений, например, логистической кривой (14.40):

$$Q = \sum \left(y_t - \frac{k}{1 + be^{at}} \right)^2.$$

Взяв производные относительно неизвестных параметров a, b и k и приравняв их нулю, получим трансцендентные уравнения, при решении которых неминуемо возникнут существенные трудности.

С аналогичной ситуацией столкнемся при попытках применения МНК для оценивания параметров модифицированной экспоненты и кривой Гомперца. Поэтому в подобных случаях следует так преобразовать уравнение кривой, чтобы оно стало линейным относительно параметров. После чего параметры этого выражения оцениваются МНК.

Простейшее преобразование модифицированной экспоненты и логистической кривой можно получить, если имеются значения их асимптот. Значение асимптоты можно в ряде случаев оценить или получить вне данного статистического наблюдения, например исходя из существа развития самого изучаемого явления и различного рода ограничений, сопутствующих ему: физические законы или известные физические постоянные, наличие природных ресурсов, законодательные акты и т. д. Наконец, это могут быть просто некоторые априорные суждения. Итак, мы полагаем, что исследователь располагает некоторой информацией о возможном значении асимптоты. Обозначим эту оценку как k^*. Тогда в выражении (14.38) заменим k на k^*. После чего нетрудно получить линейное в логарифмах уравнение модифицированной экспоненты

$$\hat{y}_t - k^* = ab^t.$$

Откуда

$$\hat{h}_t = \log(\hat{y}_t - k^*) = \log a + t \log b.$$

Параметры $\log a$ и $\log b$ теперь нетрудно оценить обычным методом наименьших квадратов, минимизируя $\sum (h_t - \hat{h}_t)^2$, где $h_t = \log(y_t - k^*)$. Соответственно получим нормальные уравнения:

$$\begin{aligned} \sum h_t &= n \log a + \log b \sum t; \\ \sum h_t t &= \log a \sum t + \log b \sum t^2. \end{aligned} \tag{14.41}$$

Уравнение логистической кривой (14.40) приводится к линейному виду следующим образом. Заменим k на k^*, после несложных преобразований получим

$$\frac{k^*}{\hat{y}_t} - 1 = be^{at},$$

и, наконец, применив натуральные логарифмы, находим

$$\hat{\beta}_t = \ln\left(\frac{k^*}{\hat{y}_t} - 1\right) = \ln b + at.$$

Параметры $\ln b$ и a этого линейного в логарифмах уравнения оцениваются с помощью МНК. Нормальные уравнения в этом случае имеют вид:

$$\begin{aligned} \sum \beta_t &= n \ln b + a \sum t; \\ \sum \beta_t t &= \ln b \sum t + a \sum t^2, \end{aligned} \tag{14.42}$$

где $\beta_t = \ln\left(\frac{k^*}{y_t} - 1\right)$.

Поскольку в только что рассмотренных случаях приведение к линейному виду уравнения кривой связано с логарифмированием

исходных уровней ряда (точнее, их преобразований), то это при ведет к некоторому смещению получаемых оценок (эта проблема кратко затрагивалась при обсуждении методов оценивания экспоненциальной кривой).

Пусть имеется ряд значений y_t, характеризующий динамику некоторого явления. Подберем для него логистическую кривую, принимая во внимание, что предельная величина этого явления равна 210 Часть необходимых для расчета параметров данных сведена в табл. 14 10 Кроме того, имеем $\Sigma t = 78$, $\Sigma t^2 = 650$. Решим систему нормальных уравнений (14 42) относительно a и $\ln b$, получим

$$a = \frac{87{,}0609 - \frac{19{,}9165}{12} \cdot 78}{650 - \frac{78^2}{12}} = -0{,}02965;$$

$$\ln b = \frac{19{,}9165}{12} - (-0{,}02965)\ \frac{78}{12} = 3{,}58682; \qquad b = 36{,}119.$$

Уравнение кривой имеет вид:

$$\hat{y}_t = \frac{210}{1 + 36{,}12e^{-0{,}2965t}}.$$

Рассчитанные по этому уравнению значения $\hat{y}_t$ показаны в последней графе таблицы.

Таблица 14.10

t	y_t	$\frac{210}{y_t} - 1$	$\beta_t = \ln\left(\frac{210}{y_t} - 1\right)$	$\beta_t t$	$\hat{y}_t$
1	7,62	26,5591	3,2794	3,2794	7,54
2	10,41	19,1729	2,9535	5,9070	10,02
3	12,75	15,4701	2,7389	8,2168	13,26
4	17,18	11,2235	2,4180	9,6720	17,45
5	22,00	8,5454	2,1454	10,7270	22,82
6	29,60	6,0946	1,8074	10,8444	29,59
7	39,88	4,2658	1,4506	10,1544	37,96
8	47,80	3,3933	1,2218	9,7744	48,06
9	62,50	2,3600	0,8587	7,7280	59,91
10	69,30	2,0303	0,7082	7,0818	73,37
11	87,30	1,4055	0,3404	3,7443	88,07
12	105,30	0,9943	—0,0057	—0,0686	103-48
Итого	511,64	—	19,9165	87,0609	511,53

Если значение асимптоты неизвестно, как это чаще всего и бывает, то оценка параметров может быть осуществлена более сложным путем опять-таки с помощью МНК. Для этого отыскивается такое тождественное преобразование уравнения кривой, которое будет линейным относительно параметров. Обратимся к уравнению модифицированной экспоненты. Напишем его для двух смежных моментов времени:

$$\hat{y}_{t+1} = k + ab^{t+1}; \quad \hat{y}_t = k + ab^t.$$

Определим прирост уровня

$$\hat{a}_t = \hat{y}_{t+1} - \hat{y}_t = ab^t(b-1).$$

Прологарифмировав последнее выражение, получим

$$\log \hat{a}_t = \log a + \log(b-1) + t \log b.$$

Пусть $g = \log a + \log(b-1)$. Тогда $\log \hat{a}_t = g + t \log b$. Задача, таким образом, состоит в оценивании параметров g и $\log b$ по данным наблюдения. Нормальные уравнения в этом случае имеют вид:

$$\begin{aligned} \sum \log u_t &= g(n-1) + \log b \sum t; \\ \sum (\log u_t) t &= g \sum t + \log b \sum t^2. \end{aligned} \tag{14.43}$$

Здесь n — число членов в исходном ряду, суммирование производится от $t=1$ до $n-1$. Имея оценки g и $\log b$, легко находим $\log a = g - \log(b-1)$ и затем параметр a. Рассмотренный метод применим лишь в тех случаях, когда ряд изменяется монотонно, т. е. тогда, когда значения приростов положительны и можно найти их логарифмы.

Что касается оценки параметров логистической кривой, то сперва с помощью регрессии оцениваются параметры k и a уравнения

$$\hat{y}_t = \frac{k}{1 + be^{at}},$$

а затем определяется параметр b. Существует целый ряд методов оценивания k и a. Рассмотрим один из них*. Прежде всего преобразуем логистическую функцию. Определим для этого разность между двумя обратными значениями соседних членов ряда, развивающихся по логистическому закону:

$$\frac{1}{\hat{y}_{t+1}} - \frac{1}{\hat{y}_t} = \frac{be^{at}(e^a - 1)}{k} = \frac{\left(\frac{k}{\hat{y}_t} - 1\right)(e^a - 1)}{k}$$

Отсюда

$$\frac{1}{\hat{y}_{t+1}} = \frac{1 - e^a}{k} + e^a \frac{1}{\hat{y}_t}.$$

Соответственно для оценки параметров $\frac{1-e^a}{k}$ и e^a определяется минимум $\sum \left(\frac{1}{y_t} - \frac{1}{\hat{y}_t}\right)^2$. Система нормальных уравнений в данном случае имеет вид:

$$\begin{aligned} \sum \frac{1}{y_{t+1}} &= \frac{1-e^a}{k}(n-1) + e^a \sum \frac{1}{y_t}; \\ \sum \frac{1}{y_{t+1}} \cdot \frac{1}{y_t} &= \frac{1-e^a}{k} \sum \frac{1}{y_t} + e^a \sum \frac{1}{y_t^2}. \end{aligned}$$

Суммирование здесь производится от $t=1$ до $t=n-1$. Решение системы относительно искомых параметров дает следующий

* Описание других методов оценивания приводится в работе [22].

