В. Ю. УРБАХ

Статистический анализ в биологических и медицинских исследованиях

АКАДЕМИЯ МЕДИЦИНСКИХ НАУК СССР

В. Ю. Урбах

СТАТИСТИЧЕСКИЙ АНАЛИЗ

в биологических и медицинских исследованиях

МОСКВА «МЕДИЦИНА» 1975

УДК 616-07 : 519.2

Издание одобрено и рекомендовано к печати редакционно-издательским советом при президиуме АМН СССР

Реферат. В книге подробно изложены вопросы статистического планирования медицинских и биологических экспериментов, предварительной статистической обработки полученного в опытах материала, оценки параметров распределения по эмпирическим данным, выявление значимости различия двух распределений; рассмотрены важные для биологических и медицинских приложений распределения (нормальное, биномиальное, пуассоновское), описаны методы регрессионного, дискриминантного, последовательного анализа и др. Особое внимание уделено детальному разбору техники вычислений.

Удачное расположение материала, необходимые математико-статистические таблицы облегчают пользование книгой.

Книга рассчитана главным образом на научных работников, ведущих исследования в различных областях медицины, физиологии и биологии.

URBAKH, V. Yu. **Statistical Analysis in Biological and Medical Research.**

This book contains statistical planning of medical and biological experiments, primary statistical preparation of empirical data, estimation of the distribution parameters on the base of sampling data. Some theoretical distributions are described important for the biological and medical applications (the normal, the binomial, and the Poissonian ones). The regression, the sequential, and the discriminant analyses are considered. The author emphasizes the details of the computation techniques.

The convenient arrangement of the material and some necessary statistical tables facilitate the use of the book.

The book is intended mainly for the restarch workers in medical, physiological, and biological areas of science.

У $\frac{50100-289}{039(01)-75}$ 95—75

ПРИНЯТЫЕ УСЛОВНЫЕ ОБОЗНАЧЕНИЯ

A — коэффициент асимметрии
$b_{y/x}$ — оценка коэффициента регрессии
d — разность рангов
E — коэффициент эксцесса
F — отношение оценок дисперсии
f — число степеней свободы
H_0 — нулевая гипотеза
H_1 — альтернативная гипотеза
h_i — относительные частоты
$M[\ ]$ — математическое ожидание
Me — медиана
Mo — мода
N — объем совокупности
n — объем выборки
n_i — частоты разрядов группировки
P — вероятность
p — доля вариант
q — априорная вероятность
R_i — ранги
r — оценка коэффициента корреляции
S — оценка ковариационной матрицы
s — оценка стандартного отклонения
s^2 — оценка дисперсии
$s_{\bar{x}}$ — оценка стандартной ошибки среднего значения
s_i — накопленные частоты
T — сумма рангов
T^{Δ} — критерий Вилкоксона для сопряженных пар
t — критерий Стьюдента
u — аргумент нормального распределения
u_P — P-процентная точка нормального распределения
v — коэффициент вариации
x_i — значения вариант
$\bar{x}$ — среднее значение
y' — пробит
y_x — условное среднее регрессионного анализа
Z — число знаков
z — преобразованный коэффициент корреляции

α — уровень значимости, вероятность ошибки I рода
β — вероятность ошибки II рода
β_{yx} — коэффициент регрессии
Δ — разность частот
δ — абсолютная неточность
ε — относительная неточность
$\theta(u)$ — площадь под кривой нормального распределения
$\theta^*(u)$ — площадь под кривой распределения Стьюдента
λ — параметр распределения Пуассона
μ — математическое ожидание
ν — параметр биномиального распределения
ν_i — частости
ξ_i — отклонения
ρ — коэффициент корреляции
ρ' — коэффициент корреляции рангов Спирмена
Σ — знак суммирования
σ — стандартное отклонение
σ^2 — дисперсия
$\sigma_{\bar{x}}$ — стандартная ошибка среднего значения
τ — критерий принадлежности варианты к совокупности
$\Phi(u)$ — интеграл вероятностей
φ — преобразованная доля вариант
χ^2 — критерий хи-квадрат Пирсона
Ψ — функция, обратная к интегралу вероятностей
$!$ — знак факториала
$|\,|$ — знак модуля (абсолютного значения)
$\langle\,\rangle$ — знак усреднения по выборкам

ПРЕДИСЛОВИЕ

Необходимость использования статистических методов в биологических и медицинских исследованиях связана в первую очередь с тем, что свойства биологических объектов обычно значительно варьируют в пределах популяций, а физиологические и другие параметры одной особи испытывают флуктуации во времени.

Настоящая книга предназначается главным образом для научных работников, ведущих исследования в различных областях биологии, физиологии, медицины. Поэтому в отличие от других пособий по биологической статистике, включая и предыдущие книги автора («Математическая статистика для биологов и медиков», 1963; «Биометрические методы», 1964), материал в этом руководстве расположен таким образом, что ключевым является решаемая исследователем конкретная задача, а не место того или иного метода в логической структуре математической статистики. Естественно, в связи с этим книга является скорее справочным, чем учебным, пособием.

Назначение книги сказалось также на отборе материала. Сравнительно много внимания уделено вопросам, которые имеют второстепенное значение в математической статистике, но являются важными для практики биологического и медицинского эксперимента. В то же время некоторые разделы, не представляющие интереса для биологов и медиков, опущены, даже если они занимают важное место в самой математической статистике. Не включен также в книгу дисперсионный анализ, так как этот раздел математической статистики имеет большое значение главным образом в агротехнических, селекционных и других сельскохозяйственных исследованиях.

В настоящем руководстве используются в основном примеры из биологии, медицины и смежных дисциплин. Следует оговориться, что цифровые данные, как правило, специально подобраны в соответствии с той целью, с которой привлекался данный пример. В некоторых случаях использованы примеры из других руководств, но и они подвергались нами тем или иным изменениям по указанным выше соображениям, и поэтому автор не счел себя вправе давать ссылки на источники, из которых взяты эти данные.

При статистическом анализе опытных данных приходится постоянно пользоваться специальными таблицами математической статистики. Таблицы, используемые во многих местах книги (напри-

мер, процентные точки нормального распределения, критические значения распределения Стьюдента, Фишера, хи-квадрат и др.), приведены в Приложениях в конце книги, причем в заголовке каждой таблицы дается ссылка на раздел текста, где эта таблица описывается. Другие специальные таблицы, имеющие «локальное» применение (критерии для проверки нормальности распределения и для отбрасывания крайних вариант, критические значения для ряда частных критериев и др.), помещены в соответствующих местах текста, но их перечень с указанием на страницы дан как в оглавлении книги, так и в Приложениях.

В книге широко используются буквы греческого алфавита. Учитывая, что в биологической и медицинской литературе эти буквы встречаются сравнительно редко, ниже мы приводим греческий алфавит (только строчные буквы):

α — альфа	ι — иота	ρ — ро
β — бэта	$\varkappa$ — каппа	σ — сигма
γ — гамма	λ — ламбда	τ — тау
δ — дельта	μ — мю	υ — ипсилон
ε — эпсилон	ν — ни	φ — фи
ξ — дзета	ξ — кси	χ — хи
η — эта	o — омикрон	ψ — пси
θ — тэта	π — пи	ω — омега

Глава первая

СТАТИСТИЧЕСКОЕ ПЛАНИРОВАНИЕ БИОЛОГИЧЕСКОГО ЭКСПЕРИМЕНТА

§ 1.1. Цели планирования эксперимента

1.1.1. Говоря здесь о планировании биологического эксперимента, мы имеем в виду не выбор биологического объекта исследования, выбор экспериментальной методики и т. п., а лишь статистическую сторону дела.

Для биологических объектов характерно то, что они в подавляющем большинстве составляют однородные популяции (виды, породы, сорта и др.), более или менее отчетливо различающиеся между собой. Именно свойства всей такой популяции интересуют исследователя даже в том случае, когда он занимается изучением отдельных особей. Собственно говоря, отдельные особи для того и изучаются, чтобы на основе данных, полученных из этого изучения, составить себе представление о свойствах популяции в целом.

Однако разные особи одной и той же популяции всегда в какой-то мере различаются между собой. Поэтому в результате измерения какого-либо признака у ряда особей будет получаться не одно, а целый ряд значений, обычно не совпадающих между собой. То же получается и при повторных экспериментах с одной и той же особью. Такой ряд значений называют *статистической совокупностью*, а каждый член этой совокупности — *вариантой*. Число вариант в совокупности называется *объемом* совокупности.

Совокупность всех вариант, которые можно было бы в принципе получить для изучаемой популяции или для выбранной постановки опыта, называют *генеральной совокупностью*. Но исследователь-биолог почти никогда не имеет в своем распоряжении генеральной совокупности: биологические популяции чаще всего чрезвычайно многочисленны, а число возможных повторений большинства экспериментов и совсем неограниченно. Поэтому обычно изучают лишь часть популяции или ставят ограниченную серию испытаний, в результате чего получается ограниченный ряд вариант, или *выборка* из генеральной совокупности. В связи с этим возникают следующие две задачи: а) выбор такого метода получения выборки, при котором эта выборка отражает свойства генеральной совокупности наиболее п р а в и л ь н о ; б) планирование объема выборки, т. е. числа наблюдений, при котором выборка отражает свойства генеральной совокупности достаточно т о ч н о.

К сожалению, чаще всего биологи-экспериментаторы пренебрегают этой важнейшей частью статистического анализа и вспоминают о статистике лишь тогда, когда эксперимент или наблюдение уже закончены и нужно заняться обработкой полученных данных. При этом нередко оказывается, что либо получено недостаточно данных (а экспериментальная установка уже разобрана или опытный материал уже израсходован на другие цели), либо проделано много лишней работы (иногда весьма трудоемкой и дорогостоящей).

Если Вы поступили именно так, то Вам незачем читать продолжение этой главы.

§ 1.2. Способы составления выборок

1.2.1. Для того чтобы свойства выборки достаточно хорошо отражали свойства генеральной совокупности, выборка должна быть составлена правильно (как принято говорить, она должна быть *репрезентативна*, т. е. «представительна»).

Существует ряд способов составления выборок, для каждого из которых имеется детально разработанная методика. Выбор того или иного способа определяется в основном конкретным характером исследования. Общее требование к составлению выборки заключается в том, чтобы в выборке были «непредвзято» представлены все возможные значения изучаемой величины, т. е. примерно в тех же пропорциях, с теми же относительными частотами, что и в генеральной совокупности.

Чаще всего предполагается, что это условие будет соблюдено, если отбирать элементы из генеральной совокупности случайно (*случайная выборка*).

Как это ни покажется парадоксальным, случайный отбор должен проводиться по определенной методике. В противном случае, как показывают опыт и специальные исследования, появляются систематические отклонения от случайности. Пусть, например, изучается распределение мышей по весу и экспериментатор, составляя выборку, руководствуется только желанием, чтобы в ней были представлены в надлежащей пропорции животные разного веса. Если ему попадутся подряд три или четыре тяжелые особи (что вполне может случиться), то он в дальнейшем невольно будет стараться «скомпенсировать» это преимущественным включением в выборку легких животных. А это внесет в составление выборки элемент, зависящий от свойств исследователя, в то время как выборка должна отражать лишь свойства генеральной совокупности. Таким же ошибочным был бы, например, отбор для опыта тех мы-

шей, которые первыми выбегут из общей клетки после открывания дверцы[1].

Этих и подобных систематических ошибок можно избежать, если пользоваться *методом случайных чисел*. Случайными числами называют последовательность чисел, выбранных из некоторой конечной (но достаточно большой) генеральной совокупности чисел при помощи какого-нибудь случайного процесса (вроде вынимания номеров при розыгрыше лотереи, но с возвратом вынутых номеров, так что отдельные номера могут попасться дважды или большее число раз). Полученную этим (или другим аналогичным) способом последовательность таких чисел записывают в виде таблицы (см. табл. I Приложений). Для удобства применения этой таблицы все числа записывают так, чтобы они имели одно и то же число цифр; если, например, таблица трехзначная, то число 6 будет записано в виде 006. Использование таблицы случайных чисел поясним на примере.

Пример 1.1. Из 146 животных, имеющихся в виварии, нужно отобрать для опыта 8. Было бы неправильно взять первые 8 по списку: может оказаться, что все они из одного помета или соседство их клеток могло как-то повлиять на их свойства и др. Поэтому мы обращаемся к таблице случайных чисел (см. табл. I Приложений). Просматривание таблицы можно начинать в любом месте и вести в произвольном направлении. Начнем, например, с начала третьей колонки (с числа 156) и будем двигаться сверху вниз; если просмотр этой колонки не даст нужного набора чисел, то мы перейдем к следующей (четвертой) колонке и т. д.

Первым числом, не превышающим 146, является 105, следующим — 002 и т. д. Продолжая таким же образом, получим набор из восьми чисел: 105, 002, 005, 098, 016, 112, 113, 119. Животных с этими номерами мы и берем в опыт.

1.2.2. Другой вид выборки — *типическая*, или *зональная*. Она состоит в том, что генеральную совокупность делят на несколько классов (типические группы, или зоны), исходя из изучаемого признака, а затем производят выборку, уже случайно, отдельно из каждого класса. Это имеет смысл делать, если в генеральной совокупности имеется какая-либо очевидная систематическая неоднородность. Пусть, например, имеется заметная уже на глаз неравномерность в густоте растений между краями поля и его серединой, между более высокой и более низкой частью и т. д. Тогда при случайной выборке из всего поля отбираемые делянки могут случайно сосредоточиться в большем количестве в одних частях поля и в меньшем количестве —

[1] Более подробно эти вопросы обсуждаются во вступительной статье В. Н. Перегудова в книге Дж. У. Снедекора (1961). Список литературы помещен в конце книги.

в других частях. Во избежание этого целесообразно разделить все поле на несколько достаточно однородных зон, а затем произвести случайную выборку в каждой зоне отдельно, с размерами выборок, пропорциональными площадям зон:

$$n_j = np_j,$$

где n_j — объем выборки в j-зоне; n — общий намеченный объем выборки; p_j — доля j-зоны (типической группы) во всей совокупности.

Пример 1.2. Поле, на котором желательно изучить выборочным методом 40 небольших делянок (чтобы определить их распределение по числу колосьев), разбито на четыре примерно однородные части. Площади этих частей составляют доли $p_j = 0{,}1$; 0,4; 0,3; 0,2 от всей площади поля. Выясним, сколько делянок должно быть выбрано на каждой из частей поля.

Самое простое — взять на каждой из четырех частей поля одно и то же число делянок, т. е. положить $n_1 = n_2 = n_3 = n_4 = 10$. Однако более близкое к истине представление о густоте колосьев на поле получится, если выбрать с каждой из частей поля соответственно: $n_1 = 40 \cdot 0{,}1 = 4$; $n_2 = 40 \cdot 0{,}4 = 16$; $n_3 = 40 \cdot 0{,}3 = 12$; $n_4 = 40 \cdot 0{,}2 = 8$ делянок — по формуле (1.1). О преимуществах зональной выборки см. также раздел 3.6.3.

Иногда применяют так называемые *механические выборки*; например, отбирают каждую десятую особь или данные на каждый 4-й день и т. д. При таком способе составления выборки нужно следить за тем, чтобы не получался «резонанс» с каким-либо периодическим процессом, могущим оказывать влияние на изучаемый признак.

§ 1.3. Планирование объема выборки

1.3.1. В этом параграфе речь будет идти о вычислении объема выборки, обеспечивающего заданную точность в оценке параметров генеральных совокупностей.

Для понимания дальнейшего требуется знать, что такое математическое ожидание, среднее значение, дисперсия, стандартная ошибка. Если Вы не знаете этого (или забыли), прочитайте сначала разделы **3.1.2**, **3.2.1**, **3.3.1**. *О специальной методике постановки опыта, при которой не требуется заранее планировать объем выборки (так называемом последовательном анализе), см. § 4.5.*

Рассмотрим два варианта задачи:

1) совокупность подчиняется нормальному распределению, оцениваемым параметром является математическое ожидание;

2) совокупность имеет альтернативное распределение, оцениваемым параметром является доля (или процент) вариант одного типа.

О нормальном распределении и его свойствах Вы можете прочитать в § 2.3, а об альтернативном распределении — в главе шестой. Если Вы имеете дело со вторым вариантом задачи, то можете не читать продолжение этого раздела, а перейти к следующему разделу **1.3.2.**

Пусть мы задаемся некоторой определенной точностью в оценке математического ожидания μ, т. е. хотим, чтобы неточность в этой оценке не превышала некоторого заданного значения δ. Указанная неточность определяется величиной $u_P\sigma_{\bar{x}}$, где u_P есть P-квантиль нормального распределения $\theta(u)$, а $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ — стандартная ошибка выборочного среднего значения $\bar{x}$. Таким образом, имеем условие $u_P\,\sigma/\sqrt{n} \leqslant \delta$, откуда:

$$n \geqslant \frac{u_P^2\sigma^2}{\delta^2}. \qquad (1.2)$$

Например, при $\sigma = 4{,}3$ мг мы хотим, чтобы с вероятностью $P = 0{,}95$ неточность в определении μ была не больше 0,8 мг. Тогда по формуле (1.2) находим, что выборка должна содержать не менее ($u_{0,95} = 1{,}96$):

$$n^* = 1{,}96^2\,\frac{4{,}3^2}{0{,}8^2} \approx 106$$

элементов. Конечно, для такого расчета надо знать величину σ, обычно неизвестную до исследования. Но если заведомо ясно, что придется сделать много повторных измерений, имеет смысл сделать предварительное, «прикидочное» исследование с небольшой выборкой для нахождения ориентировочного значения σ.

Обычно задаются не абсолютной неточностью $u_P\,\sigma_{\bar{x}}$, а относительной неточностью:

$$\varepsilon = \frac{u_P\,\sigma_{\bar{x}}}{\bar{x}},$$

где $u_P\,\sigma_{\bar{x}}$ характеризует ширину доверительного интервала для μ. Очевидно:

$$\varepsilon = \frac{u_P}{\bar{x}}\,\frac{\sigma}{\sqrt{n}} = \frac{u_P}{\sqrt{n}}\,\frac{\sigma}{\bar{x}} = \frac{u_P}{\sqrt{n}}\,v,$$

где v — коэффициент вариации. Отсюда получаем наименьшее допустимое значение n:

$$n^* = \frac{u_P^2}{\varepsilon^2} v^2, \qquad (1.3)$$

причем значение u_P определяется принятой доверительной вероятностью P.

О коэффициенте вариации говорится в разделе **3.3.4**

В некоторых случаях возможности увеличить объем выборки ограничены либо вследствие ограниченности экспериментального материала, находящегося в распоряжении исследователя, либо (при экспериментах, повторяемых во времени) вследствие необходимости разумно ограничить общее время проведения работы. В этом случае значение $\sigma_{\bar{x}}$ не может быть сделано сколь угодно малым, т. е. значение μ не может быть определено сколь угодно точно. Но тогда не имеет смысла измерять исходные значения вариант с очень большой точностью. Имеется эмпирическое правило, согласно которому положение последней значащей цифры в окончательном результате должно соответствовать положению первой значащей цифры в величине $\sigma_{\bar{x}}/3$. Пусть, например, расчет дал $\bar{x} = 2{,}3086$ кг, а $\sigma_{\bar{x}} = 0{,}16$ кг; так как $\sigma_{\bar{x}}/3 \approx 0{,}05$ кг, то $\bar{x}$ надо округлить до сотых долей килограмма, т. е. принять $\bar{x} = 2{,}31$ кг. Во избежание накопления ошибок в промежуточных расчетах целесообразно проводить эти расчеты с точностью на один порядок больше, чем точность окончательного результата. С этой же точностью, очевидно, следует производить и измерения. Так, в приведенном выше примере точность измерений должна составлять 0,001 кг = 1 г. Это надо всегда учитывать при планировании опыта во избежание лишней и неоправданной работы (достижение большей точности измерений часто связано с существенным усложнением методики).

1.3.2. Перейдем к случаю альтернативного распределения, когда оценивается доля (процент) вариант одного из двух типов.

Прежде чем читать дальше, обратитесь к § 6.2, в котором говорится о стандартной ошибке доли (процента) вариант. О доверительном интервале см. раздел **3.7.1.**

Если ожидаемое значение p не очень мало, то неточность результата лучше всего характеризуется абсолютной погрешностью $\delta = u_P \sigma_p$. При заданной δ наименьшее допустимое значение n^* определяется равенством:

$$\delta \approx u_P \sigma_p^* = u_P \sqrt{\frac{p(100-p)}{n^*}}; \quad n^* \approx \frac{u_P^2}{\delta^2} p(100-p).$$

Перед началом исследования значение p, конечно, неизвестно. Поэтому приходится при вычислении n^* брать то значение p, при котором $p\,(100-p)$ является наибольшим. Очевидно, это будет при $p = 50\%$, когда $p(100-p) = 2500$, так что:

$$n^* = \frac{u_P^2}{\delta^2}\,2500, \quad \delta \text{ в } \%. \tag{1.4}$$

Если, например, желательно, чтобы 95%-ный доверительный интервал для $\hat{p}$ составлял $\delta = 5\%$, то требуется ($u_{95\%} = 1{,}96$):

$$n > \frac{1{,}96^2}{5^2}\,2500 = 384.$$

Когда ожидаемое значение p мало, лучшей характеристикой неточности результата является относительная погрешность $\varepsilon = u_P \sigma_p / p$. Так как в этом случае $100 - p \approx 100$, то:

$$\varepsilon = \frac{u_P}{\sqrt{n^*}}\sqrt{\frac{100-p}{p}} \approx \frac{u_P}{\sqrt{n^*}}\sqrt{\frac{100}{p}},$$

откуда:

$$n^* = \frac{u_P^2}{\varepsilon^2}\,\frac{100}{p}; \tag{1.5}$$

в этом случае (т. е. при малых долях) для нахождения n^* нужно иметь какую-нибудь предварительную оценку p.

§ 1.4. Планирование регрессионного эксперимента

1.4.1. В § 1.3 рассматривались вопросы, связанные с планированием объема выборок в экспериментах, целью которых является получение количественной информации о тех или иных свойствах (параметрах) биологических объектов. Более общей задачей является планирование эксперимента с целью изучить зависимость этих параметров от каких-либо факторов или, говоря на языке математики, получить соответствующее регрессионное уравнение[1]. При этом требуется предварительно, до начала эксперимента, решить следующие вопросы:

1) выбрать модель регрессионной зависимости (линейная, квадратичная или какая-либо другая нелинейная);

б) выбрать границы интервалов, в которых мы намерены варьировать значения каждого из рассматриваемых факторов;

в) выбрать «шаг» варьирования для каждого фактора.

[1] Регрессии посвящена глава восьмая.

Для правильного выбора регрессионной модели требуется иметь хотя бы общее представление о характере изучаемой зависимости (из литературных данных, исходя из теоретических соображений и др.). Если оно отсутствует, то часто используют линейную модель, считая, что она во всяком случае может служить первым приближением.

Выбор диапазона варьирования значений факторов обычно определяется задачей исследования. При этом надо иметь в виду, что экстраполяция уравнений регрессии за пределы изученного диапазона хотя и возможна, но крайне нежелательна, так как сопровождается расширением доверительной зоны (подробнее об этом см. в разделе 8.5.1).

Что касается выбора шага варьирования в исследуемом диапазоне изменения фактора, то чаще всего значения, для которых ставится эксперимент, выбирают равноотстоящими. Однако это не всегда наилучшее решение. В общем случае выбор расположения и числа уровней фактора зависит от количества априорной информации о виде исследуемой зависимости. Например, при линейной зависимости целесообразно располагать экспериментальные точки не равномерно по диапазону варьирования, а на краях этого диапазона; обоснование такого вывода дается в разделе 8.5.4.

1.4.2. Особенно важно правильно спланировать регрессионный эксперимент, когда изучается зависимость интересующего нас параметра сразу от нескольких факторов, так как здесь может быть достигнута особенно большая экономия времени и средств.

Если первый фактор варьируют на m_1 уровнях, второй на m_2 уровнях и т. д., то при k факторах требуется произвести $m_1 \cdot m_2 \ldots m_k$ экспериментов, чтобы получить достаточно полное представление об изучаемой зависимости[1]. Такой план опыта называется *полным факторным экспериментом*. Когда число уровней выбрано для всех факторов одинаковым ($m_1 = m_2 = \ldots = m_k = m$), то требуемое число экспериментов равно m^k.

Число экспериментов может быть значительно уменьшено, если имеются основания считать, что зависимость изучаемого параметра от каждого фактора линейна. Тогда можно ограничиться лишь двумя уровнями каждого фактора, так что число экспериментов будет 2^k. Выбор уровней факторов в известной мере произволен, однако определенные преимущества имеет такой выбор, при котором диапазон варьирования фактора берется тем шире, чем меньше предполагаемое влияние этого фактора на результирующий параметр; обоснование такого выбора дается в разделе 1.5.3.

Расчеты значительно упрощаются, если численные значения фактора нормированы таким образом, что середина интервала варь-

[1] Вопрос о необходимости повторных экспериментов, позволяющих оценивать значимость получаемых оценок, рассматривается ниже.

ирования принята за начало отсчета (нулевой уровень), а в качестве единицы масштаба выбрана полуширина этого интервала. Тогда, очевидно, нижние уровни всех факторов будут иметь значения -1, а верхние уровни — значения $+1$.

При составлении плана факторного эксперимента заготавливают таблицу всех возможных комбинаций уровней факторов. Эта таблица носит название *матрицы плана*. Например, при трех факторах ($k = 3$) полный факторный эксперимент содержит $2^3 = 8$ комбинаций уровней и матрица плана может иметь вид табл. 1.1 (уровни факторов $+1$ и -1 записаны просто как $+$ и $-$). Матрица плана для четырех факторов получается, если эту матрицу для трех факторов повторить дважды: один раз для $x_4 = -1$ и другой раз для $x_4 = +1$. Так же получают матрицы планов для пяти, шести и большего числа факторов.

ТАБЛИЦА 1.1

Номер опыта	Факторы		
	x_1	x_2	x_3
1	—	—	—
2	+	—	—
3	—	+	—
4	+	+	—
5	—	—	+
6	+	—	+
7	—	+	+
8	+	+	+

В принятой линейной модели уравнение регрессии имеет следующий вид:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \qquad (1.6)$$

где β_0, β_1, β_2, ..., β_k — параметры, которые должны быть оценены по результатам эксперимента. Но при любом $k \geqslant 2$ общее число опытов в полном факторном эксперименте (т. е. 2^k) превышает число параметров линейной модели, которые нужно оценить (т. е. $k + 1$). Это обстоятельство можно использовать трояко: а) расширить модель, доведя число оцениваемых параметров до числа проведенных опытов; б) использовать остающиеся $2^k - (k + 1)$ степеней свободы[1] для нахождения стандартных ошибок[1] оцениваемых $k + 1$ параметров; в) уменьшить число опытов, спланировав разумным образом так называемой дробный факторный эк-

[1] Подробнее о степенях свободы см. в разделе 3.4.1, а о стандартных ошибках — в разделе 3.5.1.

сперимент. Ниже мы подробнее рассмотрим эти три возможности.

1.4.3. «Линейная» модель, которая при k факторах содержит 2^k параметров, имеет вид:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i,j} \beta_{ij} x_i x_j + \ldots + \beta_{12\ldots k} x_1 x_2 \ldots x_k, \quad (1.7)$$

т. е. она включает «свободный член» β_0, сами факторы x_i, все возможные произведения этих факторов по два, по три и т. д. до произведения всех факторов. Модель названа условно «линейной» потому, что она не включает вторых и более высоких степеней факторов, но она, разумеется, не является действительно линейной из-за произведений факторов.

Например, при $k = 3$ получаем модель:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3.$$

Члены $\beta_{ij} x_i x_j$ характеризуют так называемое *взаимодействие* двух факторов, а член $\beta_{123} x_1 x_2 x_3$ — тройное взаимодействие. Термин «взаимодействие факторов» не следует понимать буквально. В данном случае этот термин означает, что результат действия одного фактора зависит от того, на каком уровне варьирования находится другой фактор. Так, члены $\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$ регрессионного уравнения можно записать в любой из двух форм:

$$\beta_1 x_1 + (\beta_2 + \beta_{12} x_1) x_2$$

или

$$(\beta_1 + \beta_{12} x_2) x_1 + \beta_2 x_2.$$

В первом случае $\beta_2 + \beta_{12} x_1$ можно рассматривать как регрессионный коэффициент для x_2, и сразу видно, что он зависит от x_1; во втором случае $\beta_1 + \beta_{12} x_2$ есть регрессионный коэффициент для x_1, и он зависит от x_2. В некоторых случаях влияние одного фактора может при разных значениях другого фактора оказаться даже разных знаков; так, повышение влажности стимулирует рост растений при высоких, но замедляет его при низких температурах.

По результатам эксперимента y_ν (ν — номер опыта, принимающий значения от 1 до $N = 2^k$) можно найти выборочные оценки[1] для параметров β_0, β_i, β_{ij} и т. д.; эти оценки обозначим b_0, b_i, b_{ij}.... Вычисления можно производить, пользуясь общими методами регрессионного анализа, описанными в § 8.1 и 8.2, т. е. ре-

[1] О выборочных оценках параметров см. раздел 3.2.3.

шая так называемую систему нормальных уравнений. Общее решение этой системы имеет вид:

$$\left.\begin{aligned} b_0 &= \frac{1}{N}\sum_{\nu=1}^{N} y_\nu, \\ b_i &= \frac{1}{N}\sum_{\nu=1}^{N} y_\nu x_{i\nu}, \\ b_{ij} &= \frac{1}{N}\sum_{\nu=1}^{N} y_\nu x_{i\nu} x_{j\nu} \end{aligned}\right\} \qquad (1.8)$$

и т. д. Ввиду особой простоты системы уровней факторов (все $x_{i\nu}$ имеют значения только либо -1, либо $+1$) вычисления сводятся просто к алгебраическим суммированиям величин y_ν со знаками плюс (если $x_{i\nu}$ или $x_{i\nu}x_{j\nu}$ и т. д. равны $+1$) или минус (если $x_{i\nu}$ ... равны -1) и последующему делению результата на число опытов плана $N = \Sigma x^2_{i\nu} = 2^k$. Так, для матрицы планирования из табл. 1.1 имеем:

$$b_0 = (y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8)/8,$$

$$b_1 = (-y_1 + y_2 - y_3 + y_4 - y_5 + y_6 - y_7 + y_8)/8,$$

$$b_2 = (-y_1 - y_2 + y_3 + y_4 - y_5 - y_6 + y_7 + y_8)/8,$$

$$b_3 = (-y_1 - y_2 - y_3 - y_4 + y_5 + y_6 + y_7 + y_8)/8,$$

$$b_{12} = (y_1 - y_2 - y_3 + y_4 + y_5 - y_6 - y_7 + y_8)/8$$

и т. д. Для удобства расчетов матрицу плана обычно дополняют столбцами соответствующих взаимодействий факторов. Пример такой расширенной матрицы, обобщающей матрицу плана, задающего условия опытов из табл. 1.1, приведен в табл. 1.2. Знаки для столбцов взаимодействий получают путем перемножения знаков столбцов факторов по правилам: $(+1)\,(+1) = (-1)\,(-1) = +1$, $(+1)\,(-1) = (-1)\,(+1) = -1$ и т. д. Для полного единообразия считают, что и β_0 есть коэффициент при некоторой фиктивной переменной x_0, имеющей всегда значения $+1$, и вводят соответствующий столбец в матрицу планирования. В последнем столбце матрицы записывают результаты опытов.

ТАБЛИЦА 1.2

Номер опыта	Факторы				Взаимодействия				Результат опыта
	x_0	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$	
1	+	−	−	−	+	+	+	−	y_1
2	+	+	−	−	−	−	+	+	y_2
3	+	−	+	−	−	+	−	+	y_3
4	+	+	+	−	+	−	−	−	y_4
5	+	−	−	+	+	−	−	+	y_5
6	+	+	−	+	−	+	−	−	y_6
7	+	−	+	+	−	−	+	−	y_7
8	+	+	+	+	+	+	+	+	y_8

Пример 1.3. Пусть изучается влияние на выход продукта биохимического процесса трех факторов: температуры (уровни $x_1 = -1$ и $x_1 = +1$ соответствуют 27° и 33°), pH (уровни $x_2 = -1$ и $x_2 = +1$ соответствуют 6,8 и 7,2) и концентрации определенного фермента (уровни $x_3 = -1$ и $x_3 = +1$ соответствуют 44 и 52 в некоторых единицах). Значения выхода продукта (в условных единицах) при разных комбинациях уровней факторов (т. е. в каждом из 8 опытов полного факторного эксперимента) записаны в последнем столбце табл. 1.3.

ТАБЛИЦА 1.3

	x_0	Факторы; планирование эксперимента			Взаимодействия				Результаты опытов
		$t°$	pH	c					
		x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$	
Нулевой уровень		30	7,0	48					
Шаг варьирования		3	0,2	4					
Нижний уровень (−)		27	6,8	44					
Верхний уровень (+)		33	7,2	52					
Номер опыта:									
1	+	−	−	−	+	+	+	−	86
2	+	+	−	−	−	−	+	+	93
3	+	−	+	−	−	+	−	+	114
4	+	+	+	−	+	−	−	−	79
5	+	−	−	+	+	−	−	+	80
6	+	+	−	+	−	+	−	−	103
7	+	−	+	+	−	−	+	−	99
8	+	+	+	+	+	+	+	+	71
Коэффициенты регрессии	90,62	−4,13	0,12	−2,38	−11,62	2,88	−0,88	−1,12	

По этим данным находим, беря знаки из матрицы планирования:

$$b_0 = \frac{1}{8}(+86+93+114+79+80+103+99+71) =$$

$$= +725 : 8 = +90{,}62,$$

$$b_1 = \frac{1}{8}(-86+93-114+79-80+103-99+71) =$$

$$= -33 : 8 = -4{,}13.$$

$$b_2 = \frac{1}{8}(-86-93+114+79-80-103+99+71) =$$

$$= +1 : 8 = 0{,}125,$$

$$b_3 = \frac{1}{8}(-86-93-114-79+80+103+99+71) =$$

$$= -19 : 8 = -2{,}38,$$

$$b_2 = \frac{1}{8}(+86-93-114+79+80-103-99+71) =$$

$$= -93 : 8 = -11{,}62,$$

$$b_{13} = \frac{1}{8}(+86-93+114-79-80+103-99+71) =$$

$$= +23 : 8 = +2{,}88,$$

$$b_{23} = \frac{1}{8}(+86+93-114-79-80-103+99+71) =$$

$$= -7 : 8 = -0{,}875,$$

$$b_{123} = \frac{1}{8}(-86+93+114-79+80-103-99+71) = -1{,}125.$$

Таким образом, уравнение регрессии есть:

$$y = 90{,}62 - 4{,}13x_1 + 0{,}12x_2 - 2{,}38x_3 - 11{,}62x_1x_2 + 2{,}88x_1x_3 - 0{,}88x_2x_3 - 1{,}12x_1x_2x_3. \qquad (*)$$

Теперь можно перейти от нормированных к «натуральным» единицам, учитывая, что

$$x_1 = \frac{t - t_0}{\lambda_t}, \text{ где } t_0 = \frac{33° + 27°}{2} = 30°, \ \lambda_t = \frac{33° - 27°}{2} = 3°,$$

$$x_2 = \frac{\pi - \pi_0}{\lambda_\pi}, \text{ где } \pi_0 = \frac{7{,}2 + 6{,}8}{2} = 7{,}0, \quad \lambda_\pi = \frac{7{,}2 - 6{,}8}{2} = 0{,}2,$$

$$x_3 = \frac{c - c_0}{\lambda_c}, \text{ где } c_0 = \frac{52 + 44}{2} = 48, \quad \lambda_c = \frac{52 - 44}{2} = 4,$$

причем t, π и c обозначают соответственно температуру, pH и концентрацию. Подставляя эти значения в уравнение (*), получим:

$$y = 90{,}62 - 1{,}38\,(t - 30) + 0{,}625\,(\pi - 7{,}0) - 0{,}595\,(c - 48) -$$
$$- 19{,}38\,(t - 30)(\pi - 7{,}0) + 0{,}240\,(t - 30)(c - 48) -$$
$$- 1{,}094\,(\pi - 7{,}0)(c - 48) - 0{,}469\,(t - 30)(\pi - 7{,}0)(c - 48). \quad (**)$$

1.4.4. Представление о роли каждого из факторов и их взаимодействиях можно получить, вычислив отношение выборочной оценки соответствующего коэффициента регрессии к ее стандартной ошибке или, иначе говоря, оценив значимость этой оценки.

Прежде чем читать дальше, надо ознакомиться с понятиями стандартной ошибки (см. раздел **3.5.1***), значимости (см. раздел* **4.1.3***), t-критерия Стьюдента (см. раздел* **4.2.1***), F-критерия Фишера (см. раздел* **4.6.1***).*

Последнее можно выполнить для всех 2^k параметров модели (1.7) только тогда, когда полный факторный эксперимент (состоящий из 2^k опытов) проведен не менее двух раз[1]. Вычисление стандартных ошибок коэффициентов регрессии можно производить, пользуясь общими методами регрессионного анализа (см. § 8.5). Приведем формулу для данного частного случая: если полный k-факторный эксперимент, состоящий из $N = 2^k$ опытов ($\nu = 1, 2, \ldots, N$) повторен n раз ($m = 1, 2, \ldots, n$), то квадрат стандартной ошибки любого регрессионного коэффициента b_ν равен:

$$s^2_{b_\nu} = \sum_{\nu=1}^{N} \sum_{m=1}^{n} \left(y_{\nu m} - \bar{y}_\nu\right)^2 / Nn\,(n - 1), \quad (1.9)$$

причем

$$\bar{y}_\nu = \sum_{m=1}^{n} y_{\nu m}/n.$$

[1] В этом случае при вычислении коэффициентов регрессии используются значения: $\bar{y}_\nu = \sum_{m=1}^{n} y_{\nu m}/n$, где n—число повторностей и $m = 1, 2, \ldots, n$ — номера этих повторностей.

Коэффициент регрессии считается незначимым и исключается из модели, если $|b_\nu| < t_\alpha s_{b_\nu}$, где t_α — значение критерия Стьюдента для числа степеней свободы $f = N(n-1)$ и выбранного уровня значимости α. В некоторых случаях можно исключать те или иные факторы и взаимодействия из модели, основываясь только на сравнении соответствующих коэффициентов регрессии. Например, из модели, описываемой уравнением (*), представляется возможным исключить члены b_2x_2, $b_{23}x_2x_3$ и $b_{123}x_1x_2x_3$, т. е. принять для данных из табл. 1.3 модель:

$$y = \beta_0 + \beta_1x_1 + \beta_3x_3 + \beta_{12}x_1x_2 + \beta_{13}x_1x_3. \qquad (***)$$

Тот факт, что какой-то коэффициент регрессии (например, b_2) оказался незначимым, не всегда означает, что соответствующий фактор вообще не влияет на результат,— может оказаться, что в данном эксперименте его влияние не сказывается только потому, что для этого фактора был выбран слишком узкий интервал варьирования.

Если повторные опыты в эксперименте отсутствуют и принята «укороченная» модель вида (***), то варьирование значений y, связанное с отброшенными членами, должно рассматриваться как случайное и с ним можно сравнивать варьирование, связанное с оставшимися членами, для оценки значимости последних. Можно показать[1], что в комплексах типа 2^k сумма квадратов отклонений, связанная с каждым из факторов или взаимодействий, равна $SS_\nu = Nb_\nu^2$, где $N = 2^k$. Поэтому случайное («остаточное») варьирование можно считать равным:

$$\sum_{\nu'} SS_{\nu'} = N\sum_{\nu'} b_{\nu'}^2,$$

где ν' — те значения ν, которые не включены в модель. Число степеней свободы этой величины равно числу членов, не включенных в модель (обозначим это число через N'); тогда оценка случайной дисперсии равна:

$$s_Z^2 = N\frac{1}{N'}\sum_{\nu'} b_{\nu'}^2.$$

Дисперсия, связанная с каждым из оставленных членов модели оценивается величиной $s_\nu^2 = SS_\nu/f = Nb_\nu^2$, так как здесь число степеней свободы f равно единице. Значимость ν-го фактора или взаимодействия, учтенного в модели, оценивается по отношению:

$$F = \frac{s_\nu^2}{s_Z^2} = b_\nu^2 \Big/ \frac{1}{N'}\sum b_{\nu'}^2.$$

[1] См. *Е. В. Маркова, А. Н. Лисенков.* Планирование эксперимента в условиях неоднородностей. М., «Наука», 1973.