результат:

$$e^a = \frac{(n-1)\sum \frac{1}{y_{t+1}} \cdot \frac{1}{y_t} - \sum \frac{1}{y_{t+1}} \cdot \sum \frac{1}{y_t}}{(n-1)\sum \frac{1}{y_t^2} - \left(\sum \frac{1}{y_t}\right)^2};$$

$$\alpha = \frac{1-e^a}{k} = \frac{\sum \frac{1}{y_{t+1}} \sum \frac{1}{y_t^2} - \sum \frac{1}{y_t} \sum \frac{1}{y_{t+1}} \cdot \frac{1}{y_t}}{(n-1)\sum \frac{1}{y_t^2} - \left(\sum \frac{1}{y_t}\right)^2}. \qquad (14.44)$$

Откуда

$$k = \frac{1-e^a}{\alpha}. \qquad (14.45)$$

Теперь оценим параметр b. Поскольку параметры a и k уже оценены, то поступим следующим образом. Преобразовав уравнение логистической кривой, получим

$$\frac{k}{\hat{y}_t} - 1 = be^{at},$$

откуда

$$\ln b = -at + \ln\left(\frac{k}{\hat{y}_t} - 1\right).$$

Данное уравнение может быть записано для каждого члена ряда, т. е. для $t = 1, \ldots, n$. Усредним обе части уравнения

$$\frac{\sum \ln b}{n} = \frac{-a\sum t}{n} + \frac{\sum \ln\left(\frac{k}{\hat{y}_t} - 1\right)}{n}.$$

Имея в виду, что при $t = 1, \ldots, n$

$$\sum t = \frac{n(n+1)}{2},$$

и полагая, что

$$\sum \ln\left(\frac{k}{\hat{y}_t} - 1\right) = \sum \ln\left(\frac{k}{y_t} - 1\right),$$

получим

$$\ln b = \frac{-a(n+1)}{2} + \frac{1}{n}\sum_1^n \ln\left(\frac{k}{y_t} - 1\right). \qquad (14.46)$$

Подберем логистическую кривую для данных предыдущего примера при условии, что асимптота не задана. Найдем необходимые для составления нормальных уравнений суммы:

$$\sum_1^{n-1} \frac{1}{y_t} = 0{,}531; \quad \sum_1^{n-1} \frac{1}{y_{t+1}} = 0{,}4093; \quad \sum_1^{n-1} \frac{1}{y_t^2} = 0{,}04086; \quad \sum_1^{n-1} \frac{1}{y_{t+1}} \frac{1}{y_t} = 0{,}0311.$$

Откуда на основе (14.44) получим $e^a = 0{,}7449$, $\alpha = 0{,}01244$. Таким образом,

$$\frac{1}{\hat{y}_{t+1}} = 0{,}001244 + 0{,}7449\frac{1}{\hat{y}_t}.$$

Поскольку $e^a = 0{,}7449$, то $a = \ln 0{,}7449 = -0{,}2945$; в свою очередь $\kappa = (1 - 0{,}7449) : 0{,}001244 = 205{,}06$.

Параметр b определим по формуле (14.46):

$$\ln b = \frac{0{,}2945\,(12+1)}{2} + \frac{1}{12}\sum_{1}^{12} \ln\left(\frac{205{,}06}{y_t} - 1\right) = 3{,}6406.$$

Откуда $b = 38{,}12$ и

$$\hat{y}_t = \frac{205{,}06}{1 + 38{,}12e^{-0{,}2945t}}.$$

14.5. Доверительные интервалы тренда

Выравнивание временного ряда и отыскание тренда дают возможность для определения основной тенденции развития явления во времени. Однако изучение тенденции развития, строго говоря, на этом не заканчивается, особенно если тренд используется как инструмент экстраполяции, г. е. для продления в будущее тенденции, наблюдавшейся в прошлом. Дело в том, что оценки параметров тренда не свободны от погрешностей, связанных, во-первых, с тем, что объем информации, на основе которой производилось оценивание, ограничен и в некотором смысле эту информацию можно рассматривать как выборку. Во всяком случае смещение периода наблюдения только на один шаг во времени, добавление или устранение членов ряда (например, иногда вместо годовых уровней для увеличения числа наблюдений берут квартальные показатели) приводит к изменению численных оценок параметров. Во-вторых, каждый член ряда содержит случайную компоненту. Отсюда значения $\hat{y}_t$ несут на себе груз неопределенности, связанной с ошибками параметров. Соответственно определение тренда $\hat{y}_t$ в большинстве случаев может сопровождаться определением доверительных интервалов (подобно тому, как регрессионная модель сопровождается характеристикой доверительного интервала). Характеристика тренда при этом существенно уточняется.

Прежде чем приступить к определению ошибки тренда и его доверительного интервала, необходимо сделать оговорку об условности приводимых ниже расчетов. Рассматриваемые методы получены перенесением результатов, найденных для регрессии выборочных показателей, на анализ динамических рядов. При этом предположение регрессионного анализа о нормальности распределения отклонений вокруг линии регрессии не может, по существу, ни утверждаться, ни быть проверено при анализе динамических рядов. Дискуссии в статистической литературе пролили свет на трудности, связанные с этой проблемой. В итоге, однако, принципиально новый подход не был найден. Все предложения так или иначе связаны с определением доверительного интервала на основе оценки среднего квадратического отклонения членов ряда от тренда.

В общем виде доверительный интервал для тренда определяется как

$$\hat{y}_t \pm t_\alpha s_{\hat{y}}, \qquad (14.47)$$

где $s_{\hat{y}}$ — средняя квадратическая ошибка тренда; $\hat{y}_t$ — расчетное (трендовое) значение y_t; t_α — значение t-статистики.

Поскольку, как это будет показано, в основе $s_{\hat{y}}$ лежит среднее квадратическое отклонение членов ряда от тренда (s), т. е. отклонение фактических наблюдений от расчетных, полученных при выравнивании ряда, найдем сперва s. По определению

$$s = \sqrt{\frac{\sum (y_t - \hat{y}_t)^2}{\nu}},$$

где y_t, $\hat{y}_t$ — фактическое и расчетное значения членов ряда; ν — число степеней свободы, определяемое в зависимости от числа наблюдений (n) и числа оцениваемых параметров кривой z, $\nu = n - z$. Так, если выравнивание производится по прямой, то $\nu = n - 2$, для параболы второго порядка $\nu = n - 3$ и т. д.

При большой протяженности ряда расчет $\sum (y_t - \hat{y}_t)^2$ достаточно трудоемок, поэтому можно рекомендовать прием, не требующий расчета отклонений для каждого момента времени. Он основан на следующем. Пусть Q_1 — сумма квадратов отклонений от тренда, Q_0 — сумма квадратов отклонений от средней. Тогда

$$Q_1 = Q_0 - \sum (\hat{y}_t - \bar{y})^2,$$

где $\hat{y}_t - \bar{y}$ — разность между значением тренда в точке t и средней.

Допустим, что тренд описывается уравнением прямой $\hat{y}_t = a + bt$. Определим величину $\sum (\hat{y}_t - \bar{y})^2$ для этого тренда. Учитывая, что при переносе начала координат в середину ряда $\sum_{-p}^{p} t = 0$ и, следовательно, $a = \bar{y}$, находим

$$\hat{y}_t - \bar{y} = a + bt - \bar{y} = bt.$$

Откуда

$$\sum_{-p}^{p} (\hat{y}_t - \bar{y})^2 = \sum (bt)^2 = b^2 \sum t^2 = b \frac{(n-1)\, n\, (n+1)}{12};$$

$$Q_1 = Q_0 - b^2 \sum_{-p}^{p} t^2. \qquad (14.48)$$

Если $t = 1, 2, \ldots, n$, то

$$\hat{y}_t - \bar{y} = \frac{\sum y}{n} - b \frac{\sum t}{n} + bt - \frac{\sum y}{n} = bt - b \frac{\sum t}{n}.$$

После ряда преобразований получим

$$\sum_{1}^{n} (\hat{y}_t - \bar{y})^2 = \frac{b^2}{12} (n+1)\, n\, [2\,(2n+1) - 3\,(n+1)]. \qquad (14.49)$$

Сумму квадратов отклонений от параболического тренда при условии, что $\sum_{-p}^{p} t = 0$, находим следующим образом:

$$\sum_{-p}^{p} (\hat{y}_t - \bar{y})^2 = \sum \left(a + bt + ct^2 - \frac{\sum y}{n} \right)^2 = b^2 \sum t^2 + c^2 \sum t^4;$$

$$Q_1 = Q_0 - \left(b^2 \sum_{-p}^{p} t^2 + c^2 \sum_{-p}^{p} t^4 \right). \qquad (14.50)$$

Доверительные интервалы полиномиальных трендов. При исследовании множественной регрессии в гл. 13 было показано, что дисперсия регрессии определяется матричным выражением

$$s_{\hat{y}}^2 = s^2 \mathbf{x}_p^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_p, \qquad (14.51)$$

где $\mathbf{X}$ — матрица значений независимых переменных, а $\mathbf{x}_p$ — вектор значений независимых переменных. При выравнивании временных рядов в качестве независимых переменных выступает время и, следовательно, вместо $\mathbf{X}^T\mathbf{X}$ теперь можно записать $\mathbf{T}^T\mathbf{T}$, а вместо $\mathbf{x}_p$ запишем $\mathbf{t}_L$, $\mathbf{T}$ — матрица размерностью $n \times (\lambda + 1)$, n — число членов в ряду (число наблюдений), λ — степень выравнивающего полинома (соответственно $\lambda + 1$ — число оцениваемых параметров полинома),

$$\boldsymbol{T} = \begin{bmatrix} 1 & t & t^2 & \dots & t^\lambda \\ \dots & \dots & \dots & \dots & \dots \\ 1 & t & t^2 & \dots & t^\lambda \end{bmatrix},$$

$\boldsymbol{t}_L = (1 \; t_L t_L^2 \; \dots \; t_L^\lambda)$ — вектор размерностью $\lambda + 1$, где t_L — момент времени, для которого определяется ошибка тренда. Таким образом, вместо (14.51) можно написать уравнение, определяющее дисперсию тренда:

$$s_{\hat{y}}^2 = s^2 \mathbf{t}_L^T (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{t}_L. \qquad (14.52)$$

Найдем матрицу

$$\mathbf{T}^T\mathbf{T} = \begin{bmatrix} n & \sum t & \dots & \sum t^\lambda \\ \sum t & \sum t^2 & \dots & \sum t^{\lambda+1} \\ \dots & \dots & \dots & \dots \\ \sum t^\lambda & \sum t^{\lambda+1} & \dots & \sum t^{2\lambda} \end{bmatrix}. \qquad (14.53)$$

Рассмотрим тренды, описываемые соответственно прямой, квадратичной и кубической параболой. Матрицы $\mathbf{T}^T\mathbf{T}$ для оценки этих кривых при условии, что начало отсчета времени приходится

на середину ряда $\left(\text{т. е. } \sum_{-p}^{p} t = 0\right)$, имеют следующий вид:

$$\mathbf{T}^T\mathbf{T}_1 = \begin{bmatrix} n & 0 \\ 0 & \sum t^2 \end{bmatrix};$$

$$\mathbf{T}^T\mathbf{T}_2 = \begin{bmatrix} n & 0 & \sum t^2 \\ 0 & \sum t^2 & 0 \\ \sum t^2 & 0 & \sum t^4 \end{bmatrix};$$

$$\mathbf{T}^T\mathbf{T}_2 = \begin{bmatrix} n & 0 & \sum t^2 & 0 \\ 0 & \sum t^2 & 0 & \sum t^4 \\ \sum t^2 & 0 & \sum t^4 & 0 \\ 0 & \sum t^4 & 0 & \sum t^6 \end{bmatrix}.$$

Соответственно

$$(\mathbf{T}^T\mathbf{T})_1^{-1} = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \sum t^2 \end{bmatrix}; \tag{14.54}$$

$$(\mathbf{T}^T\mathbf{T})_2^{-1} = \begin{bmatrix} \frac{\sum t^4}{A} & 0 & \frac{-\sum t^2}{A} \\ 0 & \frac{1}{\sum t^2} & 0 \\ \frac{-\sum t^4}{A} & 0 & \frac{n}{A} \end{bmatrix}. \tag{14.55}$$

где $A = n\sum t^4 - (\sum t^2)^2$;

$$(\mathbf{T}^T\mathbf{T})_3^{-1} = \begin{bmatrix} \frac{\sum t^4}{A} & 0 & \frac{-\sum t^4}{A} & 0 \\ 0 & \frac{\sum t^6}{B} & 0 & \frac{-\sum t^4}{B} \\ \frac{-\sum t^2}{A} & 0 & \frac{n}{A} & 0 \\ 0 & \frac{-\sum t^4}{B} & 0 & \frac{\sum t^6}{B} \end{bmatrix}, \tag{14.56}$$

где A имеет то же содержание, что и выше, а

$$B = \sum t^2 \sum t^6 - (\sum t^4)^2.$$

Определим теперь величину $s_{\hat{y}}^2$ для тренда, характеризуемого уравнением прямой. Подставив в (14.52) значение $(\mathbf{T}^T\mathbf{T})^{-1}$ и $\mathbf{t}_L$,

получим

$$s_{\hat{y}}^2 = s^2 [1 \; t_L] \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum t^2} \end{bmatrix} \begin{bmatrix} 1 \\ t_L \end{bmatrix}; \qquad s_{\hat{y}}^2 = s^2 \left(\frac{1}{n} + \frac{t_L^2}{\sum t^2} \right). \qquad (14.57)$$

Здесь и далее t_L и t измеряются от середины ряда. Подставив в (14.57) значение $\sum t^2$ для случая, когда число членов ряда нечетное, имеем

$$s_{\hat{y}}^2 = s^2 \left(\frac{1}{n} + \frac{12}{n(n^2-1)} t_L^2 \right). \qquad (14.58)$$