Однако из табл. VI и IV Приложений видно, что если число степеней свободы числителя равно единице, то $F_\alpha(f_1, f_2) = t^2_\alpha(f_2)$; например, $F_{0,05}(1; 18) = 4{,}41$; $t_{0,05}(18) = 2{,}10$, причем $4{,}41 = (2{,}10)^2$. Поэтому использование записанного выше отношения можно заменить сравнением отношения

$$t = |b_\nu| \Big/ \sqrt{\frac{1}{N'} \sum_{\nu'} b^2_{\nu'}}$$

с критическим значением распределения Стьюдента (при числе степеней свободы, равном N'). Но тогда величину $\sqrt{\frac{1}{N'} \sum_{\nu'} b^2_{\nu'}}$ можно рассматривать как стандартную ошибку коэффициентов регрессии в «укороченной» модели, а выполнение неравенства

$$|b_\nu| > t_\alpha(N') \sqrt{\frac{1}{N'} \sum_{\nu'} b^2_{\nu'}} \qquad (1.10)$$

можно считать условием значимости соответствующего коэффициента регрессии.

Пример 1.4. Оценим значимость оценок коэффициентов регрессии, если для данных из табл. 1.3 принята модель (***). В данном случае:

$$\sqrt{\frac{1}{N'} \sum_{\nu'} b^2_{\nu'}} = \sqrt{\frac{1}{3}(0{,}12^2 + 0{,}88^2 + 1{,}12^2)} \approx 0{,}83.$$

Если принять уровень значимости $\alpha = 0{,}05$, то $t_{0.05}(3) = 3{,}18$, так что значимым (с вероятностью 0,95) следует считать только те b_ν, которые по абсолютной величине превышают $3{,}18 \cdot 0{,}83 = 2{,}64$. Этому условию удовлетворяют все b_ν из модели (***), кроме $b_3 = -2{,}38$. Но так как b_3 ненамного меньше требуемого 2,64, то имеет смысл оставить член $\beta_3 x_3$ в модели, удовлетворившись значимостью, несколько меньшей 0,95.

К вопросам, рассматриваемым в настоящем разделе, примыкает вопрос о проверке того, правомерно ли было исключение из модели членов, содержащих вторую и более высокие степени факторов. Такая проверка основана на том, что величина b_0, получаемая в регрессионном анализе, является, вообще говоря, оценкой не для коэффициента β_0, а для суммы $\beta_0 + \Sigma\beta_{ii} + \Sigma\beta_{iii} + \ldots$, где β_{ii}, β_{iii} и т. д.— коэффициенты при x_i^2, x_i^3 и т. д. Очевидно, величину b_0 можно рассматривать как оценку для β_0 только в модели, где $\beta_{ii} = 0$, $\beta_{iii} = 0$ и т. д. В то же время можно получить чистую оценку для β_0, не зависящую от значений β_{ii}, β_{iii} ..., если поставить дополнительный опыт в «центре эксперимента», т. е. при $\tilde{x}_1 = \tilde{x}_2 = \ldots = 0$, так как тогда $\beta_0 + \sum_i \beta_{ii} x_i^2 + \sum_i \beta_{iii} x_i^3 + \ldots = \beta_0$. Это приводит к следую-

щему критерию: члены второй и более высоких степеней нельзя исключать из модели, если:

$$\frac{|y_0 - b_0|}{s_{b_0}} > t_\alpha(f), \qquad (1.11)$$

где y_0 — результат опыта в «центре эксперимента», а f — число степеней свободы величины $s^2_{b_0}$.

1.4.5. В конце раздела 1.4.2 было отмечено, что если выбрана линейная модель регрессии, то число параметров, подлежащих оценке, т. е. $k+1$, всегда меньше числа опытов в полном факторном эксперименте с данным числом факторов, т. е. меньше 2^k. Если нет достаточной уверенности в адекватности линейной модели, то целесообразно использовать остающиеся $2^k-(k+1)$ степени свободы для оценки взаимодействий факторов, как это показано в разделе 1.4.3. Если же адекватность линейной модели не вызывает сомнений, то в таком употреблении «лишних» степеней свободы нет надобности, и тогда их можно использовать для нахождения стандартных ошибок оцениваемых $k+1$ параметров (см. раздел 1.4.4). Но при большом числе факторов таких «лишних» степеней свободы оказывается очень много (например, при $k=5$ их 26, при $k=6$ их уже 57 и т. д.), что делает эксперимент явно неэкономным. В таком случае целесообразно уменьшить число опытов, спланировав так называемый *дробный факторный эксперимент*.

Пусть, например, изучается влияние четырех факторов ($k=4$). Тогда в линейной модели требуется оценить пять параметров $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$, в то время как полный факторный эксперимент требует постановки $2^4=16$ опытов (табл. 1.4). Для полу-

ТАБЛИЦА 1.4

Номер опыта	Факторы				Взаимодействия		
	x_1	x_2	x_3	x_4	x_1x_2	$x_1x_2x_3$ . . .	$x_1x_2x_3x_4$
1	−	−	−	−	+	−	+
2	+	−	−	−	−	+	−
3	−	+	−	−	−	+	−
4	+	+	−	−	+	−	+
5	−	−	+	−	+	+	−
6	+	−	+	−	−	−	+
7	−	+	+	−	−	−	+
8	+	+	+	−	+	+	−
9	−	−	−	+	+	−	−
10	+	−	−	+	−	+	+
11	−	+	−	+	−	+	+
12	+	+	−	+	+	−	−
13	−	−	+	+	+	+	+
14	+	−	+	+	−	−	−
15	−	+	+	+	−	−	−
16	+	+	+	+	+	+	+

чения нужных оценок было бы достаточно пяти опытов, однако обычно в таких случаях планируют эксперимент так, чтобы он включал половину, или одну четверть, или одну восьмую и т. д. от числа всех опытов полного факторного эксперимента[1]. Так, при пяти факторах требуется оценить шесть параметров, и тогда планируют одну четверть от полного факторного эксперимента, включающего $2^5 = 32$ опыта, что дает опять 8 опытов; при шести факторах ($k + 1 = 7$) достаточно одной восьмой от полного факторного эксперимента, включающего $2^6 = 64$ опыта, и т. д. Набор опытов полного факторного эксперимента обычно называют *репликой*, а для дробных факторных экспериментов используются термины *полуреплика* (или $^1/_2$-реплика), *четверть-реплика* ($^1/_4$-реплика) и т. д. Кроме того, используются следующие обозначения: полуреплика от реплики 2^4 будет 2^{4-1}, $^1/_8$-реплика от реплики 2^6 будет 2^{6-3} и т. д. Скажем, для 9-факторного эксперимента требуется не менее 10 опытов, так что берем наименьшую дробную реплику вида 2^{9-p}, для которой $2^{9-p} \geqslant 10$; ясно, что это будет дробная реплика $2^{9-p} = 16 (= 2^4)$, так что $p = 5$, т. е. берем $^1/_{32}$-реплику вида 2^{9-5}.

Теперь рассмотрим вопрос о построении матрицы плана при использовании дробных реплик. Очевидно, нельзя просто механически отделить нужную долю от полной реплики. Например, если при построении полуреплики 2^{4-1} просто взять половину от полной реплики 2^4 (что дает матрицу, обведенную рамкой в табл. 1.4), то в эксперименте фактор x_4 должен будет во всех восьми опытах поддерживаться на одном уровне, что не позволит оценить влияние этого фактора. Ясно, что уровни фактора x_4 нужно тоже както варьировать, причем желательно, чтобы в четырех опытах он был на нижнем уровне и в четырех опытах — на верхнем; иначе говоря, столбец x_4 должен, как и столбцы x_1, x_2, x_3, содержать четыре минуса и четыре плюса. Желательно также, чтобы расстановка минусов и плюсов в столбце x_4 не совпадала с расстановкой их в каком-либо из столбцов x_1, x_2, x_3.

Эти два условия могут быть выполнены, если «позаимствовать» расстановку минусов и плюсов из какого-нибудь столбца взаимодействий. (Напомним, что последние не представляют реального планирования условий эксперимента.) Какое же из взаимодействий следует использовать? В принципе можно взять любое, но обычно берут взаимодействие наивысшего порядка (в данном примере — четверное взаимодействие $x_1x_2x_3x_4$). Причина состоит в том, что чаще всего (хотя и не всегда) роль взаимодействий в многофактор-

[1] В противном случае нарушается так называемая ортогональность реплики, в результате чего столбцы плана перестают быть независимыми. Подробнее об этом см. в книге В. В. Налимова и Н. А. Черновой, гл. 2, § 2.

ной регрессии убывает с ростом порядка этих взаимодействий. Поэтому, если столбцы двойных взаимодействий еще могут понадобиться для оценки соответствующих коэффициентов регрессии (напомним, что в дробном факторном эксперименте 2^{4-1} еще имеются три лишних степени свободы, так как планируются восемь опытов для оценки пяти параметров), то необходимостью проверки гипотезы $\beta_{1234} = 0$ можно пренебречь.

Если для полуреплики 2^{4-1} выбрана вторая половина реплики 2^4, где для всех опытов $x_4 = 1$, то расстановка минусов и плюсов в столбце $x_1x_2x_3x_4$ такова же, что и в столбце $x_1x_2x_3$. Поэтому можно считать, что для планирования уровней фактора x_4 выбран столбец $x_1x_2x_3$. Но тогда можно рассматривать видоизмененную (путем использования расстановки знаков из столбца $x_1x_2x_3x_4$ для планирования уровней фактора x_4) полуреплику 2^{4-1} как видоизмененную же (путем использования расстановки знаков столбца $x_1x_2x_3$ для планирования уровней добавочного фактора x_4) полную реплику 2^3. Так обычно и поступают: для дробного факторного эксперимента 2^{4-1} строят матрицу плана полного факторного эксперимента 2^3 путем введения дополнительного фактора x_4 с уровнями, взятыми из столбца $x_1x_2x_3$. В отличие от реплики 2^3, где столбец $x_1x_2x_3$ используется только для расчета коэффициента b_{123}, в полуреплике 2^{4-1} этот столбец определяет реальное варьирование фактором x_4.

При пяти факторах можно снова использовать видоизмененную реплику 2^3, рассматривая ее как четверть-реплику 2^{5-2}. Но здесь уже нужно использовать для планирования уровней двух дополнительных факторов, x_4 и x_5, расстановку знаков из двух столбцов взаимодействия (теряя при этом возможность оценить соответствующие коэффициенты регрессии). Один берется тот же, что и в полуреплике 2^{4-1}, т. е. столбец $x_1x_2x_3$. Что касается второго, то им может быть любой из столбцов x_1x_2, x_1x_3, x_2x_3. Однако целесообразно использовать тот из них, для которого гипотеза $\beta_{ij} = 0$ наиболее вероятна (из каких-либо априорных соображений). Если же для однозначного выбора нет никаких особых оснований, то можно прибегнуть к таблице случайных чисел (о пользовании которой см. раздел **1.2.1**). Очевидно, лишние степени свободы, которые отвечают взаимодействиям, не использованным для планирования уровней дополнительных факторов, могут служить либо для оценки соответствующих коэффициентов регрессии (b_{ij}), либо для нахождения стандартных ошибок коэффициентов регрессии линейных членов (b_i). Так, если в дробном факторном эксперименте типа 2^{5-2} принято $x_4 = x_1x_2x_3$ и $x_5 = x_1x_3$, то можно, помимо получения оценок b_0, b_1, b_2, b_3, b_4, b_5, получить еще либо оценки b_{12}, b_{23}, либо оценки стандартных ошибок (с двумя степенями свободы) коэффициентов b_0, b_1, ..., b_5.

Так же строятся планы и для других дробных факторных экспериментов. Более подробно об этом можно прочитать в книге В. В. Налимова и Н. А. Черновой (см. список литературы в конце книги) (гл. 2, § 1 и 3).

§ 1.5. Планирование экстремального эксперимента

1.5.1. Экстремальным называют эксперимент, целью которого является найти ту комбинацию численных значений факторов, при которой интересующая нас функция от этих факторов принимает экстремальное (максимальное или минимальное) значение. Например, в некотором биохимическом процессе нас может интересовать выход продукции в единицу времени, и если известно, что он зависит от значений температуры (t), pH (π) и концентрации определенного фермента (c), причем существует такая комбинация значений $t = t_m$, $\pi = \pi_m$ и $c = c_m$, при которых выход максимален, то требуется найти эти значения t_m, π_m, c_m. Можно поставить другую задачу: найти такую комбинацию значений t, π и c, которая при заданном уровне производительности обеспечивает минимальную себестоимость продукта.

В каждом из этих случаев речь идет о поисках оптимальной комбинации значений факторов, но критерии оптимальности в этих двух случаях различны. Величину, для которой ищут оптимальную комбинацию значений факторов, называют *параметром оптимизации*. В первой из упомянутых выше задач параметром оптимизации будет производительность процесса, а во второй задаче — себестоимость продукта. Вопрос о выборе параметра оптимизации не так прост, как может показаться на первый взгляд[1]; мы не будем его обсуждать, а будем считать этот параметр заданным.

Экстремальное значение параметра не всегда существует. Если, например, параметром оптимизации является чистота реактива (содержание примесей), а фактором, от которого зависит этот параметр,— затраты на изготовление реактива, то нельзя указать такое значение затрат, чтобы при дальнейшем увеличении затрат улучшение качества реактива сменилось его ухудшением[2]. Имея это в виду, мы будем в дальнейшем рассматривать только такие задачи, в которых экстремальное значение параметра оптимизации существует. Для определенности мы будем далее говорить только о максимуме, поскольку задачи о максимуме и минимуме совершенно аналогичны.

Изменение параметра оптимизации при изменении значений факторов (а иногда и сам параметр) принято называть *откликом*.

[1] По этому поводу см. в книге Ю. П. Адлера, гл. II.

[2] Но, конечно, параметр может принимать экстремальное значение при изменении других факторов, также влияющих на этот параметр.

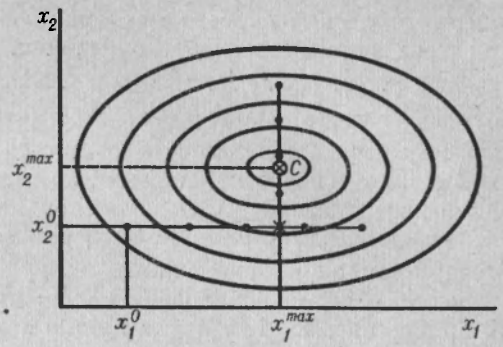


Рис. 1.1. Планирование поиска оптимальных значений факторов.

Если варьируют значения одного фактора, то отклик можно изобразить графически некоторой линией, которую назовем *линией отклика*. При варьировании двух факторов геометрическим образом отклика будет поверхность, называемая *поверхностью отклика*. Для однообразия этот термин употребляется и тогда, когда варьируют более двух факторов, хотя в этом случае отклик изображается в действительности некоторой многомерной гиперповерхностью.

Если параметр оптимизации имеет максимум, то поверхность отклика будет выпуклой, причем координаты вершины будут оптимальным набором значений факторов. Отыскание этих координат и составляет задачу *экстремального эксперимента*.

В дальнейших разделах этого параграфа будет описан один из методов решения этой задачи — метод крутого восхождения по поверхности отклика. Другой метод — метод симплекс-планирования рассматривается в § 1.6. В разделе 1.6.4 проводится сопоставление этих методов.

1.5.2. Для построения хотя бы в квадратичном приближении кривой линии отклика (для последующего нахождения положения максимума) требуется по крайней мере три экспериментальные точки. Если параметр оптимизации зависит от k факторов, то для построения квадратичной модели поверхности отклика можно использовать 3^k экспериментальных точек. Но в экстремальных задачах не нужна вся поверхность отклика, требуется лишь узнать координаты ее вершины. Тогда в простейшем случае достаточно иметь не более $3k$ экспериментальных точек. Действительно, рассмотрим для примера двухфакторную задачу. Выпуклую поверхность отклика изобразим так, как это обычно делается на топографических картах — линиями равной высоты (рис. 1.1). Поставим три опыта при трех значениях первого фактора, но при фиксированном значении второго фактора, например x_2^0. Построив по трем экспериментальным точкам кривую отклика, найдем ее вершину и тем самым координату x_1^{max} этой вершины. Теперь, фикси-

руя это значение первого фактора, поставим три новых опыта при трех значениях второго фактора. Построив новую кривую отклика $y(x_2;\ x_1 = x_1^{max})$, находим максимум этой кривой и координату x_2^{max} этого максимума. Таким образом, найдено положение максимума поверхности отклика: $x_1 = x_1^{max}$, $x_2 = x_2^{max}$.

Но такой способ нахождения оптимальных значений факторов редко дает нужный результат — только тогда, когда поверхность отклика выглядит примерно так, как на рис. 1.1. В большинстве случаев точка с координатами x_1^{max}, x_2^{max}, найденными описанным выше способом, не совпадает с положением вершины, как это можно видеть из рис. 1.2. Тогда нужно ставить новую серию опытов при $x_2 = x_2^{max}$, затем еще одну серию и т. д., приближаясь «зигзагами» к точке истинного максимума. Если факторов много, то число требуемых опытов возрастает в соответствующее число раз и становится чрезвычайно большим.

Число опытов можно резко сократить, если двигаться из начальной точки $O(x_1^0,\ x_2^0)$ не параллельно одной из координатных осей, а в направлении наибольшей крутизны поверхности отклика. Это направление (OP на рис. 1.2) задается перпендикуляром к прямой, касательной к линии уровня в точке O. Поставив несколько опытов вдоль одного направления, находим ту точку (P на рис. 1.2), где параметр оптимизации максимален, ставим в этой точке новый факторный эксперимент[1] для нахождения нового направления наибольшей крутизны (PP' на рис. 1.2), ставим несколько опытов вдоль этой прямой, находим точку максимального отклика на этой прямой (Q на рис. 1.2) и т. д. Такой *способ крутого восхождения* по поверхности отклика приводит в оптимальную область после 2—3 серий опытов даже при большом числе факторов.

Как же найти направление крутого восхождения? В случае плоскости, описываемой уравнением:

$$y = b_0 + b_1 x_1 + b_2 x_2,$$

направление наибольшей крутизны этой плоскости задается соотношением:

$$x_1 : x_2 = b_1 : b_2, \qquad (*)$$

т. е. это направление на координатной плоскости x_1, x_2 должно быть ближе к оси той переменной x_i, от которой больше всего

[1] По возможности интервалы варьирования факторов в этом новом факторном эксперименте следует выбирать меньше, чем в первоначальном эксперименте. Это объясняется тем, что чем ближе к точке оптимума, тем больше кривизна поверхности отклика. О связи кривизны поверхности отклика с выбором интервалов варьирования факторов см. последний пункт раздела 1.5.3.

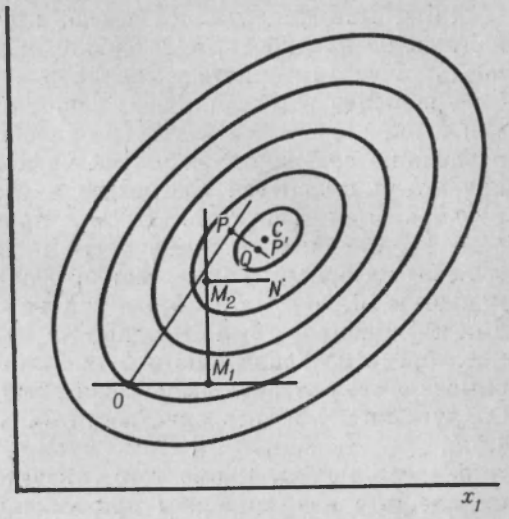


Рис. 1.2. Планирование поиска оптимума методом крутого восхождения.

зависит высота точек поверхности y (а эта зависимость определяется коэффициентом b_i). Этот результат обобщается на любое число факторов:

$$x_1 : x_2 : \ldots : x_k = b_1 : b_2 : \ldots : b_k. \qquad (**)$$

В случае криволинейной поверхности отклика нужно построить в точке O плоскость, касательную к данной поверхности, и затем двигаться в направлении наибольшей крутизны этой плоскости. На языке регрессионного анализа это означает, что строится линейное приближение уравнения регрессии (при помощи факторного эксперимента, описанного в § 1.4).

После того как найдено направление крутого восхождения, надо выбрать точки вдоль этого направления, в которых будут ставиться опыты данной серии, или, как говорят, выбрать величину шага перемещения (шага крутого восхождения). Рекомендуется выбирать величину шага таким образом, чтобы уже первый шаг привел к точке, лежащей на границе области факторного эксперимента хотя бы по одному фактору. Следовательно, величину шага желательно выбрать так, чтобы равнялось единице (по абсолютной величине) изменение переменной с наибольшим коэффициентом регрессии. Шаги по другим координатам обычно округляют, чтобы упростить условия опытов. Если характер эксперимента таков, что реализуется одновременно вся серия опытов, то целесообразно запланировать 5—6 опытов в точках факторного пространства, отстоящих одна от другой на один шаг. Если же опыты реализуются последовательно во времени, то часто оказывается целесообразным пропускать некоторые точки.

1.5.3. Как указывалось выше, крутое восхождение осуществляется по плоскости, касательной к поверхности отклика в нулевой точке факторного эксперимента. Поэтому необходимо, чтобы уравнение регрессии, полученное в результате этого эксперимента, описывало действительно плоскость. А это значит, что оно не должно содержать значимых членов взаимодействия. Последние могли появиться вследствие неудачного выбора интервалов варьирования факторов. Пусть, например, коэффициенты b_1, b_2 и b_{12} получились одинакового порядка величины. Очевидно, если бы интервалы варьирования обеих переменных x_1 и x_2 были меньше в 10 раз, то b_1 и b_2 получились бы также меньше в 10 раз, а коэффициент b_{12} был бы меньше в 100 раз, и им можно было бы пренебречь по сравнению с b_1 и b_2, так что линейную (в строгом смысле этого слова) модель можно было бы считать адекватной. О другой причине неадекватности линейной модели см. в разделе **1.5.4.**

Модель может, кроме того, оказаться несимметричной в том смысле, что коэффициенты при линейных членах имеют разный порядок. Такая модель также неудобна для крутого восхождения. Если, например, один из линейных коэффициентов во много раз больше всех остальных, то крутое восхождение фактически вырождается в движение вдоль одной оси. Причиной резкой несимметричности модели может опять-таки быть неудачный выбор интервалов варьирования факторов.

Пример 1.5. Пусть полный факторный эксперимент типа 2^3 в задаче, описанной в примере 1.3 (см. раздел **1.4.3**), дал уравнение регрессии:

$$y = 98{,}2 - 7{,}17x_1 + 0{,}43x_2 + 58{,}5x_3 + \\ + 0{,}12x_1x_2 - 8{,}36x_1x_3 + 0{,}29x_2x_3 + 0{,}74x_1x_2x_3.$$

Модель оказалась и нелинейной (велик коэффициент при x_1x_3), и несимметричной (все коэффициенты при линейных членах имеют разный порядок), поэтому принимается решение поставить повторный факторный эксперимент[1] с изменением интервалов варьирования факторов. В данном случае представляется целесообразным увеличить интервал для x_2 в 10 раз и во столько же раз уменьшить интервал для x_3. Пусть новый факторный эксперимент дает уравнение регрессии:

$$y = 109{,}4 - 6{,}81x_1 + 3{,}97x_2 + 8{,}12x_3 + \\ + 0{,}73x_1x_2 - 1{,}15x_1x_3 + 0{,}67x_2x_3 + 0{,}55x_1x_2x_3.$$

[1] При этом естественно выбрать нулевую точку нового эксперимента в точке, в которой предыдущий эксперимент дал лучшее значение; в данном примере пусть это будет точка $t° = 27°$, $pH = 7{,}2$, $c = 44$.

Эта модель уже линейна и симметрична и может служить для крутого восхождения, если считать, что этот факторный эксперимент является первым этапом более обширного эксперимента по нахождению оптимальной комбинации значений факторов x_1, x_2, x_3, т. е. такой комбинации значений температуры, pH и концентрации фермента, при которой выход продукта изучаемого биохимического процесса максимален.

Для построения схемы крутого восхождения по поверхности отклика используем только коэффициенты при линейных членах, т. е. b_1, b_2 и b_3. Таким образом, нужно поставить серию опытов (скажем, из 5 опытов) в точках, указанных в табл. 1.5. Значение Δ характеризует величину шага для перемещения в направлении крутого восхождения. Выберем это значение так, чтобы изменение переменной с наибольшим коэффициентом регрессии (в данном случае это x_3) в одном шаге равнялось единице, т. е. выберем $\Delta = 1/8{,}12 \approx 0{,}123$. Тогда изменения переменных x_1 и x_2 в одном шаге будут: $6{,}81 \cdot 0{,}123 \approx 0{,}84$ и $3{,}97 \cdot 0{,}123 \approx 0{,}49$.

ТАБЛИЦА 1.5

Номер точки	Координаты точек		
	x_1	x_2	x_3
1	0	0	0
2	—6,81Δ	3,97Δ	8,12Δ
3	—2·6,81Δ	2·3,97Δ	2·8,12Δ
4	—3·6,81Δ	3·3,97Δ	3·8,12Δ
5	—4·6,81Δ	4·3,97Δ	4·8,12Δ

Поэтому получается схема крутого восхождения, записанная в табл. 1.6.

ТАБЛИЦА 1.6

Номер точки	Координаты точек		
	x_1	x_2	x_3
1	0	0	0
2	—0,84	0,49	1
3	—1,68	0,98	2
4	—2,52	1,47	3
5	—3,36	1,96	4

Однако эта схема не совсем удобна для практического планирования опытов, так как факторы записаны в нормированных единицах. Поэтому перейдем к натуральным единицам, учитывая, что

в последнем эксперименте нулевая точка соответствует значениям $t° = 27°$, pH $= 7{,}2$, $c = 44$, а единицы масштаба равны соответственно 3°, 0,2 и 4 (и, конечно, при надлежащем округлении). Тогда получится схема, записанная в табл. 1.7.

ТАБЛИЦА 1.7

Номер точки	Значения факторов			Результат опыта
	$t°$	pH	c	
1	27	7,2	44	106
2	24	7,3	48	109
3	22	7,4	52	112
4	19	7,5	56	118
5	17	7,6	60	104

Поставив соответствующие опыты, получаем значения выхода, записанные в последнем столбце табл. 1.7. Наибольшее значение (118) получилось в точке $t° = 19°$, pH $= 7{,}5$, $c = 56$. Эту точку берем в качестве нулевой точки нового факторного эксперимента, который даст новое направление крутого восхождения.

1.5.4. Может оказаться, что неадекватность линейной модели сохраняется и после повторения факторного эксперимента с измененными интервалами варьирования факторов. Это означает, что центр (нулевая точка) эксперимента находится в области, где кривизна поверхности отклика уже очень велика, т. е. находится в области оптимума. Эту область обычно называют «почти стационарной областью».

Исследование почти стационарной области требует использования квадратичных моделей, сложного планирования опытов и большой вычислительной работы. В то же время повышение параметра оптимизации, которое может быть достигнуто, обычно невелико именно вследствие того, что область почти стационарна. Поэтому применение указанных сложных методов[1] оправдано только там, где важно даже небольшое увеличение параметра оптимизации, например в производствах с многотонной или особо ценной продукцией. В большинстве биологических приложений можно удовлетвориться достижением почти стационарной области. Впрочем, некоторое дополнительное увеличение параметра оптимизации может быть получено, если поставить добавочный факторный эксперимент с центром в точке, давшей в предыдущем эксперименте (по результатам которого и был сделан вывод о достижении почти стационарной области) наилучшее значение.

[1] Описание их см. в книгах Ю. П. Адлера, гл. VI, и В. В. Налимова и Н. А. Черновой, гл. 4. и 5.

§ 1.6. Симплекс-планирование

1.6.1. *Симплекс-планирование* есть один из методов планирования экстремального эксперимента.

> *О сущности экстремального эксперимента говорится подробно в разделе* **1.5.1.** *Необходимо прочесть этот раздел, прежде чем читать дальше. Другой метод планирования экстремального эксперимента — метод крутого восхождения — описан в § 1.5. Сопоставление этих двух методов будет приведено в разделе* **1.6.4.**

Симплексом называется простейшая, т. е. с минимальным числом элементов, геометрическая фигура, не содержащая криволинейных элементов. Очевидно, в пространстве двух измерений (т. е. на плоскости) симплексом является треугольник, а в пространстве трех измерений — тетраэдр, т. е. треугольная пирамида (имеющая четыре вершины, в то время как у куба их шесть).

Симплекс-планирование начинается с серии опытов, каждый из которых ставится при комбинации значений факторов, соответствующей одной из вершин правильного симплекса (правильного треугольника, правильного тетраэдра и др.); разумеется, значения факторов должны измеряться не в натуральных единицах, а в нормированных переменных, чтобы быть безразмерными. Очевидно, в k-мерном пространстве симплекс имеет $k + 1$ вершин[1].

После реализации серии из $k + 1$ опытов в вершинах симплекса отмечают ту вершину, в которой получен наихудший результат (наименьшее значение параметра оптимизации), и ставят новый опыт в точке, являющейся зеркальным отражением этой худшей точки относительно противолежащей грани[2]. Например, в симплексе, изображенном на рис. 1.3, худшей оказалась точка 2_0, и тогда новый опыт ставится в точке 2_1. Эта новая точка, совместно с точками 1_0, 3_0 и 4_0, составляющими «отражающую» грань первоначального симплекса, образует новый правильный симплекс $1_0 2_1 3_0 4_0$. Теперь выбирают наихудшую точку этого нового симплекса (допустим, на этот раз ею оказалась точка 4_0) и строят ее зеркальное отражение относительно грани $1_0 2_1 3_0$ — точку 4_1, в которой ставят следующий опыт. Пример такого планирования приведен в табл. 1.8; в последней строке записано планирование очередного опыта (в точке 3_1), но опыт еще не реализован и результат не получен.

---

[1] О нахождении координат этих вершин будет сказано в разделе **1.6.2.**

[2] О том, как найти это зеркальное отражение, см. раздел **1.6.3.**

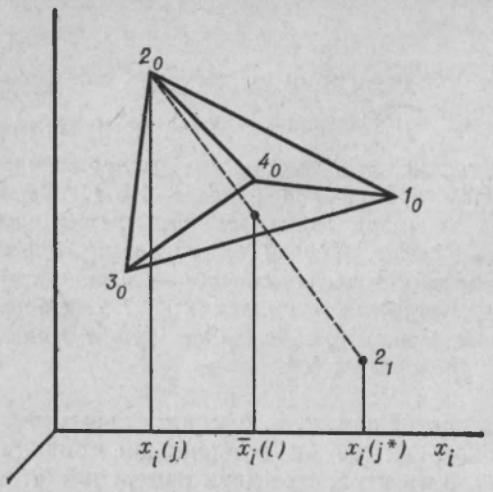


Рис. 1.3. Симплекс-планирование поиска оптимума.

Поскольку на каждом шаге симплекс-планирования исключается худшая точка, весь симплекс постепенно перемещается в область оптимума. Заметим, что при каждом шаге симплекс не перекатывается (т. е. не поворачивается вокруг одного из ребер), а как бы выворачивается наизнанку — все вершины, составляющие основание симплекса, противолежащее худшей вершине, остаются на месте.

ТАБЛИЦА 1.8

Вершина	Результат опыта	Симплекс
1_0	43,8	Начальный
2_0	23,9	симплекс
3_0	40,4	$1_0 2_0 3_0 4_0$
4_0	34,7	
2_1	38,2	$1_0 2_1 3_0 4_0$
4_1	56,3	$1_0 2_1 3_0 4_1$
2_2	61,5	$1_0 2_2 3_0 4_1$
3_1		$1_0 2_2 3_1 4_1$

Может получиться, что новая точка опять оказывается худшей в новом симплексе. Тогда нужно вернуться к предыдущему симплексу и отразить вторую (с конца) по качеству вершину.

Чем ближе к оптимуму, тем чаще приходится совершать такие «рыскания». Наконец, в самой окрестности оптимума все новые опыты в точках, симметричных ко всем вершинам последнего симплекса, дают результаты, худшие чем в наилучшей вершине этого симплекса. Поэтому такую ситуацию можно считать приз-

наком достижения оптимума, а наибольшее значение параметра оптимизации в последнем симплексе следует принять за его максимум. Однако в области оптимума рекомендуется повторить опыты по нескольку раз, чтобы получить возможность исключить случайные флуктуации значений параметра оптимизации.

1.6.2. Если поместить начало координат в центре симплекса, а длину ребра выбрать равной единице, то координаты вершин симплекса будут определяться матрицей, приведенной в табл. 1.9.

ТАБЛИЦА 1.9

j \ i	1	2	3	4	...	$k-1$	k
1	r_1	r_2	r_3	r_4	...	r_{k-1}	r_k
2	$-R_1$	r_2	r_3	r_4	...	r_{k-1}	r_k
3	0	$-R_2$	r_3	r_4	...	r_{k-1}	r_k
4	0	0	$-R_3$	r_4	...	r_{k-1}	r_k
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
k	0	0	0	0	...	$-R_{k-1}$	r_k
$k+1$	0	0	0	0	...	0	$-R_k$

Здесь j есть номер вершины симплекса, принимающий значения от 1 до $k+1$, а i есть номер координаты этой вершины, принимающий значения от 1 до k. Величины r_i и R_i равны:

$$r_i = \sqrt{\frac{1}{2i(i+1)}}, \qquad R_i = \sqrt{\frac{i}{2(i+1)}}; \tag{1.12}$$

они являются соответственно радиусами вписанной и описанной i-мерных сфер (на рис. 1.4 для примера приведен двумерный правильный симплекс, т. е. равносторонний треугольник).

При планировании эксперимента удобнее выражать координаты вершин симплекса в натуральных единицах. Пусть для фактора φ_i выбран начальный интервал варьирования, верхний и нижний концы которого обозначим соответственно φ_i^+ и φ_i^-. Переход к безразмерным единицам (нормирование) состоял в том, что полагалось:

$$x_i = \frac{\varphi_i - \varphi_i^0}{\lambda_{\varphi_i}},$$

т. е. начало отсчета переносилось в некоторую точку φ_i^0 и применялась единица измерения λ_{φ_i}. В факторном эксперименте (см. раздел 1.4.2) экспериментальные точки располагаются в вершинах гиперкуба со стороной 2 (от -1 до $+1$), а симплекс, опре-

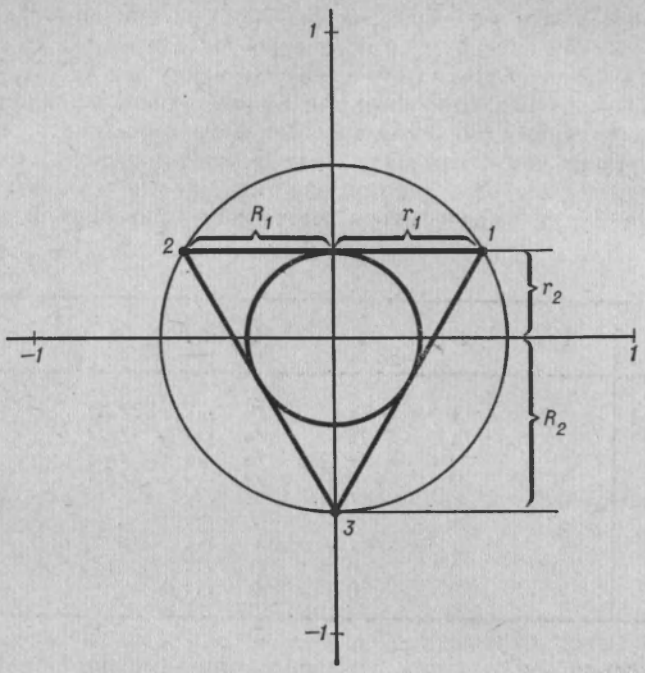


Рис. 1.4. Двумерный правильный симплекс.

деляемый табл. 1.9 и формулой (1.12), имеет сторону 1. Чтобы получить в симплекс-планировании охватываемую точками плана экспериментальную область таких же габаритов (но не объема!), что и в факторном эксперименте, положим:

$$\lambda_{\varphi_i} = \varphi_i^+ - \varphi_i^-,$$

так что:

$$x_i = \frac{\varphi_i - \varphi_i^0}{\varphi_i^+ - \varphi_i^-}.$$

Верхняя и нижняя границы для x_i равны r_i и $-R_i$, поэтому:

$$\frac{\varphi_i^+ - \varphi_i^0}{\varphi_i^+ - \varphi_i^-} = r_i, \quad \frac{\varphi_i^- - \varphi_i^0}{\varphi_i^+ - \varphi_i^-} = -R_i$$

и

$$\frac{\varphi_i^- - \varphi_i^0}{\varphi_i^+ - \varphi_i^0} = -\frac{R_i}{r_i} = -\frac{\sqrt{\frac{i}{2(i+1)}}}{\sqrt{\frac{1}{2i(i+1)}}} = -i;$$

отсюда:

$$\varphi_i^0 = \frac{i\varphi_i^+ + \varphi_i^-}{i+1}. \qquad (1.13)$$

Следовательно, вершины начального симплекса имеют натуральные координаты, определяемые матрицей из табл. 1.10, причем значения φ_i^0 вычисляют по формуле (1.13).

ТАБЛИЦА 1.10

j \ i	1	2	3	...	$k-1$	k
1	φ_1^+	φ_2^+	φ_3^+	...	φ_{k-1}^+	φ_k^+
2	φ_1^-	φ_2^+	φ_3^+	...	φ_{k-1}^+	φ_k^+
3	φ_1^0	φ_2^-	φ_3^+	...	φ_{k-1}^+	φ_k^+
4	φ_1^0	φ_2^0	φ_3^-	...	φ_{k-1}^+	φ_k^+
.	.	.	.	...	.	.
.	.	.	.	...	.	.
.	.	.	.	...	.	.
k	φ_1^0	φ_2^0	φ_3^0	...	φ_{k-1}^-	φ_k^+
$k+1$	φ_1^0	φ_2^0	φ_3^0	...	φ_{k-1}^0	φ_k^-

Пример 1.6. Запишем координаты вершин симплекса для данных из примера 1.3 (раздел 1.4.3), учитывая, что границы интервалов варьирования факторов таковы:

	Нижняя граница	*Верхняя граница*
$t°$	27,0°	33,0°
pH	6,8	7,2
c	44,0	52,0

Пользуясь формулой (1.13), вычисляем значения φ_i^0 для $t°$ и pH:

$$\frac{1\cdot 33{,}0 + 27{,}0}{1+1} = 30{,}0,$$

$$\frac{2\cdot 7{,}2 + 6{,}8}{2+1} = 7{,}06\ldots \approx 7{,}1.$$

Поэтому начальный симплекс будет:

Вершины	*Координаты вершин*		
	$t°$	pH	c
1_0	33,0	7,2	52,0
2_0	27,0	7,2	52,0
3_0	30,0	6,8	52,0
4_0	30,0	7,1	44,0

1.6.3. Перед каждым новым опытом надо найти точку j^*, симметричную к «худшей» вершине j симплекса относительно противолежащей грани, т. е. фактически относительно центра этой грани. Пусть $l \neq j$ (l пробегает k значений) обозначает вершины этой грани. Тогда $x_i(l)$ есть i-я координата центра грани, а i-я координата точки j^*, как видно на рис. 1.3, равна:

$$x_i(j^*) = x_i(l) + [\bar{x}_i(l) - x_i(j)] = 2\bar{x}_i(l) - x_i(j).$$

Это справедливо при любом расположении точек j и j^*; если бы потребовалось отразить точку j^*, то для i-й координаты точки j получилось бы:

$$x_i(j) = \bar{x}_i(l) - [x_i(j^*) - \bar{x}_i(l)] = 2\bar{x}_i(l) - x_i(j^*),$$

т. е. та же формула. Учитывая, что:

$$\bar{x}_i(l) = \frac{1}{k} \sum_{l \neq j} x_i(l),$$

получаем окончательно:

$$x_i(j^*) = \frac{2}{k} \sum_{l \neq j} x_i(l) - x_i(j). \qquad (1.14)$$

Пример 1.7. Пусть в симплексе из примера 1.6 (раздел 1.6.2) «худшей» оказалась вершина 3_0. Тогда нужно поставить опыт в точке 3_1, симметричной к точке 3_0 относительно грани $1_0 2_0 4_0$. Пользуясь формулой (1.14), найдем координаты точки 3_1:

$$t^\circ(3_1) = \frac{2}{3}(33{,}0 + 27{,}0 + 30{,}0) - 30{,}0 = 30{,}0,$$

$$\text{pH}(3_1) = \frac{2}{3}(7{,}2 + 7{,}2 + 7{,}1) - 6{,}8 = 7{,}5,$$

$$c(3_1) = \frac{2}{3}(52{,}0 + 52{,}0 + 44{,}0) - 52{,}0 = 46{,}7.$$

1.6.4. Симплекс-планирование имеет ряд преимуществ перед методом крутого восхождения:

1. Численные значения параметра оптимизации не входят в какие-либо расчеты, поэтому симплекс-планирование может применяться для поисков условий получения оптимальных качественных характеристик продукта.

2. Поиск области оптимума при симплекс-планировании не опирается на какую-либо модель поверхности отклика в окрестности эксперимента, что часто приводит к значительной экономии опытов. Дело в том, что число опытов в методе крутого восхождения

иногда значительно увеличивается, если полученная по результатам полного факторного эксперимента линейная модель оказывается неадекватной.