Определим теперь $s_{\hat{y}}^2$ для некоторых значений t_L. Для начала положим, что $t_L = 0$ (это значит, что определяется значение дисперсии в точке, соответствующей середине ряда), в этом случае

$$s_{\hat{y}}^2 = \frac{s^2}{n}.$$

Если $t_L = \frac{n-1}{2}$, т. е. $s_{\hat{y}}^2$ определяется для момента, которым заканчивается интервал наблюдения, то, подставив t_L в (14.58), получим

$$s_{\hat{y}}^2 = s^2 \left[\frac{1}{n} + \frac{3(n-1)}{n(n+1)} \right].$$

Пусть теперь $t_L = \frac{n-1}{2} + L$ и $L > 0$, т. е. дисперсия тренда определяется для момента времени, лежащего вне интервала наблюдения. В этом случае

$$s_{\hat{y}}^2 = s^2 \left\{ \frac{1}{n} + \frac{3(n-1)}{n(n+1)} + \frac{12}{n(n^2-1)} \left[(n-1)L + L^2 \right] \right\}. \qquad (14.59)$$

Так, если $L = n$ (практически предельный случай), то

$$s_{\hat{y}}^2 = s^2 \left[\frac{1}{n} + \frac{3(n-1)}{n(n+1)} + \frac{12(2n-1)}{n^2-1} \right].$$

Проиллюстрируем сказанное числовым примером. Пусть ряд охватывает 25 членов. Тогда $s_{\hat{y}}^2 = 0{,}04 s^2$ для $t_L = 0$, $s_{\hat{y}}^2 = 0{,}1508 s^2$ для $t_L = 12$, наконец, $s_{\hat{y}}^2 = 1{,}0931 s^2$ для $t_L = 12 + 25$.

В целом же квадратическая ошибка увеличивается с ростом t_L, соответственно расширяются доверительные границы тренда.

В примере на с. 294 к ряду был подобран линейный тренд. Определим теперь доверительные интервалы для этого тренда. Прежде всего найдем дисперсию членов ряда вокруг тренда. Для этого воспользуемся соотношением (14.48). Выше было рассчитано: $\Sigma t^2 = 408$; $b = 1{,}288$; $\Sigma y = 951{,}6$; найдем $\Sigma y^2 = 55463{,}6$;

$$Q_0 = \sum y^2 - \left(\frac{\sum y}{n} \right)^2 n = 2196{,}4;$$

$$Q_1 = 2196{,}4 - 1{,}288^2 \cdot 408 = 1519{,}1.$$

При числе степеней свободы, равном 17 – 2 = 15, находим

$$s = \sqrt{\frac{1519}{15}} = 10{,}07.$$

В соответствии с (14.57)

$$s_y = 10{,}07 \sqrt{\frac{1}{17} + \frac{t_L^2}{408}}.$$

Определим теперь доверительный интервал для расчетного $\hat{y}_t$ в момент времени $t_L = 10$. В этом случае $\hat{y}_{10} = 55.943 + 1{,}228 \cdot 10 = 68{,}85$ и $s\hat{y} = 5{,}55$. Допустим, что доверительная вероятность равна 0,8 (в этом случае $t_\alpha = 1{,}341$). Доверительный интервал для этого момента охватывает следующий диапазон значений: $68{,}85 \pm 1{,}341 \cdot 5{,}55 = 61{,}41 \div 76{,}29$.

Определим теперь дисперсию тренда, следующего квадратичной параболе. Подставим в общее выражение дисперсии вектор $\mathbf{t}_L$ и матрицу $(\mathbf{T}^T\mathbf{T})^{-1}$, соответствующие параболе второй степени:

$$s_{\hat{y}}^2 = s^2 \begin{bmatrix} 1 & t_L & t_L^2 \end{bmatrix} \begin{bmatrix} \frac{\sum t^4}{A} & 0 & -\frac{\sum t^2}{A} \\ 0 & \frac{1}{\sum t^2} & 0 \\ -\frac{\sum t^2}{A} & 0 & \frac{n}{A} \end{bmatrix} \begin{bmatrix} 1 \\ t_L \\ t_L^2 \end{bmatrix} =$$

$$= s^2 \left[\frac{\sum t^4}{A} - \frac{\sum t^2}{A} t_L^2 + \frac{1}{\sum t^2} t_L^2 + \left(-\frac{\sum t^2}{A} + \frac{n}{A} t_L^2 \right) t_L^2 \right].$$

После ряда преобразований получим

$$s_{\hat{y}} = s \sqrt{\frac{1}{\sum t^2} t_L^2 + \frac{\sum t^4 - (2 \sum t^2) t_L^2 + n t_L^4}{n \sum t^4 - (\sum t^2)^2}}. \qquad (14.60)$$

Аналогично для кубической параболы находим

$$s_{\hat{y}} = s \sqrt{\frac{\sum t^4 - (2 \sum t^2) t_L^2 + n t_L^4}{n \sum t^4 - (\sum t^2)^2} + \frac{(\sum t^6) t_L^2 - (2 \sum t^4) t_L^4 + (\sum t^2) t_L^6}{\sum t^2 \sum t^6 - (\sum t^4)^2}}. \qquad (14.61)$$

Доверительные интервалы трендов, приводимых к линейному виду. Как было показано выше, уравнения ряда кривых нетрудно привести к линейному виду. Параметры полученного уравнения прямой и подлежат оценке. Данный подход можно распространить и на определение доверительных интервалов трендов. Соответственно сказанному сперва определяются доверительные интервалы прямой, а затем производится обратный переход — расчет доверительных интервалов для тренда, представленного исходной кривой. Например, оценка параметров экспоненты сводится к оценке параметров прямой в логарифмах. Следовательно, доверительные интервалы этой прямой можно определить как $\log \hat{y}_t \pm t_\alpha s_y$ (где $s_{\hat{y}}$ — средняя квадратическая

ошибка линейного в логарифмах тренда). Наконец, потенцируя полученные результаты, находим доверительные интервалы для тренда, представленного экспонентой.

Выше был оценен экспоненциальный тренд: $\hat{y}_t = 92{,}247 \cdot 1{,}0415^t$ ($t = 1, 2, \ldots$) или $\ln \hat{y}_t = 4{,}52447 + 0{,}04065t$. По данным табл. 14.9 нетрудно найти $\sum (\ln y_t - \ln \hat{y}_t)^2 = 0{,}04476$. Откуда при числе средней свободы $\nu = 12 - 2 = 10$ получим

$$s = \sqrt{\frac{0{,}04476}{10}} = 0{,}0669$$

Найдем теперь $s_{\hat{y}}$ для линейного в логарифмах тренда. Напомним, что при расчете $s_{\hat{y}}$ по формуле (14.57) отсчет времени ведется от середины ряда, поэтому в данном случае $t = -5{,}5; -4{,}5; \ldots; -0{,}5; 0{,}5; 0{,}5; \ldots; 4{,}5; 5{,}5$. Соответственно $\Sigma t^2 = 143$ и

$$s_{\hat{y}} = 0{,}0669 \sqrt{\frac{1}{12} + \frac{t_L^2}{143}},$$

Допустим, что необходимо определить значение уровня на два шага вперед, т. е. $\hat{y}_{14}$. В этом случае $t_L = 14 - \bar{t} = 14 - 6{,}5 = 7{,}5$, $\ln \hat{y}_{14} = 5{,}09357$, $\hat{y}_{14} = 162{,}97$. Отсюда

$$s_{\hat{y}} = 0{,}0669 \sqrt{\frac{1}{12} + \frac{7{,}5^2}{143}} = 0{,}0462.$$

При условии, что доверительная вероятность равна 80% ($t_\alpha = 1{,}397$),

$$\operatorname{antln} (\ln \hat{y}_{14} \pm t_\alpha s_{\hat{y}}) = \operatorname{antln} (5{,}09357 \pm 1{,}397 \cdot 0{,}0462) = 152{,}8 \div 173{,}8.$$

Заметим, что доверительный интервал тренда здесь несимметричен относительно трендового значения уровня.

Рассмотренный метод дает результаты при $b > 1$.

Аналогичным путем можно получить доверительные интервалы и для иных видов тренда, приводимых к линейному виду.

14.6. Регрессия

В гл. 12 и 13 зависимости между явлениями изучались с помощью регрессионного и корреляционного анализа, причем данные были представлены главным образом в виде выборок. Там же было показано, что эти же методы анализа могут быть распространены и на временные ряды, в частности при оценке производственных функций. Регрессию между показателями, представленными динамическими рядами, можно получить разными путями: непосредственно между уровнями исследуемых рядов (y_t) или между отклонениями от трендов соответствующих рядов.

Если находить регрессию первым способом, то весьма вероятно, что будет наблюдаться автокорреляция остатков.

Различный подход возможен и при коррелировании временных рядов (коррелирование y_t или ε_t). Разницу в содержании коэффициентов корреляции легко понять из следующих рассуждений. Пусть имеются два ряда y_t и x_t, $t = 1, \ldots, n$. Для начала допустим, что оба эти ряда строго соответствуют некоторым функциям времени, например $\hat{y}_t = 2t$ и $\hat{x}_t = 0{,}5t$. Отсюда следует, что между $\hat{y}_t$ и $\hat{x}_t$ существует функциональная зависимость $\hat{y}_t = 4\hat{x}_t$.

И следовательно, коэффициент корреляции между ними равен 1. Наложим на тренд колебания. Пусть последние будут постоянными для каждого ряда ($\varepsilon_y = 0{,}5$ и $\varepsilon_x = 0{,}05$), меняющими свой знак на каждом шаге во времени, причем знаки у отклонений ε_y и ε_x противоположны в каждый момент времени. Очевидно, что в этом случае между отклонениями наблюдается полная отрицательная корреляция ($r = -1$). Вместе с тем если подсчитать коэффициент корреляции между y_t и x_t (где $y_t = \hat{y}_t \pm 0{,}5$, $x_t = \hat{x}_t \pm 0{,}05$), то при $n = 10$ $r_{yx} = 0{,}993$. Таким образом, при корелировании уровней получим указание на то, что между переменными наблюдается высокая положительная взаимосвязь, в то время как отклонения от трендов характеризуются отрицательной взаимосвязью. Нетрудно догадаться, что на значение r_{yx} оказало преобладающее влияние то, что оба тренда имеют одинаковое направление в развитии и одну форму (линейную). Корреляция трендов не является убедительным свидетельством именно взаимосвязи соответствующих показателей (это необходимо, но недостаточно).

Включение времени в регрессию. В уравнение регрессии можно в явном виде ввести в качестве независимой переменной член, характеризующий время. Обычно время вводится в виде линейного члена. Уравнение регрессии в этом случае будет иметь вид: $y = f(x_1, x_2, \ldots, t)$. Включение времени наряду с другими независимыми переменными позволяет выделить регрессию на неучтенные в явном виде факторы, связанные со временем. Заметим, что включение времени равносильно расчету регрессии отклонений от тренда. Покажем это. Возьмем три переменные y, x и z, где y — зависимая от x и z переменная. Пусть развитие каждой из них можно описать линейным трендом, т. е.

$$\hat{y}_t = a_1 + b_1 t; \quad \hat{x}_t = a_2 + b_2 t; \quad \hat{z}_t = a_3 + b_3 t.$$

Отклонения фактических уровней от трендов составляет

$$y_t - \hat{y}_t, \quad x_t - \hat{x}_t, \quad z_t - \hat{z}_t,$$

где y_t, x_t, z_t — фактические, а $\hat{y}_t$, $\hat{x}_t$, $\hat{z}_t$ — расчетные значения соответствующих переменных. Найдем теперь уравнение регрессии отклонений зависимой переменной на отклонения независимых переменных:

$$y_t - \hat{y}_t = k_0 + k_1 (x_t - \hat{x}_t) + k_2 (z_t - \hat{z}_t) + u_t,$$

где u_t — ошибка уравнения регрессии. Подставив в эту регрессию развернутые выражения для отклонений, находим

$$y_t - a_1 - b_1 t = k_0 + k_1 (x_t - a_2 - b_2 t) + k_2 (z_t - a_3 - b_3 t) + u_t.$$

Проведя ряд преобразований, получим

$$y_t = (k_0 + a_1 - k_1 a_2 - k_2 a_3) + k_1 x_t + k_2 z_t + (b_1 - k_1 b_2 - k_2 b_3) t + u_t.$$

Многочлены, взятые в скобки, содержат линейные комбинации параметров и, следовательно, сами являются некоторыми пара-

метрами. Обозначим их соответственно как c и g, тогда

$$y_t = c + k_1 x_t + k_2 z_t + gt + u_t,$$

т. е. получили линейную регрессию на три переменные, включая время. Обнаруживается, что включение времени в регрессию y_t на x_t и z_t оказывается равнозначным (в отношении коэффициентов k_1, k_2 и погрешности u_t) определению регрессии отклонений от трендов. Заметим, что такие характеристики, как коэффициент корреляции и ошибки коэффициентов регрессии, будут, естественно, различаться для двух рассмотренных способов анализа взаимосвязей. Включение времени в регрессию обычно существенно снижает автокорреляцию.

Выше мы исследовали проблему при условии, что каждая переменная характеризуется линейным трендом. Продемонстрированный метод приемлем и для случаев, когда тренды представлены многочленами более высоких степеней.

Время в регрессию может включаться и другим образом. Иногда есть основание предполагать, что коэффициент при независимой переменной устойчиво изменяется вместе с ходом времени. Допустим, это изменение равномерное, тогда соответствующий коэффициент регрессии можно представить в виде произведения $a_j = a_j^* t$, а сам член регрессии предстанет в виде $a_j^* t x_j$. В этом случае оценивание можно осуществить опять-таки с помощью МНК, предварительно преобразовав значения переменной x_j; иначе говоря, в матрицу $\mathbf{X}^T\mathbf{X}$ вместо x_{tj} теперь вводятся значения tx_{tj}.

Авторегрессия. При анализе динамических рядов в большинстве случаев приходится констатировать, что соседние члены ряда находятся в некоторой зависимости друг от друга, т. е. значение переменной y в момент t является функцией от значений в предшествующие моменты $y_t = f(y_{t-1}, y_{t-2}, \ldots)$. В этом случае для характеристики взаимосвязи показателей можно прибегнуть к разработке авторегрессионной модели. Приведем линейный ее вариант:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_m y_{t-m} + \varepsilon_t, \qquad (14.62)$$

где ε_t — несистематическая составляющая y_t; m — число членов, которое охватывается авторегрессией (порядок авторегрессии). Обычно предполагается, что $M(\varepsilon_t) = 0$, $\sigma^2(\varepsilon_t) = \sigma = \text{const}$, наконец, ε_t и ε_{t-1} — независимы. Это дает возможность оценивать параметры (14.62) с помощью МНК. Заметим, однако, что авторегрессия не входит в класс обычных линейных множественных регрессий, так как там предполагается, что переменные в правой части фиксированы и неслучайны, а в (14.62) y_{t-1}, y_{t-2}, ... суть случайные величины.

В ряде случаев переменная y_t характеризуется не только зависимостью от предшествующих значений этой переменной, но и испытывает влияние ряда факторов, которые могут быть представлены другими переменными x_j. В таких случаях регрессию

можно охарактеризовать в виде уравнения смешанной авторегрессии:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ \dots \ + \alpha_m y_{t-m} + \beta_1 x_1 + \beta_2 x_2 + \ \dots \ + \beta x + \varepsilon_t.$$

Здесь опять предполагаем $M(\varepsilon_t) = 0$, $\sigma^2(\varepsilon_t) = \sigma^2 = \text{const}$, ε_t и ε_{t+1} — независимы. Оценки для α_i и β_j находим МНК.