3. Правила симплекс-планирования (построение нового симплекса путем отражения наихудшей вершины; возвращение к предыдущему симплексу и отражение другой вершины, если новая точка оказалась худшей в новом симплексе; прекращение поиска оптимума, если подобная процедура не дает улучшения ни для одной вершины) легко поддаются полной формализации, что позволяет поручить поиск оптимума вычислительной машине, в которую, естественно, надо после каждого опыта вводить полученное значение параметра оптимизации. Если такой ввод может быть осуществлен автоматически с датчика, то весь поиск оптимума можно полностью автоматизировать.

Вместе с тем метод симплекс-планирования имеет определенные недостатки, ограничивающие возможности его применения:

1. Если не считать серии опытов начального симплекса, которые могут быть выполнены одновременно, все дальнейшие опыты можно производить только последовательно во времени, т. е. новый опыт можно ставить только после того, как получены результаты предыдущего опыта. Ясно, что, когда каждый опыт требует много времени (например, при изучении отдаленных последствий действия препарата или при изучении каких-либо показателей сельскохозяйственной культуры), применять симплекс-планирование нецелесообразно.

2. Обязательным этапом в методе крутого восхождения является факторный эксперимент, позволяющий построить уравнение регрессии для параметра оптимизации. Это уравнение, выполняя в процессе поиска оптимальной области свое прямое назначение — определение направления крутого восхождения, дает одновременно возможность судить о влиянии на параметр оптимизации отдельных факторов и их взаимодействий. В симплекс-планировании для получения по точкам плана коэффициентов линейной модели пришлось бы делать специальные дополнительные расчеты.

3. Хотя при симплекс-планировании непосредственно не используется информация, даже приближенная, о форме поверхности отклика, это не значит, что метод совсем безразличен к этой форме. Например, движение симплекса вверх по поверхности округлого холма будет менее зигзагообразным, чем движение по гребню сплющенного с боков холма, и, следовательно, потребует меньшего числа шагов (т. е. опытов). Вообще поиск оптимума идет тем эффективнее, чем симметричнее поверхность отклика. Но форма этой поверхности (напомним, что речь идет все время о форме поверхности в нормированном факторном пространстве) зависит от выбора интервалов варьирования каждого из факторов:

при изменении интервала варьирования фактора поверхность отклика в нормированном пространстве растягивается или сжимается в соответствующем направлении. Это значит, что неудачная (в указанном выше смысле) форма поверхности отклика отражает неудачный выбор интервалов варьирования факторов (вернее, соотношения этих интервалов). В методе крутого восхождения, где эта проблема также существует, есть возможность проконтролировать и исправить положение: при неудачном выборе интервалов варьирования факторов уравнение регрессии получается несимметричным в коэффициентах при линейных членах, и сразу видно, как надо изменить интервалы варьирования (см. подробнее в разделе 1.5.3).

С методами решения других задач планирования экспериментов можно ознакомиться по книгам, указанным в списке литературы. Кроме того, см. статьи А. Н. Лисенкова «Применение математических методов планирования эксперимента в медико-биологических исследованиях» (Вестн. АМН СССР, 1973, № 5, с. 78—84) и «Основные принципы и методы планирования многофакторных медико-биологических экспериментов» (сб. «Применение математических методов в медико-биологических исследованиях». Тр. Ин-та полиомиелита и вирусных энцефалитов АМН СССР, Т. 20, 1972, с. 10—36).

Глава вторая

ПРЕДВАРИТЕЛЬНАЯ СТАТИСТИЧЕСКАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

§ 2.1. Основные задачи статистической обработки результатов биологического исследования

2.1.1. Современные биологические и медицинские исследования большей частью таковы, что их результаты выражаются количественно. Но развитие живого организма определяется очень многими и весьма разнообразными условиями внутреннего и внешнего порядка, поэтому как анатомические и физиологические характеристики, так и ответы на внешние воздействия у особей одной и той же популяции всегда в той или иной мере варьируют. Это и вызывает необходимость в применении статистических методов для описания и анализа результатов биологических и медицинских исследований. Прежде всего требуется многократное повторение эксперимента или наблюдения, чтобы быть уверенным, что результат отражает достаточно хорошо свойства всей популяции, а не случайные свойства одной особи.

Понятия «многократное повторение» и «отражает достаточно хорошо» уточняются и конкретизируются в § 1.3 и в § 3.5 и 3.7.

В результате однородной серии экспериментов или наблюдений получается так называемая статистическая совокупность (см. раздел 1.1.1), которая и подлежит «статистической обработке» (или статистическому анализу).

Задачи статистического анализа, вообще говоря, весьма многообразны. В этой книге будут рассмотрены лишь некоторые из них, в частности:

1. По данным выборки дать описание генеральной совокупности, построив доверительные интервалы для параметров распределения. Эти вопросы изложены в главах третьей, шестой (§ 6.2) и седьмой (§ 7.2).

2. Определить значимость различия между двумя совокупностями (чаще всего речь идет о сравнении опыта с контролем). Этому посвящены главы четвертая и пятая (см. также § 6.3 и 6.4 и § 7.3).

3. Изучить статистическую связь между двумя совокупностями (регрессионный и корреляционный анализ см. в главе восьмой).

Некоторые понятия и термины, употребляемые в этом параграфе, могут оказаться для Вас незнакомыми при первом чтении. Не следует этим смущаться — они разъясняются в дальнейшем, в соответствующих местах.

Решению любой из этих задач должна предшествовать некоторая общая процедура, включающая три этапа: а) упорядочение эмпирической совокупности, б) выбор математической модели распределения и в) отбрасывание «выскакивающих» вариант. Изложение этой процедуры составляет содержание настоящей главы.

Упорядочение статистических совокупностей, полученных в результате исследования, включает классификацию или группировку вариант, построение статистических распределений и их графическое изображение. Эти приемы изложены в § 2.2.

После этого, исходя из предварительной информации о свойствах изучаемого объекта или характера явления, выбирают определенную теоретическую модель статистической совокупности. Точнее, речь идет о выборе типа распределения вариант в генеральной совокупности. Так, при решении вопросов, связанных с измерениями размеров, весов и тому подобных свойств живых объектов, распределения вариант можно предполагать нормальными; при изучении редких явлений (например, появление бактерий в ячейках цитометра) чаще всего употребляется в качестве модели распределения Пуассона; если интересуются числом доминантных и рецессивных форм при скрещивании, то подходящей моделью будет альтернативное распределение и т. д.

Выбор модели нужен потому, что он определяет всю дальнейшую статистическую обработку. При альтернативном и пуассоновском распределениях эта обработка производится способами, изложенными соответственно в главах шестой и седьмой. Большинство методов, изложенных в остальных главах, применимо лишь при нормальном (или не очень отклоняющемся от нормального) распределении.

Как уже было сказано, выбор модели определяется прежде всего существом дела (впрочем, известное значение имеет и требование математической простоты модели). Этому выбору помогает также графический анализ: построение гистограммы или полигона частот, использование вероятностной бумаги, составление корреляционного поля, различные преобразования координат и др. Но правильность выбора модели можно проверить и при помощи статистических критериев.

Проверка нормальности распределения описана в § 2.5, а проверка распределения Пуассона — в разделе **7.1.4.** Однако такая проверка не является во всех случаях обязательной — ее следует производить только тогда, когда возникает какое-ли-

бо сомнение. Например, рассматривают распределение длины волокна хлопка и, как обычно при распределении линейных размеров, считают естественным принять модель нормального распределения; между тем гистограмма частот указывает на заметную асимметрию распределения.

Если выбрана модель нормального распределения (и в случае необходимости показано статистически, что такой выбор не противоречит эмпирическим данным), то можно проверить принадлежность к изучаемой генеральной совокупности «выскакивающих» вариант (если таковые подозреваются в выборке); соответствующая методика изложена в § 2.6.

После этих предварительных этапов (т. е. классификация или группировка вариант, выбор математической модели распределения и отбрасывание «выскакивающих» вариант) можно приступить к решению статистических задач, сформулированных выше.

§ 2.2. Классификация и группировка вариант. Графическое представление распределения

2.2.1. Подлежащий статистической обработке «сырой» материал представляет собой ряд значений, которые, вообще говоря, не совпадают друг с другом. Такой ряд значений называют *статистической совокупностью*, а каждый член этой совокупности — *вариантой*. Число вариант в совокупности называют *объемом* совокупности; это число будет обозначено буквой n.

Вначале совокупность представляет собой ряд значений, записанных в той последовательности, в какой эти значения были получены. Первой задачей статистической обработки этого материала является наведение определенного порядка в полученном ряде. Этой цели может служить расположение вариант в какой-либо заранее выбранной последовательности. Характер этой последовательности существенно зависит от характера изучаемого признака. В этом отношении принято различать три вида признаков — *количественные*, *порядковые* и *качественные*, которым соответствуют три принципа расположения вариант.

К первому виду относят признаки, которые могут быть охарактеризованы количественно — вес животного, процент гемоглобина крови, число инъекций и др. В этом случае первоначальное упорядочение совокупности состоит в том, что варианты располагаются в порядке возрастания или убывания их численных значений.

К порядковым признакам принято относить те признаки, для которых точная количественная характеристика либо невозможна, либо нецелесообразна, но в то же время имеется возможность расположить варианты в определенном порядке. Например, до-

вольно трудно охарактеризовать строго количественно ответы учащихся на экзамене, но их можно оценить условными баллами, после чего возможна расстановка ответов в порядке убывания (или возрастания) этих баллов. Другим примером являются таблицы спортивных соревнований, где имена участников или названия команд располагаются в определенном порядке, например в соответствии с набранным количеством очков. Последовательные места, занимаемые вариантами при таком упорядочении, носят названия *рангов*, а сам процесс приписывания каждой варианте определенного ранга называется *ранжированием*.

Наконец, признаки третьего типа — качественные. Это такие признаки, при которых нет не только количественной оценки, но и ранжирования. Примерами могут служить разный цвет волос, разные виды болезней, разные сельскохозяйственные культуры и др. Особенно важным и часто встречающимся в биологических исследованиях является частный случай, когда имеется только два возможных класса группировки (две альтернативы). Например, если изучить летальное действие различных доз облучения, то в каждом опыте совокупность разбивается только на две части — животные погибшие и животные выжившие. Такое распределение принято называть *альтернативным*. Примерами альтернативного распределения могут служить: наличие или отсутствие генетических мутаций, появление пли непоявление какого-либо рефлекса, рождение особи того или иного пола, расщепление гибридов (по одному признаку) на две формы и др.

В данной главе рассматриваются только количественные признаки. Порядковым и качественным (точнее, альтернативным) признакам посвящены соответственно § 8.6 и глава шестая.

2.2.2. Первым шагом при упорядочении совокупности с количественным признаком может быть расположение вариант в порядке возрастания или убывания их численных значений. Однако такой способ упорядочения пригоден только при очень малом числе вариант (один—два десятка). Если же число вариант велико, то производится та или иная *группировка* вариант.

Пусть количественный признак является *дискретным*, т. е. может принимать лишь некоторые значения, отстоящие друг от друга на конечные интервалы (например, число больных может быть только целочисленным). Тогда подсчитывают, сколько раз встречается каждое из значений, и в результате получается два ряда чисел: первый ряд содержит все несовпадающие значения вариант, расположенные в каком-либо порядке (возрастания или убывания), а числа второго ряда указывают число вариант, имею-

щих соответствующее значение. Пример такой группировки приведен в табл. 2.1.

ТАБЛИЦА 2.1

Число деревьев на делянке	11	12	13	14	15	16	Всего
Число делянок с данным числом деревьев	3	6	16	19	14	2	60

Числа первого (верхнего) ряда суть *значения* вариант: они будут обозначаться x_i, причем индекс i указывает порядковый номер значения (в данном случае i пробегает значения 1, 2, 3, 4, 5, 6, так как здесь имеется шесть различных значений вариант). Числа второго (нижнего) ряда называются *численностями*, или *частотами* (так как они указывают, насколько *часто* встречаются соответствующие значения) и обозначаются n_i. Указанный ряд пар чисел составляет *статистическое распределение* — распределение частот n_i по значениям x_i.

Очевидно, сумма частот равна объему совокупности. Мы запишем это так:

$$\sum_{i=1}^{k} n_i = n \tag{2.1}$$

(читается: «сумма n_i от 1 до k равна n»). Например, выражение $\sum_{i=1}^{6} n_i$ есть сокращенная запись суммы $n_1 + n_2 + n_3 + n_4 + n_5 + n_6$. Часто пишут упрощенно Σn_i, не указывая пределов суммирования, если нет опасения, что это может привести к недоразумениям.

Наряду с частотами иногда удобно пользоваться *относительными частотами*, или так называемыми *частостями* ν_i. Каждая частость указывает долю общего объема совокупности, приходящуюся на данное значение признака, так что:

$$\nu_i = \frac{n_i}{n}. \tag{2.2}$$

Очевидно:

$$\sum_i \nu_i = \sum \frac{n_i}{n} = \frac{\Sigma n_i}{n} = \frac{n}{n} = 1. \tag{2.3}$$

Если количественный признак *непрерывен*, т. е. может принимать любые значения в некотором интервале, то группировка заключается в том, что диапазон вариаций делят на определенное

число частей (*разрядов*), а затем подсчитывают число вариант, попадающих в каждый из разрядов. При выборе числа разрядов (это число будем в дальнейшем обозначать буквой k) обычно руководствуются тем, чтобы характерные особенности распределения не были завуалированы, а нехарактерные, случайные колебания были бы сглажены. Поэтому при большом объеме совокупности ($n > 100$) число разрядов выбирают больше (например, 9—12), а при малом объеме меньше (6—9). Пример группировки непрерывных вариант показан в табл. 2.2 (распределение зерен пшеницы по длине).

ТАБЛИЦА 2.2

Границы разрядов, мм	Частоты
5,175—5,225	1
5,225—5,275	4
5,275—5,325	7
5,325—5,375	11
5,375—5,425	16
5,425—5,475	30
4,475—5,525	14
5,525—5,575	8
5,575—5,625	6
5,625—5,675	2
5,675—5,725	1
Всего . . .	100

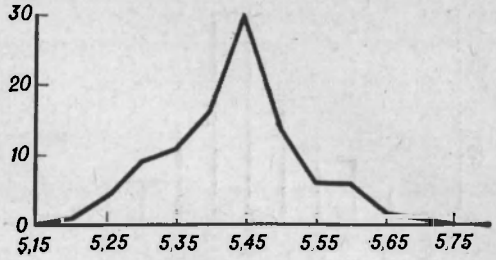
В данной главе мы будем заниматься только непрерывными совокупностями. Однако следует заметить, что если совокупность дискретна, но число несовпадающих значений велико, то отличие этой дискретной совокупности от непрерывной становится несущественным и ее можно описывать как непрерывную.

Существенно дискретные распределения, имеющие значения для биологической практики, рассматриваются отдельно в главах шестой (альтернативное распределение) и седьмой (распределение Пуассона).

2.2.3. После того как произведена группировка совокупности по разрядам, характер распределения более или менее проясняется. Однако он выступает еще более выпукло и особенно наглядно при графическом изображении этого распределения.

Среди многих способов графического изображения распределений чаще всего применяются два способа: построение *полигона*

Рис. 2.1. Полигон частот.



(т. е. многоугольника) *частот* и построение *гистограммы* (сетчатой диаграммы).

В первом случае все значения, лежащие в данном разряде, «стягиваются» к середине этого разряда. Например, в разряд 5,375—5,427 (см. табл. 2.2) попадает 16 зерен, которые, вообще говоря, имеют разную длину (5,38; 5,39, 5,40; 5,41; 5,42); между тем мы условно считаем, что все 16 зерен имеют длину 5,40 мм, соответствующую середине разряда. То же относится и к остальным разрядам. После этого строится график так, как это показано на рис. 2.1. Точки, отвечающие каждому из разрядов, отстоят от горизонтальной оси (которую называют осью абсцисс) на расстояния, пропорциональные соответствующим частотам. Разумеется, масштабы могут быть по обеим осям произвольные, но удобно и наиболее привычно выбирать их так, чтобы соотношение ширины и высоты графика было близко к 2 : 1.

На гистограмме каждый разряд изображается прямоугольником с шириной, пропорциональной ширине разряда, и с высотой, пропорциональной частоте данного разряда. Для распределения, приведенного в табл. 2.2, например получается картина, представленная на рис. 2.2.

Изображение распределения при помощи гистограммы представляет собой другой крайний случай идеализации: если в случае полигона частот все значения, лежащие внутри разряда, «стягиваются» к середине разряда, то в случае гистограммы они считаются распределенными равномерно по всему разряду. Поэтому в принципиальном отношении оба способа изображения следует считать равноценными, и выбор между ними определяется чаще всего привычкой или вкусом исследователя. Впрочем, иногда отмечают как преимущество, что площадь, ограниченная гистограммой, пропорциональна объему совокупности, в то время как площадь, ограниченная полигоном частот и осью абсцисс, не имеет такой простой интерпретации. В то же время, если совокупность существенно дискретна, естественно изображать ее полигоном частот.

2.2.5. Рассмотрим теперь непрерывную совокупность неограниченно большого объема. В этом случае можно в принципе взять

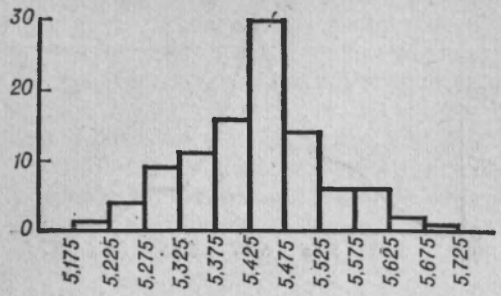


Рис. 2.2. Гистограмма.

число разрядов группировки настолько большим и соответственно ширину каждого разряда настолько малой, что графическое изображение распределения будет очень мало отличаться от непрерывной и гладкой кривой. Эту кривую можно описать аналитически, т. е. в виде формулы, некоторой функцией $y = f(x)$, указывающей, чему равна ордината y, соответствующая заданному значению абсциссы x. Вид функции $f(x)$ зависит, конечно, от формы кривой.

В интервал абсцисс от x до $x + \Delta x$ попадет примерно $\Delta n = f(x)\Delta x$ вариант (рис. 2.3), а в единицу длины — примерно:

$$\frac{\Delta n}{\Delta x} = \frac{f(x)\,\Delta x}{\Delta x} = f(x)$$

вариант. Поэтому функцию $f(x)$ можно назвать *плотностью распределения вариант*. При группировке вариант в разряды мы заменяем истинную плотность (значения которой различны для разных x) некоторой средней, в пределах разряда, плотностью (одинаковой для всех x внутри данного разряда). Делается это с целью упрощения дальнейших вычислений.

Иногда приходится (или удобнее) выбирать интервалы неодинаковыми по ширине. Чтобы это не привело к искажению графика, нужно откладывать по ординатам не частоты разрядов, а плотности частот (т. е. частное от деления частоты на ширину ин-

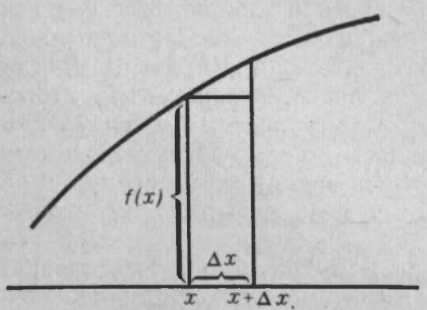


Рис. 2.3. Плотность распределения.

тервала). Очевидно, при одинаковой ширине всех интервалов плотности относятся между собой так же, как частоты, поэтому переходить от частот к плотностям нет надобности.

§ 2.3. Нормальное распределение

2.3.1. Упорядочение эмпирической статистической совокупности, состоящее в классификации или группировке вариант и в некоторых случаях также в графическом изображении получившегося распределения, есть лишь первый шаг статистической обработки материала. Для дальнейшего анализа требуется выбрать определенную модель распределения вариант в генеральной совокупности. Если существом дела подсказывается, что распределение является непрерывным, то чаще всего в качестве модели берут так называемое *нормальное распределение*. Выбор именно этого распределения в качестве модели диктуется многими соображениями, которые станут ясными из дальнейшего. Поскольку ссылки на свойства нормального распределения будут в последующем встречаться неоднократно, мы считаем целесообразным дать в этом месте краткое изложение этих свойств.

Впрочем, если Вы считаете, что достаточно знакомы с этим вопросом, то можете пропустить § 2.3. Если же, наоборот, Вы хотите познакомиться не только со свойствами нормального распределения, но и с его происхождением, обратитесь к § 6.1, в частности к разделу 6.1.4.

Нормальное распределение (или распределение Гаусса) имеет вид, изображенный на рис. 2.4. Ординаты этой кривой, т. е. плотности вероятности, описываются уравнением:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \qquad (2.4)$$

Здесь μ и σ — математическое ожидание и стандартное отклонение, характеризующие положение центра распределения и степень рассеяния вариант относительно этого центра (эти понятия разъясняются подробнее в § 3.1 и 3.3). Величина $e \approx 2{,}718\ldots$ есть основание натуральных логарифмов; выбор именно этой величины в качестве основания показательной функции диктуется соображениями чисто математического порядка.

Из рис. 2.4 видно, что нормальное распределение является симметричным, т. е. положительные и отрицательные отклонения равной величины встречаются одинаково часто. Далее, мы видим,

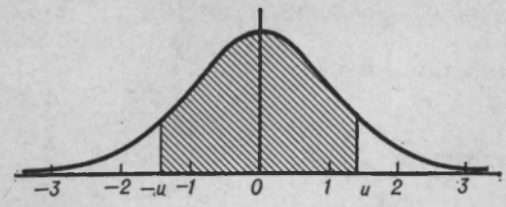


Рис. 2.4. Нормальная плотность распределения.

что кривая убывает по мере удаления от середины распределения; это означает, что большие отклонения бывают реже, чем малые. Но кривая не пересекает ось абсцисс, а приближается к ней асимптотически при неограниченном увеличении отклонений; значит, в принципе могут иметь место сколь угодно большие отклонения (хотя вероятность очень больших отклонений чрезвычайно мала).

Для функции $\varphi(x)$ имеются подробные таблицы, в которых в качестве аргумента принято безразмерное *относительное отклонение*:

$$u = \frac{x - \mu}{\sigma},$$

а сами значения функции выражены в единицах $1/\sigma$; иными словами, речь идет о таблице для функции

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}.$$

Эта таблица может быть использована для построения гауссовой кривой (т. е. графика функции плотности нормального распределения). Однако чаще всего можно обойтись следующими значениями:

u	0	±0,5	±1,0	±1,5	±2,0	±2,5	±3,0
$\varphi(u)$	0,399	0,352	0,242	0,130	0,054	0,018	0,004

Очевидно, на форму гауссовой кривой оказывает существенное влияние значение стандартного отклонения σ. Заметим прежде всего, что в каждой из половин кривой (левой и правой) можно различить две части: в первой, ближе к середине, кривая выгнута кверху, а во второй, дальше от середины, она выгнута книзу; это значит, что в первой части темп убывания ординат все убыстряется, а во второй части он замедляется. Между этими частями находится так называемая точка перегиба; вблизи этой точки кривая имеет наиболее крутой наклон. Можно показать, что абсцисса этой точки перегиба равна σ. Следовательно, чем больше σ, тем «шире» кривая. Расширение кривой при увеличнии σ

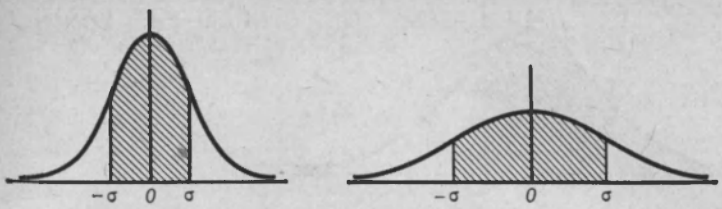


Рис. 2.5. Форма нормальной плотности распределения при разных стандартных отклонениях.

сопровождается понижением ее вершины. В самом деле, вершина соответствует абсциссе $x = \mu$. Но из (2.4) следует, что при $x = \mu$:

$$\varphi(x = \mu) = \frac{1}{\sqrt{2\pi}\,\sigma} \approx \frac{0{,}4}{\sigma},$$

так как $e^0 = 1$. Значит, чем больше σ, тем меньше $\varphi(x = \mu)$.

Таким образом, при малых σ кривая плотности нормального распределения «стягивается» к середине, а при больших σ она «расплывается» в стороны (рис. 2.5).

При этом мы подразумеваем, что в обоих случаях площади, ограничиваемые кривой и осью абсцисс, одинаковы[1].

2.3.2. Можно поставить вопрос о том, какая часть отклонений заключена в пределах от $-\sigma$ до $+\sigma$, т. е. какая часть всех вариант отклоняется от среднего значения не более чем на σ. Геометрически эта часть выражается заштрихованной площадью на рис. 2.5. Расчет показывает, что ее величина составляет $\sim$0,683 всей площади. Следовательно, в среднем 68,3% (или примерно $^2/_3$) всех вариант отклоняются от среднего значения не больше, чем на величину среднего квадратического отклонения. Аналогично можно получить, что в пределах от -2σ до $+2\sigma$ лежит 95,5% всех вариант, в пределах от -3σ до $+3\sigma$ лежит 99,7% и т. д. Имеются подробные таблицы, в которых указывается доля вариант, лежащих в пределах от $-u\sigma$ до $+u\sigma$ (заштрихованная часть на рис. 2.4), с шагом $\Delta u = 0{,}1$ или даже 0,01; аргумент u выбран для удобства безразмерным. Эта доля вариант обозначается $\Theta(u)$. Табл. II Приложений содержит значения $\Theta(u)$ с шагом $\Delta u = 0{,}01$. В левой части столбца указаны целые и десятые, а в верхней строке — сотые доли аргумента u; например, $\Theta(2{,}13) = 0{,}9668$. Для экономии места в таблице даются только десятичные знаки вероятностей, а нуль (целых) и запятая опу-

[1] Вычисления показывают, что площадь под кривой $e^{-u^2/2}$ равна $\sqrt{2\pi}$. Так как для удобства расчетов желательно, чтобы площадь под гауссовой кривой была равна единице, то в уравнение этой кривой (2.4.) введен «нормирующий множитель» $1/\sqrt{2\pi}$.

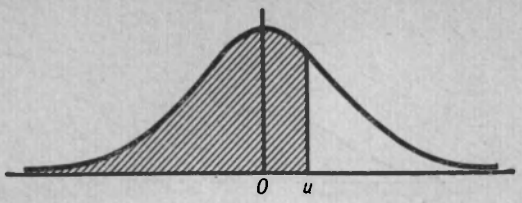


Рис. 2.6. Геометрический смысл интеграла вероятностей.

щены. Как увидим далее, табл. II Приложений приходится пользоваться довольно часто.

Во многих случаях удобно пользоваться таблицей, указывающей долю вариант, лежащих левее заданной абсциссы (заштрихованная часть на рис. 2.6). Такая величина обозначается $\Phi(u)$ и называется *интегралом вероятностей*; название этой функции связано с тем, что вычисление площадей, ограниченных кривыми, сводится к математической операции интегрирования (вычисления интеграла). Значения $\Phi(u)$ приведены в табл. III Приложений, а график этой функции представлен на рис. 2.7. Разумеется, доля вариант, лежащих правее абсциссы u, равна $1 - \Phi(u)$.

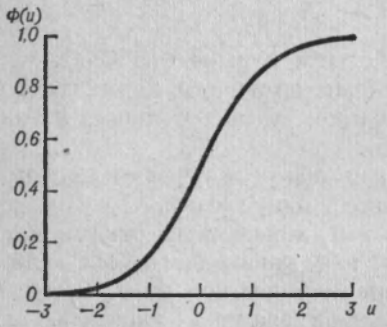


Рис. 2.7. Функция нормального распределения (интеграл вероятностей).

Пример 2.1. Из 500 студентов у 27 рост превышает средний рост для всей совокупности более чем на 10 см. Каково примерно стандартное отклонение распределения студентов по росту (предполагаемого нормальным)?

Так как минимальное отклонение равно здесь 10 см, то величина u на рис. 2.7 равна $10/\sigma$. По условию задачи правый незаштрихованный «хвост» на графике приближенно составляет $\frac{27}{500} = 0{,}054$ всей площади под кривой. Значит:

$$0{,}054 = 1 - \Phi\left(\frac{10}{\sigma}\right)$$

или

$$\Phi\left(\frac{10}{\sigma}\right) = 1 - 0{,}054 = 0{,}946.$$

Из табл. III Приложений получаем $10/\sigma = 1{,}61$, так что $\sigma = 10/1{,}61 \approx 6{,}2$ см.

Из сравнения рис. 2.5 и 2.6 следует:

$$\Phi(u) = \frac{1}{2} + \frac{\theta(u)}{2} = \frac{1}{2} - [1 + \theta(u)], \tag{2.5}$$

$$1 - \Phi(u) = \frac{1}{2}[1 - \theta(u)]. \tag{2.6}$$

Это позволяет находить значение $\Phi(u)$, пользуясь табл. II Приложений, и значения $\theta(u)$, пользуясь табл. III Приложений.

§ 2.4. Асимметрия распределения

2.4.1. Распределение Гаусса обычно получается при совместном воздействии ряда малых независимых (значит, случайно сочетающихся) факторов, число которых неограниченно велико. Это условие (одновременное воздействие большого числа малых по сравнению с общей суммой факторов) выполняется в природе очень часто. Поэтому гауссово распределение и принято называть нормальным. В частности, очень многие статистические совокупности, встречающиеся в биологической практике, имеют нормальное или почти нормальное распределение. Вместе с тем нередки случаи, когда распределение не является нормальным даже приблизительно. Это прежде всего заметно асимметричные распределения. Действительно, нормальное распределение, как было показано выше, является симметричным, поэтому распределение заведомо нельзя рассматривать как нормальное, если асимметрия достаточно велика.

О том, какая асимметрия «достаточно велика», и вообще о проверке нормальности распределения сказано в § 2.5.

Не вдаваясь пока в исследование причин асимметрии распределения, посмотрим, каким образом можно дать ее количественное описание.

В § 2.3 мы уже пользовались понятием математического ожидания, которое обсуждается более обстоятельно в § 3.1. Введем теперь еще понятие *отклонения*:

$$\xi = x - \mu.$$

Нетрудно показать (это сделано в том же § 3.1), что среднее отклонение[1]

$$\langle \xi \rangle = \frac{1}{n} \sum n_i \xi_i$$

[1] Угловыми скобками мы будем обозначать усреднение.

всегда равно нулю. Но средний куб отклонений

$$\langle \xi^3 \rangle = \frac{1}{n} \Sigma n_i \xi_i^3 = \frac{1}{n} \Sigma n_i (x_i - \mu)^3 \qquad (2.7)$$

в общем случае отличен от нуля. Действительно, пусть некоторое заданное распределение содержит, помимо прочих, отклонения $\xi_1 = -5$ с частотой $n_1 = 12$ и $\xi_2 = 2$ с частотой $n_2 = 30$. При вычислении $\langle \xi \rangle$ отвечающие этим разрядам слагаемые будут $n_1 \xi_1 = 12(-5) = -60$ и $n_2 \xi_2 = 30 \cdot 2 = 60$, так что они компенсируются. В то же время при вычислении $\langle \xi \rangle^3$ соответствующие слагаемые будут $n_1 \xi_1^3 = 12(-125) = -1500$ и $n_2 \xi_2^3 = 30 \cdot 8 = 240$ и компенсации не будет. Величина $\langle \xi^3 \rangle$ равна нулю в том случае, если распределение симметрично: так как при таком распределении расположенные симметрично (по отношению к центру распределения) частоты равны между собой, то они дадут равные по величине, но противоположные по знаку (так как значения возводятся в нечетную степень) вклады в сумме $\Sigma n_i \xi_i^3$.

Очевидно, величина $\langle \xi^3 \rangle$ тем больше, чем сильнее выражена асимметрия распределения; кроме того, знак величины $\langle \xi^3 \rangle$ однозначно связан с направлением асимметрии: если распределение вытянуто в сторону положительных значений (центр распределения принимаем за нуль), то $\langle \xi^3 \rangle > 0$; в противном случае $\langle \xi^3 \rangle < 0$.

По этим причинам естественно принять $\langle \xi^3 \rangle$ в качестве характеристики асимметрии распределения. Однако численное значение характеристики асимметрии не должно меняться при изменении масштаба измерения величин x. Это условие будет соблюдено, если разделить $\langle \xi^3 \rangle$ на σ^3, так как параметр:

$$\rho_3 = \frac{\langle \xi^3 \rangle}{\sigma^3} \qquad (2.8)$$

будет безразмерным. Этот параметр называют *коэффициентом асимметрии*.

О несмещенной выборочной оценке коэффициента асимметрии см. раздел **3.4.2.** *О стандартных ошибках и доверительном интервале для* ρ_3 *см. разделы* **3.5.2** и **2.5.1.**

2.4.2. Рассмотрим теперь некоторые причины, приводящие к отклонению эмпирического распределения от нормального.

Одной из причин асимметрии может быть то, что совокупность является неоднородной; последнее означает, что в одну совокупность сведены две или большее число нормальных совокупностей, каждая из которых характеризуется своим набором основных параметров μ и σ. Так, если мы будем строить распределение по

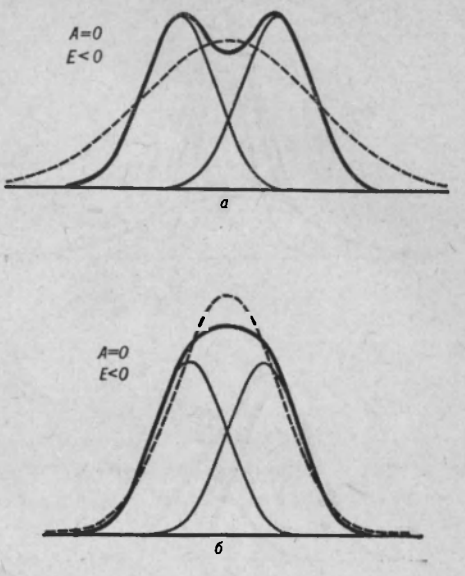


Рис. 2.8. Смеси двух нормальных совокупностей с одинаковыми стандартными отклонениями и разными математическими ожиданиями.

весу для групп каких-либо животных без различия пола, то заведомо не получится нормальное распределение, если даже для каждой из подсовокупностей — отдельно для самцов и для самок — распределение является нормальным.

Если разница между средними значениями двух подсовокупностей больше, чем стандартное отклонение каждой из них, то кривая распределения будет двухвершинной (рис. 2.8, *а*); при небольшом различии средних значений кривая имеет одну, но тупую вершину (рис. 2.8, *б*). На рис. 2.9 показаны два других частных случая, приводящих к негауссову распределению (тонкие линии — графики плотностей смешиваемых нормальных подсовокупностей; жирные линии — графики плотностей суммарных неоднородных совокупностей; пунктирные линии — графики нормальных совокупностей, имеющих такое же стандартное отклонение, что и суммарные совокупности).

Разумеется, вряд ли кто-либо станет сводить в одну статистическую совокупность самцов и самок. Но вполне возможны случаи, когда в исследуемую группу животных попадают особи из партий, имеющих несколько различное происхождение или развивавшихся в несколько различных условиях. Поэтому отклонение распределения от нормального всегда наталкивает на мысль о том, что совокупность не является однородной. При этом надо учесть, что неоднородность совокупности не обязательно досадное следствие методической ошибки. Например, в селекционной практике неоднородность совокупности может отражать появление в популяции под действием какого-либо фактора группы особей

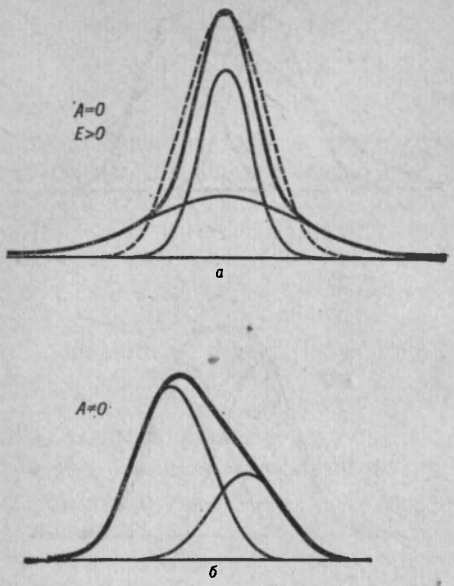


Рис. 2.9. Смеси двух нормальных совокупностей: (*а*) — с одинаковыми математическими ожиданиями и разными стандартными отклонениями; *б* — с разными математическими ожиданиями и разными объемами.

с определенным сдвигом по исследуемому признаку; ясно, что в этом случае неоднородность совокупности будет весьма желаемым результатом. Изучение характера распределения позволяет выявить это расщепление популяции — выделение нового сорта, породы и др.— на очень ранней стадии и тем самым верно выбрать направление дальнейших поисков.

В ряде случаев можно даже вычислить параметры составляющих совокупностей. В частности, это можно сделать, если предположить, что данная совокупность состоит только из двух подсовокупностей (в том смысле, что если имеется примесь и других подсовокупностей, то она мала и при заданной точности может не учитываться)[1].

2.4.3. Неоднородность совокупности — лишь одна из возможных причин отклонений распределения от нормального. Появление асимметрии может быть связано также с особенностями выбора признака, по которому изучается распределение.

Пример 2.2. В табл. 2.3 приведено распределение по весу зерен пшеницы: были взяты пробы со 100 делянок и для каждой из этих проб определен средний вес зерна. На рис. 2.10, изображающем полигон частот этого распределения, видно, что оно симметрично. Расчет подтверждает это: коэффициент асимметрии $\rho_3 = 0{,}06$, т. е. весьма мал. Возьмем теперь в качестве признака,

[1] О способах решения этой задачи см. в книге автора «Биометрические методы», с. 74-76.

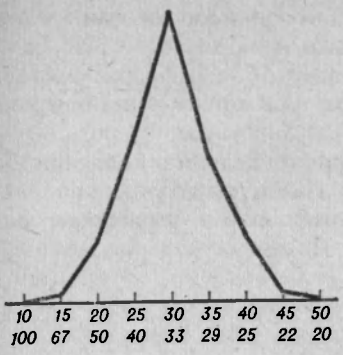


Рис. 2.10. Симметричное распределение веса зерен.

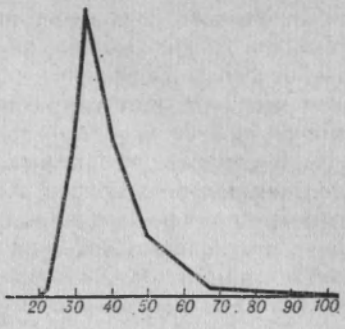


Рис. 2.11. Асимметричное распределение числа зерен в 1 г.

по которому происходит распределение, не вес одного зерна, а число зерен, приходящихся на 1 г. Тогда получим числа

$$x_i' = \frac{1}{x_i},$$

выписанные в столбце 3 табл. 2.3 и на рис. 2.10 под соответствующими числами x_i. Естественно, шкала значения x_i' получилась неравномерной. Если мы хотим взять за основу числа x_i', то следует сделать шкалу этих значений равномерной. Это может быть достигнуто соответствующей деформацией графика, что приводит к рис. 2.11. Вновь получившееся распределение уже асимметрично, что подтверждается и расчетом: $\rho_3 = 1{,}34$.

ТАБЛИЦА 2.3

Средний вес зерна (мг) на делянке	Число делянок	Число зерен в 1 г
15	1	67
20	8	50
25	22	40
30	39	33
35	20	29
40	9	25
45	1	22
Сумма	100	

Конечно, в данном случае можно было бы сказать, что вес зерна является более фундаментальным признаком, чем число зерен, приходящихся на 1 г, и что поэтому асимметрию во втором случае

следует считать искусственной; однако вряд ли все сочтут такое рассуждение достаточно убедительным. Например, можно ли сразу ответить на вопрос, какой признак более фундаментален — время между двумя ударами пульса или число ударов пульса в минуту?

Если считать, что нормальное распределение является более «естественным», то можно пытаться в каждом случае превращать асимметричное распределение в симметричное, применяя надлежащее преобразование аргумента. Помимо уже рассмотренного преобразования:

$$x' = -\frac{1}{x},$$

можно использовать также преобразование:

$$x' = \sqrt{x}$$

или какое-нибудь другое. Чаще всего применяют преобразование:

$$x' = \lg x.$$

Следует, однако, отметить, что далеко не всегда удается достаточно простым способом объяснить происхождение асимметрии и тем самым обосновать разумность того или иного преобразования. Часто бывает так, что хотя и удается чисто эмпирически подобрать подходящее преобразование, превращающее заданное асимметрическое распределение в симметричное, но дать какую-нибудь более или менее ясную интерпретацию этого преобразования невозможно.

§ 2.5. Критерий нормальности распределения

2.5.1. Теперь, познакомившись с понятием и количественной характеристикой асимметрии, мы можем перейти к рассмотрению критерия нормальности распределения.