14.7. Экстраполяция временных рядов

Один из наиболее распространенных методов прогнозирования заключается в экстраполяции, т. е. в продлении в будущее тенденции, наблюдавшейся в прошлом. Экстраполяция тенденций динамических рядов сравнительно широко применяется в практике в силу ее простоты, возможности осуществления на основе относительно небольшого объема информации, наконец, ясности принятых допущений. Отсутствие иной информации помимо отдельно рассматриваемого временного ряда часто оказывается решающим при выборе этого метода прогнозирования.

При таком подходе к прогнозированию предполагается, что размер признака, характеризующего явление, формируется под воздействием множества факторов, причем не представляется возможным выделить порознь их влияние. В связи с этим ход развития связывается не с какими-либо конкретными факторами, а с течением времени. Экстраполяция базируется на следующих допущениях: 1) развитие явления может быть с достаточным основанием охарактеризовано плавной (эволюторной) траекторией — трендом; 2) общие условия, определяющие тенденцию развития в прошлом, не претерпят существенных изменений в будущем. Таким образом, экстраполяция дает описание некоторого общего будущего развития объекта прогнозирования, в случае если есть основания полагать существование известной инерционности в развитии явления и, следовательно, гипотеза о будущем развитии может в значительной мере базироваться на анализе прошлого.

При определении прогностических значений того или иного явления с помощью экстраполяции тренда наибольший интерес представляет, по-видимому, не сама экстраполяция — это более или менее механический прием. В самом деле, экстраполяция временного ряда дает возможность получить точечную прогностическую оценку. Однако поскольку экономические переменные, как правило, непрерывны, то указание их точечных значений, строго говоря, лишено содержания, поскольку «попадание в точку» имеет нулевую вероятность. Прогноз должен быть выдан в виде некоторой «вилки» значений. Один из путей получения такой «вилки» заключается в определении доверительных интервалов прогноза.

Доверительные интервалы могут быть определены двояко: формально и неформально. Что касается последнего, то это дело экспертного суждения, которое выносится при качественном осмыслении результатов прогноза, сопоставлении их с другими

имеющимися у эксперта данными и т. д. Формальный же доверительный интервал учитывает лишь ту неопределенность, которая связана с ограниченностью числа наблюдений и соответствующей неточностью найденных оценок параметров. Основной вопрос — в какой мере в будущем сохранится найденная тенденция — естественно, не может быть решен с помощью таких доверительных интервалов. Это дело содержательного экономического анализа и, вероятно, экспертной оценки.

Доверительные интервалы прогнозов, определяемых по полиномиальным трендам. В § 14.5 было показано, как определять доверительные интервалы для трендов. Поскольку фактические данные колеблются вокруг трендов, то, очевидно, доверительный интервал прогноза должен быть шире доверительного интервала тренда и учитывать указанный дополнительный источник колебаний. Соответственно интервал определяется теперь как

$$\hat{y}_{t+L} \pm t_\alpha s_p,$$

где s_p — средняя квадратическая ошибка прогноза. Для линейного тренда находим (сравните с (14.57))

$$s_p = s\sqrt{1+\frac{1}{n}+\frac{t_L^2}{\sum t^2}},$$

где t и t_L измеряются от середины ряда.

Аналогично для квадратичной параболы имеем (сравните с (14.60))

$$s_p = s\sqrt{1+\frac{t_L^2}{\sum t^2}+\frac{\sum t^4-\left(2\sum t^2\right)t_L^2+nt_L^4}{n\sum t^4-\left(\sum t^2\right)^2}}.$$

Экстраполяция по экспоненциальной средней. Для краткосрочного прогнозирования в ряде случаев может быть применена экспоненциальная скользящая средняя. Если прогнозирование ведется на один шаг вперед, то

$$y_{t+1} = Q_t, \tag{14.63}$$

где Q_t — экспоненциальная средняя. Естественно, что прогноз, выраженный точечной оценкой (14.63) не удовлетворяет нас, поскольку не учитывает возможных отклонений от этой оценки. Для определения доверительных интервалов прогноза будем рассуждать следующим образом. Пусть σ_Q — средняя квадратическая ошибка экспоненциальной средней. Напомним, что $\sigma_Q = \sigma\sqrt{\frac{\alpha}{2-\alpha}}$ (где σ — среднее квадратическое отклонение от Q_t; α — параметр сглаживания). Тогда интервал, полученный как $Q_t \pm t_\alpha\sigma_Q$, характеризует границы, в которых с заданной вероятностью следует ожидать значение средней. Переход от точечного прогноза (экспоненциальной средней) к интервалу, в котором находится оценка

средней, увеличит надежность прогноза. Однако при этом остается в силе предположение о том, что прогнозируемый показатель равен средней, т. е. при таком подходе не учитывается то, что отдельные значения исследуемого показателя варьировали вокруг средней в прошлом и, несомненно, будут варьировать в будущем. Поэтому доверительный интервал для прогностической оценки должен учитывать и этот фактор. Отсюда общая дисперсия (связанная как с колеблемостью средней, так и с варьированием индивидуальных значений вокруг средней) составит величину $\sigma^2 + \sigma_Q^2$. Таким образом, доверительный интервал для прогностической оценки определяется как

$$Q_t \pm t_\alpha \sigma \sqrt{1 + \frac{\alpha}{2 - \alpha}}. \qquad (14.64)$$

Заметим, что применение экспоненциальной средней в качестве основы для прогностической оценки, вообще говоря, имеет смысл лишь при относительно небольшой колеблемости уровней. Если же уровни имеют значительную флуктуацию во времени, то значения среднего квадратического отклонения будут весьма высокими и, следовательно, доверительный интервал окажется очень широким.

Возьмем для примера данные табл. 14.5, в которой сглаживались данные об урожайности. Пусть прогноз определяется как $y_{t+1} = Q_t$. При условии, что Q_t определяется при $\alpha = 0{,}3$, прогноз для $t = 16$ составит 17,9 ц/га. Среднее квадратическое отклонение равно 3,37 ц/га. Откуда квадратическая ошибка средней равна: $3{,}37 \sqrt{\frac{0{,}3}{2 - 0{,}3}} = 1{,}416$ ц/га, а доверительный интервал для прогноза при $t_\alpha = 1{,}753$ (90%-ная доверительная вероятность, число степеней свободы равно 15) составит

$$17{,}9 \pm 1{,}753 \cdot 3{,}37 \sqrt{1 + \frac{0{,}3}{2 - 0{,}3}} = 17{,}9 \pm 6{,}4 \text{ ц/га}.$$

Таким образом, результат настолько неопределен, что вряд ли имеет какую-либо практическую ценность.

Сглаживание и экстраполяция по экспоненциальной средней — формальные приемы, которые тем не менее имеют обширную область применения в краткосрочном прогнозировании, а именно там, где следует осуществить прогноз большого числа однородных показателей на основе изолированных рядов динамики и нет необходимости прибегать к более тонким методам (с выделением и анализом трендов), например при краткосрочных прогнозах остатков большой номенклатуры товаров.

Адаптивное прогнозирование. Под адаптивным прогнозированием будем понимать процесс разработки прогноза с «самообучением». Мы уже имели дело с «самообучающимися» моделями простейших видов — экспоненциальной средней. Как было показано выше, эти средние с каждым шагом во времени (получением новой информации) изменяются с учетом наблюдавшейся на предыдущем шаге ошибки прогноза. Адаптивные возможности такого метода ограничены. Вообще говоря, учет

новой информации на каждом шаге процесса возможен и при оценивании параметров кривых, выравнивающих ряд. В этом случае оценивание параметров осуществляется заново каждый раз при получении новых данных. Однако такой прием весьма трудоемок и к нему прибегают исключительно редко.