Напоминаем еще раз, что проверку нормальности предпринимают только тогда, когда имеются реальные основания усомниться в таком характере распределения.

Как мы уже знаем, нормальное распределение симметрично, так что условием нормальности является равенство $\rho_3 = 0$. Но при этом надо учесть, что эмпирическая совокупность, с которой мы обычно имеем дело, всегда есть выборка из некоторой генеральной совокупности. Очевидно, повторные выборки из одной и той же генеральной совокупности не тождественны между собой, так что и значения коэффициентов асимметрии у них окажутся разными. Следовательно, если даже генеральная совокупность

симметрична, то выборки, как правило, будут несколько асимметричными. Можно показать, что при этом выборочные коэффициенты асимметрии сами распределены симметрично (и даже почти нормально) со стандартным отклонением, которое приблизительно обратно пропорционально квадратному корню из объема выборки. Это приводит к тому, что случайное появление не очень больших значений выборочных ρ_3 не противоречит предположению о том, что в генеральной совокупности $\rho_3 = 0$, а появление значений ρ_3, сильно превышающих свою стандартную ошибку (о стандартной ошибке см. раздел **3.5.1**), маловероятно. На этом и основана проверка того, что в генеральной совокупности $\rho_3 = 0$: при данном объеме выборки n выборочное значение ρ_3 не должно превышать определенного критического значения.

Идеи, на которых основывается применение статистических критериев, рассматриваются более обстоятельно в § 4.1.

ТАБЛИЦА 2.4

Критерии для проверки нормальности распределения; ноль целых и запятая опущены (из книги Л. Н. Большева и Н. В. Смирнова, с. 320)

n	A_α		c_α	
	5%	1%	5%	1%
30	661	982	739—863	710—884
35	621	921	743—859	716—878
40	587	869	746—855	721—873
45	558	825	749—852	725—869
50	533	787	752—849	729—866
60	492	723	755—844	734—859
70	459	673	758—840	739—855
80	432	631	761—838	743—852
90	409	596	763—835	746—848
100	389	567	764—834	749—846
150	321	464	770—827	758—837
200	280	403	774—823	763—832
300	230	329	778—818	769—826
400	200	285	781—816	773—822
500	179	255	782—814	776—820
600	163	233	784—812	778—818
700	151	215	785—811	779—816
800	142	202	786—810	780—815
900	134	190	787—809	781—814
1000	127	180	787—809	782—813

Если $|\rho_3| < A_\alpha$, а $c = |\xi|/s$ находится в пределах, указанных в столбце c_α, то принимается гипотеза о нормальности распределения. Если же нарушается хотя бы одно из этих условий, то гипотеза о нормальности распределения отвергается.

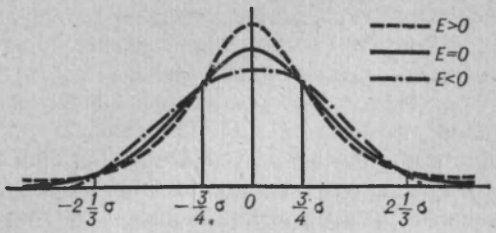


Рис. 2.12. Нормальное, островершинное и туповершинное распределения.

Критические значения ρ_3 даны в табл. 2.4. Каждому значению n (объема выборки) отвечают два значения, обозначаемые обычно как $A_{0,05}$ и $A_{0,01}$. Если $|\rho_3| \geqslant A_{0,05}$, то вероятность того, что в генеральной совокупности $\rho_3 = 0$, не превышает 0,05 (или 5%); если же $|\rho_3| \geqslant A_{0,01}$, то соответствующая вероятность не превышает 0,01 = 1%. Очевидно, вероятность того, что в генеральной совокупности $\rho_3 \neq 0$, будет соответственно не меньше, чем 0,95 = 95%, или 0,99 = 99%.

Таким образом, если, по данным выборки, получается $|\rho_3| > A_\alpha$ (где $\alpha = 0{,}05$ или 0,01), то гипотеза о том, что в генеральной совокупности $\rho_3 = 0$, может быть отвергнута с вероятностью $1 - \alpha$ (т. е. 0,95 или 0,99). О выборе того или другого уровня вероятности см. раздел 4.1.1.

2.5.2. Однако симметричность распределения в генеральной совокупности еще не означает, что это распределение нормально. Имеется еще одно свойство, по которому распределение может отличаться от нормального. Это свойство называется *крутостью*, или *эксцессом*, и иллюстрируется рис. 2.12. Средняя кривая на этом рисунке изображает *нормальное* распределение, а остальные два распределения называют соответственно *островершинным* и *туповершинным*.

Крутость распределения можно характеризовать различными показателями, но в данном случае практически наиболее удобен показатель

$$c = \frac{|\bar{\xi}|}{s}, \tag{2.9}$$

где

$$|\bar{\xi}| = \frac{1}{n} \Sigma n_i |\xi| = \frac{1}{n} \Sigma n_i |x_i - \bar{x}| \tag{2.10}$$

есть *среднее абсолютное отклонение*, вычисленное по выборочным данным. В этих формулах $\bar{x}$ и s являются выборочными оценками математического ожидания μ и стандартного отклонения σ; об этих оценках см. § 3.2 и 3.4.

Расчет показывает, что при нормальном распределении $c = \sqrt{\frac{2}{\pi}} \approx 0{,}798$. Поэтому основанием для отказа от гипотезы о

нормальном распределении может служить такое отличие выборочного значения c от указанного 0,798, вероятность которого меньше, чем принятый уровень α. Соответствующие критические значения c_α (для $\alpha = 0{,}05$ и $\alpha = 0{,}01$) приведены в табл. 2.4. Поскольку распределение выборочных c несколько асимметрично, нижние и верхние критические значения c_α расположены на различных расстояниях от 0,798; поэтому в табл. 2.4 приведены *критические интервалы* для c.

Гипотеза о нормальности распределения вариант в генеральной совокупности принимается, если $|\rho_3| \leqslant A_\alpha$, а c находится в пределах, указанных в столбце c_α. Если же $|\rho_3| > A_\alpha$ или c находится вне пределов, указанных в столбце c_α, то гипотеза о нормальности распределения отвергается. С целью сокращения в таблице везде опущен нуль целых и запятая.

Пример 2.3. Проверим нормальность распределения из табл. 2.2 (раздел 2.2.3). Расчет дает $\rho_3 = 0{,}04$, $s = 1{,}91$, $|\xi| = 1{,}45$, $c = 0{,}760$.

Обращаясь к табл. 2.4, видим, что $\rho_3 = 0{,}04$ много меньше 5%-ного критического значения $A_{0.05} = 0{,}389$, а значение $c = 0{,}760$ лишь немного выходит из 5%-ного критического интервала (0,764÷÷ 0,834) и во всяком случае гораздо ближе к нижней границе этого интервала, чем в нижней границе 1%-ного критического интервала (т. е. к значению 0,749). Поэтому можно считать, что нет оснований отвергнуть нулевую гипотезу.

§ 2.6. Исключение сильно отклоняющихся вариант

2.6.1. При статистической обработке биологического материала часто сталкиваются с наличием в исследуемой совокупности некоторого количества вариант, значения которых довольно резко отличаются от основной массы наблюдений. Появление таких вариант может объясняться естественной вариабельностью случайной величины, вследствие чего большие отклонения от центра распределения не исключены. Но они могут также свидетельствовать о неоднородности статистической совокупности: если вариантами являются значения какого-либо показателя для разных особей некоторой популяции, то неоднородность будет следствием попадания в совокупность особей из другой биологической популяции; если вариантами являются результаты каких-то повторных замеров на одном и том же объекте, то неоднородность будет следствием спорадических нарушений стандартных условий эксперимента.

Конечно, появление резко выделяющихся вариант указывает прежде всего на необходимость тщательной проверки исследуемой популяции или обстановки эксперимента. При этом варианты, ус-

ловия получения которых противоречили стандарту, должны отбрасываться независимо от их значений. Однако во многих случаях не удается получить прямых указаний о неоднородности изучаемой совокупности. Тогда приходится прибегать к статистическим критериям.

Применение последних основано на том, что если распределение вариант в генеральной совокупности нормально или близко к нормальному, то появление в выборке вариант, далеко отклоняющихся от центра распределения, хотя и возможно, но очень маловероятно.

Дальнейшие рассуждения основаны на свойствах нормального распределения. Если Вы их забыли, прочтите § 2.3.

Известно, что при нормальном распределении вероятность появления вариант, отстоящих от математического ожидания дальше, чем на $u_\alpha\sigma$, равна $1 - \Theta(u_\alpha) = \alpha$; например, в единичном опыте вероятность появления варианты на расстоянии $2{,}58\sigma$ и больше от μ равна 0,01. Отсюда часто делается вывод, что при выбранном уровне значимости α можно отбрасывать варианты, для которых $|u| = \frac{|x - \mu|}{\sigma} > u_\alpha$ Однако такое заключение ошибочно. Ведь ясно, что как бы ни была мала относительная вероятность появления каких-либо вариант, абсолютная вероятность их появления может оказаться большой при достаточно большом объеме выборки. Очевидно, при построении критерия исключения (соответствующее критическое значение обозначим ω_α) надо исходить из условия, что в выборке данного объема n из нормальной генеральной совокупности не должно содержаться, с определенной вероятностью P, ни одной варианты, отклоняющейся от μ больше, чем на $\omega_\alpha\sigma$. Выбранный уровень вероятности $\alpha = 1 - P$ имеет здесь тот смысл, что если выбрать из нормальной генеральной совокупности большое число выборок, объема n каждая, то в среднем лишь в 100α процентах из них будут попадаться варианты вне пределов $\mu \pm \omega_\alpha\sigma$, а $100(1 - \alpha) = 100\ P$ процентов выборок не будут содержать вариант вне этих пределов. Из сказанного ясно, что критические значения ω_α должны зависеть как от принятого уровня вероятности α, так и от объема выборки n.

Табл. 2.5 содержит эти критические значения $\omega_\alpha\ (n)$. При построении критерия принято во внимание, что значения μ и σ обычно неизвестны и заменяются их оценками $\bar{x}$ и s, так что:

$$\omega_{max} = \frac{x_{max} - \bar{x}}{s}; \qquad \omega_{min} = \frac{\bar{x} - x_{min}}{s}. \tag{2.11}$$

ТАБЛИЦА 2.5

Критерий ω_α для отбрасывания крайних вариант

n	ω_α		n	ω_α		n	ω_α	
	5%	1%		5%	1%		5%	1%
5	1,92	1,97	21	2,80	3,11	80	3,33	3,70
6	2,07	2,16	22	2,82	3,13	90	3,37	3,74
7	2,18	2,31	23	2,84	3,16	100	3,40	3,77
8	2,27	2,43	24	2,86	3,18	120	3,46	3,83
9	2,35	2,53	25	2,88	3,20	150	3,53	3,90
10	2,41	2,62	26	2,90	3,22	200	3,61	3,98
11	2,47	2,69	27	2,91	3,24	300	3,73	4,09
12	2,52	2,75	28	2,93	3,26	400	3,80	4,17
13	2,56	2,81	29	2,94	3,28	500	3,87	4,24
14	2,60	2,86	30	2,96	3,29	600	3,92	4,28
15	2,64	2,90	35	3,02	3,36	700	3,96	4,32
16	2,67	2,94	40	3,08	3,42	800	3,99	4,35
17	2,70	2,98	45	3,12	3,48	900	4,02	4,38
18	2,73	3,02	50	3,16	3,52	1000	4,05	4,41
19	2,75	3,05	60	3,22	3,58	1500	4,14	4,50
20	2,78	3,08	70	3,28	3,64	2000	4,21	4,56

Величины x_{max} и x_{min} представляют собой то наибольшее или наименьшее значение вариант в выборке, принадлежность которых к рассматриваемой генеральной совокупности вызывает подозрение.

Что касается упомянутых здесь оценок математического ожидания μ и стандартного отклонения σ, то о них следует прочитать в § 3.2 и 3.4.

Если окажется, что $\omega > \omega_\alpha$, то сомнительные варианты можно квалифицировать как «артефакты» и исключить из дальнейшей обработки. В отношении выбора уровня вероятности α можно рекомендовать следующий подход: если $\omega > \omega_{0,01}$, то варианта отбрасывается, а если $\omega \leqslant \omega_{0,05}$, то она безусловно оставляется; если же ω окажется между $\omega_{0,05}$ и $\omega_{0,01}$, то при решении вопроса об отбрасывании варианты должна быть проявлена определенная осторожность.

2.6.2. При большом объеме выборки вопрос об исключении «артефактов» не стоит особенно остро, так как относительный «вес» нескольких сомнительных вариант при вычислении усредненных показателей сравнительно невелик. Если же выборка мала, то даже одно неправильное значение может заметно исказить результат усреднения.

Для случая малых выборок можно указать упрощенный способ оценки принадлежности варианты к заданной совокупности,

хотя и менее точный. Этот способ основан на замене выражения

$$\omega = \frac{x - \bar{x}}{s} \qquad (*)$$

другим, более просто вычисляемым.

Пусть мы имеем выборку $x_1, x_2, \ldots, x_n$. Обозначим через $x_{(1)}, x_{(2)}, \ldots x_{(n)}$ те же варианты, но расположенные в порядке возрастания. Если, например, задана совокупность:

$$x_1 = 21, \quad x_2 = 17, \quad x_3 = 6, \quad x_4 = 24, \quad x_5 = 18,$$

то в этих обозначениях будем иметь:

$$x_{(1)} = 6, \quad x_{(2)} = 17, \quad x_{(3)} = 18, \quad x_{(4)} = 21, \quad x_{(5)} = 24.$$

Мы хотим проверить, не отклоняются ли слишком сильно крайние варианты, т. е. $x_{(1)}$ и $x_{(n)}$. Считая, что принадлежность к данной совокупности остальных вариант, в частности $x_{(2)}$ и $x_{(n-1)}$, не подвергается сомнению, можно характеризовать абсолютные отклонения крайних вариант от совокупности не величинами $x_{(n)} - \bar{x}$ и $\bar{x} - x_{(1)}$, а величинами $x_{(n)} - x_{(n-1)}$ и $x_{(2)} - x_{(1)}$; конечно, критические значения должны быть при этом иные.

Что касается величины σ, оценкой которой служит величина s в знаменателе (*), то ее можно оценить при помощи размаха варьирования $x_{(n)} - x_{(1)}$. Но так как именно крайние значения сомнительны, то целесообразно не связывать оценку значимости отклонения $x_{(n)}$ с сомнительной величиной $x_{(1)}$, а оценку значимости отклонения $x_{(1)}$ — с сомнительной величиной $x_{(n)}$. Поэтому будем при оценке значимости величины $x_{(n)} - x_{(n-1)}$ оценивать σ через $x_{(n)} - x_{(2)}$, а при оценке значимости величины $x_{(2)} - x_{(1)}$ — через $x_{(n-1)} - x_{(1)}$.

Таким образом, приходим к следующим величинам для решения вопроса о принадлежности крайних вариант к совокупности:

$$\tau' = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}; \quad \tau'' = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}. \qquad (2.12)$$

Критические значения для τ (в случае нормальной генеральной совокупности они, конечно, одинаковы для τ' и τ'') зависят как от принятого уровня вероятности α, так и от объема выборки n. Значения $\tau_\alpha(n)$ даются в табл. 2.6.

ТАБЛИЦА 2.6

Критические значения τ'_α и τ''_α для отбрасывания крайних вариант

n	τ_α		n	τ_α		n	τ_α	
	0,01	0,05		0,01	0,05		0,01	0,05
4	0,991	0,955	13	0,520	0,410	22	0,414	0,320
5	0,916	0,807	14	0,502	0,395	23	0,407	0,314
6	0,805	0,689	15	0,486	0,381	24	0,400	0,309
7	0,740	0,610	16	0,472	0,369	25	0,394	0,304
8	0,683	0,554	17	0,460	0,359	26	0,389	0,299
9	0,635	0,512	18	0,449	0,349	27	0,383	0,295
10	0,597	0,477	19	0,439	0,341	28	0,378	0,291
11	0,566	0,450	20	0,430	0,334	29	0,374	0,287
12	0,541	0,428	21	0,421	0,327	30	0,369	0,283

Пример 2.4. В приведенной выше выборке из 5 вариант

21 17 6 24 18

может вызвать сомнения варианта $x_{(1)} = 6$. По формуле (2.12) имеем:

$$\tau'' = \frac{17-6}{21-6} = \frac{11}{15} \approx 0{,}73;$$

это значение меньше, чем $\tau_{0,05}(5) = 0{,}81$, поэтому варианту $x_3 = 6$ отбросить нельзя.

Хотя второй критерий проще первого (требуется меньше вычислений), он не может полностью заменить его, так как хорош только для малых выборок. Действительно, при большом объеме выборки может оказаться много «несобственных» вариант, что потребует многократного повторения расчета величины (2.12) и тем самым сведет на нет преимущество простоты. Кроме того, в большой выборке «несобственные» варианты могут сами составлять компактную группу (примесь элементов другой генеральной совокупности), и тогда критерий (2.12) не обнаружит неоднородности.

Конечно, надо помнить, что оба критерия основаны на предположении о нормальности распределения вариант в генеральной совокупности. Поэтому если есть основания сомневаться в таком характере распределения, то лучше избегать применения этих критериев.

Глава третья

ОПИСАНИЕ НЕПРЕРЫВНОЙ СТАТИСТИЧЕСКОЙ СОВОКУПНОСТИ

§ 3.1. Параметры распределения. Математическое ожидание, медиана и мода

3.1.1. После упорядочения, описанного в главе второй, статистическая совокупность предстает в виде некоторого распределения. Замена совокупности отдельных значений небольшим упорядоченным рядом несовпадающих значений (или разрядов) и отвечающих им частот приводит к заметному «уплотнению» информации. Однако это «уплотнение информации» можно продолжить, если характеризовать распределение совсем небольшим числом каких-то обобщенных показателей. Эти обобщенные показатели, вычисленные по данным эмпирической совокупности (т. е. по данным выборки), называются *статистиками*. Они могут служить оценками соответствующих показателей, характеризующих генеральную совокупность и называемых *параметрами* этой генеральной совокупности (точнее — параметрами распределения генеральной совокупности).

Сколько параметров и какие именно характеризуют распределение достаточно полно, это зависит от вида распределения. Например, альтернативное распределение можно описать одним единственным параметром — долей вариант одного из типов.

Альтернативному распределению посвящена специальная глава — шестая.

3.1.2. Большинство симметричных распределений достаточно полно описывается всего двумя параметрами, указывающими:

1) положение центра, вокруг которого группируются варианты, и

2) рассеяние вариант относительно этого центра.

Для описания умеренно асимметричных распределений используется еще третий параметр — коэффициент асимметрии; об этом параметре см. § 2.4. Если же Вы имеете дело с существенно асимметричным распределением редких событий (распределение Пуассона), то обратитесь к главе седьмой, целиком посвященной этому распределению.

В качестве параметра, указывающего положение центра распределения, чаще всего выбирают *математическое ожидание*

$$\mu = \mathrm{M}(x) = \Sigma p_j x_j, \tag{3.1}$$

где x_j — несовпадающие значения и p_j — соответствующие им вероятности. Эта формула относится к дискретным распределениям. Если же распределение непрерывно, то математическое ожидание определяется как предел суммы $\Sigma x \cdot f(x) \cdot \Delta x$ [где $f(x)$ есть плотность распределения] при бесконечном уменьшении Δx и соответственно бесконечном увеличении числа слагаемых; такой предел суммы называется интегралом.

Разность:

$$\xi_j = x_j - \mu$$

назовем *отклонением* варианты x_j от математического ожидания. Можно показать следующее:

1. Среднее отклонение (т. е. математическое ожидание отклонения) равно нулю:

$$\mathrm{M}(\xi) = \Sigma p_j \xi_j = 0. \tag{3.2}$$

2. Средний квадрат отклонений вариант от μ меньше, чем средний квадрат их отклонений от любого другого значения x; иначе говоря, μ есть то значение x, для которого сумма $\Sigma p_j (x_j - x)^2$ имеет минимальное значение:

$$\Sigma p_j (x_j - \mu)^2 < \Sigma p_j (x_j - x)^2 \tag{3.3}$$

при всех $x \neq \mu$.

Последнее свойство и является причиной того, что центр распределения чаще всего характеризуют математическим ожиданием.

3.1.3. Нетрудно показать, что если к каждому значению x прибавить одно и то же постоянное число a, т. е. заменить все значения x значениями $x^* = x + a$, то математическое ожидание увеличится на то же число a, т. е.

$$\mathrm{M}\{x + a\} = \mathrm{M}\{x\} + a; \tag{3.4}$$

в самом деле:

$$\mathrm{M}\{x^*\} = \mathrm{M}\{x + a\} = \Sigma p_j (x_j + a) = \Sigma p_j x_j + a \Sigma p_j = \mathrm{M}\{x\} + a,$$

так как $\Sigma p_j = 1$. Аналогично доказывается, что

$$\mathrm{M}\{ax\} = a\mathrm{M}\{x\}. \tag{3.5}$$

Эти два свойства математического ожидания позволяют получать характеристики величин, найденных путем косвенных измерений.

Если связь между x^* и x нелинейна, то $M(x^*)$ выражается через $M(x)$ более сложным образом и зависит также от дисперсии $\sigma^2(x)$ значений x (о дисперсии см. § 3.3). Например[1]:

$$M(e^x) = e^{M(x)}\left\{1 + \frac{1}{2}\sigma^2(x)\right\}, \tag{3.6}$$

$$M(\lg x) = \lg M(x) - \frac{\sigma^2(x)}{4{,}6\,[M(x)]^2}, \tag{3.7}$$

$$M(x^k) = [M(x)]^k\left\{1 + \frac{k(k-1)}{2}\cdot\frac{\sigma^2(x)}{[M(x)]^2}\right\}. \tag{3.8}$$

В частности, из последней формулы имеем:

$$M(x^2) = [M(x)]^2 + \sigma^2(x), \tag{3.9}$$

$$M(\sqrt{x}) = \sqrt{M(x)}\left\{1 - \frac{1}{8}\frac{\sigma^2(x)}{[M(x)]^2}\right\}, \tag{3.10}$$

$$M\left(\frac{1}{x}\right) = \frac{1}{M(x)}\left\{1 + \frac{\sigma^2(x)}{[M(x)]^2}\right\}. \tag{3.11}$$

3.1.4. Интересующая нас величина z может быть функцией от двух (или большего числа) величин x, y, ... и при этом может возникнуть необходимость выразить $M(z)$ через $M(x)$, $M(y)$, ... Нетрудно показать, что

$$M(x \pm y) = M(x) \pm M(y). \tag{3.12}$$

Очевидно, это равенство немедленно обобщается на любое число варьирующих величин.

Можно также показать, что

$$M(xy) = M(x)\,M(y) \tag{3.13}$$

при условии, что x и y варьируют независимо одна от другой. Если, например, $M(x)$ есть среднее (за месяц) число болевших работников предприятия, а $M(y)$ — среднее число дней нетрудоспособности на одного болевшего, то $M(x)M(y)$ будет средним (за месяц) числом рабочих дней, пропущенных работниками предприятия

---

[1] Для тех, кто знаком с дифференциальным исчислением, поясняем происхождение этих формул. Функцию $x^* = F(x)$ разлагают в ряд Тейлора около точки $x_0 = M(x)$ и ограничиваются первыми тремя членами разложения: $x^* = F\,[M(x)] + F'[M(x)]\,\{x - M(x)\} + \frac{1}{2}F''\,[M(x)]\,\{x - M(x)\}^2$. Если затем произвести усреднение и учесть, что $M\{x - M(x)\} = 0$, $M\,(\{x - M(x)\}^2) = \sigma^2(x)$, то получится: $M(x^*) = F\,[M(x)] + \frac{1}{2}F''\,[M(x)]\sigma^2\,(x)$. Из этих выкладок, кстати, видно, что все приводимые далее формулы, за исключением (3.9), являются приближенными.

по болезни. Но при эпидемии гриппа, когда длительность болезни возрастает «параллельно» возрастанию числа заболевающих, равенство (3.13) нарушится. В этом случае:

$$M(xy) = M(x)\,M(y) + cov(x, y), \qquad (3.13')$$

где $cov(x, y)$ есть так называемая ковариация величин x и y (см. раздел 8.3.1).

3.1.5. В некоторых случаях применение математического ожидания для характеристики центра распределения невозможно или нецелесообразно. Пусть, например, имеется так называемая незамкнутая совокупность, когда не указано начало или конец ряда (или оба вместе);

Площадь пашни, га	<500	500—1000	1000—2000	2000—3000	3000—5000	>5000	Итого
Число хозяйств	12	35	63	103	187	26	428

Здесь нельзя вычислить среднее значение, потому что неизвестны середины первого и последнего разрядов. Тогда центр распределения характеризуют *медианой* (обозначается *Me*), которая делит совокупность на две части таким образом, что половина всех вариант меньше этой величины, а другая половина — больше. Например, в совокупности с вариантами:

16 19 21 26 27 31 32 35 39 41 45 47 48

медианой будет варианта 32, так как шесть значений (16, 19, 21, 26, 27 и 31) меньше 32 и столько же значений (35, 39, 41, 45, 47 и 48) больше 32. Если бы варианта 16 отсутствовала, так что общее число вариант было бы четным, то в качестве медианы следовало бы принять полусумму двух средних вариант:

$$Me = \frac{32 + 35}{2} = 33{,}5.$$

Если объем совокупности велик, то вычисление медианы производится следующим образом. После группировки совокупности в разряды составляется так называемый ряд накопленных частот s_i:

$$s_1 = n_1, \quad s_2 = n_1 + n_2, \quad s_3 = n_1 + n_2 + n_3 \text{ и т. д.}$$

Для распределения из табл. 2.2 (см. раздел 2.2.3) ряд накопленных частот будет иметь вид табл. 3.1.

ТАБЛИЦА 3.1

x_i	5,20	5,25	5,30	5,35	5,40	5,45	5,50	5,55	5,60	5,65	5,70
n_i	1	4	7	11	16	30	14	8	6	2	1
s_i	1	5	12	23	39	69	83	91	97	99	100

При практическом вычислении s_i нет надобности каждый раз производить суммирование всех частот, так как $s_i = s_{i-1} + n_i$; например (см. табл. 3.1), $s_6 = s_5 + n_6 = 39 + 30 = 69$, $s_7 = s_6 + n_7 = 69 + 14 = 83$ и т. д.

Из рассмотрения этого ряда видно, что медиана лежит между значениями 5,40 (такое и меньшее значения имеют 39% всех вариант) и 5.45 (такое и меньшее значения имеют 69% всех вариант). Для более точного указания положения медианы надо знать распределение вариант внутри этого интервала. Обычно при нахождении медианы принимается условно, что внутри интервала варианты распределены равномерно (напомним, что так же поступают и при построении гистограммы). Тогда задача сводится к элементарной пропорции:

$$\begin{array}{r|l} 5{,}40 \sim 39\% & Me = 5{,}40 + \dfrac{5{,}45 - 5{,}40}{69 - 39}(50 - 39) = \\ 5{,}49 \sim 69\% & \\ Me \sim 50\% & = 5{,}40 + \dfrac{0{,}05 \cdot 11}{30} \approx 5{,}42. \end{array}$$

Расчет основан на том, что если увеличению накопленной частоты на $69-39 = 30$ единиц соответствует сдвиг значений на $5{,}45-5{,}40 = 0{,}05$ мм, то увеличению s_i на $50-39 = 11$ единиц будет соответствовать сдвиг во столько раз меньше, чем 0,05, во сколько раз 11 меньше, чем 30.

Центр распределения можно также характеризовать *модой* (обозначается *Mo*) — серединой того разряда группировки, в котором содержится наибольшее число вариант (этот разряд называется *модальным*). Мода является единственно возможной характеристикой «центра» при группировке по качественному признаку, для количественных же совокупностей более предпочтительно математическое ожидание в силу свойств (3.2) и (3.3), которым не обладает мода. Однако иногда понятие моды может оказаться полезным и для совокупностей с количественной градацией. Так, на рис. 3.1 изображено распределение по возрасту заболевших дифтерией (на 10 тыс. населения соответствующего возраста). Очевидно, знание среднего возраста заболевающих дифтерией (в данном

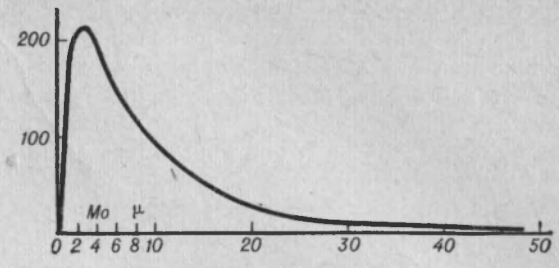


Рис. 3.1. Математическое ожидание и мода в асимметричном распределении.

случае 7,75 года) менее интересно, чем знание возраста, в котором чаще всего происходит заболевание (в данном случае от 2 до 4 лет), в частности, при решении вопроса о том, где должны быть сосредоточены главные профилактические условия: в школах или в дошкольных учреждениях.

В самом грубом приближении в качестве моды можно принять середину разряда, на который приходится наибольшая частота.

Если распределение более или менее симметрично, т. е. по мере удаления от вершины кривая распределения убывает примерно одинаково быстро в обе стороны, то мода и математическое ожидание близки между собой. Поэтому в таких случаях мода, находимая более просто, может служить приближенной оценкой математического ожидания.

§ 3.2. Среднее значение как оценка математического ожидания

3.2.1. Для вычисления математического ожидания μ нужно знать вероятности p_j значений x_j в генеральной совокупности. Обычно эти вероятности неизвестны, но, изучив выборку из интересующей нас генеральной совокупности, мы можем получить относительные частоты h_j в этой выборке или абсолютные частоты $n_j = nh_j$, где n — объем выборки. В случае непрерывной генеральной совокупности величины h_j заменяют усредненные (для конечных интервалов группировки) плотности вероятности.

Используя значения h_j или n_j, можно вычислить величину:

$$\bar{x} = \Sigma h_j x_j = \frac{1}{n} \Sigma n_j x_j, \tag{3.14}$$

которую называют *средним значением*; если варианты в выборке не сгруппированы, то

$$\bar{x} = \frac{1}{n} \Sigma x_i, \tag{3.15}$$

где x_i — значения отдельных вариант.

3.2.2. Легко убедиться, что свойства (3.2) и (3.3) математического ожидания присущи и среднему значению. Это позволяет упростить его вычисление, которое даже после группировки вариант все еще остается достаточно громоздким. Покажем это на том же примере из табл. 2.2 из раздела **2.2.3.** Очевидно, каждую из вариант этой таблицы можно представить в виде:

$$x_i = 5{,}00 + y_i,$$

и тогда:

$$\bar{x} = 5{,}00 + \bar{y}.$$

Дальнейшее упрощение достигается тем, что вводится условная шкала значений, а именно ширина разряда Δx принимается за единицу, так что номера разрядов можно рассматривать как новые значения (в единицах $l = \Delta x$). Тогда вместо табл. 2.2 имеем:

x_i	1	2	3	4	5	6	7	8	9	10	11
n_i	1	4	7	11	16	30	14	8	6	2	1

(начало отсчета $x_0 = 5{,}15$, масштаб $l = 0{,}05$).

Вычисления можно еще больше упростить, если выбрать в качестве начала отсчета разряд с наибольшей частотой (в нашем примере значение 5,45):

x_i	−5	−4	−3	−2	−1	0	1	2	3	4	5
n_i	1	4	7	11	16	30	14	8	6	2	1

(*)

Тогда

$$\bar{x} = 5{,}45 + \frac{0{,}05}{100}(-1\cdot 5 - 4\cdot 4 - 7\cdot 3 - 11\cdot 2 - 16\cdot 1 + 14\cdot 1 +$$

$$+ 8\cdot 2 + 6\cdot 3 + 2\cdot 4 + 1\cdot 5) = 5{,}45 + \frac{0{,}05}{100}(-80 + 61) =$$

$$= 5{,}45 - \frac{0{,}05}{100}\,19 \approx 5{,}44.$$

Переход от истинных значений 5,20; 5,25; 5,30 ...; 5,65; 5,70 к условным значениям −5; −4; −3 ...; 4; 5 иногда называют *кодированием*.

Нетрудно показать, что для среднего значения $\bar{x}$ справедливы формулы (3.4) — (3.5), (3.6) — (3.11) и (3.12) — (3.13), выведенные для математического ожидания μ.

3.2.3. Среднее значение есть характеристика выборки (т. е. статистика), а не параметр генеральной совокупности. Но $\bar{x}$ мож-

но рассматривать как *оценку* для величины μ. Разберем смысл этого утверждения.

Значения $\bar{x}$, полученные для разных выборок из одной генеральной совокупности, обычно не совпадают. Это можно иллюстрировать следующим простым примером. Пусть мы имеем генеральную совокупность, состоящую из пяти вариант ($N = 5$):

$$x_i : 8 \quad 16 \quad 20 \quad 24 \quad 32$$

(числа могут обозначать, скажем, высоту каких-либо растений в сантиметрах). Математическое ожидание равно:

$$\mu = \frac{8 + 16 + 20 + 24 + 32}{5} = \frac{100}{5} = 20 \text{ см}.$$

Заменим, однако, изучение всей генеральной совокупности изучением выборки из нее объемом $n = 4$. При случайном составлении выборки в нее может попасть с равной вероятностью любое из возможных сочетаний из $N = 5$ элементов по $n = 4$. Число таких сочетаний, как известно, равно

$$C_5^4 = \frac{5 \cdot 4 \cdot 3 \cdot 2}{1 \cdot 2 \cdot 3 \cdot 4} = 5.$$

Вот эти сочетания:

№ 1	8	16	20	24
№ 2	8	16	20	32
№ 3	8	16	24	32
№ 4	8	20	24	32
№ 5	16	20	24	32

Вычисляя для каждой такой выборки среднее арифметическое, получаем значения:

$$\bar{x}_j : 17 \quad 19 \quad 20 \quad 21 \quad 23$$

Среднее арифметическое из этих выборочных средних равно, конечно, математическому ожиданию:

$$\mu = \frac{17 + 19 + 20 + 21 + 23}{5} = 20.$$

Но такой способ нахождения математического ожидания не имеет никакого смысла, так как проще непосредственно обработать генеральную совокупность. Ведь необходимость в изучении выборки потому и возникает, что, поскольку биологические популяции, как правило, весьма многочисленны, нужно избежать рассмотрения всей генеральной совокупности, не говоря уже о

совокупности всех возможных выборок, число которых обычно несравненно больше, чем число самих членов генеральной совокупности. Достаточно сказать, что даже при $N = 20$ число возможных выборок объемом $n = 5$ составляет:

$$C_{20}^{5} = \frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 15\,504,$$

а при $N = 100$ можно составить 17310309456440 различных выборок по $n = 10$ вариант в каждой; понятно, что при обычных для биологических популяций значениях N число возможных выборок совершенно необозримо.

В связи с этим возникает вопрос: можно ли по результатам одной лишь из такого огромного числа выборок судить о свойствах всей генеральной совокупности? На первый взгляд кажется, что это невозможно, так как приведенный выше пример показал, что среднее значение $\bar{x}$, полученное по одной выборке, не совпадает, как правило, с математическим ожиданием μ. Однако можно показать, что чем больше объем выборки, тем меньше вероятность того, что $\bar{x}$ будет значительно отличаться от μ. Это утверждение имеет следующий смысл. Когда число возможных несовпадающих выборок велико (как это и бывает в реальных условиях), то выборочные средние образуют некоторое практически непрерывное статистическое распределение; это распределение таково, что значения $\bar{x}$ концентрируются в основном около μ, причем эта концентрация тем теснее, чем больше были объемы выборок. Наличие концентрации значений $\bar{x}$ означает, что распределение величин $\bar{x}$ имеет вид одновершинной кривой — с максимальной частотой посередине и убыванием частот к краям распределения. Это можно легко объяснить тем, что каждое из крайних значений $\bar{x}$ (самое малое и самое большое) может получиться только в одной выборке, включающей либо n самых малых вариант, либо n самых больших, в то время как средние по величине значения $\bar{x}$ могут получиться многими способами. Например, пусть имеется генеральная совокупность из 11 вариант со значениями 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Будем образовывать выборки из $n = 2$ вариант. Очевидно, самое малое $\bar{x}$ будет иметь выборка (0; 1) — для нее $\bar{x} = 0{,}5$, причем ясно, что другой выборки с таким же $\bar{x}$ образовать нельзя. Существует также лишь одна выборка с наибольшим $\bar{x} = 9{,}5$ — это выборка (9; 10). Но выборок с $\bar{x} = 5$ можно образовать несколько: (4; 6), (3; 7), (2; 8), (1; 9) и (0; 10); выборок же с $\bar{x} = 2$ только две: (1; 3) и (0; 4). Это связано с тем, что чем ближе к краю генеральной совокупности, тем больше ограничены возможности «раздвигания» вариант в паре для сохранения заданного $\bar{x}$.

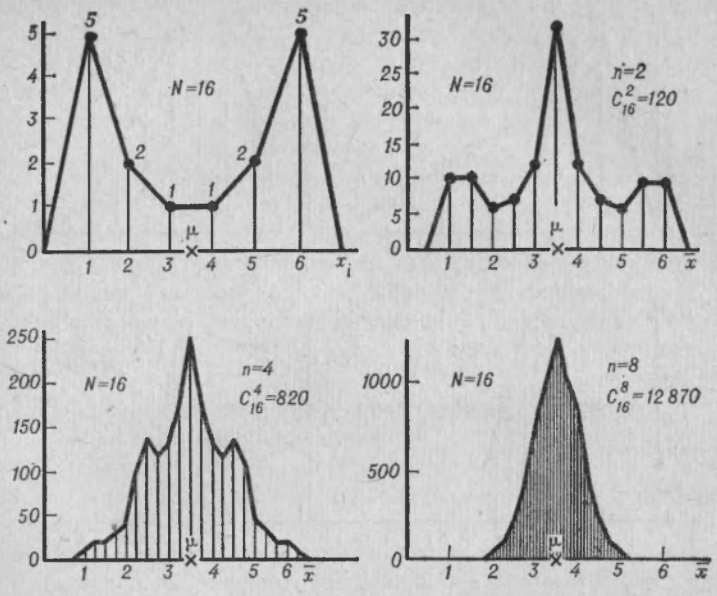


Рис. 3.2. Двухвершинное распределение вариант в генеральной совокупности и три распределения выборочных средних значений.

Характерно, что при достаточно большом n распределение выборочных средних $\bar{x}$ оказывается одновершинным даже в том случае, когда распределение вариант в генеральной совокупности имеет в середине провал. На рис. 3.2 изображены распределение вариант в генеральной совокупности объемом $N = 16$ и три распределения выборочных средних, полученных из выборок с $n = 2$, $n = 4$ и $n = 8$.

Так же можно убедиться, что распределение выборочных средних становится при больших n симметричным, если даже распределение вариант в генеральной совокупности явно асимметрично. Это показано на рис. 3.3.

Центром распределения выборочных средних значений $\bar{x}$ является математическое ожидание μ. Но отсюда следует важный вывод: хотя выборочное среднее значение $\bar{x}$, полученное по результатам одной только выборки, и не равно математическому ожиданию генеральной совокупности, оно все же указывает значение, вблизи которого находится μ. Поэтому выборочное среднее значение $\bar{x}$ называют *оценкой* математического ожидания μ.

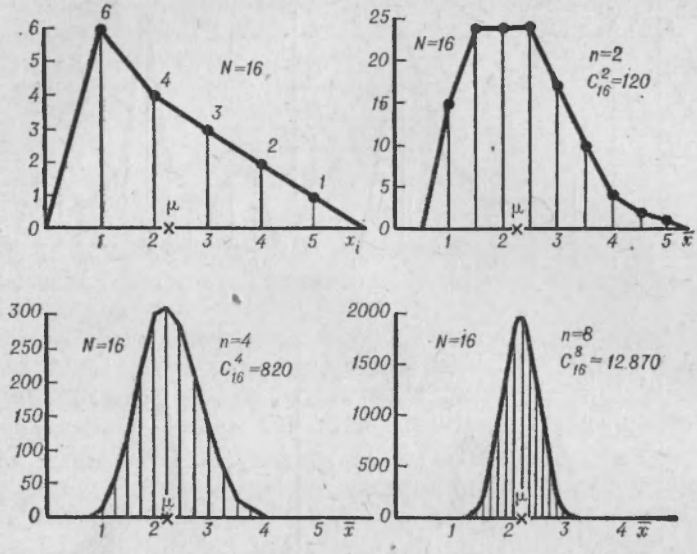


Рис. **3.3.** Ассиметричное распределение вариант в генеральной совокупности и три распределения выборочных средних значений.

Среднее значение выборки будет неправильной оценкой математического ожидания генеральной совокупности, если в выборку почему-либо попали элементы, не принадлежащие к этой генеральной совокупности, или просто «артефакты». О том, как избавиться от таких искажений, сказано в § 2.6. Вообще же о правильном составлении выборок см. § 1.2. Иногда имеется возможность оценить μ *по данным нескольких выборок из изучаемой генеральной совокупности; об этом см. § 3.6.*

Однако информация о математическом ожидании μ, доставляемая средним значением $\bar{x}$, весьма неполна. В самом деле, утверждение, что μ находится вблизи от $\bar{x}$, остается довольно неопределенным, пока не уточнено понятие «вблизи». Необходимость такого уточнения приводит к понятиям «стандартной ошибки среднего значения» и «доверительного интервала для математического ожидания», которые обсуждаются в § 3.5 и 3.7.

При статистической обработке результатов эксперимента или наблюдений указание доверительного интервала для параметров является обязательным. Поэтому, если Вы имеете дело с непрерывным (еще лучше — нормальным) распределением, не забудьте обратиться к § 3.5 и 3.7.

Величина стандартной ошибки среднего значения (и связанного с ней доверительного интервала для математического ожидания) зависит от рассеяния вариант в распределении. К описанию характеристик этого рассеяния мы и переходим.

§ 3.3. Характеристики рассеяния вариант. Дисперсия и коэффициент вариации

3.3.1. Сравним изображения распределения длины красных бобов для двух совокупностей одинакового объема. В первом случае (рис. 3.4, *а*) для посева использовались семена генетически чистой линии, а во второй (рис. 3.4, *б*) — рядовые семена. Несмотря на то что среднее значение в обоих случаях одинаково, различие между этими двумя кривыми очевидно. Это различие состоит в том, что в случае *а* большинство вариант тесно группируется вблизи середины распределения, а в случае *б* довольно большое число вариант отклоняется сравнительно далеко от середины. Принято говорить, что в первом случае *рассеяние* вариант мало, а во втором случае оно велико.

Нашей ближайшей задачей будет получить количественную характеристику этого рассеяния. Естественно характеризовать рассеяние вариант около середины распределения при помощи величины, полученной усреднением всех их отклонений от μ. Однако эта величина не может быть средним арифметическим из отклонений, так как, согласно (3.2), это среднее арифметическое всегда тождественно равно нулю. Это является следствием того, что положительные и отрицательные отклонения в общем взаимно компенсируются.

Поэтому вычисление среднего отклонения должно производиться таким образом, чтобы избежать указанной компенсации поло-

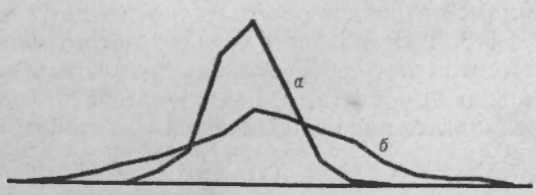


Рис. 3.4. Распределения длины красных бобов: *а* — генетически чистая линия; *б* — рядовые семена.

жительных и отрицательных отклонений. Это в свою очередь можно сделать разными способами.

Можно прежде всего отвлечься от знака отклонения, учитывая величину отклонения независимо от того, в какую сторону оно произошло. Принято говорить, что в этом случае отклонения берутся *по модулю* или *по абсолютной величине* (фактически это означает, что все отклонения считаются положительными); когда величина a берется по абсолютной величине (т. е. без учета ее знака), то это обозначают так: $|a|$. Таким образом, мы можем определить среднее отклонение из выражения:

$$\langle |\xi| \rangle = \sum_{i=1}^{k} p_i |\xi_i|. \tag{3.16}$$

Величина $\langle |\xi| \rangle$ называется *средним абсолютным отклонением*.

Можно избежать компенсации отрицательных и положительных отклонений другим способом, беря не абсолютные значения, а квадраты отклонений (потому что при возведении в квадрат как положительные, так и отрицательные числа дают положительные числа). Тогда мы в результате усреднения получим средний квадрат отклонения:

$$\langle \xi^2 \rangle = \sum_{i=1}^{k} p_i \xi_i^2 = \sum_{i=1}^{k} p_i (x_i - \mu)^2, \tag{3.17}$$

который называют *дисперсией*. Средним отклонением в этом случае следует считать, очевидно, квадратный корень из среднего квадрата отклонения, т. е. величину:

$$\sigma = \sqrt{\langle \xi^2 \rangle} = \sqrt{\Sigma p_i (x_i - \mu)^2}; \tag{3.18}$$

ее называют *средним квадратическим отклонением*, или *стандартным отклонением*, а иногда просто *стандартом* распределения. Величину σ будем считать положительной (т. е. берется всегда положительное значение квадратного корня).

Несмотря на то что вычисление σ более сложно, чем вычисление $\langle |\xi| \rangle$, в качестве характеристики рассеяния используется обычно именно σ. Это связано с тем, что среднее квадратическое отклонение имеет ряд свойств, делающих очень удобным оперирование с ним; подробнее об этом будет сказано ниже.

3.3.2. Вычисление стандартного отклонения по формуле (3.18) представляет собой несколько трудоемкую задачу. Однако расчет можно упростить, если использовать формулу, получающуюся в результате ряда элементарных преобразований. Именно, так как

$$(x_i - \mu)^2 = x_i^2 - 2x_i\mu + \mu^2,$$

то

$$\sigma^2 = \Sigma p_i (x_i - \mu)^2 = \Sigma p_i x_i^2 - 2\Sigma p_i x_i \mu + \Sigma p_i \mu^2.$$

Но μ и μ^2 не зависят от индекса i и поэтому их можно выносить за знак суммы, тогда:

$$2\Sigma p_i x_i \mu = 2\mu \Sigma p_i x_i = 2\mu \cdot \mu = 2\mu^2;$$

$$\Sigma p_i \mu^2 = \mu^2 \Sigma p_i = \mu^2.$$

Поэтому окончательно:

$$\sigma^2 = \Sigma p_i (x_i - \mu)^2 = \Sigma p_i x_i^2 - \mu^2. \qquad (3.19)$$

При использовании выборки математическое ожидание μ заменяется средним значением $\overline{x}$, а вероятности p_i — относительными частотами.

Применим теперь эту формулу к вычислению стандартного отклонения, используя условную шкалу (*) из раздела 3.2.2. Имеем:

$$\frac{1}{n}\Sigma n_i x_i^2 = \frac{1}{100}(1 \cdot 25 + 4 \cdot 16 + 7 \cdot 9 + 11 \cdot 4 + 16 \cdot 1 + 14 \cdot 1 +$$

$$+ 8 \cdot 4 + 6 \cdot 9 + 2 \cdot 16 + 1 \cdot 25) = 3{,}69;$$

так как $(\overline{x})^2 = (-0{,}19)^2 \approx 0{,}04$, то:

$$\sigma^2 = 3{,}69 - 0{,}04 = 3{,}65; \quad \sigma = \sqrt{3{,}65} = 1{,}91$$

в единицах $l = 0{,}05$ мм, т. е.:

$$\sigma = 1{,}91 \cdot 0{,}05 \approx 0{,}095 \approx 0{,}1 \text{ мм}.$$

3.3.3. В разделах 3.1.3 и 3.1.4 были даны формулы для математических ожиданий функций от варьирующих величин:

$$\langle \alpha x \rangle = \alpha \langle x \rangle; \qquad (3.5)$$

$$\langle x + a \rangle = \langle x \rangle + a; \qquad (3.4)$$

$$\langle x - y \rangle = \langle x \rangle - \langle y \rangle \qquad (3.12)$$

$$\langle x + y \rangle = \langle x \rangle + \langle y \rangle. \qquad (3.12')$$

Теперь найдем выражения для соответствующих дисперсий:

$$\sigma^2\{\alpha x\}, \quad \sigma^2\{x + a\}, \quad \sigma^2\{x - y\} \text{ и } \sigma^2\{x + y\};$$

фигурные скобки означают, что σ^2 рассматривается как функция от всей совокупности значений аргументов.

Имеем по определению:

$$\sigma^2\{\alpha x\} = \Sigma p_i (\alpha x_i - \langle \alpha x \rangle)^2.$$

Учитывая (3.5), можно переписать это в виде:

$$\sigma^2\{\alpha x\} - \Sigma p_i (\alpha x_i - \alpha\mu)^2.$$

Так как α можно вынести за скобки (конечно, возведя в квадрат), а затем и за знак суммы, то мы получаем

$$\sigma^2\{\alpha x\} = \alpha^2 \Sigma p_i (x_i - \mu)^2 = \alpha^2 \sigma^2\{x\}. \qquad (3.20)$$

Далее,

$$\sigma^2\{x + a\} = \Sigma p_i [(x_i + a) - (\mu + a)]^2 = \Sigma p_i (x_i - \mu)^2,$$

так что:

$$\sigma^2\{x + a\} = \sigma^2\{x\}. \qquad (3.21)$$

Выполнив аналогичные выкладки, можно показать, что если x и y варьируют независимо, то:

$$\sigma^2\{x - y\} = \sigma^2\{x\} + \sigma^2\{y\}, \qquad (3.22)$$

$$\sigma^2\{x + y\} = \sigma^2\{x\} + \sigma^2\{y\}. \qquad (3.22')$$

Следовательно, при независимом варьировании x и y оказывается:

$$\sigma^2\{x - y\} = \sigma^2\{x + y\}.$$

Рассмотрение простых численных примеров показывает, что если $\sigma\{x\}$ и $\sigma\{y\}$ сильно различаются по величине, то $\sigma\{x \pm y\}$ слабо зависит от меньшей дисперсии и сильно зависит от большей дисперсии. Пусть, например $\sigma\{x\} = 6$ и $\sigma\{y\} = 2$; тогда:

$$\sigma\{x - y\} = \sqrt{6^2 + 2^2} = \sqrt{36 + 4} = \sqrt{40} \approx 6{,}33.$$

Пусть теперь, усовершенствовав методику эксперимента, мы сумеем уменьшить вдвое одно из стандартных отклонений. Если это усовершенствование относится к $\sigma\{y\}$, то $\sigma\{x - y\}$ почти не изменится:

$$\sigma'\{x - y\} = \sqrt{6^2 + 1^2} = \sqrt{36 + 1} = \sqrt{37} \approx 6{,}08.$$

Если же вдвое уменьшится $\sigma\{x\}$, то получится:

$$\sigma''\{x - y\} = \sqrt{3^2 + 2^2} = \sqrt{9 + 4} = \sqrt{13} \approx 3{,}61.$$

Следовательно, при планировании и выполнении эксперимента надо стремиться к уменьшению большей дисперсии.

3.3.4. Значение σ не всегда достаточно полно характеризует вариабельность рассматриваемой величины. В самом деле, для

зерен пшеницы (средняя длина 5,4 мм) стандартное отклонение $\sigma = 1{,}8$ мм означало бы наличие весьма значительного разброса вариант, в то время как для огурцов со средней длиной 129 мм то же значение $\sigma = 1{,}8$ мм указывало бы на высокую степень одинаковости этих огурцов в отношении их длины. Поэтому вводят понятие относительного среднего отклонения:

$$v = \frac{\sigma}{\mu}; \qquad (3.23)$$

эту величину, выраженную в процентах, называют *коэффициентом вариации*. В приведенных примерах:

$$\frac{1{,}8}{5{,}4} = 33{,}3\%; \quad \frac{1{,}8}{129} = 1{,}4\%.$$

Понятие коэффициента вариации полезно еще в том отношении, что позволяет сравнивать вариабельности совокупностей, значения которых имеют различную размерность. Например, не имеет смысла спрашивать, в какой совокупности рассеяние больше — в распределении деревьев по толщине со стандартным отклонением $\sigma = 8{,}6$ см или в распределении клубней картофеля по весу с $\sigma = 14{,}1$ г. Но если мы отнесем эти значения σ к соответствующим значениям средних, то получим безразмерные (или выраженные в процентах) величины, которые уже можно сравнивать между собой. Так, если средняя толщина деревьев равна 28,3 мм, а средний вес клубней картофеля — 94,7 г, то в первом случае:

$$v = \frac{8{,}6}{28{,}3} = 30{,}4\%,$$

а во втором

$$v = \frac{14{,}1}{94{,}7} = 14{,}9\%,$$

т. е. во втором случае относительное рассеяние варианта примерно вдвое меньше, чем в первом.

Однако коэффициент вариации не имеет смысла употреблять для величин, которые могут принимать как положительные, так и отрицательные значения. Например, совершенно бессмысленно вычислять коэффициент вариации для колебаний среднесуточных температур (в пределах, скажем, от $-8°$ до $+11°$) при среднемесячной температуре $+1°$. В данном случае величина σ более адекватно отразит характер явления.

§ 3.4. Выборочная оценка дисперсии, стандартного отклонения и коэффициента ассимметрии

3.4.1. Ранее (см. раздел **3.2.3**) мы нашли, что $\bar{x}$ может служить оценкой для μ. Можно ли утверждать, что и выборочная дисперсия $\frac{1}{n}\Sigma n_i\,(x_i - \bar{x})^2$ может служить оценкой генеральной дисперсии σ^2?

Нетрудно убедиться, что среднее значение из выборочных дисперсий не совпадает с генеральным σ^2, т. е. что при вычислении выборочных дисперсий проявляется не только случайный разброс, но имеет место также систематическая ошибка. Это связано со следующим обстоятельством. Как было указано в разделе **3.1.2**, среднее значение любой совокупности обладает тем свойством, что сумма квадратов отклонений вариант от этого среднего меньше, чем сумма квадратов отклонений от любой другой величины. Иными словами, величина $\Sigma_i n_i(x_i - x)^2$ имеет наименьшее значение в том случае, если в качестве величины x взято среднее значение совокупности. Значит, для каждой из возможных выборок сумма квадратов отклонений вариант x_i от своего выборочного среднего $\bar{x}$ меньше, чем сумма квадратов отклонений вариант x_i от любого другого значения x — значит, в том числе и от математического ожидания μ. Следовательно, при вычислении дисперсии на основании выборки по обычной формуле (3.17)

$$\sigma^2_{\text{выб}} = \frac{1}{n}\Sigma n_i\,(x_i - \bar{x})^2$$

мы всегда получаем заниженную оценку.

Это можно подтвердить следующим простым расчетом. Перепишем величину $\Sigma n_i(x_i - \bar{x})^2$ в виде:

$$\Sigma n_i\left[(x_i - \mu) - (\bar{x} - \mu)\right]^2$$

(прибавив и вычтя в скобках μ). Проведя простые выкладки, получим:

$$\Sigma n_i\,(x_i - \bar{x})^2 = \Sigma n_i\,(x_i - \mu)^2 - n\,(\bar{x} - \mu)^2, \qquad (*)$$

откуда сразу видно, что

$$\Sigma n_i\,(x_i - \bar{x})^2 < \Sigma n_i\,(x_i - \mu)^2.$$

Если произвести усреднение всех членов равенства (*) по всем возможным выборкам, то получится:

$$\left\langle \frac{1}{n}\Sigma n_i\,(x_i - \bar{x})^2 \right\rangle = \left\langle \frac{1}{n}\Sigma n_i\,(x_i - \mu)^2 \right\rangle - \left\langle (\bar{x} - \mu)^2 \right\rangle. \qquad (**)$$

Первый член правой части этого равенства есть, по определению, дисперсия генеральной совокупности, т. е. σ^2. Что касается второго члена, то он представляет собой дисперсию выборочных средних значений в их распределении около математического ожидания генеральной совокупности. Но эта величина, как доказывается в § 3.5 (см. раздел **3.5.1**), равна σ^2/n. Поэтому равенство (**) можно переписать в виде:

$$\left\langle \frac{1}{n} \sum n_i (x_i - \bar{x})^2 \right\rangle = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2$$

или:

$$\sigma^2 = \left\langle \frac{1}{n-1} \sum n_i (x_i - \bar{x})^2 \right\rangle.$$

Отсюда следует, что усреднение по всем выборкам даст правильное значение генеральной дисперсии σ^2 лишь в том случае, если в качестве выборочной оценки этой дисперсии мы возьмем величину

$$s^2 = \frac{1}{n-1} \sum n_i (x_i - \bar{x})^2, \tag{3.24}$$

а не величину:

$$\sigma^2_{\text{выб}} = \frac{1}{n} \sum n_i (x_i - \bar{x})^2. \tag{3.24'}$$

Принято говорить, что s^2 является *несмещенной* оценкой дисперсии σ^2 (в то время как $\sigma^2_{\text{выб}}$ будет *смещенной* оценкой, так как после усреднения по всем выборкам мы получим смещенное значение генеральной дисперсии $\frac{n-1}{n}\sigma^2$).

Можно сказать, что для того, чтобы устранить упомянутую выше систематическую ошибку, нужно при вычислении выборочной оценки дисперсии делить сумму квадратов отклонений $\sum n_i (x_i - \bar{x})^2$ не на число всех отклонений n, а на число отклонений, являющихся н е з а в и с и м ы м и. Дело в том, что отклонения связаны условием:

$$\sum n_i (x_i - \bar{x}) = 0, \tag{***}$$

так что независимо могут быть заданы только $n - 1$ отклонений, а n-ое отклонение должно по необходимости быть таким, чтобы выполнялось условие (***). Число независимых величин, участвующих в образовании той или иной статистики, называется *числом степеней свободы* этой статистики и обозначается через f. Оно равно общему числу величин, по которым вычисляется статистика, минус число условий, связывающих эти величины. Дисперсия выборки вычисляется по n отклонениям, связанным одним условием

(***), так что число степеней свободы выборочной дисперсии равно $f = n - 1$; среднее значение вычисляется по n вариантам, не связанным какими-либо условиями, а поэтому число степеней свободы среднего значения есть $f = n$.

Таким образом, несмещенную оценку (т. е. отклоняющуюся от соответствующего параметра генеральной совокупности лишь случайно, но не систематически) среднего значения надо вычислять по формуле:

$$\bar{x} = \frac{1}{f} \Sigma n_i x_i = \frac{1}{n} \Sigma n_i x_i,$$

а несмещенную оценку дисперсии — по формуле

$$s^2 = \frac{1}{f} \Sigma n_i (x_i - \bar{x})^2 = \frac{1}{n-1} \Sigma n_i (x_i - \bar{x})^2.$$

В пояснение этих формул можно привести еще следующее соображение. Представим себе, что получено только одно значение x (т. е. $n = 1$), так что нам приходится оценивать параметры генеральной совокупности только по одной варианте. Тогда в качестве оценки среднего значения генеральной совокупности мы, естественно, примем величину $\bar{x} = \frac{x}{1} = x$. Что же касается дисперсии вариант в генеральной совокупности, то мы о ней ничего не можем знать, если имеется только одно значение x. Но такой же ответ дает и формула (3.24):

$$s^2 = \frac{1}{1-1} (x - x)^2 = \frac{0}{0} = \text{неопределенность}.$$

Между тем по формуле (3.24′) мы бы имели:

$$\sigma^2_{\text{выб}} = \frac{1}{1} (x - x)^2 = 0,$$

т. е. заключение об отсутствии дисперсии в генеральной совокупности, что явно неверно.

Однако если математическое ожидание генеральной совокупности известно, то полученное одно значение x уже позволяет судить о рассеянии вариант. Этому отвечает и написанная выше формула. Действительно, в этом случае $f = n$, так что при $n = 1$:

$$s^2 = \frac{1}{1} (x - \mu)^2 = (x - \mu)^2, \quad s = |x - \mu|.$$

Пример 3.1. Для проверки качества pH-метра были проведены измерения восьми образцов воды (бидистиллята). Результаты реализованного эксперимента приведены в табл. 3.2.

ТАБЛИЦА 3.2

x_i	7,24	7,03	6,88	7,15	6,69	6,92	6,74	7,19	55,84
ξ_i	0,24	0,03	—0,12	0,15	—0,31	—0,08	—0,26	0,19	—0,16
ξ_i^2	0,0576	0,0009	0,0144	0,0225	0,0961	0,0064	0,0676	0,0361	0,3016

(в последнем столбце записаны строчные суммы). Так как математическое ожидание $\mu = 7{,}00$ здесь известно, то несмещенной оценкой генеральной дисперсии σ^2 будет величина:

$$s^2 = \frac{0{,}3016}{8} = 0.0377,$$

чему соответствует оценка стандартного отклонения

$$s = \sqrt{0{,}0378} = 0{,}194.$$

Если бы не было известно, что измерялся стандартный образец, то расчет был бы иным; применяя формулы (3.19) и (3.24), мы бы получили:

$$s^2 = \frac{0{,}3016 - \dfrac{(-0{,}16)^2}{8}}{7} = \frac{0{,}2984}{7} = 0{,}0426, \quad s = 0{,}206.$$

Из (3.24) имеем оценку стандартного отклонения:

$$s = \sqrt{\frac{\Sigma n_i (x_i - \bar{x})^2}{n-1}}. \qquad (3.25)$$

Величины s^2 и s будут неправильными оценками соответствующих параметров, если в выборку попали «посторонние» элементы. Об исключении таких «артефактов» сказано в § 2.6. О стандартных ошибках оценок s^2 и s см. раздел **3.5.2**, *а о доверительном интервале для σ см. раздел* **3.7.3.**

3.4.2. Подобно тому, как $\sigma^2_{\text{выб}}$ есть смещенная оценка генеральной дисперсии σ^2, величина $\rho_3 = \langle \xi^3 \rangle / \sigma^3$, вычисленная по данным выборки, также представляет собой смещенную оценку соответствующего генерального параметра. Для получения несмещенной оценки коэффициента асимметрии нужно ввести надлежащую поправку, что приводит к формуле:

$$A = \frac{\sqrt{n(n-1)}}{n-2}\,\rho_3. \qquad (3.26)$$

О том, что такое коэффициент асимметрии, говорится в § 2.4. О стандартной ошибке и о доверительном интервале для ρ^3 *см. разделы* **3.5.2.** *и* **2.5.1.**

§ 3.5. Стандартные ошибки статистик непрерывного распределения

3.5.1. Как было сказано в разделе **3.2.3**, все возможные выборочные средние значения образуют некоторое распределение, центром которого является математическое ожидание генеральной совокупности. Упоминалось также, что концентрация значений $\bar{x}$ около μ тем теснее (т. е. их рассеяние тем меньше), чем больше были объемы выборок. Имея теперь объективный и удобный показатель концентрации (или рассеяния) в распределении, а именно дисперсию или стандартное отклонение, разберем вопрос о распределении значений $\bar{x}$ около μ более подробно.

Дисперсию выборочных средних относительно центра их распределения, т. е. относительно μ, можно найти следующим образом. Пусть мы имеем выборку:

$$x_1, x_2, \ldots, x_n$$

из n независимых вариант. Поскольку

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

то

$$n\bar{x} = x_1 + x_2 + \cdots + x_n,$$

так что

$$\sigma^2\{n\bar{x}\} = \sigma^2\{x_1 + x_2 + \cdots + x_n\}. \qquad (*)$$

Но согласно (3.20),

$$\sigma^2\{n\bar{x}\} = n^2\sigma^2\{\bar{x}\}, \qquad (**)$$

а согласно (3.22),

$$\sigma^2\{x_1 + x_2 + \cdots + x_n\} = \sigma^2\{x_1\} + \sigma^2\{x_2\} + \cdots + \sigma^2\{x_n\},$$

так как все $\sigma^2\{x_i\}$ одинаковы, то можно написать:

$$\sigma^2\{x_1 + x_2 + \cdots + x_n\} = n\sigma^2\{x\}. \qquad (***)$$

Заменив левую и правую части в (*) в соответствии с (**) и (***), получим:

$$n^2\sigma^2\{\bar{x}\} = n\sigma^2\{x\},$$

откуда:

$$\sigma^2\{\bar{x}\} = \frac{\sigma^2\{x\}}{n}. \qquad (3.27)$$

Это можно переписать в виде:

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

Величину $\sigma\{\bar{x}\}$ можно назвать *стандартным отклонением среднего значения*; обычно ее обозначают через $\sigma_{\bar{x}}$, так что имеем:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}; \qquad (3.28)$$

в биологической и медицинской литературе величину $\sigma_{\bar{x}}$ иногда обозначают буквой m.

Уменьшение $\sigma_{\bar{x}}$ при увеличении объема выборки можно себе представить наглядно следующим образом. Пусть мы имеем выборку объема n, а затем начинаем прибавлять к ней новые варианты, вычисляя каждый раз заново $\bar{x}$. Очевидно, размах колебаний отдельных значений будет в среднем оставаться на постоянном уровне, а колебания в значениях $\bar{x}$ будут затухать по мере увеличения n. Это объясняется тем, что чем больше число вариант участвовало в образовании выборочного среднего, тем меньше прибавление еще одной варианты может его сместить (поскольку $\bar{x}$ определяется как средневзвешенное).

Очевидно, если $n = 1$, т. е. «выборками» являются отдельные варианты, то $\sigma_{\bar{x}} = \sigma$, как это и должно быть.

Величину $\sigma_{\bar{x}}$ принято называть также *стандартной ошибкой среднего значения*, так как она характеризует ошибку, которая в среднем допускается, когда рассматривают $\bar{x}$ в качестве μ.

Если желательно, чтобы стандартная ошибка среднего значения не превышала определенной заданной величины, то нужно до начала исследования спланировать соответствующим образом объем выборки. О том, как это сделать, сказано в § 1.3.

Обычно величина σ неизвестна и также оценивается по выборке, причем используется несмещенная оценка:

$$s = \sqrt{\frac{\Sigma n_i (x_i - \bar{x})^2}{n-1}}$$

(см. раздел **3.4.1**). Тогда:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{\Sigma n_i (x_i - \bar{x})^2}{n(n-1)}}. \qquad (3.29)$$

3.5.2. Приведем (без вывода) формулы, по которым вычисляются стандартные ошибки несмещенных выборочных оценок дисперсии, стандартного отклонения, коэффициента вариации и коэффициента асимметрии:

$$\left.\begin{aligned} s_{s^2} &= s^2\sqrt{\frac{2}{n}}, \qquad s_s = \frac{s}{\sqrt{2n}}, \\ s_v &= \frac{v}{\sqrt{n-1}}\sqrt{\frac{1}{2}+\left(\frac{v}{100}\right)^2}, \qquad s_A = \sqrt{\frac{6}{n+3}}. \end{aligned}\right\} \qquad (3.30)$$

Эти формулы — приближенные, вернее асимптотические для больших n. Уточнение стандартной ошибки оценки стандартного отклонения см. в разделе **3.7.3**, а коэффициента асимметрии — в разделе **2.5.1.**

3.5.3. Часто величина, интересующая исследователя, является некоторой функцией от величины (или ряда величин), непосредственно определяемой в опыте. В § 3.1 и 3.3 было показано, как вычисляются математическое ожидание и дисперсия для некоторых функций от случайных величин: $z = f(x)$, $z = f(x, y)$.

Поскольку обычно изучается выборка, то получаемое значение $\bar{x}$ есть лишь оценка математического ожидания μ, причем рассеяние выборочных $\bar{x}$ относительно μ характеризуется величиной $\sigma_{\bar{x}}$. Так же значения[1] $\tilde{z} = f(\bar{x})$ рассеяны вокруг генерального среднего $\hat{z}$ с некоторым стандартным отклонением $\sigma_{\tilde{z}}$.

Оценку $s_{\tilde{z}}$ стандартной ошибки $\sigma_{\tilde{z}}$ можно вычислить по обычной формуле (3.28), подставляя соответствующее значение σ. Так, если $z = ax$, то, согласно (3.20):

$$\sigma\{z\} = \sigma\{ax\} = |a|\,\sigma\{x\},$$

[1] Обозначение $f(\bar{x}) = \bar{z}$ употреблено потому, что в общем случае [когда функция $f(x)$ нелинейна] значение $f(\bar{x})$ не равно $\overline{f(x)} = \bar{z}$ [в частности, $(\bar{x})^2 \neq \overline{x^2}$].

так что:

$$\sigma_{\tilde{z}} = |a|\sigma_{\bar{x}}, \quad s_{\tilde{z}} = |a|s_{\bar{x}}; \tag{3.31}$$

если $z = x + a$, то в соответствии с (3.21) получаем:

$$\sigma_{\tilde{z}} = \sigma_{x+a} = \sigma_{\bar{x}}; \quad s_{\tilde{z}} = s_{\bar{x}}. \tag{3.32}$$

Однако в практике приходится встречаться с более сложными случаями, когда случайная величина входит в показатель степени или находится под знаком логарифма. Тогда имеют место следующие приближенные формулы (которые даем без вывода):

$$\left.\begin{aligned} &\text{если } z = \alpha e^{\beta x}, \text{ то } \sigma_{\tilde{z}} = |\beta|\tilde{z}\sigma_{\bar{x}}; \\ &\text{если } z = \alpha \lg \beta x, \text{ то } \sigma_{\tilde{z}} = \frac{|\alpha|}{\bar{x}}\sigma_{\bar{x}} \end{aligned}\right\} \tag{3.33}$$

(α и β — численные множители);

$$\text{если } z = x^m, \quad \text{то} \quad \sigma_{\tilde{z}} = |m|\bar{x}^{m-1}\sigma_{\bar{x}}. \tag{3.34}$$

3.5.4. Перейдем к случаю, когда z зависит от двух случайных величин. Если $z = x \pm y$, то в соответствии с (3.21) и (3.22) получим (когда x и y варьируют независимо одна от другой):

$$s_{\bar{x}\pm\bar{y}} = \sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}. \tag{3.35}$$

Если $z = xy$ или $z = \frac{x}{y}$, то $s_{\tilde{z}}$ вычисляется по формуле (которую даем без вывода):

$$s_{\tilde{z}} = |z|\sqrt{\left(\frac{s_{\bar{x}}}{\bar{x}}\right)^2 + \left(\frac{s_{\bar{y}}}{\bar{y}}\right)^2}. \tag{3.36}$$

§ 3.6. Объединение выборок

3.6.1. Пусть мы имеем w выборок разного объема $n^{(1)}, n^{(2)}, \ldots, n^{(w)}$ из одной генеральной совокупности. По каждой выборке мы можем найти оценки $\bar{x}^{(1)}, \bar{x}^{(2)}, \ldots, \bar{x}^{(w)}$ для математического ожидания μ и несмещенные оценки $s_1^2, s_2^2, \ldots, s_w^2$ для генеральной дисперсии σ^2. Так как стандартные ошибки оценок генеральных параметров зависят от объема выборки, то, очевидно, мы по-

лучим более точные оценки для μ и σ^2, если объединим все w выборок в одну выборку суммарного объема:

$$n = n^{(1)} + n^{(2)} + \cdots + n^{(w)}$$

и найдем $\bar{x}$ и s^2 для объединенной выборки.

Конечно, это не значит, что мы должны фактически свести все выборки в одну общую совокупность и произвести заново группировку по разрядам и всю дальнейшую обработку. Можно найти $\bar{x}$ и s^2 объединенной выборки, усреднив надлежащим образом частные значения $\bar{x}^{(j)}$ и $s^2_{(j)}$. Можно ожидать, что наилучший результат получится, если при этом усреднении мы припишем каждой выборке тем больший «вес», чем более точную оценку генерального параметра она дает. Но точность оценки, как мы знаем, характеризуется объемом выборки. Поэтому наилучшей оценкой для μ будет:

$$\bar{x} = \frac{n^{(1)}\bar{x}^{(1)} + n^{(2)}\bar{x}^{(2)} + \cdots + n^{(w)}x^{(w)}}{n^{(1)} + n^{(2)} + \cdots + n^{(w)}} = \frac{1}{n}\sum n^{(j)}\bar{x}^{(j)}. \qquad (3.37)$$

Пример 3.2. В табл. 3.3 приведены данные о размерах эритроцитов крови, полученные на четырех мазках от одного человека; в столбцах 2—5 записаны соответствующие частоты. Пользуясь формулой (3.37) найдем усредненное значение $\bar{x}$:

$$\bar{x} = \frac{1}{1549}(392 \cdot 8{,}33 + 364 \cdot 8{,}26 + 345 \cdot 8{,}30 + 448 \cdot 8{,}28) =$$

$$= \frac{1}{1549}(3260 + 3010 + 2860 + 3710) = \frac{12840}{1549} = 8{,}29.$$

ТАБЛИЦА 3.3

Диаметр эритроцита, мк	Первый мазок	Второй мазок	Третий мазок	Четвертый мазок	Сводная выборка
1	*2*	*3*	*4*	*5*	*6*
6	12	9	11	17	49
7	54	48	32	62	196
8	183	204	191	219	797
9	96	66	74	97	333
10	31	24	29	36	120
11	16	13	8	17	54
$n^{(j)}$	392	364	345	448	1549
$\bar{x}^{(j)}$	8,33	8,26	8,30	8,28	8,29
$s^2_{(j)}$	0,850	0,681	0,706	0,861	0,796

3.6.2. При объединении выборок возникает также необходимость найти усредненную оценку дисперсии. Формула имеет вид:

$$s^2 = \frac{1}{n-w}\sum_{j=1}^{w}\left(n^{(j)}-1\right)s^2_{(j)} + \frac{1}{n-1}\sum_{j=1}^{w} n^{(j)}\left(\bar{x}^{(j)}-\bar{x}\right)^2. \quad (3.38)$$

Первый член в формуле (3.38) есть средневзвешенная из выборочных дисперсий; при этом в качестве «веса» для каждой из $s^2_{(j)}$ берется ее число степеней свободы[1] $f^{(j)} = n^{(j)} - 1$. Необходимость второго члена видна из следующего: если бы все частные дисперсии были равны нулю (т. е. в пределах каждой совокупности все варианты были бы одинаковы), но частные средние значения различны, то в первой совокупности имелось бы $n^{(1)}$ отклонений $\bar{x}^{(1)} - \bar{x}$ от общего среднего, во второй совокупности — $n^{(2)}$ отклонений $\bar{x}^{(2)} - \bar{x}$ и т. д., так что вычисление дисперсии по формуле (3.24) дало бы как раз выражение, совпадающее со вторым членом формулы (3.38). Следовательно, смысл этой формулы состоит в том, что дисперсия суммарной совокупности равна средней частной дисперсии плюс дисперсия частных средних. Например, для выборки из табл. 3.3. имеем:

$$s^2 = \frac{1}{1545}(391\cdot 0{,}850 + 363\cdot 0{,}681 + 344\cdot 0{,}706 + 447\cdot 0{,}861) +$$
$$+ \frac{1}{1548}(392\cdot \underset{0{,}04}{0{,}02^2} + 364\cdot \underset{0{,}03}{0{,}05^2} + 345\cdot \underset{0{,}01}{0{,}01^2} + 448\cdot \underset{0{,}01}{0{,}03^2}) =$$
$$= \frac{1207}{1545} + \frac{1{,}5}{1548} = 0{,}781.$$

В данном случае поправка, вносимая вторым членом формулы (3.38), очень мала, так как значения $\bar{x}^{(j)}$ близки между собой.

После того как найдена усредненная оценка s^2, можно получить оценку для $\sigma_{\bar{x}}$ на основании объединенной выборки по формуле:

$$s_{\bar{x}} = s/\sqrt{n}.$$

---

[1] Если объединяемые совокупности рассматриваются не как выборки, а как самостоятельные генеральные совокупности, то при сведении их в одну совокупность число степеней свободы дисперсии для каждой из совокупностей равно $n^{(j)}$, а не $n^{(j)} - 1$. Например, если найдены $\mu^{(j)}$ и $\sigma^2_{(j)}$ для каких-нибудь показателей (скажем, удойности коров) в отдельных хозяйствах, а хотят узнать соответствующие показатели в среднем по району, то совокупность для каждого хозяйства должна считаться генеральной совокупностью, так как мы рассматриваем $\mu^{(j)}$ и $\sigma^2_{(j)}$ как характеристики именно этой совокупности, а не как оценки для всей популяции данного биологического вида.

Для данных из табл. 3.5 имеем:

$$s_{\bar{x}} = \sqrt{\frac{0,784}{1549}} = 0,022.$$

3.6.3. Объединение выборок является обязательной составной частью вычислений, когда пользуются так называемой типической (зональной) выборкой (см. раздел 1.2.2.). В этом случае стандартную ошибку среднего значения вычисляют следующим образом:

а) при произвольных объемах n_j подвыборок — по формуле:

$$\sigma_{\bar{x}} = \sqrt{\sum_j \frac{p_j^2 \sigma_j^2}{n_j}}; \qquad (3.39)$$

б) если объемы подвыборок n_j пропорциональны объемам типических групп (зон), то подстановка в формулу (3.39) значений $n_j = np_j$ дает:

$$\sigma_{\bar{x}} = \sqrt{\frac{\Sigma p_j \sigma_j^2}{n}}. \qquad (3.40)$$

Пример 3.3. Вычислим по указанным формулам значения $\sigma_{\bar{x}}$ для данных из примера 1.2; будем считать, что значения σ_j для частей поля равны:

$$\sigma_j = 2,\ 3,\ 1,\ 4.$$

Учитывая также, что $p_j = 0,1;\ 0,4;\ 0,3;\ 0,2$, имеем:

а) $$\sigma_{\bar{x}} = \sqrt{\frac{0,01 \cdot 4}{10} + \frac{0,16 \cdot 9}{10} + \frac{0,09 \cdot 1}{10} + \frac{0,04 \cdot 16}{10}} = 0,466;$$

б) $$\sigma_{\bar{x}} = \sqrt{\frac{0,1 \cdot 4 + 0,4 \cdot 9 + 0,3 \cdot 1 + 0,2 \cdot 16}{40}} = 0,433.$$

Очевидно, чем меньше для данной выборки значение $\sigma_{\bar{x}}$, тем выборочное среднее ближе к математическому ожиданию. Из рассмотренного примера видно, что способ «б» составления типической выборки лучше, чем способ «а».

Формула (3.38) позволяет понять, почему зональная (типическая) выборка имеет преимущество по сравнению с простой (не зональной) случайной выборкой. Простая случайная выборка эквивалентна зональной выборке со случайным набором чисел n_j. В этом случае формула (3.39) будет давать для $\sigma_{\bar{x}}$, как правило, даже большие значения, чем при равенстве всех n_j. Уменьшение $\sigma_{\bar{x}}$ при зональной выборке объясняется тем, что такое построение

выборки позволяет вычислить полную дисперсию как усредненную внутризональную дисперсию, между тем как в случае простой случайной выборки в полную дисперсию вошла бы также межзональная дисперсия, отражающая дисперсию зональных средних значений.

§ 3.7. Доверительный интервал для математического ожидания и дисперсии нормального распределения

3.7.1. Зная не только выборочное среднее значение, но и его стандартную ошибку, мы получаем более определенное представление о математическом ожидании генеральной совокупности. Это представление можно сделать еще определеннее, если построить *доверительный интервал* для μ, к описанию которого мы и переходим.

Методы построения доверительных интервалов, которые будут описаны ниже, основаны на свойствах так называемого нормального распределения. Оно описано в § 2.3. Советуем предварительно прочесть (или хотя бы просмотреть) § 2.3.

Как было показано в разделе **3.5.1**, выборочные средние значения $\bar{x}$ распределены вокруг μ со стандартным отклонением $\sigma_{\bar{x}}$. Можно также доказать, что при увеличении объема выборок распределение выборочных средних приближается к нормальному независимо от того, как распределены варианты в генеральной совокупности. Действительно, величина выборочной средней зависит от случайного сочетания многих (n) вариант, вклад каждой из которых примерно одинаков и относительно мал (он пропорционален «весу» этой варианты в выборке, т. е. $\frac{1}{n}$); но это как раз условия, ведущие к образованию нормального распределения (см. раздел **6.1.4**), причем эти условия реализуются тем полнее, чем больше число n вариант в выборке.

Из сказанного выше следует, что относительное отклонение выборочного среднего $\bar{x}$ от математического ожидания μ, т. е. величина

$$\tau = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}, \tag{3.41}$$

распределено так же, как относительные отклонения вариант x_i от μ, т. е. величины:

$$u = \frac{x_i - \mu}{\sigma}$$

в нормально распределенной генеральной совокупности.

Поэтому вероятность того, что $\bar{x}$ будет находиться в пределах $\mu \pm \tau\sigma_{\bar{x}}$, можно описывать функцией $\Theta(u)$, значения которой даны в табл. II Приложений,— нужно только вместо u подставлять τ.

Отсюда мы, в частности, получаем, что ~68,3% всех выборочных $\bar{x}$ находятся в пределах $\mu \pm \sigma_{\bar{x}}$; иными словами, имеется вероятность 0,683, что $\bar{x}$ отличается от μ не более чем на $\pm \sigma_{\bar{x}}$. Это имеет следующий смысл: пусть взято 100 выборок объемом n каждая и, следовательно, получено 100 интервалов $(\bar{x} - \sigma_{\bar{x}}, \bar{x} + \sigma_{\bar{x}})$; они все будут несколько различаться между собой положениями своих центров $\bar{x}$, но ~ 68 из этих интервалов покроют μ (или, иными словами, ~68 этих интервалов будут содержать μ).

По этой причине интервал $(\bar{x} - \sigma_{\bar{x}}, \bar{x} + \sigma_{\bar{x}})$ называют *доверительным интервалом для математического ожидания.*

Из сказанного ясно, что указание доверительного интервала заключает в себе некоторое утверждение о математическом ожидании генеральной совокупности, а не о средних значениях других возможных выборок.