Адаптивное прогнозирование осуществляют с помощью ряда методов *. Рассмотрим наиболее распространенный из них. Однако до этого необходимо ввести новое понятие, связанное с экспоненциальной средней. Итак, назовем величину (14.15) экспоненциальной средней первого порядка (первичное сглаживание). Обозначим ее как $Q_t^{(1)}$. Как будет показано далее, при разработке статистических прогностических моделей полезно прибегнуть к расчету экспоненциальных средних более высоких порядков, т. е. средних, получаемых путем многократного экспоненциального сглаживания. Общая запись экспоненциальной средней порядка k имеет вид:

$$Q_t^{(k)} = \alpha \sum (1-\alpha)^j Q_{t-1}^{(k)}. \tag{14.65}$$

Преобразуя формулу (14.65), получим рекуррентное выражение

$$Q_t^{(k)} = \alpha Q_t^{(k-1)} + (1-\alpha) Q_{t-1}^{(k)}. \tag{14.66}$$

Таким образом, экспоненциальная средняя порядка k на момент t есть линейная комбинация экспоненциальной средней порядка $k-1$ на момент t и средней порядка k на момент $t-1$. На основе этой формулы получим выражения для экспоненциальных средних первого, второго и третьего порядков:

$$Q_t^{(1)} = \alpha y_t + (1-\alpha) Q_{t-1}^{(1)};$$
$$Q_t^{(2)} = \alpha Q_t^{(1)} + (1-\alpha) Q_{t-1}^{(2)};$$
$$Q_t^{(3)} = \alpha Q_t^{(2)} + (1-\alpha) Q_{t-1}^{(3)}.$$

Теперь можно перейти к методам адаптивного прогнозирования, базирующимся на экспоненциальных средних первого и более высокого порядков. Пусть прогностическое значение переменной y_t на L шагов вперед, обозначим его как $\hat{y}_{t+L}$, оценивается с помощью полинома степени k:

$$\hat{y}_{t+L} = a_t + b_t L + c_t L^2 + \ldots + g_t L^k. \tag{14.67}$$

Таким образом, прогноз в этом выражении представлен в виде функции продолжительности периода упреждения. Доказано, что параметры a_t, b_t, ... могут быть получены как линейные комбинации экспоненциальных средних соответствующих порядков. В частности, если предполагается существование линейной зависимости

$$\hat{y}_{t+L} = a_t + b_t L, \tag{14.68}$$

* Анализ методов адаптивного прогнозирования и некоторые новые подходы к этой проблеме рассматриваются в [11].

то

$$a_t = 2Q_t^{(1)} - Q_t^{(2)};$$

$$b_t = \frac{\alpha}{1-\alpha}\left(Q_t^{(1)} - Q_t^{(2)}\right).$$

Для квадратичной зависимости

$$\hat{y}_{t+L} = a_t + b_t L + c_t L^2 \qquad (14.69)$$

оценки параметров получим с помощью следующих формул:

$$a_t = 3\left[Q_t^{(1)} - Q_t^{(2)}\right] + Q_t^{(3)};$$

$$b_t = \frac{\alpha}{2(1-\alpha)^2}\left[(6-5\alpha)Q_t^{(1)} - 2(5-4\alpha)Q_t^{(2)} + (4-3\alpha)Q_t^{(3)}\right];$$

$$c_t = \frac{\alpha^2}{2(1-\alpha)^2}\left(Q_t^{(1)} - 2Q_t^{(2)} + Q_t^{(3)}\right).$$

Подставив в (14.68) и (14.69) соответствующие значения параметров и несколько преобразовав их, получим

для линейной модели:

$$\hat{y}_{t+L} = \left(2 + \frac{\alpha L}{1-\alpha}\right)Q_t^{(1)} - \left(1 + \frac{\alpha L}{1-\alpha}\right)Q_t^{(2)}; \qquad (14.70)$$

для квадратичной модели

$$\hat{y}_{t+L} = \left[6(1-\alpha)^2 + (6-5\alpha)\alpha L + \alpha^2 L^2\right]\frac{Q_t^{(1)}}{2(1-\alpha)^2} -$$

$$-\left[6(1-\alpha)^2 + 2(5-4\alpha)\alpha L + 2\alpha^2 L^2\right]\frac{Q_t^{(2)}}{2(1-\alpha)^2} +$$

$$+\left[2(1-\alpha)^2 + (4-3\alpha)\alpha L + \alpha^2 L^2\right]\frac{Q_t^{(3)}}{2(1-\alpha)^2}. \qquad (14.71)$$

Выражения (14.70) и особенно (14.71) громоздки; их можно существенно упростить, подставив числовые значения α и L. Так, для прогноза по линейной модели на один шаг вперед ($L = 1$) имеем:

для $\alpha = 0{,}1$ $\hat{y}_{t+1} = 2{,}111Q_t^{(1)} - 1{,}111Q_t^{(2)}$;

для $\alpha = 0{,}2$ $\hat{y}_{t+1} = 2{,}25Q_t^{(1)} - 1{,}25Q_t^{(2)}$;

для $\alpha = 0{,}3$ $\hat{y}_{t+1} = 2{,}429Q_t^{(1)} - 1{,}429Q_t^{(2)}$.

При прогнозе на два шага вперед ($L = 2$) по той же модели получим:

для $\alpha = 0{,}1$ $\hat{y}_{t+2} = 2{,}222Q_t^{(1)} - 1{,}222Q_t^{(2)}$;

для $\alpha = 0{,}2$ $\hat{y}_{t+2} = 2{,}5Q_t^{(1)} - 1{,}5Q_t^{(2)}$;

для $\alpha = 0{,}3$ $\hat{y}_{t+2} = 2{,}857Q_t^{(1)} - 1{,}857Q_t^{(2)}$.

После подстановки значений α в квадратическую модель (14.71) получим для $L=1$:

для $\alpha=0{,}1$ $\hat{y}_{t+1}=3{,}346Q_t^{(1)}-3{,}58Q_t^{(2)}+1{,}235Q_t^{(3)}$;

для $\alpha=0{,}2$ $\hat{y}_{t+1}=3{,}813Q_t^{(1)}-4{,}375Q_t^{(2)}+1{,}563Q_t^{(3)}$;

для $\alpha=0{,}3$ $\hat{y}_{t+1}=4{,}469Q_t^{(1)}-5{,}51Q_t^{(2)}+2{,}041Q_t^{(3)}$.

Как было показано выше, для расчета экспоненциальных средних применяются рекуррентные формулы. В связи с этим возникает проблема определения некоторых начальных значений этих средних, т. е. $Q_0^{(1)}$, $Q_0^{(2)}$, $Q_0^{(3)}$. Выбор начальных условий обычно не сказывается на значении экспоненциальных средних после 15—20 шагов при условии, что колеблемость уровней ряда не очень велика. В этих случаях достаточно в качестве грубых начальных значений средних принять первый уровень ряда. Однако если ряд короток и прогностические значения необходимо оценить, скажем, после 7—10 шагов, то к выбору начальных условий для сглаживания следует подойти более осторожно. В частности, рекомендуется начальные значения определять по следующим формулам:

для линейной модели:

$$Q_0^{(1)}=a-\frac{1-\alpha}{\alpha}b;$$

$$Q_0^{(2)}=a-2\frac{1-\alpha}{\alpha}b;$$

для квадратичной модели:

$$Q_0^{(1)}=a-\frac{1-\alpha}{\alpha}b+\frac{(1-\alpha)(2-\alpha)}{2\alpha^2}c;$$

$$Q_0^{(2)}=a-\frac{2(1-\alpha)}{\alpha}b+\frac{(1-\alpha)(3-2\alpha)}{\alpha^2}c;$$

$$Q_0^{(3)}=a-\frac{3(1-\alpha)}{\alpha}b+\frac{3(1-\alpha)(4-3\alpha)}{2\alpha^2}c,$$

где a, b и c — параметры уравнения прямой и параболы, подобранной при выравнивании ряда динамики с помощью МНК.