Вероятность 68,3% того, что интервал $(\bar{x} - \sigma_{\bar{x}}, \bar{x} + \sigma_{\bar{x}})$ содержит μ, сравнительно невысока. С большей уверенностью можно утверждать, что μ покрывается интервалом $(\bar{x} - 2\sigma_{\bar{x}}, \bar{x} + 2\sigma_{\bar{x}})$, так как вероятность этого равна 95,5%; в том, что μ содержится в интервале $(\bar{x} - 3\sigma_{\bar{x}}, \bar{x} + 3\sigma_{\bar{x}})$, можно быть почти уверенным: вероятность этого равна 99,7%.

Чем выше требование к вероятности вывода об интервале, содержащем μ, тем шире должен быть интервал, могущий обеспечить такую вероятность. Уровень этой вероятности принято называть *доверительной вероятностью*, или *доверительным уровнем*; выбор той или иной доверительной вероятности определяет ширину доверительного интервала, который с принятой вероятностью содержит μ. Так, при доверительной вероятности $P = 95{,}5\%$ границы доверительного интервала для μ составляют $\bar{x} \pm 2\sigma_{\bar{x}}$, при доверительной вероятности $P = 99{,}7\%$ границами доверительного интервала будут $\bar{x} \pm 3\sigma_{\bar{x}}$ и т. д.

Обычно для доверительной вероятности выбирают более «круглые» числа: 95% или 99% (иногда также 99,9%). Тогда границы доверительных интервалов будут соответственно равны $\bar{x} \pm 1{,}96\ \sigma_{\bar{x}}$ и $\bar{x} \pm 2{,}58\ \sigma_{\bar{x}}$; числа 1,96 и 2,58 найдены по табл. II Приложений: это те значения $\tau = u$, при которых $\Theta(u)$ имеет заданные значения 0,95 и 0,99 (рис. 3.5).

В биологических применениях 95%-ный доверительный интервал считается достаточно надежным. Однако при исследованиях, уточняющих предыдущие результаты, а также в случаях,

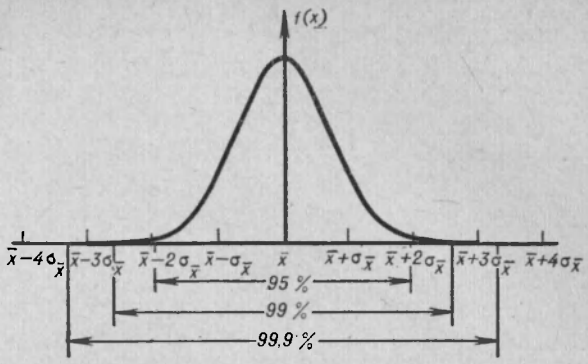


Рис. 3.5. Доверительные интервалы для математического ожидания при нормальном распределении.

когда от результатов исследования зависит принятие или непринятие решения о затрате средств и труда (применение добавочных удобрений или подкормок, реорганизация производства и др.), целесообразно применять более высокий доверительный уровень $P = 99\%$ или даже $P = 99{,}9\%$.

3.7.2. Очевидно, относительное отклонение выборочного среднего от математического ожидания, т. е.

$$\tau = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}},$$

будет оцениваться величиной:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}, \qquad (3.42)$$

где $s_{\bar{x}}$ определяется формулой (3.29). При больших объемах выборок величины t, как и величины τ, распределены нормально. Это значит, что доля интервалов $(\bar{x} - ts_{\bar{x}}, \bar{x} + ts_{\bar{x}})$, покрывающих μ, дается функцией $\Theta(t)$. Если же объемы выборок малы, то распределение величины t отличается от нормального и тем сильнее, чем меньше объем выборок. Тогда доля интервалов $(\bar{x} - ts_{\bar{x}}$ $\bar{x} + ts_{\bar{x}})$, покрывающих μ, определяется не функцией $\Theta(t)$, а некоторой другой функцией $\Theta^*(t)$. Вид этой функции зависит как от объема выборок n, так и от характера распределения вариант в генеральной совокупности.

Распределение величин t при разных n (или при разном числе степеней свободы $f = n - 1$) для случая, когда варианты в генеральной совокупности распределены нормально, было найдено Стьюдентом (псевдоним В. Госсета). Графики этого *распределения Стьюдента* для $f = 1$ и $f = 5$ приведены на рис. 3.6; там же для сравнения изображена нормальная кривая, являющаяся предельной (при $f = \infty$) для кривых Стьюдента.

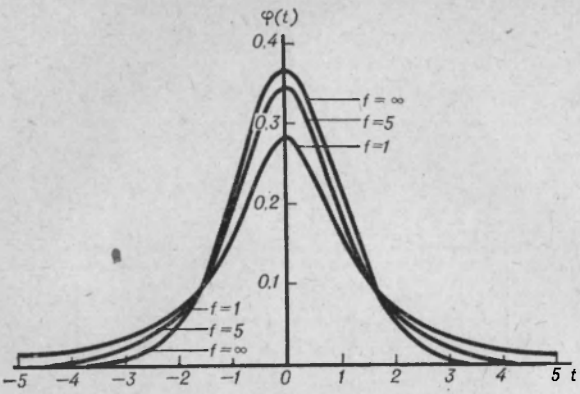


Рис. 3.6. Распределения Стьюдента при разном числе степеней свободы.

Анализ распределения Стьюдента показывает, что при малых f значения $\Theta^*(t)$ меньше, чем $\Theta(t)$, поэтому доверительные интервалы для принятых доверительных уровней 95% и 99% должны быть шире, чем получившиеся ранее на основе нормального распределения $\bar{x} \pm 1{,}96\, s_{\bar{x}}$ и $\bar{x} \pm 2{,}58\, s_{\bar{x}}$. Значения t, обеспечивающие принятый доверительный уровень, должны теперь зависеть не только от этого уровня, но и от объема совокупности n (или от числа степеней свободы f). Табл. IV Приложений дает значения t_P при доверительных уровнях 95, 99 и 99,9% для разных f.

Пример 3.4. При изучении в 10 опытах образования у собаки условного рефлекса под действием ранее индифферентного раздражителя были получены результаты (время между моментом включения условного раздражителя и моментом начала слюноотделения): $\bar{x} = 8{,}47$ с, $s_{\bar{x}} = 1{,}19$ с. Надо найти 95%-ный доверительный интервал для μ, характеризующего данное животное.

Для $P = 95\%$ и $f = n - 1 = 9$ (число степеней свободы дисперсии) находим в табл. IV Приложений значение $t = 2{,}26$. Поэтому границы доверительного интервала будут:

$$8{,}47 - 2{,}26 \cdot 1{,}19 = 8{,}47 - 2{,}69 = 5{,}78,$$

$$8{,}47 + 2{,}26 \cdot 1{,}19 = 8{,}47 + 2{,}69 = 11{,}16.$$

Результаты записывают обычно в одной из следующих двух форм:

$$5{,}78 \div 11{,}16, \quad \text{или} \quad 8{,}47 + 2{,}69.$$

Из табл. IV Приложений видно, что значения t_P зависят особенно резко от f при малых f. Поэтому увеличение малых n приводит к сужению доверительного интервала (определяемого величиной $t_P s_{\bar{x}} = \frac{t_P s}{\sqrt{n}}$ не только за счет уменьшения множителя $1/\sqrt{n}$, но в еще большей степени за счет уменьшения t_P. Так, при

$P = 95\%$ изменение n с двух опытов до трех уменьшает множитель $t_p/\sqrt{n}$ с $12{,}71/\sqrt{2} = 9{,}0$ до $4{,}30/\sqrt{3} = 2{,}5$, т. е. доверительный интервал сужается в $9{,}0 : 2{,}5 = 3{,}6$ раза; при $P = 99\%$ ширина доверительного интервала уменьшается даже примерно в 8 раз ($63{,}66/\sqrt{2} = 45{,}0$; $9{,}93/\sqrt{3} = 5{,}7$; $45{,}0/5{,}7 = 7{,}9$). При больших значениях n увеличение n на одну единицу сказывается на ширине доверительного интервала гораздо меньше.

Построение доверительного интервала описанным способом законно, если распределение вариант в исходной генеральной совокупности нормально (так как в противном случае $(\bar{x} - \mu)/s_{\bar{x}}$ не подчиняется распределению Стьюдента). Когда имеются сомнения в нормальности исходного распределения, надо воспользоваться критериями, описанными в § 2.5. О построении доверительных интервалов для параметров, играющих роль математического ожидания в случаях распределения Пуассона и альтернативного распределения, см. соответственно § 7.2 и 6.2. Если Вы хотите, чтобы ширина доверительного интервала для математического ожидания не превышала определенной заданной величины, то Вы должны до начала исследования спланировать соответствующим образом объем выборки; о том, как это сделать, сказано в § 1.3.

3.7.3. Иногда приходится определять доверительные интервалы для стандартного отклонения. Проще всего это можно делать, используя формулу (3.30) для s_s и считая распределение величин s приближенно нормальным. Однако такой способ дает более или менее правильные результаты только при достаточно больших n (> 30). Поэтому предпочтительнее находить доверительные интервалы для σ при помощи специальной таблицы, основанной на более точном распределении выборочных s. Эта табл. 3.4 дает значения $q_P'(f)$ и $q_P''(f)$, определяющие границы P%-ного доверительного интервала (предполагая, что распределение вариант в генеральной совокупности нормально):

$$q_P' s < \sigma < q_P'' s. \qquad (3.43)$$

ТАБЛИЦА 3.4

Значения $q'_P(f)$ и $q''_P(f)$ для построения доверительного интервала для стандартного отклонения (из книги Л. Н. Большева и Н. С. Смирнова, с. 297)

f	99%		95%		f	99%		95%	
	q'	q''	q'	q''		q'	q''	q'	q''
3	0,483	6,47	0,566	3,73	21	0,712	1,617	0,769	1,429
4	0,516	4,39	0,599	2,87	22	0,717	1,595	0,773	1,416
5	0,546	3,48	0,624	2,45	23	0,722	1,576	0,777	1,402
6	0,569	2,98	0,644	2,202	24	0,726	1,558	0,781	1,391
7	0,588	2,66	0,661	2,035	25	0,730	1,541	0,784	1,380
8	0,604	2,440	0,675	1,916	26	0,734	1,526	0,788	1,371
9	0,618	2,227	0,688	1,826	27	0,737	1,512	0,791	1,361
10	0,630	2,154	0,699	1,755	28	0,741	1,499	0,794	1,352
11	0,641	2,056	0,708	1,698	29	0,744	1,487	0,796	1,344
12	0,651	1,976	0,717	1,651	30	0,748	1,475	0,799	1,337
13	0,660	1,910	0,725	1,611	40	0,774	1,390	0,821	1,279
14	0,669	1,854	0,732	1,577	50	0,793	1,336	0,837	1,243
15	0,676	1,806	0,739	1,548	60	0,808	1,299	0,849	1,217
16	0,683	1,764	0,745	1,522	70	0,820	1,272	0,858	1,198
17	0,690	1,727	0,750	1,499	80	0,829	1,250	0,866	1,183
18	0,696	1,695	0,756	1,479	90	0,838	1,233	0,873	1,171
19	0,702	1,666	0,760	1,460	100	0,845	1,219	0,878	1,161
20	0,707	1,640	0,765	1,444	200	0,887	1,15	0,912	1,11

Пример 3.5. В примере 3.4 мы имели $s_x^2 = 1{,}19$, $n = 10$. Этому соответствует $s = 1{,}19\sqrt{10} = 3{,}76$. Если нас интересует 95% доверительный интервал для σ, то по табл. 3.4 $q_{95}'(9) = 0{,}688$ и $q_{95}''(9) = 1{,}826$. Тогда границы доверительного интервала для σ будут:

$$3{,}76 \cdot 0{,}688 \approx 2{,}59 \text{ и } 3{,}76 \cdot 1{,}826 \approx 6{,}87.$$

Глава четвертая

ВЫЯВЛЕНИЕ РАЗЛИЧИЯ МЕЖДУ ПАРАМЕТРАМИ ДВУХ НОРМАЛЬНЫХ РАСПРЕДЕЛЕНИЙ

§ 4.1. Понятие статистической значимости различия

4.1.1. Одной из важнейших задач биологического исследования является получение данных о результатах действия внешних факторов на живой объект. В статистическом плане эти результаты часто проявляются как изменение значений параметров соответствующих распределений: увеличение среднего веса животных или средней урожайности культуры при улучшении рациона или применении удобрения, снижение физиологической функции под действием химического или физического агента и т. д. Однако при этом возникает следующее осложнение: как подопытная, так и контрольная группы объектов представляют собой выборки из некоторых генеральных совокупностей, а поэтому средние значения, дисперсии и другие показатели, вычисленные для этих групп, носят на себе отпечаток выборочного варьирования. Это приводит к тому, что, например, средние значения в опытной и контрольной группах будут обязательно различными даже в том случае, если никакого реального действия фактора нет — ведь средние значения двух выборок из одной и той же генеральной совокупности всегда различны. Таким образом, встает задача: в каждом конкретном случае уметь ответить на вопрос, является ли различие между средними значениями опытной и контрольной групп просто различием между средними значениями двух выборок из одной и той же генеральной совокупности или же оно отражает различие между математическими ожиданиями двух различных генеральных совокупностей. Такой же вопрос можно поставить о дисперсиях и других параметрах распределения.

Общий метод решения этой задачи таков. Начинают с предположения о том, что обе эмпирические совокупности являются выборками из одной и той же генеральной совокупности (или из двух одинаково распределенных, т. е. с одинаковыми параметрами, генеральных совокупностей); это предположение называют *нулевой гипотезой* и обозначают H_0. Далее вычисляют вероятность того, что при условии справедливости нулевой гипотезы расхождение между выборочными оценками параметров, связанное только с выборочным варьированием, может достигнуть фактически наблюденной величины; если эта вероятность окажется очень малой, то нулевая гипотеза отвергается (т. е. маловероятно, что

расхождение вызвано случайными причинами, а не реальным различием).

Пример 4.1. Выборочное распределение оказалось асимметричным с коэффициентом асимметрии $A = 0{,}3$. Значит ли это, что и генеральная совокупность, из которой взята эта выборка, также асимметрична?

Нулевая гипотеза будет состоять в том, что генеральная совокупность симметрична ($\rho_3 = 0$), а асимметрия выборочного распределения объясняется случайностью вхождения вариант в выборку. Пусть вычисление показало, что вероятность образования выборки объема n с коэффициентом асимметрии $A = 0{,}3$ или больше из симметричной генеральной совокупности равна $P = 0{,}001$. Тогда мы скажем: не приходится рассчитывать на то, что при извлечении одной выборки получилась именно одна из тех, которые имеют столь малую вероятность; вернее всего нулевая гипотеза неправильна. Если бы указанная вероятность получилась равной, например, $P' = 0{,}15$, то случайное образование выборки с данным значением A нельзя было бы считать невозможным и нулевую гипотезу нельзя было бы отвергнуть.

Предельно допустимое значение вероятности, начиная с которого вероятность можно считать малой, называют *уровнем значимости* — различие считается значимым (т. е. реальным), если вероятность того, что нулевая гипотеза верна, меньше уровня значимости (его обозначают буквой α). Таким образом, если вероятность нулевой гипотезы меньше α, то она отвергается; если же эта вероятность больше α, то нулевая гипотеза принимается. Очевидно, уровень значимости характеризует, в какой мере мы рискуем ошибиться, отвергая нулевую гипотезу.

Значение α можно задавать любое, но обычно выбирают одно из значений $\alpha = 0{,}05 = 5\%$, $\alpha = 0{,}01 = 1\%$ или $\alpha = 0{,}001 = 0{,}1\%$. Выбор того или иного конкретного значения α определяется конкретными задачами исследования. Например, если исследуется новый лечебный препарат и нужно показать, что его побочное действие не опасно для жизни, то даже уровень значимости 0,001 должен считаться слишком высоким. Наоборот, если речь идет об улучшении продуктивности стада за счет недорогого изменения рациона, то достаточно и небольшой уверенности в положительном результате. При этом, разумеется, не исключаются дальнейшие уточнения экспериментальных данных (например, путем постановки дополнительных опытов, последующих наблюдений и т. д.).

4.1.2. Надо всегда иметь в виду, что утверждение о том, что нет достаточных оснований отвергнуть гипотезу об отсутствии различия, совсем не равносильно утверждению, что отсутствие различия доказано. Иными словами, можно лишь утверждать, что данные наблюдений не противоречат предположению

об отсутствии различия, но нельзя утверждать, что эти данные доказывают отсутствие такого различия. Так, в рассмотренном выше примере мы при $P' = 0{,}15$ не отвергли бы нулевую гипотезу. Но если мы вместо того, чтобы сказать, что имеющиеся данные не противоречат предположению о симметричности генеральной совокупности, станем утверждать, что последняя действительно симметричная, то мы рискуем впасть в ошибку: наблюденное значение $A = 0{,}3$ могло получиться и при генеральном значении ρ_3, отличном от нуля.

Такая ошибка, которая допускается, когда не отвергают гипотезу H_0, в действительности неверную, носит название *ошибки II рода*, в отличие от рассмотренной выше *ошибки I рода*, когда отвергают гипотезу H_0, на самом деле верную. Вероятность ошибки *II* рода обозначают β.

4.1.3. Вернемся теперь к примеру 4.1. Вычисление вероятности того, что выборка объема n из симметричной генеральной совокупности будет иметь коэффициент асимметрии $A \geqslant 0{,}3$, довольно сложно. Поэтому имеет смысл составить таблицу, в которой были бы приведены готовые результаты таких вычислений (очевидно, такая таблица должна будет иметь два аргумента — A и n). Однако для решения вопроса о возможности отвергнуть нулевую гипотезу требуется лишь знать, меньше ли эта вероятность, чем $0{,}01 = 1\%$, но не надо знать, насколько она меньше. Поэтому более целесообразно составить такую таблицу, в которой для каждого значения n указывалось бы то значение $A_{0.01}$, которое удовлетворяет условию: вероятность того, что $A > A_{0.01}$, равна $0{,}01 = 1\%$; эта таблица будет иметь только один аргумент: n. Величину $A_{0,01}$ назовем 1%-ным верхним *критическим значением* коэффициента асимметрии (для данного объема выборки n).

Численное значение величины $A_{0,01}$ (при данном n) существенно зависит от того, как формулируется нулевая гипотеза H_0. Часто эта гипотеза просто утверждает, что $\rho_3 = 0$. В этом случае H_0 будет отвергаться тогда, когда выборочный коэффициент асимметрии окажется либо больше $A_{0.01}$, либо меньше $A_{0,01}$ (рис. 4.1, *а*); величина $A_{0.01}$ должна здесь удовлетворять тому условию, что площади, заштрихованные на рис. 4.1, *а*, составляют в сумме 1% от всей площади под кривой распределения выборочных A (при достаточно больших n это распределение нормально).

Но в соответствии с целями исследования нулевая гипотеза может формулироваться более определенно: в одних случаях $\rho_3 \geqslant 0$, в других случаях $\rho_3 \leqslant 0$. Тогда в первом случае H_0 будет отвергаться при $A < -A_{0.01}$, а во втором случае — при $A > A_{0,01}$ (рис. 4.1, *б* и *в*), причем здесь $A_{0,01}$ опять удовлетворяет условию, что заштрихованная площадь на рис. 4.1, *б* и *в* составляет 1% от площади под кривой распределения (на рис. 4.1 мас-

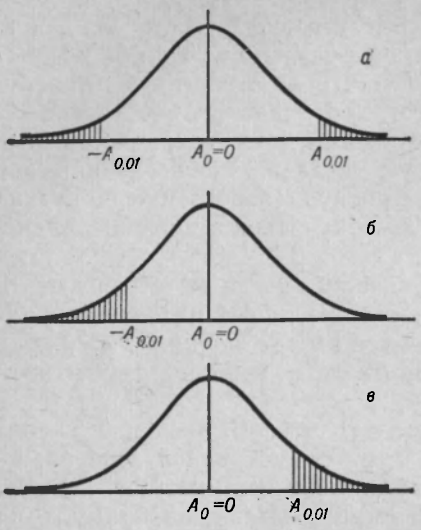


Рис. 4.1. Двусторонняя и односторонние критические области.

штаб для наглядности не соблюден); но ясно, что $A_{0,01}$ из рис. 4.1, *б* и *в* отличается от $A_{0,01}$ из рис. 4.1, *а*.

Критерии, проверяющие нулевую гипотезу типа $\rho_з = 0$, называются *двусторонними*, а критерии, проверяющие нулевую гипотезу типа $\rho_з \geqslant 0$ или $\rho_з \leqslant 0$, называются *односторонними*; происхождение этих названий ясно из рис. 4.1.

Вопрос о значимости коэффициента асимметрии использован здесь просто в качестве примера. Реальное значение этот вопрос имеет при проверке нормальности распределения (см. § 2.5).

§ 4.2. Сравнение двух средних значений (критерий Стьюдента)

4.2.1. Разбираемый в этом параграфе вопрос является одним из основных в биологических приложениях математической статистики.

Пример 4.2. Необходимо выяснить эффективность применения некоторого препарата (или какого-то комплекса мероприятий), имеющего целью повысить сопротивляемость организма животных по отношению к определенной инфекции.

Опыт может быть поставлен так: берут две группы животных (например, мышей) одного пола и возраста — не обязательно одинаковой численности. Мышам одной группы вводят исследуемый препарат, мышам другой группы этот препарат не вводят; первую группу будем называть опытной, вторую группу — кон-

трольной. Затем мышам обеих групп вводят инфекцию и наблюдают, сколько дней переживают мыши опытной и контрольной групп. Пусть при этом получились результаты, приведенные в табл. 4.1. Прежде всего очевидно, что как в опытной, так и в контрольной группе надо было использовать именно группу животных, а не по одному животному: ведь некоторые из мышей контрольной группы пережили некоторых мышей из опытной группы.

ТАБЛИЦА 4.1

Число дней	3	4	5	6	7	8	9	n	$\bar{x}$	s	$s_{\bar{x}}$
Опыт	1	1	6	11	8	4	1	32	6,25	1,25	0,22
Контроль	1	4	9	7	2			23	5,22	0,97	0,20

Из табл. 4.1 видно, что средние значения для опытной и контрольной групп не совпадают. Однако это еще не дает основания считать доказанной эффективность препарата. В самом деле, ведь каждая из групп животных представляет собой лишь случайную выборку из генеральной совокупности. Если бы мы взяли другую контрольную группу (т. е. другую случайную выборку из той же генеральной совокупности) животных, не подвергшихся действию препарата (т. е. незащищенных) мышей, то получилось бы другое среднее значение. Так можно ли считать исключенным, что случайная выборка из незащищенной генеральной совокупности могла бы дать $\bar{x} = 6{,}25$? Следовательно, вопрос сводится к тому, не является ли расхождение между средними значениями в опытной и контрольной группах просто расхождением между двумя выборочными средними двух выборок, взятых из одной и той же генеральной совокупности. Последнее означало бы, что мыши из опытной группы принадлежат к той же самой генеральной совокупности, что и мыши контрольной группы, а именно к генеральной совокупности незащищенных животных. Это в свою очередь означает, что исследуемый препарат не обладает защитными свойствами (не переводит получившее этот препарат животное в другую генеральную совокупность).

Прежде чем читать дальше, просмотрите разделы **3.1.2**, **3.2.1**, **3.3.1**, **3.4.1**, **3.5.1**, *а также раздел* **3.7.2** *о распределении Стьюдента. Кроме того, обязательно нужно усвоить понятия, разбираемые в § 4.1.*

В разделе **3.7.2** уже говорилось, что если распределение вариант в генеральной совокупности нормально, то величина

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

имеет распределение Стьюдента. С вероятностью P эта величина не превышает значения t_P, даваемого табл. IV Приложений. Обратно, вероятность того, что из-за случайностей выборки она превысит t_P, равна $1 - P$. Если выбрать значение $1 - P$ достаточно малым, то в случае $t > t_P$ можно будет (с малой вероятностью $1 - P$ ошибиться) отвергнуть гипотезу о том, что выборка со средним значением $\bar{x}$ и стандартной ошибкой $s_{\bar{x}}$ взята из генеральной совокупности с математическим ожиданием μ.

Из сказанного ясно, что в данном случае вероятность $1 - P$ по смыслу есть не что иное, как уровень значимости α. Поэтому условие отказа от сформулированной выше нулевой гипотезы о математическом ожидании можно записать в виде:

$$\frac{|\bar{x} - \mu|}{s_{\bar{x}}} > t_\alpha, \qquad (4.1)$$

где t_α — критическое значение t. Так как $\alpha = 1 - P$ (и обратно, $P = 1 - \alpha$), то табл. IV Приложений для доверительных граничных значений t_P есть одновременно таблица для критических значений[1] t_α. Например, критическое значение для 0,01, или 1%, равно доверительному граничному значению для 0,99, или 99%. Указанный критерий называется *критерием Стьюдента*.

Пример 4.3. При определении pH раствора было получено значение $7{,}48 \pm 0{,}21$ (на $n = 10$ пробах). Можно ли считать реакцию раствора щелочной?

Имеем:

$$t = \frac{7{,}48 - 7{,}00}{0{,}21} = 2{,}28.$$

Так как $t_{0,05}(9) = 2{,}26$, $t_{0.01}(9) = 3{,}25$, то значимость щелочной реакции сомнительна: t лишь немного превышает $t_{0,05}$.

4.2.2. Если дисперсия распределения σ^2 известна, то легко найти точное значение стандартной ошибки среднего значения: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Как указывалось в разделе 3.7.1, величина

$$\tau = \frac{|\bar{x} - \mu|}{\sigma_{\bar{x}}}$$

имеет нормальное распределение. Поэтому критерием принадлежности данной выборки к генеральной совокупности с математичес-

[1] Строго говоря, доверительные граничные значения и критические значения следовало бы обозначать разными буквами, так как для критических значений имеем $t_\alpha = t_{1-P}$.

ким ожиданием μ будет:

$$\frac{|\bar{x}-\mu|}{\sigma_{\bar{x}}} > u_\alpha, \qquad (4.1')$$

где u_α есть α-процентная точка нормального распределения (в частности, $u_{0,05}=1{,}96$, $u_{0,01}=2{,}58$, $u_{0,001}=3{,}29$).

4.2.3. При решении вопроса о равенстве математических ожиданий двух совокупностей задача сводится к определению значимости отношения:

$$t=\frac{(\bar{x}-\hat{x})-(\bar{y}-\hat{y})}{s_{\bar{x}-\bar{y}}}, \qquad (*)$$

где $\hat{x}$ и $\hat{y}$ обозначают соответствующие математические ожидания, а $s_{\bar{x}-\bar{y}}$ есть оценка стандартной ошибки

$$\sigma_{\bar{x}-\bar{y}}=\sqrt{\sigma_{\bar{x}}^2+\sigma_{\bar{y}}^2}=\frac{\sigma^2\{x\}}{n_x}+\frac{\sigma^2\{y\}}{n_y}. \qquad (**)$$

Величина (*) имеет распределение Стьюдента, если варианты обеих совокупностей распределены нормально и их дисперсии $\sigma^2\{x\}$ и $\sigma^2\{y\}$ одинаковы. Последнее условие (т. е. равенство $\sigma^2\{x\}=\sigma^2\{y\}$) выполняется автоматически в том случае, когда нулевая гипотеза гласит, что обе выборки взяты из одной генеральной совокупности. Тогда

$$\sigma_{\bar{x}-\bar{y}}^2=\sigma^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right),$$

где σ^2— общая для обеих совокупностей дисперсия.

Коль скоро предположение о равенстве $\sigma^2\{x\}$ и $\sigma^2\{y\}$ сделано, то величины

$$s^2\{x\}=\frac{1}{n_x-1}\sum_i(x_i-\bar{x})^2, \qquad s^2\{y\}=\frac{1}{n_y-1}\sum_j(y_j-\bar{y})^2, \qquad (***)$$

вычисленные по выборочным данным, должны рассматриваться как две оценки одной и той же дисперсии σ^2. Чтобы найти наилучшую оценку последней, усредняют оценки, полученные по данным каждой из выборок. Усреднение производится с учетом «веса» каждой из выборочных оценок $s^2\{x\}$ и $s^2\{y\}$, причем «весом» является в данном случае число степеней свободы[1] $f_x=n_x-1$ и $f_y=n_y-1$. Поэтому наилучшей оценкой дисперсии σ^2 будет величина:

$$s^2=\frac{(n_x-1)\,s^2\{x\}+(n_y-1)\,s^2\{y\}}{(n_x-1)+(n_y-1)}=\frac{\Sigma(x_i-\bar{x})^2+\Sigma(y_j-\bar{y})^2}{n_x+n_y-2}, \qquad (4.2)$$

[1] О числе степеней свободы см. раздел 3.4.1.

так что:

$$s_{\bar{x}-\bar{y}} = \sqrt{\frac{\Sigma(x_i-\bar{x})^2+\Sigma(y_j-\bar{y})^2}{n_x+n_y-2}\left(\frac{1}{n_x}+\frac{1}{n_y}\right)} =$$

$$= \sqrt{\frac{\Sigma(x_i-\bar{x})^2+\Sigma(y_j-\bar{y})^2}{n_x+n_y-2}\cdot\frac{n_x+n_y}{n_x n_y}}. \qquad (4.3)$$

Далее, в условиях нулевой гипотезы $\hat{x} = \hat{y}$, так что числитель в (*) имеет вид $\bar{x} - \bar{y}$. Поэтому условие отказа от нулевой гипотезы будет:

$$|t| = \frac{|\bar{x}-\bar{y}|}{s_{\bar{x}-\bar{y}}} > t_\alpha, \qquad (4.4)$$

где $s_{\bar{x}-\bar{y}}$ дается формулой (4.3). При отыскании величины t_α в табл. IV Приложений принимается, что число степеней свободы равно:

$$f = n_x + n_y - 2, \qquad (4.5)$$

так как $n_x + n_y$ значений x_i и y_j связаны двумя условиями, из которых определялись $\bar{x}$ и $\bar{y}$.

Для данных из табл. 4.1 получаем:

$$s_{\bar{x}-\bar{y}} = \sqrt{\frac{48{,}3+20{,}7}{32+23-2}\cdot\frac{32+23}{32\cdot 23}} = 0{,}312;$$

$$t = \frac{6{,}25-5{,}22}{0{,}312} = 3{,}30;$$

$$f = 32+23-2 = 53 \approx 50; \quad t_{0.01}(50) = 2{,}68.$$

Так как $t > t_{0,01}$, то нулевая гипотеза отвергается с вероятностью 0,99. Следовательно, расхождение между опытом и контролем можно считать значимым, т. е. препарат определенно обладает защитным действием.

Описанная выше методика относится к случаю, когда варианты одной совокупности варьируют независимо от вариант другой. Критерий для случая попарно связанных вариант описан в § 4.4.

Кроме того, напомним еще раз, что критерий Стьюдента применим только тогда, когда распределение вариант в генеральной совокупности нормально (или близко к нор-

мальному). Если есть основания опасаться, что распределение сильно отличается от нормального (о строгом критерии для проверки этого см. § 2.5), то следует сравнивать центры двух эмпирических совокупностей при помощи порядкового критерия Вилкоксона (§ 4.3). Этот критерий менее чувствителен, но зато он не требует никаких предположений о форме распределения вариант в генеральной совокупности.

4.2.4. Рассмотрим еще один пример применения критерия Стьюдента, приводящий к иной, чем в разделе 4.2.3, формулировке нулевой гипотезы.

Пример 4.4. Каждый из двух сортов ячменя высевался на пяти делянках. В табл. 4.2 приведены урожаи в килограммах (урожай сорта II на одной из делянок был поврежден и поэтому не включен в дальнейшую обработку).

ТАБЛИЦА 4.2

Номера делянок	1	2	3	4	5	n	$\bar{x}$	$\Sigma (x_i - \bar{x})^2$
Сорт I	9,1	7,3	8,0	7,9	9,4	5	8,34	3,09
Сорт II	7,6	8,2	5,7	6,1	—	4	6,90	4,26

Так как средний урожай сорта I ($\bar{x}_{\text{I}} = 8{,}34$) выше, чем для сорта II ($\bar{x}_{\text{II}} = 6{,}90$), то напрашивается вывод, что вообще сорт I более урожайный. Но такой вывод может оказаться ложным: ведь каждое из чисел 8,34 и 6,90 есть не генеральное, а выборочное среднее. Другая случайная выборка, также состоящая из 5 делянок, наверняка дала бы для $\bar{x}_{\text{I}}$ значение, отличное от 8,34; в частности, это новое значение могло бы оказаться меньше, чем 8,34. В то же время и для $\bar{x}_{\text{II}}$ выборка дала бы значение, отличное от 6,90, в частности, могло бы получиться число большее, чем 6,90. Поэтому повторение всего опыта с обоими сортами могло бы дать результат $\bar{x}_{\text{I}} < \bar{x}_{\text{II}}$. Очевидно, вероятность такого исхода тем меньше, чем больше отношение $\bar{x}_{\text{I}} - \bar{x}_{\text{II}}$ к $s_{\bar{x}_{\text{I}} - \bar{x}_{\text{II}}}$.

Таким образом, опять приходим к критерию Стьюдента для оценки значимости того, что μ_{I} превышает μ_{II}. Однако данная задача в биологическом отношении отличается от задачи в примере 4.2.

Действительно, в примере 4.2 сравнивались две группы мышей, взятых из одной популяции. Нулевая гипотеза состояла в том, что препарат, который давали животным одной группы, не про-

изводит никакого действия. Если эта гипотеза верна, то мыши, получившие препарат, остаются в той же популяции, что и мыши, не получавшие препарата. Но тогда обе выборки должны относиться к одной и той же генеральной совокупности. Очевидно, в этом случае нулевая гипотеза означает, что должны одновременно выполняться два условия: $\mu_I = \mu_{II}$ и $\sigma_I^2 = \sigma_{II}^2$.

В примере 4.4 с самого начала ясно, что растения принадлежат к двум различным популяциям: если даже урожайности обоих сортов одинаковы, то эти сорта, вероятнее всего, различаются по большинству других признаков, например по высоте соломы, полеганию, устойчивости к болезням, продолжительности вегетации и др. Но если обе группы объектов относятся к разным популяциям, то совсем не обязательно, чтобы в их распределении по урожайности выполнялось условие $\sigma_I^2 = \sigma_{II}^2$, если даже $\mu_I = \mu_{II}$. Поэтому нулевая гипотеза утверждает здесь только то, что $\mu_I = \mu_{II}$.

Таким образом, приходим к задаче о проверке такой нулевой гипотезы: хотя генеральные совокупности, к которым принадлежат обе выборки, и различны (коль скоро у них разные дисперсии), но они имеют одинаковые математические ожидания.

Когда $\sigma^2\{x\} \neq \sigma^2\{y\}$, распределение величины (*) зависит не только от числа степеней свободы, но и от отношения $\sigma^2\{x\}/\sigma^2\{y\}$, которое в данном случае неизвестно (ведь сами дисперсии $\sigma^2\{x\}$ и $\sigma^2\{y\}$ неизвестны — можно только найти их оценки $s^2\{x\}$ и $s^2\{y\}$). Однако оказывается возможным применять и в этом случае t-критерий, если при отыскании t_α в таблице критических значений пользоваться измененным числом степеней свободы:

$$f' = (n_x + n_y - 2)\left(\frac{1}{2} + \frac{s^2\{x\}\, s^2\{y\}}{s^4\{x\} + s^4\{y\}}\right). \qquad (4.6)$$

Когда $s^2\{x\} = s^2\{y\}$, то второй множитель в формуле (4.6) равен $^1/_2 + {}^1/_2 = 1$, так что для числа степеней свободы получается обычное значение формулы (4.5). Но если $s^2\{x\} \gg s^2\{y\}$ или $s^2\{x\} \ll s^2\{y\}$, то

$$\frac{s^2\{x\}\, s^2\{y\}}{s^4\{x\} + s^4\{y\}} \ll 1.$$

и тогда число степеней свободы уменьшается примерно вдвое; последнее означает, что если меньшая дисперсия не влияет на общую дисперсию, то она не должна влиять и на число степеней свободы.

Что касается величины $s_{\bar{x}-\bar{y}}$, то она в соответствии с (**) вычисляется в этом случае по формуле:

$$s_{\bar{x}-\bar{y}} = \sqrt{\frac{s^2\{x\}}{n_x} + \frac{s^2\{y\}}{n_y}}. \qquad (4.7)$$

где $s^2\{x\}$ и $s^2\{y\}$ находят по формулам (***). Следовательно, при $\sigma^2\{x\} \neq \sigma^2\{y\}$ условие отказа от нулевой гипотезы гласит:

$$|t'| = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s^2\{x\}}{n_x} + \frac{s^2\{y\}}{n_y}}} > t_\alpha, \qquad (4.8)$$

причем для нахождения t_α в табл. IV Приложений берут число степеней свободы из формулы (4.6).

Применим этот критерий к данным из примера 4.4. Получаем:

$$s_{\mathrm{I}}^2 = \frac{3{,}09}{4} = 0{,}772; \quad s_{\mathrm{II}}^2 = \frac{4{,}26}{3} = 1{,}42;$$

$$s_{\bar{x}_{\mathrm{I}} - \bar{x}_{\mathrm{II}}} = \sqrt{\frac{0{,}772}{5} + \frac{1{,}42}{4}} = \sqrt{0{,}509} = 0{,}714;$$

$$t' = \frac{8.34 - 6.90}{0{,}714} = \frac{1.44}{0{,}714} \approx 2{,}02;$$

$$f' = (5 + 4 - 2)\left(\frac{1}{2} + \frac{0{,}772 \cdot 1{,}42}{0{,}772^2 + 1{,}42^2}\right) = 7\,(0{,}50 + 0{,}42) = 6{,}44.$$

Учитывая, что $t_{0{,}05}(6) = 2{,}45$ и $t_{0{,}05}(7) = 2{,}37$, можно принять условно $t_{0{,}05}(6{,}44) = 2{,}41$. Поскольку $t < t_{0{,}05}$, нулевая гипотеза не отвергается.

4.2.5. Когда $\frac{\bar{x} - \bar{y}}{s_{\bar{x} - \bar{y}}} < t_\alpha$, то значимость различий между $\hat{x}$ и $\hat{y}$ остается недоказанной. Так как $s_{\bar{x} - \bar{y}}$ уменьшается при увеличении объема выборки n, то может показаться, что, взяв достаточно большое n, можно всегда добиться выполнения условия $t > t_\alpha$ (и тем самым доказать значимость любого различия). Это, конечно, не так. Если различия объективно нет, то при увеличении n разность между выборочными средними значениями $\bar{x}$ и $\bar{y}$ будет уменьшаться в таком же темпе, что и $s_{\bar{x} - \bar{y}}$ (т. е. как $1/\sqrt{n}$) — ведь, согласно нулевой гипотезе, наблюдаемое различие появилось именно вследствие случайностей в образовании выборок, влияние которых уменьшается при увеличении объема выборки[1].

Тем не менее если t оказалось близко к t_α, имеет смысл попытаться доказать значимость различия, увеличив объем выборок. Однако при этом невозможно указать определенно, насколько именно нужно увеличить этот объем — ведь полученное значение

[1] Впрочем, если сравнивать две группы из разных популяций, то нулевая гипотеза $x - y$ представляет собой лишь идеализацию: в действительности разность $x - y$ никогда в точности не равна нулю. Тогда при достаточно больших объемах выборок эта разность всегда может быть выявлена как значимая, как бы она ни была мала.

t (зависящее в значительной мере от величины разности $\bar{x} - \bar{y}$) характеризует главным образом данный конкретный опыт, а не объект как таковой.

4.2.6. Если разность между $\hat{x}$ и $\hat{y}$ оказалась значимой, то встает вопрос о нахождении доверительного интервала для этой разности. Очевидно, ширина этого интервала определяется величиной $t_P s_{\bar{x}-\bar{y}}$, причем если неизвестно отношение $\sigma^2\{x\}/\sigma^2\{y\}$, то $s_{\bar{x}-\bar{y}}$ надо вычислять по формуле:

$$s_{\bar{x}-\bar{y}} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n_x(n_x-1)} + \frac{\Sigma(y_j - \bar{y})^2}{n_y(n_y-1)}}, \qquad (4.9)$$

ведь формула (4.3) соответствовала нулевой гипотезе о том, что оба распределения одинаковы. Таким образом, если сравниваются две группы, бывшие до опыта в одной популяции, то при $\sigma^2\{x\} = \sigma^2\{y\}$ применяется t-критерий с $s_{\bar{x}-\bar{y}}$ из формулы (4.3) [при неизвестном отношении дисперсий — с $s_{\hat{x}-\hat{y}}$ из (4.9)], а затем, если отличие оказалось значимым, вычисляется доверительный интервал для $\hat{x} - \hat{y}$ с соответствующим $s_{\bar{x}-\bar{y}}$ [из формулы (4.3) или из формулы (4.9)]. При этом для отыскания табличного значения t используется в первом случае число степеней свободы $f = n_x + n_y - 2$, а во втором случае f' из формулы (4.6).

4.2.7. Иногда возникает необходимость сравнить не две, а большее число эмпирических совокупностей. Эта задача может быть решена при помощи t-критерия путем попарного сравнивания всех совокупностей. Однако такое решение требует большого количества вычислений, так как при увеличении числа сравниваемых совокупностей число пар, которые необходимо сравнить, быстро растет. Поэтому такая задача решается обычно специально разработанным для этой цели методом, который называется дисперсионным анализом. Этот метод излагается почти во всех руководствах по биологической статистике (см. литературу в конце книги).

§ 4.3. Критерий Вилкоксона

4.3.1. Как указывалось в конце раздела **4.2.3**, применение критерия Стьюдента предполагает нормальность распределения в генеральных совокупностях. Если есть основания считать, что это условие нарушается (строгий критерий для проверки этого описан в § 2.5), то сравнение центров двух эмпирических распределений может производиться по критерию Вилкоксона, к описанию которого мы переходим.

Этот критерий не использует численные значения вариант, а лишь их ранги (т. е. их место в ряду при расположении вариант в порядке возрастания или убывания). Поэтому его называют *ранговым*, или *порядковым*, *критерием*. Иногда его еще называют *непараметрическим*, так как применение его не требует оценки параметров распределений (математических ожиданий, дисперсий и др.)[1].

Из сказанного ясно, что применение критерия Вилкоксона (как и других порядковых критериев) не требует никаких предположений о характере распределения вариант в генеральных совокупностях.

Пусть мы имеем два ряда значений, которые хотим сравнить: $x_1, x_2, \ldots, x_l$ и $y_1, y_2, \ldots, y_m$, числа $l = n_x$ и $m = n_y$ могут быть неодинаковыми. Будем считать, что ряды x и y переписаны так, что числа x_i и y_j расположены в порядке возрастания, и объединим их в один общий ряд. Если, например, заданы ряды:

x: 11 14 15 26 33 38

y: 19 24 31 42 45

то общий ряд будет:

$$x_1\ x_2\ x_3\ y_1\ y_2\ x_4\ y_3\ x_5\ x_6\ y_4\ y_5$$

т. е.

11 14 15 19 24 26 31 33 38 42 45

Когда числа x_i и y_j представляют собой некоторые численные значения вариант, то различие обоих рядов по центральной тенденции будет определяться разностью средних значений:

$$\frac{x_1 + x_2 + \cdots + x_l}{n_x} \text{ и } \frac{y_1 + y_2 + \cdots + y_m}{n_y}. \qquad (*)$$

Если ранжировать варианты так, чтобы большему численному значению соответствовал больший ранг, то сумма рангов будет, как правило, больше у того ряда, у которого больше сумма значений. Поэтому можно в первом приближении судить о суммах значений $x_1 + x_2 + \ldots + x_l$ и $y_1 + y_2 + \ldots + y_m$ по соответствующим суммам рангов:

$$R_{x_1} + R_{x_2} + \cdots + R_{x_l} = T_x;$$

$$R_{y_1} + R_{y_2} + \cdots + R_{y_m} = T_y.$$

[1] Подробнее о непараметрических критериях см. книги Б. Л. Ван дер Вардена и Е. В. Гублера и А. А. Генкина.

В данном случае построение критерия различия упрощается вследствие того, что все ранги, как правило, представляют собой числа натурального ряда, причем сумма их $T_x + T_y$ равна:

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2},$$

где $n = n_x + n_y$. Это приводит к тому, что при заданных значениях n_x и n_y значимость различия между центрами обоих рядов полностью характеризуется одной только величиной T_x. Действительно, данное значение T_x (при заданных n_x и n_y) однозначно определяет $\frac{T_x}{n_x}$, $T_y = \frac{n(n+1)}{2} - T_x$ и $\frac{T_y}{n_y}$, а тем самым и разность $\frac{T_x}{n_x} - \frac{T_y}{n_y}$. Поэтому для каждой пары чисел n_x, n_y можно указать определенное критическое значение T_α, отвечающее выбранному уровню значимости α.

Значения $T_{0,05}$ и $T_{0,01}$ приведены в табл. 4.3. С T_α должна сравниваться меньшая из сумм T_x и T_y, причем ранжирование должно быть произведено таким образом, чтобы меньшее из чисел T_x и T_y отвечало меньшему же из объемов n_x и n_y. Если, например, заданы ряды

x: 4 7 9 16 18

y: 23 26 32

то ряды рангов будут:

R_x: 1 2 3 4 5

R_y 6 7 8

ТАБЛИЦА 4.3

Критические значения T_α (критерия Вилкоксона) (из книги Л. Н. Большева и Н. С. Смирнова, с. 419—421)

n_y \ n_x	4	5	6	7	8	9	10
4	10	11	12 **10**	13 **10**	14 **11**	14 **11**	15 **12**
5		17 **15**	18 **16**	20 **16**	21 **17**	22 **18**	23 **19**
6			26 **23**	27 **24**	29 **25**	31 **26**	32 **27**
7				36 **32**	38 **34**	40 **35**	42 **37**
8					49 **43**	51 **45**	53 **47**
9						62 **56**	65 **58**
10							78 **71**

Число для $\alpha = 0{,}05$ напечатано обычным шрифтом, а для $\alpha = 0{,}01$ — жирным шрифтом.

Нулевая гипотеза принимается при $T \geqslant T_\alpha$ и отвергается при $T < T_\alpha$.

Более обширная таблица (до $n = 30$) имеется в книге Снедекора, с. 126.

Тогда

$$T_x = 1 + 2 + 3 + 4 + 5 = 15,$$

$$T_y = 6 + 7 + 8 = 21,$$

в то время как $n_x = 5$, $n_y = 3$; в данном случае $T_x < T_y$, а $n_x > n_y$. Чтобы устранить это несоответствие, надо изменить порядок ранжирования, т. е. перенумеровать варианты не в порядке возрастания, а в порядке убывания. Тогда ряды рангов будут:

R_x: 4 5 6 7 8
R_y: 1 2 3

Теперь получится

$$T_x = 4 + 5 + 6 + 7 + 8 = 30;$$

$$T_y = 1 + 2 + 3 = 6,$$

т. е.

$$T_y < T_x \text{ и } n_y < n_x.$$

Смысл данного критерия состоит в том, что в условиях нулевой гипотезы суммы T_x и T_y не должны слишком сильно отклоняться от их среднего значения

$$\frac{T_x + T_y}{2} = \frac{n(n+1)}{4},$$

так что меньшая из этих сумм не должна быть слишком малой.

Пример 4.5. В биохимическом исследовании, проведенном методом меченых атомов, измерялась скорость счета радиоактивных препаратов — 9 препаратов опытной серии и 5 препаратов контрольной серии. Полученные значения (в импульсах в минуту) записаны в табл. 4.4.

ТАБЛИЦА 4.4

Опыт	340	343	322	349	332	320	313	304	329
Контроль	318	321	318	301	312				

Проверим значимость различия при помощи критерия Стьюдента, предположив, что распределение вариант нормально. Для упрощения вычислений применяем кодирование, уменьшая все числа на 300. Тогда расчет выглядит так, как в табл. 4.5. Поскольку $t_{0,05}(12) = 2{,}18$, то различие незначимо ($t < t_{0,05}$).

ТАБЛИЦА 4.5

Опыт			Контроль			
x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	y_j	$y_j-\bar{y}$	$(y_j-\bar{y})^2$	
40	12	144	18	4	16	$\bar{x}=\frac{252}{9}=28;$
43	15	225	21	7	49	$\bar{y}=\frac{70}{5}=14;$
22	—6	36	18	4	16	
49	21	441	1	—13	169	$\bar{x}-\bar{y}=28-14=14;$
32	4	16	12	—2	4	$s_{\bar{x}-\bar{y}}=\sqrt{\frac{1728+254}{9+5-2}}\times$
20	—8	64	70	0	254	$\times\sqrt{\frac{9+5}{9\cdot5}}=$
13	—15	225				$=\sqrt{\frac{1982\cdot14}{12\cdot9\cdot5}}=7{,}18;$
4	—24	576				$t=\frac{14}{7{,}18}=1{,}95,\ f=12$
29	1	1				
252	0	1728				

Проверка при помощи критерия Вилкоксона оказывается значительно проще. Прежде всего располагаем все данные в один упорядоченный ряд и проставляем их ранги:

x_i		304		313			320		322	329	332	340	343	349
y_j	301		312		318	318		321						
R_x		2		4			7		9	10	11	12	13	14
$_yR$	1		3		5	6		8						

Так как мы имеем здесь дело с числами, а не с буквами x и y, то во избежании путаницы записываем их не в одну, а в две строки. Теперь подсчитываем сумму рангов для каждого из рядов:

$$T_x=2+4+7+9+10+11+12+13+14=82;$$

$$T_y=1+3+5+6+8=23.$$

Мы видим, что действительно:

$$T_x+T_y=\frac{n(n+1)}{2},\ \text{т. е. } 82+23=\frac{14\cdot15}{2}=105.$$

Из табл. 4.3 имеем для $n_x=5$, $n_y=9$ (ввиду того что T_y оказалось меньше, чем T_x, мы изменили обозначения рядов x и y) $T_{0,05}=22$; $T_{0,01}=18$. Поскольку $T>T_{0,05}$, нулевая гипотеза не отвергается.

В данном случае можно было не вычислять T_x, так как сразу видно, что сумма пяти рангов будет меньшей.

4.3.2. Когда значения n_x и n_y выходят за пределы табл. 4.3, можно воспользоваться тем обстоятельством, что при достаточно больших объемах выборок распределение величин T_x и T_y приближается к нормальному с математическими ожиданиями:

$$\widehat{T}_x = \frac{n_x(n+1)}{2}; \quad \widehat{T}_y = \frac{n_y(n+1)}{2}$$

и дисперсиями:

$$\sigma^2\{T_x\} = \sigma^2\{T_y\} = \frac{n_x n_y (n+1)}{12}.$$

Поэтому значимость отклонения T_x от $\widehat{T}_x$ можно оценивать по u-критерию, вычислив:

$$u = \frac{\widehat{T}_x - T_x}{\sigma\{T_x\}} = \frac{\frac{n_x(n+1)}{2} - T_x}{\sqrt{\frac{n_x n_y (n+1)}{12}}}.$$

Это можно переписать в виде:

$$u = \sqrt{3}\,\frac{n_x(n+1) - 2T_x}{\sqrt{n_x n_y (n+1)}}.$$

Удобнее пользоваться величиной:

$$\omega = \frac{u}{\sqrt{3}} = \frac{n_x(n+1) - 2T_x}{\sqrt{n_x n_y (n+1)}}, \tag{4.10}$$

сравнивая ее с критическими значениями:

$$\omega_{0,05} = \frac{u_{0,05}}{\sqrt{3}} = \frac{1,96}{\sqrt{3}} = 1,13; \qquad \omega_{0,01} = \frac{u_{0,01}}{\sqrt{3}} = \frac{2,58}{\sqrt{3}} = 1,49.$$

Для примера 4.5 имеем:

$$\frac{5(14+1) - 2\cdot 23}{\sqrt{5\cdot 9(14+1)}} = \frac{75-46}{\sqrt{675}} = \frac{29}{26} = 1,115;$$

так как $\omega < \omega_{0,05}$, то нулевая гипотеза не отвергается — результат, полученный ранее по табл. 4.3. Однако в данном случае применение формулы (4.10) не совсем правомерно, так как числа n_x и n_y недостаточно велики.

§ 4.4. Сравнение совокупностей с попарно связанными вариантами

Об упоминавшихся выше понятиях математического ожидания, дисперсии и и-критерия см. разделы 3.1.1., 3.3.1 *и* 4.2.2. *О применении критерия Вилкоксона к случаю сопряженных пар см. раздел* 4.4.3.

4.4.1. Стандартная ошибка разности $\sigma_{\bar{x}-\bar{y}}$ выражается через стандартные ошибки $\sigma_{\bar{x}}$ и $\sigma_{\bar{y}}$ в виде

$$\sigma_{\bar{x}-\bar{y}} = \sqrt{\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2}$$

лишь в том случае, если варианты одной совокупности варьируют независимо от вариант другой совокупности. Если же между вариантами обеих совокупностей имеется статистическая связь, то эта формула неприменима.

В практике биологического эксперимента этот последний случай встречается довольно часто. Например, воздействие того или иного агротехнического мероприятия на растения всегда происходит на фоне сильно варьирующих условий погоды, рельефа местности, почвы и др., имеющих для развития растений первостепенное значение; поэтому при сравнении двух рядов урожайностей за несколько лет всегда имеется сопряженность между значениями из обоих рядов, относящихся к одинаковым годам. То же имеет место, если сравнивают урожайности, полученные в одном году на ряде делянок, с урожайностями, полученными на этих же делянках в другом году. Аналогичное положение возникает при сравнении результатов двух серий опытов, поставленных на одном и том же животном для двух разных физиологических состояний, и др.

Пример 4.6. В табл. 4.6 приведены данные опыта по изучению влияния определенной предпосевной обработки семян пшеницы на урожайность. В соответствии с погодными условиями урожайность меняется год от года как в опыте, так и в контроле; поэтому значения опытной группы и значения контрольной группы нельзя считать взаимно независимыми — они попарно связаны тем, что значения из каждой пары относятся к одному и тому же году и тем самым к одним и тем же условиям погоды (которые, как это видно из табл. 4.6, оказывают на урожайность гораздо большее влияние, чем предпосевная обработка семян).

Здесь набор разностей Δ_i в колонке 4 вполне однозначен — в отличие от случая, когда варианты в каждом из рядов варьируют независимо: в последнем случае можно было бы произвольно переставлять варианты в колонках 2 и 3, что приводило бы к различ-

ТАБЛИЦА 4.6

Год	Опыт x_i	Контроль y_i	Разность $\Delta_i = x_i - y_i$	Δ_i^2
1947	22,9	19,4	3,5	12,25
1948	20,2	16,2	4,0	16,00
1949	19,5	16,9	2,6	6,76
1950	30,5	29,3	1,2	1,44
1951	35,6	31,4	4,2	17,64
1952	31,9	28,5	3,4	11,56
1953	27,7	25,7	2,1	4,41
Сумма	188,3	167,3	21,0	70,06
Среднее	26,9	23,9	3,0	

ным наборам разностей. Поэтому мы можем числа в колонке 4 рассматривать как некоторый вариационный ряд, характеризующийся определенным средним значением $\bar{\Delta}$, дисперсией $\sigma^2\{\Delta\}$, стандартной ошибкой среднего $\sigma_{\bar{\Delta}}$ и т. д. Нулевая гипотеза в этом случае гласит, что $\hat{\Delta} = 0$, так что, используя обычный t-критерий, ее отвергают, если

$$|t_{\bar{\Delta}}| = \frac{|\bar{\Delta}|}{s_{\bar{\Delta}}} > t_\alpha; \quad s_{\bar{\Delta}} = \sqrt{\frac{\Sigma(\Delta_i - \bar{\Delta})^2}{n(n-1)}}. \qquad (4.11)$$

Число степеней свободы равно здесь, конечно, $f = n - 1$.

По данным табл. 4.6 имеем:

$$\bar{\Delta} = \frac{21,0}{7} = 3,0 \,(= 26,9 - 23,9);$$

$$\Sigma(\Delta_i - \bar{\Delta})^2 = \Sigma\Delta_i^2 - \frac{1}{n}(\Sigma\Delta_i)^2 = 70,06 - \frac{21,0^2}{7} = 7,06,$$

так что:

$$s_{\bar{x}-\bar{y}} = s_{\bar{\Delta}} = \sqrt{\frac{7,06}{7 \cdot 6}} = 0,41.$$

Таким образом,

$$t_{\bar{\Delta}} = \frac{3,0}{0,41} = 7,32.$$

В данном случае $f = n - 1 = 7 - 1 = 6$; так как $t_{0,01}(6) = 3{,}71$, то различие следует считать значимым.

4.4.2. Если бы к данным из табл. 4.6 применялся обычный критерий (т. е. не учитывающий сопряженность пар), то различие не было бы обнаружено: большой разброс урожайности из-за сильной вариабельности погодных условий привел бы к завышенной дисперсии разности и, следовательно, к заниженной величине t. Правильный метод расчета позволил исключить этот фактор, действовавший одинаково на оба элемента каждой пары, и тем самым выявить «чистый» эффект предпосевной обработки.

Когда наличие сопряженности в парах не вытекает непосредственно из существа дела, формула (4.11) все равно дает правильное значение t. Но в этом случае применение критерия $t_{\overline{\Delta}}$ может не обнаружить имеющееся на самом деле различие, так как при этом используется вдвое меньшее число степеней свободы — $(n - 1)$ вместо $2n - 2 = 2(n - 1)$ — и тем самым большее критическое значение t_α.

Отсюда следует, что t-критерием в форме (4.11) целесообразно пользоваться только тогда, когда сопряженность пар несомненна; в этом случае влияние уменьшения числа степеней свободы перекроется влиянием уменьшения дисперсии разности.

4.4.3. Описанный выше критерий, как и обычный вариант критерия Стьюдента (§ 4.2), пригоден только при нормальности распределения в генеральных совокупностях. Если о распределении ничего неизвестно или имеются основания предполагать, что оно сильно отличается от нормального, следует пользоваться порядковым критерием Вилкоксона, но видоизмененным для случая сопряженных пар.

О критерии Вилкоксона (и вообще о понятии порядкового критерия) см. § 4.3.

Пользуются им следующим образом. Каждой разности приписывается определенный ранг в зависимости от ее величины, причем знак этой разности не принимается во внимание; чем больше разность, тем больше считается ее ранг. Если нулевая гипотеза справедлива, то сумма рангов, отвечающих разностям одного знака, должна равняться сумме рангов, отвечающих разностям другого знака. Вследствие случайностей выборки могут наблюдаться отступления от этого равенства, но вероятность больших отступлений мала. Поэтому, если указанные две суммы рангов сильно различаются, то это может послужить основанием для того, чтобы отвергнуть нулевую гипотезу.

Если подсчитана одна сумма рангов, то при заданном общем числе пар вариант n вторая сумма определяется однозначно, так как сумма всех рангов (обоих типов) равна

$$1+2+\cdots+n=\frac{n(n+1)}{2}.$$

Поэтому вместо того, чтобы сопоставлять между собой две суммы рангов T_1^{Δ} и T_2^{Δ}, достаточно сопоставить одну из сумм (удобно, чтобы это была меньшая сумма) с числом пар n. Как правило, меньшей оказывается та сумма рангов T^{Δ}, которая отвечает разностям со знаком, представленным в меньшем числе. Критические значения T_{α}^{Δ} для разных n даются в табл. 4.7.

Нулевая гипотеза принимается при $T^{\Delta} \geqq T_{\alpha}^{\Delta}$ и отвергается при $T^{\Delta} < T_{\alpha}^{\Delta}$.

Если какие-либо разности равны нулю, то они просто исключаются из рассмотрения, причем соответственно уменьшается n.

ТАБЛИЦА 4.7

Критические значения T_{α}^{Δ} — критерия Вилкоксона для сопряженных пар (составлено по книге Д. Б. Оуэна, с. 326—328)

n	5%	1%	n	5%	1%	n	5%	1%
6	3	—	11	14	8	16	36	24
7	4	1	12	18	10	17	42	28
8	6	2	13	22	13	18	48	33
9	9	4	14	26	16	19	54	38
10	11	6	15	31	20	20	61	44

Пример 4.7. Сравнивалось действие двух экстрактов вируса табачной мозаики. Для этого каждую из половин листа натирали соответствующим препаратом. Число мест поражений записано в табл. 4.8. Поскольку два числа в каждой строке относятся к двум половинам одного и того же листа, то варианты можно считать попарно сопряженными.

ТАБЛИЦА 4.8

Ряд I	Ряд II	Разности	Ранги разностей	(T^{Δ})
20	31	—11	6	6
39	22	17	8	
43	45	—2	1	1
13	6	7	4,5	
28	21	7	4,5	
26	13	13	7	
17	17	0	—	
49	46	3	2	
36	31	5	3	

В одной из пар значения одинаковы, так что разность равна нулю. Эта пара исключается и остается $n = 8$. Поскольку $T^{\Delta} =$

$= 6 + 1 = 7$ превышает значение $T^{\Delta}_{0,05}(8) = 6$, указанное в табл. 4.7, то нулевая гипотеза не отвергается.

Если $n > 25$ (так что табл. 4.7 пользоваться нельзя), то можно применить u-критерий, так как при больших n значения T^{Δ} распределены нормально со средним значением и дисперсией:

$$\hat{T}^{\Delta} = \frac{n(n+1)}{4}, \qquad \sigma^2_{T^{\Delta}} = \frac{n(n+1)(2n+1)}{24};$$

нулевая гипотеза принимается, если

$$u_{T^{\Delta}} = \frac{\hat{T}^{\Delta} - T^{\Delta}}{\sigma_{T^{\Delta}}} \tag{4.12}$$

меньше или равно u_{α}, и отвергается, если $u_{T\Delta} > u_{\alpha}$ (причем $u_{0.05} = 1{,}96$ и $u_{0.01} = 2{,}58$). Пусть, например, $n = 46$ и $T^{\Delta} = 263$. Тогда:

$$\hat{T}^{\Delta} = \frac{46 \cdot 47}{4} = 540.5; \qquad \sigma_{T^{\Delta}} = \sqrt{\frac{46 \cdot 47 \cdot 93}{24}} = 91{,}5,$$

так что:

$$u_{T^{\Delta}} = \frac{540{,}5 - 263}{91{,}5} = \frac{277{,}5}{91{,}5} = 3{,}04.$$

Поскольку $u_{T\Delta} > u_{0,01}$, то нулевая гипотеза отвергается.

§ 4.5. Последовательный анализ

В этом параграфе используются понятия: математическое ожидание (см. раздел 3.1.1*), дисперсия (см. раздел* 3.3.1*) и ее выборочная оценка (см. раздел* 3.3.1*), нулевая и альтернативная гипотезы (см. раздел* 4.1.1*), критерий значимости (см. раздел* 4.1.1*), ошибки I и II рода (см. раздел* 4.1.2*). О нормальности распределения см. § 2.3.*

4.5.1. Гипотеза H_1, конкурирующая с нулевой гипотезой H_0: $\hat{\pi} = \pi_0$ (π — любой параметр совокупности), чаще всего имеет вид $\hat{\pi} \neq \pi_0$. Иногда H_1 гласит: $\hat{\pi} > \pi_0$ (или $\hat{\pi} < \pi_0$).

Однако могут быть случаи, когда требуется еще более определенная формулировка альтернативной гипотезы. Пусть, например, предлагается новое агротехническое мероприятия, направленное к повышению урожайности некоторой культуры. Поскольку проведение этого мероприятия требует известных затрат, оно ока-

жется целесообразным только в том случае, если увеличение урожайности будет не меньше некоторой определенной величины δ. Поэтому если нулевая гипотеза имеет здесь вид $\mu \leqslant x_{(0)}$ (где $x_{(0)}$ — урожайность в контроле), то альтернативная гипотеза будет: $\mu \geqslant x_{(0)} + \delta = x_{(1)}$.

Указанная задача может быть решена обычными методами с применением t-критерия. Однако более удобным оказывается так называемый «*метод последовательного анализа*» (А. Вальд).

При обычном статистическом анализе математическая обработка результатов производится после завершения серии наблюдений, объем которой (т. е. число наблюдений) было оценено заранее в соответствии с принятым уровнем значимости (и предварительной оценкой дисперсии вариант). При последовательном же анализе число наблюдений заранее не фиксируется; математическая обработка (впрочем, как мы сейчас увидим, совершенно элементарная) производится после каждого наблюдения, причем в результате этой обработки выясняется, можно ли принять одну из конкурирующих гипотез (и какую именно) или же следует продолжить испытания. Как показывает опыт, число требующихся при этом наблюдений оказывается в среднем примерно вдвое меньше, чем при классическом статистическом анализе.

Обозначим через $P_{0,n}$ плотность вероятности (в дискретном случае — вероятность) получить значения $x_1, x_2, \ldots, x_n$ при условии, что выборка взята из генеральной совокупности с математическим ожиданием $\mu \leqslant x_{(0)}$, и через $P_{1,n}$ — плотность вероятности получить эти значения при условии, что выборка взята из генеральной совокупности с $\mu \geqslant x_{(1)}$.

Когда $P_{1,n} \leqslant \alpha$, гипотеза $\mu \geqslant x_{(1)}$ может быть отвергнута; при этом риск ошибиться (если в действительности эта гипотеза верна) не превышает α. Но отвергнуть гипотезу $\mu \geqslant x_{(1)}$ еще не значит, что надо обязательно принять гипотезу $\mu \leqslant x_{(0)}$; можно ограничиться утверждением, что $\mu < x_{(1)}$. Если же мы сделаем более определенное утверждение $\mu \leqslant x_{(0)}$, то мы рискуем сделать ошибку II рода (приняв гипотезу в действительности неверную). Вероятность этой ошибки II рода не будет превышать β, если $P_{0,n} \geqslant 1 - \beta$. Это неравенство можно объединить с неравенством $P_{1,n} \leqslant \alpha$ в одно неравенство:

$$\frac{P_{1,n}}{P_{0,n}} \leqslant \frac{\alpha}{1-\beta}. \qquad (*)$$

Действительно, если $a < b$ и $c > d$, то тем более $a/c < b/d$.

Рассуждая совершенно аналогично, находим условие возможности отвергнуть гипотезу $\mu \leqslant x_{(0)}$ и принять гипотезу $\mu \geqslant x_{(1)}$:

$$\frac{P_{0,n}}{P_{1,n}} \leqslant \frac{\beta}{1-\alpha};$$

это неравенство можно заменить равносильным неравенством:

$$\frac{P_{1,n}}{P_{0,n}} \geqslant \frac{1-\alpha}{\beta} \qquad (**)$$

(например, если $2 < 3$, то $1/2 > 1/3$).

Таким образом, если выполняется неравенство (*), то принимается гипотеза $\mu \leqslant x_{(0)}$, а если выполняется неравенство (**), то принимается гипотеза $\mu \geqslant x_{(1)}$; если же окажется, что

$$\frac{\alpha}{1-\beta} < \frac{P_{1,n}}{P_{0,n}} < \frac{1-\alpha}{\beta}, \qquad (***)$$

то это будет означать, что испытания надо продолжать.

4.5.2. Если распределение вариант в генеральной совокупности нормально, то в соответствии с (2.4)

$$P_{0,n} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - x_{(0)})^2},$$

$$P_{1,n} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - x_{(1)})^2};$$

здесь также использовано известное из теории вероятностей утверждение: если события A, B, C, ..., Z имеют вероятности соответственно P_A, P_B, P_C, ..., P_Z и независимы, то вероятность одновременного осуществления всех этих событий равна произведению $P_A P_B P_C \ldots P_Z$.

Тогда:

$$\frac{P_{1,n}}{P_{0,n}} = e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(x_i - x_{(1)})^2 - (x_i - x_{(0)})^2]}.$$

Логарифмируя это равенство, получаем:

$$ln\,\frac{P_{1,n}}{P_{0,n}} = \frac{x_{(1)} - x_{(0)}}{s^2}\left(\sum_{i=1}^{n} x_i - \frac{x_{(1)} + x_{(0)}}{2}\,n\right)$$

(после элементарного преобразования выражения в квадратных скобках и замены дисперсии σ^2 ее оценкой s^2).

Если подставить это выражение для $ln\,\frac{P_{1,n}}{P_{0,n}}$ в неравен-

ства (***), то можно получить равносильные неравенства:

$$\frac{x_{(1)}+x_{(0)}}{2}n+\frac{s^2}{x_{(1)}-x_{(0)}}ln\frac{\beta}{1-\alpha}<\sum_{i=1}x_i<$$

$$<\frac{x_{(1)}+x_{(0)}}{2}n+\frac{s^2}{x_{(1)}-x_{(0)}}ln\frac{1-\beta}{\alpha}. \quad (4.13)$$

Это значит, что испытания надо продолжать до тех пор, пока сумма вариант Σx_i лежит в пределах от

$$L_0(n)=\frac{x_{(1)}+x_{(0)}}{2}n-\frac{s^2}{x_{(1)}-x_{(0)}}ln\frac{1-\alpha}{\beta}=an-b' \quad (4.14)$$

до

$$L_1(n)=\frac{x_{(1)}+x_{(0)}}{2}n+\frac{s^2}{x_{(1)}-x_{(0)}}ln\frac{1-\beta}{\alpha}=an+b''; \quad (4.15)$$

когда же сумма Σx_i достигнет или пересечет одну из этих границ, то испытания можно прекратить и принять ту или иную гипотезу (в зависимости от того, какая из границ будет достигнута).

Удобно изображать результаты графически, построив прямые (4.14) и (4.15). Очевидно, эти прямые имеют угловой коэффициент

$$a=\frac{x_{(1)}+x_{(0)}}{2} \quad (4.16)$$

и отсекают на оси ординат соответственно:

$$b'=\frac{s^2}{x_{(1)}-x_{(0)}}ln\frac{1-\alpha}{\beta}, \quad b''=\frac{s^2}{x_{(1)}-x_{(0)}}ln\frac{1-\beta}{\alpha} \quad (4.17)$$

(рис. 4.2); выбирая $\alpha=0{,}01$ и $\beta=0{,}01$, имеем:

$$ln\frac{1-\alpha}{\beta}=ln\frac{1-\beta}{\alpha}=ln\frac{0{,}99}{0{,}01}=4{,}59.$$

Значения $\sum_{i=1}^{n}x_i$ наносим на этот график в виде точек, абсциссами которых служат соответствующие значения n; точки соединяют отрезками прямой, которые образуют ломаную линию. Испытания продолжаются до тех пор, пока эта ломаная не выйдет из центральной полосы.

Из (4.17) видно, что эта полоса будет тем шире (и потребует тем больше испытаний), чем меньше $\delta=x_{(1)}-x_{(0)}$ и чем больше s; это вполне понятно: должен быть накоплен тем больший экспериментальный материал, тем тоньше различие, которое мы хотим обнаружить, и чем больше рассеяние вариант.

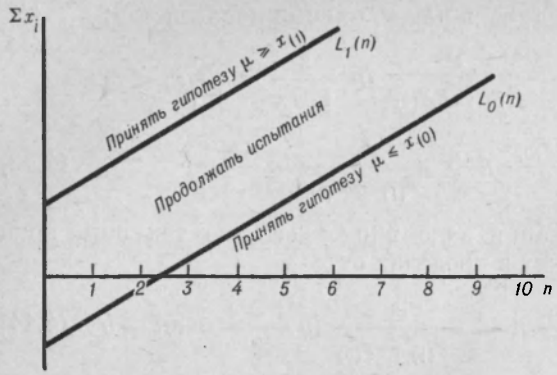


Рис. 4.2. Общий вид графика для последовательного анализа нормальных совокупностей.

Пример 4.8. Было установлено, что затраты на предпосевную обработку семян пшеницы окупятся, если она даст средний прирост урожайности не менее $\delta = 3$ ц с 1 га. Средняя урожайность необработанных семян — 20,8 ц с 1 га. Многолетние наблюдения показали, что в норме $s_0 = 1{,}81$, т. е. $v = s_0/x_{(0)} = 1{,}81 : 20{,}8 = 0{,}087 = 8{,}7\%$. Если принять это значение v и для обработанных семян, то при предполагаемой урожайности $x_{(1)} = x_{(0)} + \delta = 20{,}8 + 3{,}0 = 23{,}8$ можно считать разумной оценкой σ (во всяком случае незаниженной) величину:

$$s_1 = x_{(1)} v = 23{,}8 \cdot 0{,}087 = 2{,}07, \quad s_1^2 = 4{,}3.$$

При этих данных имеем согласно (4.16) и (4.17):

$$a = \frac{20{,}8 + 23{,}8}{2} - 22{,}3, \qquad b = \frac{4{,}3}{3{,}0}\, 4{,}59 = 6{,}6.$$

График будет в данном случае иметь более удобный вид, если наносить не значения Σx_i, а $\Sigma x_i'$, где $x_i' = x_i - 20$; соответственно вместо $a = 22{,}3$ примем $a' = a - 20 = 2{,}3$. Линии $L_0(n)$ и $L_1(n)$ проведем через точки (см. рис. 4.3):

$$L_0(0) = -6{,}6; \quad L_0(10) = -6{,}6 + 2{,}3 \cdot 10 = 16{,}4;$$

$$L_1(0) = +6{,}6; \quad L_1(10) = +6{,}6 + 2{,}3 \cdot 10 = 29{,}6.$$

В нашем распоряжении имеются данные за 10 лет, приведенные в табл. 4.9.

ТАБЛИЦА 4.9

Год	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
x_i	26,7	21,0	24,1	27,1	25,1	23,0	26,2	19,4	21,8	23,4
x_i'	6,7	1,0	4,1	7,1	5,1	3,0	6,2	−0,6	1,8	3,4

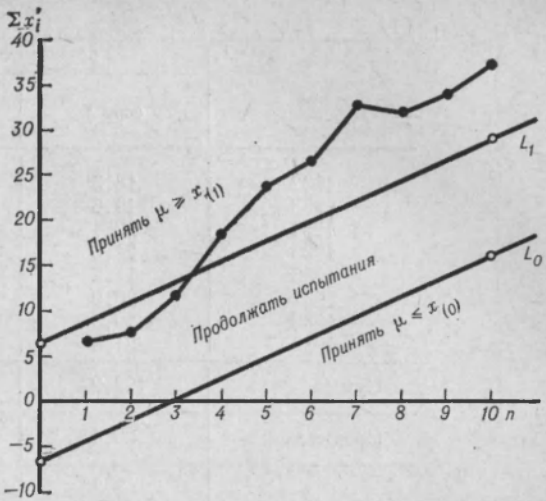


Рис. 4.3. Последовательный анализ прироста урожайности.

Нанесем на график точки с абсциссами 1; 2; 3, ... и ординатами 6,7; 6,7 + 1,0 = 7,7; 7,7 + 4,1 = 11,8 и т. д. (см. рис. 4.3). Мы видим, что уже четвертая точка ($n = 4$, $\sum_{i=1}^{4} x_i = 18{,}9$) лежит вне центральной полосы; это значит, что гипотезу $\mu \geqslant x_{(1)}$ можно было бы принять на основании только четырехлетних наблюдений.

§ 4.6. Сравнение дисперсий (*F*-критерий)

Рекомендуем просмотреть разделы, в которых говорится о дисперсии (**3.3.1**) *и стандартном отклонении* (**3.3.1**), *их выборочных оценках* (**3.4.1**), *числе степеней свободы* (**3.4.1**), *стандартных ошибках* (**3.5.1**), *нулевой гипотезе и критериях значимости — односторонних и двусторонних* (**4.1.1**, **4.1.3**).

4.6.1. Две выборочные совокупности, не различаясь значимо по своим средним значениям, могут различаться по стандартным отклонениям (или дисперсиям).

Пример 4.9. Два сорта пшеницы (табл. 4.10) имеют почти одинаковую среднюю урожайность ($x_1 = 20{,}4$ ц/га; $x_2 = 20{,}3$ ц/га), но один из них (сорт А) менее подвержен влиянию изменений погодных условий от года к году, чем другой сорт ($s_1^2 = 4{,}92$; $s_1 = 2{,}22$; $s_2^2 = 16{,}9$; $s_2 = 4{,}11$).

ТАБЛИЦА 4.10

Год	Урожайность, ц/га	
	сорт А	сорт Б
1932	18,3	16,8
1933	19,6	17,2
1934	22,1	23,7
1935	24,0	26,1
1936	17,2	15,4
1937	20,9	21,3
1938	19,3	17,4
1939	21,8	24,5
Сумма	163,2	162,4
Среднее	20,4	20,3
s	2,22	4,11

Если объемы выборок велики, то значимость этого различия можно оценить при помощи u-критерия, считая, что величина $u_{s_1-s_2} = (s_1 - s_2)/\sigma_{s_1-s_2}$ распределена нормально. При малых же выборках нормальное распределение должно быть заменено другим, более сложным распределением.

Вместо того, чтобы искать это распределение, используют изученное Р. Фишером распределение отношения $F = s_1^2/s_2^2$, принимая в качестве критерия значимости различия двух дисперсий отношение их оценок:

$$F = \frac{s_1^2}{s_2^2} \tag{4.18}$$

(*F-критерий Фишера*), которое нужно в каждом случае сравнивать с критическим значением F_α. Это критическое значение зависит, помимо выбранного уровня значимости α, еще от объемов выборок, по которым определялись оценки s_1^2 и s_2^2, точнее, от соответствующих чисел степеней свободы $f_1 = n_1 - 1$ и $f_2 = n_2 - 1$. Таблица критических значений $F_\alpha(f_1, f_2)$ приводится в Приложениях (табл. V) для двух уровней значимости: 5% и 1%. При пользовании этой таблицей надо иметь в виду, что

$$F_\alpha(f_1, f_2) \neq F_\alpha(f_2, f_1).$$

В соответствии с обычными условиями применения F-критерия при составлении таблицы критических значений F использовался односторонний критерий. Так как эта таблица содержит только значения F, бо́льшие единицы, то при вычислении величины F

надо всегда делить большую оценку на меньшую, соответственно изменив обозначения.

В нашем примере мы обозначим $16{,}9 = s_1^2$, $4{,}92 = s_2^2$, так что

$$F = \frac{s_1^2}{s_2^2} = \frac{16{,}9}{4{,}92} = 3{,}44.$$

В табл. V Приложений находим, что при числе степеней свободы числителя $f_1 = 8-1 = 7$ и числе степеней свободы знаменателя $f_2 = 8-1 = 7$ критические значения $F_\alpha(f_1, f_2)$ равны:

$$F_{0,05}(7;\ 7) = 3{,}79, \quad F_{0,01}(7;\ 7) = 7{,}00.$$

Поскольку фактическое значение $F = 3{,}44$ меньше 5%-ного критического значения, то нулевая гипотеза не отвергается.

4.6.2. В примере 4.4 из раздела **4.2.4** были получены оценки дисперсий $s_I^2 = 0{,}772$ и $s_{II}^2 = 1{,}42$ при $f_I = 4$ и $f_{II} = 3$. Отношение этих оценок

$$F = \frac{s_{II}^2}{s_I^2} = \frac{1{,}42}{0{,}772} = 1{,}84$$

меньше критического значения $F_{0.05}\ (3;\ 4) = 6{,}59$.

Если бы в формулировку нулевой гипотезы входило условие $\sigma_I^2 = \sigma_{II}^2$, то результат $F < F_{0.05}$ означал бы, что опыт не отвергает справедливость этого равенства; тогда при пользовании t-критерием нужно было бы вычислять $s_{\bar{x}-\bar{y}}$ по формулам (4.3) и (4.5). Но в примере 4.4 не было оснований выдвигать гипотезу $\sigma_I^2 = \sigma_{II}^2$, так как выборки относились к заведомо разным популяциям. Поэтому здесь нет также оснований применять F-критерий; любое различие между s_I^2 и s_{II}^2, как бы оно ни было мало, должно считаться значимым. Поэтому использование формул (4.7) и (4.6) для $s_{\bar{x}-\bar{y}}$ и f в примере 4.4 было оправдано.

Этими же формулами надо пользоваться и при сравнении средних значений для групп, взятых из одной популяции, если окажется, что их дисперсии различны. Но в этом случае различие дисперсий должно быть проверено по F-критерию.

4.6.3. В некоторых случаях бывает необходимо сравнить сразу несколько дисперсий. Если все выборки имеют одинаковый объем, то эта задача решается при помощи *критерия Кочрена* (см. книгу автора «Биометрические методы», 1964). Если же объемы выборок различны, то приходится пользоваться более сложным *критерием Бартлета* (см. книги В. В. Налимова и Дж. У. Снедекора).

Глава пятая

ПРОВЕРКА ГИПОТЕЗ О ФОРМЕ РАСПРЕДЕЛЕНИЯ (КРИТЕРИЙ ХИ-КВАДРАТ)

§ 5.1. Сравнение эмпирического распределения с теоретическим

5.1.1. В главе четвертой были указаны критерии, позволяющие установить совпадение или различие параметров двух распределений — их математических ожиданий и дисперсий. По этой причине эти критерии называют *параметрическими*.

При пользовании параметрическими критерями обычно предполагается, что сравниваемые распределения в общем однотипны и могут отличаться лишь значениями своих параметров.

Наряду с этим часто приходится проверять гипотезу о самом виде распределения. Это может быть либо гипотеза о том, что данное эмпирическое распределение есть выборочный вариант распределения определенного теоретического вида, либо гипотеза о том, что два эмпирических распределения являются двумя выборочными вариантами одного и того же генерального распределения.