Несколько слов о доверительных интервалах прогнозов, получаемых по модели (14.67). Дисперсия прогноза определяется приближенным выражением [11]:

для линейной модели:

$$s_{\hat{y}}\approx(1{,}25\alpha_1+\alpha_1^2L)\sigma^2;$$

для квадратичной модели:

$$s_{\hat{y}}\cong(2\alpha_1+3\alpha_1^2L+3\alpha_1^3L^2)\sigma^2.$$

Здесь $\alpha_1=1-(1-\alpha)^m$, где m — число параметров модели.

В чисто иллюстративных целях вновь обратимся к данным примера на с. 283 и определим прогностические значения $\hat{y}_{t+L}$ для 11-го и следующих моментов времени, воспользовавшись линейной моделью (14.68). Начальные значения

экспоненциальных средних первого и второго порядков определим по формулам (14.72) исходя из значений параметров, полученных при выравнивании первых девяти членов ряда: $a = 43{,}7$ и $b = 1{,}5$, откуда при $\alpha = 0{,}2$ $Q_0^{(1)} = 37{,}69$ и $Q_0^{(2)} = 31{,}63$. Расчет значений $Q_t^{(1)}$ и $Q_t^{(2)}$ сведен в таблицу (см. табл. 14.11). В этой же таблице приведены прогностические значения $\hat{y}_{t+L}$ на один и два шага вперед, за базу взяты первые десять уровней. Например,

$$\hat{y}_{10+1} = 2{,}25Q_{10}^{(1)} - 1{,}25Q_{10}^{(2)} = 2{,}25 \cdot 48{,}55 - 1{,}25 \cdot 45{,}59 = 52{,}25;$$

$$\hat{y}_{10+2} = 2{,}222Q_{10}^{(1)} - 1{,}222Q_{10}^{(2)} = 2{,}222 \cdot 48{,}55 - 1{,}222 \cdot 45{,}59 = 52{,}17.$$

Таблица 14.11

t	y_t	$Q_t^{(1)}$	$Q_t^{(2)}$	Прогноз на один шаг вперед	Прогноз на два шага вперед	Ошибки прогноза	
						$L=1$	$L=2$
0	—	37,69	31,63	—	—	—	—
1	54,1	40,97	33,5	—	—	—	—
...	...	...	...	—	—	—	—
10	36,3	48,55	45,59	—	—	—	—
11	75,0	53,84	47,24	52,25	—	22,75	—
12	57,2	54,51	48,69	62,09	52,17	—4,89	5,03
13	69,0	57,41	50,44	61,78	61,90	7,22	7,9
14	55,5	57,03	51,76	66,12	61,62	—10,62	—6,12
15	73,3	60,28	53,46	63,62	65,93	9,68	7,97
16	64,1	61,04	54,98	68,64	63,47	—4,54	0,63
17	60,0	60,84	—	68,64	68,14	—8,84	—8,14

Как видно из таблицы, более точным в данном примере оказывается прогноз на два шага вперед. Это объясняется тем, что изучаемое явление имеет «знакопеременную» закономерность в своем развитии, т. е. почти везде в наблюдаемом интервале времени после некоторого снижения уровня наблюдается его рост.

Высокая погрешность прогноза объясняется тем, что исходный ряд имеет большую колеблемость, в то же время он короток для того, чтобы прогнозные значения удовлетворительно адаптировались к новой информации.

Литература

1. Айвазян С. А. Статистические исследования зависимостей. М., Металлургия, 1968.
2. Венцель Е. С., Овчаров Л. А. Теория вероятностей. М., Наука, 1969.
3. Гнеденко Б. В. Курс теории вероятностей. М., Наука, 1969.
4. Демиденко Е. З. *Линейная и нелинейная регрессия.* М., Финансы и статистика, 1981.
5. Джонстон Дж. Эконометрические методы/Пер. с англ. М., Статистика, 1980.
6. Дрейпер Н., Смит Г. Прикладной регрессионный *анализ.*/Пер. с англ. М., Статистика, 1973.
7. Закс Л. Статистическое оценивание/Пер. с англ. М., Статистика, 1976.
8. Кандидаров Г., Димитров А. Таблици за изравнянане на статистически рядове за изчисляване темпове на прироста. София, 1963.
9. Кендэл М., Ранговые корреляции/Пер. с англ. М., Статистика, 1975.
10. Лизер С. Эконометрические методы и задачи/Пер. с англ. М., Статистика, 1971.
11. Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования. М., Статистика, 1979.
12. Мот Ж. Статистические предвидения и решения на предприятии/Пер. с фр. М., Прогресс, 1966.
13. Налимов В. В., Чернова О. А. Планирование экспериментов. М., Наука, 1971.
14. Пугачев В. С. Теория вероятностей и математическая статистика. М., Наука, 1979.
15. Снедекор Д. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии/Пер. с англ. М., Сельхозиздат, 1961.
16. Смирнов Н. В., Дунин-Барковский И. В. Курс теории вероятностей и математической статистики для технических приложений. М., Наука, 1965.
17. Тюрин Ю. Н. Непараметрические методы статистики. М., Знание, 1978.
18. Феллер В. *Введение в теорию вероятностей и ее приложения*/Пер. с англ. М., Изд-во иностр. лит., 1964.
19. Финни Д. Введение в *теорию планирования* экспериментов/Пер. с англ. М. Наука, 1970.
20. Хан Г., Шапиро С. *Статистические модели* в инженерных задачах/Пер с англ. М., Мир, 1969.
21. Хикс Ч. Основные принципы планирования эксперимента/Пер. с *англ.* М. Мир, 1967.
22. Четыркин Е. М. Статистические методы прогнозирования. 2-е изд. М. Статистика, 1977.
23. Янко Я. Математико-статистические таблицы/Пер. с чеш. М., Госстатиздат 1961.

Евгений Михайлович Четыркин,
Исаак Липович Калихман

ВЕРОЯТНОСТЬ И СТАТИСТИКА

Рецензент **А. М. Дубров**
Зав. редакцией **Р. А. Казьмина**
Редактор **К. С. Исаева**
Мл. редактор **Н. Е. Константинова**
Техн. редактор **Л. Г. Челышева**
Корректоры **Я. Б. Островский, Т. М. Васильева, Т. М. Иванова**
Художественный редактор **Э. А. Смирнов**
Переплет художника **В. К. Бисенгалиева**
ИБ № 1180

Сдано в набор 27.07.81. Подписано в печать 20.05.82. А07945. Формат $60\times90^1/_{16}$. Бум. тип. № 1. Гарнитура «Литературная». Печать высокая. П. л. 20, усл. п. л. 20, уч.-изд. л. 20,92. Тираж 10 000 экз. Заказ 1239. Цена 1 р. 50 к.
Издательство «Финансы и статистика»,
Москва, ул. Чернышевского, 7.

Ленинградская типография № 2 головное предприятие ордена Трудового Красного Знамени Ленинградского объединения «Техническая книга» им. Евгении Соколовой Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 198052, г. Ленинград, Л-52, Измайловский проспект, 29.