Проверить нормальность совокупности, из которой взята выборка, можно при помощи параметрических критериев (см. § 2.5). Параметрическим критерием можно также проверить гипотезу о распределени Пуассона (см. раздел **7.1.4**). Однако в обоих случаях используются определенные частные свойства соответствующих распределений ($\rho_3 = 0$ и $\langle|\xi|\rangle/\sigma = \sqrt{\frac{2}{\pi}}$ в первом случае и $\sigma^2 = \mu$ во втором).

Этот способ неудобен при более сложных распределениях, описываемых многими параметрами, и вообще неприменим, если проверяемая гипотеза состоит в том, что две выборки взяты из двух генеральных совокупностей с одинаковым, но неизвестным распределением. Для этой цели служат критерии, использующие непосредственно значения наблюденных вариант и не требующие в качестве промежуточной ступени вычисления параметров. В силу последнего обстоятельства эти критерии называют *непараметрическими*. Большим преимуществом этих критериев является то, что использование их не нуждается в каких-либо предположениях о характере распределения вариант в генеральной совокупности. Это имеет особое значение, когда эмпирические данные относятся к какой-либо новой области исследования, где о виде распределения нельзя судить на основе предыдущего опыта.

Кроме того, в некоторых случаях о распределении заведомо известно, что оно не нормальное и не пуассоново.

Наиболее часто употребляемым непараметрическим критерием является так называемый *критерий хи-квадрат* (К. Пирсон). Правда, применение его требует, чтобы выборки были не слишком малы (не менее 20—30 вариант). Если это условие не соблюдается, то следует пользоваться другими непараметрическими критериями (см. список литературы в конце книги).

Прежде чем читать дальше, обязательно прочтите § 4.1, в котором рассмотрены общие основы применения статистических критериев и разъясняются понятия, которые будут встречаться в этой главе: нулевая гипотеза, значимость различия, уровень значимости, критическое значение, односторонний и двусторонний критерии.

5.1.2. Прежде чем переходить к описанию критерия хи-квадрат, напомним еще раз, что когда в статистике говорят о совпадении эмпирического и теоретического распределений, то имеется в виду, что фактически имеющееся различие между ними можно отнести за счет различия между выборочным и генеральным распределениями (поскольку при данном объеме выборки вероятность такого или даже еще большего расхождения между последними не так уж мала).

Если нулевая гипотеза верна и заданная эмпирическая совокупность является выборкой из генеральной совокупности с определенным распределением частот $\hat{n}_i$, то частоты эмпирической совокупности n_i варьируют около соответствующих теоретических значений $\hat{n}_i$ лишь случайно. Это значит, что попадание варианты в любой из разрядов группировки можно рассматривать как случайное явление, которому отвечает характерная для этого разряда вероятность $\hat{p}_i = \hat{n}_i/n$. Это явление вполне аналогично таким явлениям, как попадание ионизирующих частиц в счетчик, телефонные вызовы и др., которые описываются статистически распределением Пуассона (см. § 7.1). Но для этого распределения $\sigma^2 = \mu$ и $\sigma = \sqrt{\mu}$ (см. раздел **7.1.3**). В данном случае будем иметь для каждого i-го разряда $\sigma_i = \sqrt{\hat{n}_i}$, так что нормированное отклонение (т. е. отнесенное к стандартному отклонению) числа вариант в каждом разряде будет равно $(n_i - \hat{n}_i)/\sqrt{\hat{n}_i}$. Различие между эмпирическим и предполагаемым теоретическим распределениями можно характеризовать суммой этих отклонений или, чтобы отклонения с разными знаками не компенсировались, суммой их квадратов. Это приводит к величине:

$$\chi^2 = \sum_i \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \tag{5.1}$$

(читается «хи-квадрат») в качестве меры различия между распределениями. Если в рассматриваемой задаче эта величина окажется «слишком большой», то различие надо будет считать значимым. Методы теории вероятностей позволяют при не слишком малом количестве наблюдений найти такие критические значения χ^2_α, которые при справедливости нулевой гипотезы могут превышаться не более чем в $\alpha = 5\%$ случаев или в $\alpha = 1\%$ случаев. Как обычно, нулевая гипотеза отвергается, если $\chi^2 > \chi^2_\alpha$, и принимается, если $\chi^2 \leqslant \chi^2_\alpha$.

Как и для других критериев, критические значения χ^2_α зависят от числа степеней свободы f. Таблица критических значений χ^2_α (f) для уровней значимости 5% и 1% приведена в Приложениях (табл. VI).

О понятии числа степеней свободы см. раздел **3.4.1.** *Ниже используются также понятия и-критерия (см. раздел* **4.2.2**), *одностороннего и двустороннего критериев (см. раздел* **4.1.3**) *и интеграла вероятностей (см. раздел* **2.3.2**).

Когда количество классов группировки велико, число степеней свободы f может оказаться больше предусмотренного в табл. VI Приложений. В таких случаях можно воспользоваться тем, что при большом числе степеней свободы f величина $\sqrt{2\chi^2}$ распределена приближенно нормально с математическим ожиданием $\sqrt{2f-1}$ и стандартным отклонением 1. Поэтому значимость величины χ^2 можно оценить, сравнив величину

$$u = \sqrt{2\chi^2} - \sqrt{2f-1} \tag{5.2}$$

с u_α. В данном случае критерий должен быть односторонним, так как по смыслу χ^2-критерия нулевая гипотеза может отвергаться только тогда, когда значение χ^2 оказывается «слишком большим». Поэтому здесь u_α должно удовлетворять не условию $\Theta(u_\alpha) = 1 - \alpha$, как при двустороннем критерии, а условию $\Phi(u_\alpha) = 1 - \alpha$, где Φ — интеграл вероятностей. Тогда:

$$u^*_{0,05} = 1{,}64, \quad u^*_{0,01} = 2{,}33$$

(см. табл. III Приложений).

Некоторым неудобством формулы (5.2) является то, что значение u получается как малая разность двух больших величин, а

это предъявляет повышенные требования к точности вычислений. Однако если f весьма велико (не менее 200—300), это затруднение можно обойти. Дело в том, что при столь больших значениях f величина χ^2 сама распределена почти нормально с математическим ожиданием f и стандартным отклонением $\sqrt{2f}$. Поэтому можно принять:

$$u = \frac{\chi^2 - f}{\sqrt{2f}}. \tag{5.3}$$

Критерий χ^2 может привести к неправильному результату, если имеются очень малые теоретические частоты: слагаемые в формуле (5.1), в которые $\hat{n}_i$ входят делителями, особенно чувствительны к неточностям в этих делителях именно при малых значениях последних. Практически принимается, что ни одно из значений $\hat{n}_i$ не должно быть меньше 3. Если это условие не выполняется, то приходится прибегать к объединению разрядов.

Так как точность нахождения величины χ^2 значительно зависит от точности в значениях $\hat{n}_i$ (особенно при не очень больших $\hat{n}_i$), то нужно пользоваться неокругленными значениями $\hat{n}_i$. То, что эти частоты не являются целыми числами, не должно нас смущать: это ведь не реальные численности каких-то объектов, а некие аналоги средних значений.

Формулу (5.1) для вычисления χ^2 можно несколько упростить. Именно, если раскрыть скобки в числителе и произвести сокращения, то получится:

$$\chi^2 = \sum_i \frac{n_i^2}{\hat{n}_i} - 2\sum_i n_i + \sum_i \hat{n}_i,$$

но $\sum n_i = \sum \hat{n}_i = n$, так что

$$\chi^2 = \sum_i \frac{n_i^2}{\hat{n}_i} - n. \tag{5.4}$$

Это делает ненужным вычисление разностей $n_i - \hat{n}_i$. Но эта формула имеет тот недостаток, что здесь χ^2 получается как сравнительно малая разность двух больших величин (в то время как при обычном способе вычисления χ^2 является суммой ряда малых величин). Это предъявляет повышенные требования к точности промежуточных вычислений и, в частности, делает невозможным использование логарифмической линейки. Но при наличии арифмометра или клавишной вычислительной машины этот способ вычисления χ^2 предпочтительнее.

ТАБЛИЦА 5.1

Год	n_x	$\hat{n}_x$	$n_x - \hat{n}_x$	(χ^2)
1891	18	23	—5	1,09
1892	23	23	0	0,00
1893	13	22	—9	3,68
1894	25	23	2	0,17
1895	33	24	9	3,38
1896	15	23	—8	2,78
1897	20	23	—3	0,39
1898	24	23	1	0,04
1899	29	24	5	1,04
1900	32	24	8	2,66
Сумма	232	232	0	15,23

Пример 5.1. В табл. 5.1 содержатся данные за 10 лет о числе рождений троен в Швеции. В столбце 3 записаны «теоретические» числа, полученные в предположении, что число рождений троен составляет во все годы одну и ту же часть всех рождений (из-за вариации по годам общего числа рождений значения в этом столбце также несколько варьируют). Расчет дает: $\chi^2 = 15{,}23$.

Теперь найдем число степеней свободы для χ^2. Как мы знаем (см. раздел **3.4.1**), число степеней свободы какой-либо статистики равно числу независимых величин, использованных при вычислении этой статистики, т. е. общему числу таких величин минус число условий, связывающих эти величины. При вычислении χ^2 используются величины разрядов частот n_i, число которых равно k. Но эти частоты связаны здесь условием $\Sigma n_i = n$, так что $f = k - 1 = 10 - 1 = 9$. Поскольку $\chi^2_{0,05}(9) = 16{,}92$, то $\chi^2 < \chi^2_{0,05}$; поэтому гипотеза о постоянстве доли рождений троен не отвергается.

Замена числа разрядов k числом степеней свободы f имеет здесь следующий смысл. Приписывая теоретическому распределению значения характеристик, вычисленные для заданного эмпирического распределения, мы искусственно сближаем оба распределения. Понятно, что это сближение будет тем больше, чем больше параметров эмпирического распределения мы используем для описания теоретического распределения. Например, в случае нормального распределения частоты связаны тремя условиями:

$$\sum n_i = n; \quad \frac{1}{n}\sum n_i x_i = \bar{x}; \quad \frac{1}{n-1}\sum n_i (x_i - \bar{x})^2 = s^2,$$

определяющими те значения характеристик (n, μ и σ), по которым строилось теоретическое нормальное распределение (и вычислялись теоретические частоты $\hat{n}_i$). Поэтому если в качестве теоретического берется нормальное распределение с неизвестными параметрами μ и σ, то $f = k - 3$. (*О нормальном распределении см §. 2.3*).

Пример 5.2. Во втором поколении дигибридного скрещивания у *Primula* получено расщепление по фенотипу 338 *AB*, 122 *Ab*, 67 *aB*, 33 *ab* (*A* — плоские листья, *a* — сморщенные листья; *B* — нормальный глазок, *b* — розовый глазок). Соответствует ли это ожидаемому соотношению 9 : 3 : 3 : 1?

Найдя теоретические численности:

$$9 \cdot \frac{338 + 122 + 67 + 33}{9 + 3 + 3 + 1} = 9 \cdot \frac{560}{16} = 315;$$

$$3 \cdot \frac{560}{16} = 105; \quad 1 \cdot \frac{560}{16} = 35,$$

составляем табл. 5.2. Расчет дает $\chi^2 = 18{,}29$ при $f = 4 - 1 = 3$ (4 численности связаны одним условием, что сумма равна 560).

ТАБЛИЦА 5.2

Признаки	n_i	$\hat{n}_i$	$n_i - \hat{n}_i$	(χ^2)
Плоские листья:				
нормальный глазок	338	315	23	1,68
розовый глазок	122	105	17	2,75
Сморщенные листья:				
нормальный глазок	67	105	—38	13,75
розовый глазок	33	35	—2	0,11
Сумма	560	560	0	18,29

По табл. VI Приложений находим $\chi^2_{0,01}(3) = 11{,}3$; так как $\chi^2 > \chi^2_{0,01}$, то нулевая гипотеза отвергается.

Поскольку оказалось, что фактическое соотношение численностей значимо расходится с ожидаемым соотношением 9 : 3 : 3 : 1, то естественно попытаться выяснить причину такого расхождения. Можно, в частности, предположить, что одна из модификаций какого-либо из признаков связана с меньшей жизнеспособностью растений. Если, например, снижение жизнеспособности растений связано с той или иной окраской глазка, то распределение по этому признаку будет отличаться от соотношения 3 : 1. Объединяя поэтому все растения с одинаковыми окрасками глазка (независимо от вида листьев), мы получаем две группы, записанные в табл. 5.3.

Здесь $f = 2-1 = 1$ и $\chi^2_{0,05}(1) = 3{,}84$. Мы видим, что $\chi^2 < \chi^2_{0,05}$, так что нулевая гипотеза (о том, что окраска глазка не влияет на жизнеспособность растения) не отвергается.

ТАБЛИЦА 5.3

Признаки	n_i	$\widehat{n}_i$	$n_i - \widehat{n}_i$	(χ^2)
Нормальный глазок	405	420	—15	0,54
Розовый глазок	155	140	15	1,61
Сумма	560	560	0	2,15

Остается другое предположение: жизнеспособность растений связана с видом листьев. Для проверки этого предположения мы группируем отдельно растения с плоскими листьями и отдельно растения со сморщенными листьями независимо от окраски глазка. Это дает табл. 5.4.

ТАБЛИЦА 5.4

Признаки	n_i	$\widehat{n}_i$	$n_i - \widehat{n}_i$	(χ^2)
Плоские листья	460	420	40	3,8
Сморщенные листья	100	140	—40	11,4
Сумма	560	560	0	15,2

Полученное число $\chi^2 = 15{,}2$ превышает критическое значение $\chi^2_{0,01}(1) = 6{,}63$, так что нулевая гипотеза отвергается. Следовательно, можно утверждать, что растения со сморщенными листьями отличаются меньшей жизнеспособностью. Это и было причиной (или, во всяком случае, одной из причин) отступления от ожидаемого соотношения 9 : 3 : 3 : 1.

5.1.3. Было бы, однако, ошибкой думать, что чем меньше значение χ^2 получится, тем лучше. Слишком малое значение χ^2, указывающее на «слишком хорошее» согласие, может быть следствием предвзятости при записи результатов (вольной или невольной). Например, при $f = 26$ величина χ^2 лишь в 1% случаев может превысить значение 54,1 (см. табл. VI Приложений); но, как показывает расчет, точно так же лишь в 1% случаев χ^2 может из-за случайных выборочных вариаций оказаться меньше 12,2.

Аналогичные нижние критические значения χ^2_α (как для $\alpha = 1\%$, так и для $\alpha = 5\%$) можно указать и для других чисел степеней свободы. Ввиду того что такая задача встречается сравнительно редко, мы не даем здесь соответствующей таблицы; ее

можно найти, например, в сборниках таблиц Л. Н. Большева и Н. В. Смирнова (с. 228—233) и Я. Янко (с. 115).

Еще раз подчеркнем, что в формулы для χ^2 следует подставлять только частоты, а не величины, получаемые при измерении, взвешивании, отсчете по шкале и т. д. В противном случае можно было бы получать любые значения χ^2 простой заменой масштаба: ведь числитель χ^2 в общем пропорционален второй степени n_i, а знаменатель — первой степени n_i, так что в целом величина χ^2 пропорциональна значениям n_i. В связи с этим может показаться, что и реальное увеличение частот n_i (т. е. увеличение объема выборки) должно вести к росту χ^2 и тем самым к тому, что различие всегда будет становиться значимым. Но это не так — если различие объективно отсутствует, то при увеличении объема выборки разности $n_i - \hat{n}_i$ (которые при объективном отсутствии различия возникают только из-за выборочных случайностей) будут расти медленнее, чем сами n_i, так что порядок величины χ^2 не будет меняться. Аналогичный вопрос уже освещался в разделе 4.2.5.

§ 5.2. Сравнение двух эмпирических распределений

5.2.1. Задача о сравнении двух эмпирических распределений (не просто их средних значений или других статистик, но всего хода кривых распределения) возникает обычно тогда, когда хотят проверить однородность эмпирического материала: если окажется, что две эмпирические совокупности распределены одинаково, то их можно будет считать выборками из одной и той же генеральной совокупности. Тогда их можно объединить в одну общую выборку большего объема, что приведет к сужению доверительных интервалов для параметров. Такая проверка особенно важна, если желают объединить данные разных авторов.

Эта задача также решается χ^2-критерием. Но нахождение величины χ^2 должно в данном случае производиться несколько иначе, чем при сравнении эмпирического распределения с теоретическим. Дело в том, что два выборочных распределения могут по-разному отклоняться от генерального распределения. Поэтому требование к близости двух эмпирических распределений должно быть «примерно вдвое» менее жестким, чем к близости эмпирического и теоретического распределений; это означает, что метод расчета χ^2 должен быть таким, чтобы при тех же расхождениях между частотами в первом случае получалось меньшее значение χ^2, чем во втором.

Пример 5.3. В табл. 5.5 (столбцы n_i и n_i') приведены два эмпирических распределения, в отношении которых предполагается, что они относятся к выборкам из одной и той же генеральной совокупности.

ТАБЛИЦА 5.5

x_i	n_i'	n_i''	$\hat{n}_i$
2	5	9	7
4	6	4	5
6	8	10	9
8	10	6	8
10	6	6	6
Сумма	35	35	35

Частоты генерального распределения, от которого могут случайно отклоняться частоты обоих эмпирических распределений, нам неизвестны. Однако в качестве первого приближения можно принять полусумму эмпирических частот, относящихся к одинаковым значениям вариант (т. е. средние значения этих частот); это даст числа, записанные в четвертом столбце таблицы.

Здесь рассматривается случай, когда объемы двух сравниваемых выборок одинаковы. О выборках разного объема см. § 5.3.

Теперь мы можем составить табл. 5.6. В первом столбце этой таблицы мы запишем все имеющиеся в нашем распоряжении эмпирические данные (частоты) из обоих заданных рядов, а во втором — соответствующие теоретические частоты; после этого обычным образом вычисляется χ^2.

ТАБЛИЦА 5.6

n_i	$\hat{n}_i$	$n_i - \hat{n}_i$	(χ^2)
5	7	−2	0,57
9	7	+2	0,57
6	5	+1	0,20
4	5	−1	0,20
8	9	−1	0,11
10	9	+1	0,11
10	8	+2	0,50
6	8	−2	0,50
6	6	0	0,00
6	6	0	0,00
70	70	0	2,76

Отметим теперь, что менее жесткое требование к близости двух эмпирических распределений (по сравнению со случаем близости между эмпирическим и теоретическим распределением, о чем упо-

миналось выше) сказывается в следующем. С одной стороны, в χ^2 входят не разности двух эмпирических частот, а полуразности (ибо входят отклонения частот от полусумм); так как в χ^2 входят не сами отклонения, а их квадраты, то в рассматриваемом случае значение χ^2 оказывается в среднем в 4 раза меньше, чем при сравнении эмпирического и теоретического распределений. В то же время число слагаемых, из которых состоит χ^2, вдвое больше, чем число разрядов группировки, т. е. каждое слагаемое $(n_i - \hat{n}_i)^2/\hat{n}_i$ входит в χ^2 дважды. Поэтому в общем значение χ^2 получается примерно вдвое меньше, если сравнивают не эмпирическое и теоретическое, а два эмпирических распределения.

5.2.2. Вычисление χ^2 можно несколько упростить. Дело в том, что величина χ^2 состоит из k слагаемых вида:

$$\frac{(n_i' - \hat{n}_i)^2}{\hat{n}_i} + \frac{(n_i'' - \hat{n}_i)^2}{\hat{n}_i}, \qquad (*)$$

где $\hat{n}_i = \frac{1}{2}(n_i' + n_i'')$. Если подставить это значение $\hat{n}_i$ в (*), то мы получим после несложных преобразований:

$$\frac{(n_i' - n_i'')^2}{2(n_i' + n_i'')} + \frac{(n_i'' - n_i')^2}{2(n_i' + n_i'')} = \frac{(n_i' - n_i'')^2}{n_i' + n_i''}.$$

Поэтому:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i' - n_i'')^2}{n_i' + n_i''}. \qquad (5.5)$$

Расчет χ^2 по этой формуле для примера 5.3 показан в табл. 5.7.

ТАБЛИЦА 5.7

n_i'	n_i''	$n_i' - n_i''$	$n_i' + n_i''$	$\frac{(n_i' - n_i'')^2}{n_i' + n_i''}$
5	9	—4	14	1,14
6	4	2	10	0,40
8	10	—2	18	0,22
10	6	4	16	1,00
6	6	0	12	0,00
35	35	0	70	2,76

Теперь определим число степеней свободы. На пять чисел (частот) последнего столбца в табл. 5.5 наложено одно условие, что сумма их должна равняться 35; это дает нам $5-1 = 4$ степени свободы.

Таким образом, $\chi^2 = 2{,}76$, $f = 4$. В табл. V Приложений находим $\chi^2_{0,05}(4) = 9{,}49$. Так как $\chi^2 < \chi^2_{0,05}$, то нулевая гипотеза не отвергается — обе эмпирические совокупности можно считать выборками из одной генеральной совокупности.

Как уже указывалось в начале настоящей главы, критерий хи-квадрат применять нельзя, если объемы выборок недостаточно велики (они должны содержать не менее 20—30 вариант).

§ 5.3. Сравнение выборок разного объема

5.3.1. В предыдущем параграфе было показано, как сравниваются эмпирические совокупности, имеющие одинаковый объем. На практике последнее условие выполняется далеко не всегда — чаще всего опытная и контрольная группы содержат различное число вариант n' и n''.

В этом случае в качестве «теоретических» частот нельзя брать полусуммы $\frac{1}{2}(n'_i + n''_i)$, так как сумма этих частот, т. е. объем «теоретической» совокупности, будет равна $\frac{1}{2}(n' + n'')$, что отличается и от n', и от n''; между тем сравниваемые эмпирическое и теоретическое распределения должны иметь одинаковый объем. Поэтому поступают так: для каждого из рядов «теоретическими» частотами считают величины $n_i + n_i$, умноженные на такой множитель, чтобы после суммирования получалась численность данного эмпирического ряда. Поясним это на примере.

Пример 5.4. В табл. 5.8 приведены два распределения (смысл значений x нас сейчас не интересует), из которых одно имеет объем $n' = 30$, а второе — объем $n'' = 40$; нужно проверить гипотезу, что это выборки из одной генеральной совокупности.

ТАБЛИЦА 5.8

x_i	n'_i	n''_i	$n'_i + n''_i$	$\widehat{n}'_i$	$\widehat{n}''_i$
(1)	(2)	(3)	(4)	(5)	(6)
1	5	6	11	4,7	6,3
2	8	7	15	6,4	8,6
3	6	10	16	6,9	9,1
4	11	17	28	12,0	16,0
Сумма	30	40	70	30,0	40,0

В столбце 4 записаны построчные суммы частот: $5 + 6 = 11$; $8 + 7 = 15$; $6 + 10 = 16$; $11 + 17 = 28$. Их сумма равна 70, т. е. она в 70/30 раз больше численности первого ряда и в 70/40 раз больше численности второго ряда. Если для первого эмпирического ряда считать «теоретическими» частотами:

$$11 \cdot \frac{30}{70} = 4{,}7; \quad 15 \cdot \frac{30}{70} = 6{,}4; \quad 16 \cdot \frac{30}{70} = 6{,}9; \quad 28 \cdot \frac{30}{70} = 12{,}0,$$

записанные в столбце 5, то их сумма 30,0 совпадает с суммой частот из столбца 2. Аналогично величины

$$11 \cdot \frac{40}{70} = 6{,}3; \quad 15 \cdot \frac{40}{70} = 8{,}6; \quad 16 \cdot \frac{40}{70} = 9{,}1; \quad 28 \cdot \frac{40}{70} = 16{,}0,$$

записанные в столбце 6, дадут в сумме 40,0, что совпадает с суммой частот из столбца 3.

Таким образом, числа внутри каждого из столбцов (5) и (6) относятся между собой как числа в столбце (4), т. е. как 11 : 15 : 16 : 28, и в то же время суммы их (30 и 40) равны суммам эмпирических частот в столбцах (2) и (3).

Проверка вычислений состоит в том, что в каждой строке сумма чисел из столбцов (5) и (6) должна равняться числу из столбца (4), а именно: $4{,}7 + 6{,}3 = 11$; $6{,}4 + 8{,}6 = 15$; $6{,}9 + 9{,}1 = 16$, $12{,}0 + 16{,}0 = 28$.

Теперь мы составляем новую таблицу (табл. 5.9), выписывая рядом эмпирические и «теоретические» частоты, а затем, как обычно, вычисляем χ^2.

ТАБЛИЦА 5.9

n_i	$\widehat{n}_i$	$n_i - \widehat{n}_i$	(χ^2)
5	4,7	0,3	0,02
6	6,3	—0,3	0,01
8	6,4	1,6	0,40
7	8,6	—1,6	0,30
6	6,9	—0,9	0,12
10	9,1	0,9	0,09
11	12,0	—1,0	0,08
17	16,0	1,0	0,06
70	70,0		1,08

5.3.2. Как и в случае равных объемов, здесь можно получить упрощенную формулу для вычислений χ^2. Подставив в выражение (*) из § 5.2 значения n_i и n_i, которые имеют вид:

$$\widehat{n}_i' = n' \frac{n_i' + n_i''}{n' + n''}, \qquad \widehat{n}_i'' = n'' \frac{n_i' + n_i''}{n' + n''},$$

получим после простых преобразований формулу:

$$\chi^2 = \frac{1}{n'n''} \sum_{i=1}^{k} \frac{\left(n'_i n'' - n''_i n'\right)^2}{n'_i + n''_i}; \qquad (5.6)$$

очевидно, при $n' = n''$ эта формула переходит в формулу (5.5).

В табл. 5.10 показано вычисление χ^2 по формуле (5.6) для примера 5.4.

ТАБЛИЦА 5.10

n'_i	n''_i	$n'_i + n''_i$	$n'_i n''$	$n''_i n'$	$n'_i n'' - n''_i n'$	$\frac{\left(n'_i n'' - n''_i n'\right)^2}{n'_i + n''_i}$
5	6	11	200	180	20	36
8	7	16	320	210	110	807
6	10	16	240	300	—60	225
11	17	28	440	510	—70	175
30	40					1243

Результат

$$\chi^2 = \frac{1243}{30 \cdot 40} = 1.04$$

практически совпадает с полученным ранее значением $\chi^2 = 1{,}08$; небольшое расхождение объясняется округлениями, допущенными в табл. 5.8 и 5.9.

Как и в случае выборок одинакового объема, применение критерия хи-квадрат требует, чтобы объемы выборок не были очень малы (они должны содержать не менее 20—30 вариант).

Для специального случая, когда оба сравниваемых распределения содержат только две группы вариант (так называемые альтернативные распределения), имеются более простые критерии (см. § 6.4).

5.3.3. Если имеется несколько совокупностей, о которых предполагается, что все они являются выборками из одной генеральной совокупности, то для проверки этой гипотезы нет необходимости сравнивать их попарно во всех сочетаниях — можно произвести однократную совместную проверку. Такой случай мы имеем в табл. 3.5, рассматривавшейся в § 3.6 (эти данные повторены в табл. 5.11).

Прежде всего надо найти «теоретические» численности, отвечающие заданным эмпирическим частотам. Если бы все четыре

выборки были взяты из одной генеральной совокупности и выборочные вариации отсутствовали бы, то соотношение частот в каждой из них было бы одинаковым — таким же, как в генеральной совокупности. Поскольку соотношение частот в генеральной совокупности нам неизвестно, мы будем считать, что наилучшей его оценкой является соотношение в сводной выборке,— исходя из того, что поскольку она больше по объему, то она репрезентативнее.

ТАБЛИЦА 5.11

Диаметр эритроцита, мк	Первый мазок	Второй мазок	Третий мазок	Четвертый мазок	Сводная выборка
6	12	9	11	17	49
7	54	48	32	62	196
8	183	204	191	219	797
9	96	66	74	97	333
10	31	24	29	36	120
11	16	13	8	17	54
Сумма	392	364	345	448	1549

Поэтому, например, число эритроцитов с диаметром 9 мк в первом мазке должно было бы так относиться к объему этой выборки 392, как число 333 из сводной выборки относится к ее общей численности 1549:

$$\frac{\widehat{n}_9^{(1)}}{392} = \frac{333}{1549},$$

откуда

$$\widehat{n}_9^{(1)} = 392\,\frac{333}{1549} = 84.$$

Аналогично находим все остальные $n_i^{(j)}$; в общем виде

$$\widehat{n}_i^{(j)} = n^{(j)}\,\frac{n_i}{\sum n_i}. \qquad (5.7)$$

Теперь составляем табл. 5.12. В каждой клетке записано три числа. Первое есть эмпирическая численность $n_i^{(j)}$; второе — «теоретическая» численность $\widehat{n}_i^{(j)}$, найденная по формуле (5.7); третье число — разность частот $n_i^{(j)} - \widehat{n}_i^{(j)}$.

ТАБЛИЦА 5.12

x_i	$n_i^{(1)}$	$n_i^{(2)}$	$n_i^{(3)}$	$n_i^{(4)}$	n_i
6	12 12 0	9 12 —3	11 11 0	17 14 3	49 0
7	54 50 4	48 46 2	32 44 —12	62 56 6	196 0
8	183 202 —19	204 187 17	191 177 14	219 231 —12	797 0
9	96 84 12	66 78 —12	74 74 0	97 97 0	333 0
10	31 30 1	24 28 —4	29 27 2	36 35 1	120 0
11	16 14 2	13 13 0	8 12 —4	17 15 2	54 0
Сумма	392	364	345	448	1549

Величину χ^2 вычисляем обычным образом по формуле (5.1):

$$0^2:12=0{,}00 \quad 4^2:50=0{,}32 \quad (-19)^2:202=1{,}79 \quad 12^2:84=1{,}72 \quad 1^2:30=0{,}03 \quad 2^2:14=\frac{0{,}29}{4{,}15}$$

$$(-3)^2:12=0{,}75 \quad 2^2:46=0{,}09 \quad 17^2:187=1{,}54 \quad (-12)^2:78=1{,}85 \quad (-4)^2:28=0{,}57 \quad 0^2:13=\frac{0{,}00}{4{,}80}$$

$$0^2:11=0{,}00 \quad (-12)^2:44=3{,}28 \quad 14^2:177=1{,}11 \quad 0^2:74=0{,}00 \quad 2^2:27=0{,}15 \quad (-4)^2:12=\frac{1{,}33}{5{,}87}$$

$$3^2:14=0{,}64 \quad 6^2:56=0{,}64 \quad (-12)^2:231=0{,}62 \quad 0^2:97=0{,}00 \quad 1^2:35=0{,}03 \quad 2^2:15=\frac{0{,}27}{2{,}20}$$

$$\chi^2=4{,}15+4{,}80+5{,}87+2{,}20=17{,}02.$$

Что касается числа степеней свободы, то оно находится здесь следующим образом. Очевидно, все частоты в $k-1$ первых строках и в $m-1$ первых столбцах могут быть произвольными; оставшаяся же в каждой строке частота будет определяться одно-

значно из итоговой частоты в этой строке, а оставшаяся в каждом столбце — из итоговой частоты по столбцу. Поэтому число степеней свободы равно:

$$f = (k - 1)(m - 1). \tag{5.8}$$

То же получится из подсчета связей между частотами. Действительно, каждая строка и каждый столбец налагают по одной связи (сумма частот должна равняться итоговой частоте), причем число связей $k + m$ надо уменьшить на единицу, поскольку сумма итогов по строкам равна сумме итогов по столбцам (т. е. общему итогу n). Таким образом:

$$f = km - (k + m - 1) = (k - 1)(m - 1).$$

При сравнении двух совокупностей $m = 2$, так что:

$$f = k - 1,$$

что мы имели и ранее.

В нашем примере $k = 6$, $m = 4$, поэтому:

$$f = (6 - 1)(4 - 1) = 5 \cdot 3 = 15.$$

Так как $\chi^2_{0,05}(15) = 25{,}0$, а фактически получилось $\chi^2 = 17{,}0$, то нулевая гипотеза не отвергается.

Когда количество классов группировки и число сравниваемых рядов велико, число степеней свободы (5.8) может оказаться больше предусмотренного в табл. VI Приложений. В таких случаях применяют приближения, описанные в разделе 5.1.2.

Глава шестая

АЛЬТЕРНАТИВНОЕ РАСПРЕДЕЛЕНИЕ

§ 6.1. Статистические задачи при альтернативном распределении. Биномиальное распределение

6.1.1. *Альтернативное распределение* есть распределение элементов совокупности на две части (две альтернативы) по какому-либо признаку — чаще всего по качественному. Очевидно, в случае качественной классификации невозможно ввести такие количественные параметры, как математическое ожидание, дисперсия и др. Однако, можно указать тем не менее определенный численный параметр, имеющий вполне точный и объективный смысл: долю вариант одного из двух типов. Если, например, имеется совокупность из 18 заболевших и 26 незаболевших животных, то доля заболевших животных будет:

$$p = \frac{18}{18+26} = 0{,}41, \text{ или } 41\%.$$

Вообще:

$$p = \frac{n_1}{n_1 + n_2} = \frac{n_1}{n}, \qquad (6.1)$$

где n_1 и n_2 — численности альтернатив, а $n = n_1 + n_2$ есть численность всей совокупности.

В отношении доли вариант в альтернативном распределении возникают те же статистические задачи, что и для параметров непрерывных распределений:

1) оценка доли $\hat{p}$ в генеральной совокупности по выборочным данным, нахождение доверительного интервала для $\hat{p}$;

2) проверка гипотез о величине $\hat{p}$, т. е. сравнение эмпирической доли p с какой-либо предполагаемой теоретической $\hat{p}$;

3) выявление различия между двумя генеральными долями $\hat{p}_1$ и $\hat{p}_2$ по выборочным данным, т. е. сравнение двух выборочных долей вариант.

Более подробно о сущности этих задач говорится в разделах **3.7.1**, **4.1.1** *и* **4.2.1.** *Настоятельно рекомендуем просмотреть эти разделы, прежде чем читать дальше.*

6.1.2. Решение перечисленных задач основывается на использовании свойств так называемого *биномиального распределения* (или *распределения Бернулли*) и различных его аппроксимаций (т. е. приближенных представлений). Биномиальное распределение тесно связано с альтернативным распределением, точнее — эти два распределения описывают разные стороны одного и того же явления.

Пусть в урне содержатся в определенной пропорции элементы двух типов, например белые и черные шары, и пусть доля элементов одного типа (например, белых шаров) равна $\hat{p}$. Очевидно, мы имеем здесь альтернативное распределение:

альтернатива:	*белые*	*черные*
доля:	$\hat{p}$	$1-\hat{p}$

Пусть далее из урны извлечены случайно n шаров[1], и в этой выборке оказалось m_1 белых шаров. Затем случайно извлечены еще n шаров, среди которых оказалось m_2 белых шаров, и т. д. Заранее очевидно, что в принципе число извлекаемых белых шаров (в серии из n извлеченных шаров) может быть любым от 0 до n. (Однако ясно также, что более вероятно появление $\hat{p}n$ белых шаров, чем какого-либо другого их числа: например, если $\hat{p} = 0{,}5$ и $n = 10$, то более вероятно появление пяти белых шаров, чем ни одного или десяти, хотя и такие возможности не исключены.)

Статистические задачи альтернативного распределения, которые были сформулированы выше, состоят здесь в том, чтобы по о д н о й выборке оценить $\hat{p}$ (найти его доверительный интервал), оценить значимость отличия этой оценки от некоторой гипотетической доли p_0 или от другой оценки, полученной в аналогичном эксперименте с другой урной. Нетрудно понять, что решению этих задач может помочь знание распределения чисел m_1, m_2, ... в последовательных выборках. Это распределение и есть *биномиальное* (происхождение этого названия сейчас станет ясным).

К описанию биномиального распределения мы и переходим. Выясним сначала, какова вероятность того, что белыми окажутся о п р е д е л е н н ы е m извлеченных шаров, т. е. с заранее заданными номерами в порядке извлечения всех n шаров. Чтобы такое событие произошло, должны одновременно осуществиться следующие n событий: m шаров с заданными номерами должны оказаться белыми (вероятность каждого такого события равна $\hat{p}$),

[1] Предполагается, что после того, как шар извлечен и замечен его цвет, этот шар возвращается в урну, так что доля $\hat{p}$ белых шаров в урне восстанавливается в точности.

а остальные $n - m$ шаров должны оказаться черными (вероятность каждого такого события равна $1 - \hat{p}$). Так как все последовательные извлечения независимы, то вероятность того, что заданные m шаров будут белыми, а остальные $n - m$ черными, равна $\hat{p}^m(1 - \hat{p})^{n-m}$. Так, если из урны, в которой белые шары составляют $^1/_4$ всех шаров, извлекается серия из 10 шаров, то вероятность того, что белыми окажутся шары с номерами 2, 7 и 9, а остальные 7 шаров (с номерами 1, 3, 4, 6, 8 и 10) окажутся черными, будет

$$\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^7 = \frac{3^7}{4^{10}} = \frac{2187}{1048576} \approx 0{,}002.$$

Теперь изменим немного задачу: нас будет интересовать появление белых шаров не обязательно под заданными номерами, а любых m штук. В рассмотренном примере — не обязательно номеров 2, 7 и 9, а любой тройки шаров. Очевидно, вероятность такого события будет больше во столько раз, сколько различных троек шаров можно составить из 10 шаров. Как известно, число таких троек есть число сочетаний из 10 элементов по три, обозначаемое C_{10}^3 и равное:

$$C_{10}^3 = \frac{10 \cdot 9 \cdot 8}{1 \cdot 2 \cdot 3} = 120.$$

Поэтому вероятность того, что при извлечении из урны 10 шаров какие-нибудь (любые) три из них окажутся белыми, будет равна:

$$\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^7 C_{10}^3 = 0{,}002 \cdot 120 = 0{,}24.$$

В общем виде формула для вероятности того, что при извлечении n шаров m из них будут белыми, имеет вид:

$$w_m = C_n^m \hat{p}^m \left(1 - \hat{p}\right)^{n-m} = C_n^m \hat{p}^m \hat{q}^{n-m}, \qquad (6.2)$$

где $\hat{q} = 1 - \hat{p}$.

Суммирование по всем возможным значениям m (очевидно, от 0 до n) должно дать полную вероятность, т. е. 1. И действительно,

$$\sum_{m=0}^{n} C_n^m p^m q^{n-m} = 1,$$

так как эта сумма есть не что иное, как развернутая формула бинома Ньютона:

$$\left(\hat{p} + \hat{q}\right)^n = \sum_{m-0}^{n} C_n^m p^m q^{n-m}, \qquad (*)$$

но левая часть равенства (*) равна 1, так как $\hat{p} + \hat{q} = 1$.

Если произведено N серий извлечения шаров из урны (по n шаров в каждой серии), то следует ожидать, что число таких серий, в которых окажется m белых шаров, будет близко к величине:

$$N_m = Nw_m = NC_n^m \hat{p}^m \hat{q}^{n-m}, \tag{6.3}$$

Распределение вероятностей (6.2) или частот (6.3) называется биномиальным. Сопоставление формул (*) и (6.2) делает понятным происхождение этого названия. Биномиальные коэффициенты в общем виде записывают так:

$$C_n^m = \frac{n(n-1)(n-2)\ldots(n-[m-1])}{m!} = \frac{m!\,(n-m)!}{n!}. \tag{6.4}$$

Знак «!» есть так называемый факториал: величина $k!$ есть произведение вида:

$$1 \cdot 2 \cdot 3 \ldots (k-2)(k-1)k.$$

Например, $3! = 1 \cdot 2 \cdot 3 = 6$, $5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$ и т. д. Легко видеть, что при возрастании k число $k!$ растет очень быстро — быстрее, чем показательная функция. Очевидно,

$$k! = (k-1)!\,k,$$

откуда:

$$(k-1)! = \frac{k!}{k}. \tag{**}$$

Так как $1! = 1$, то в соответствии с (**) естественно считать

$$0! = \frac{1!}{1} = \frac{1}{1} = 1,$$

хотя это и выглядит несколько парадоксально.

6.1.3. Особый интерес представляет случай, когда вероятности двух альтернативных исходов одинаковы: $\hat{p} = \hat{q} = 1/2$. Этот случай имеет место, например, при бросании монеты, при рождении особи мужского или женского пола (в первом приближении) и т. д.

Если $\hat{p} = \hat{q} = 1/2$, то:

$$\hat{p}^m \hat{q}^{n-m} = \left(\frac{1}{2}\right)^m \left(\frac{1}{2}\right)^{n-m} = \left(\frac{1}{2}\right)^n = \frac{1}{2^n},$$

так что:

$$N_m = \frac{N}{2^n} C_n^m. \tag{6.5}$$

Пример 6.1. В колонках 1 и 2 табл. 6.1 содержатся данные о числе петушков в 80 выводках по 12 цыплят в каждом. Так, в одном выводке совсем не было петушков (число петушков равно нулю), в шести выводках было по 3 петушка, в 16 выводках — по 7 петушков и т. д.

Посмотрим, можно ли это распределение считать биномиальным. Для этого нужно по формуле (6.4) найти биномиальные коэффициенты C_n^m для разных m при $n = 12$, после чего по формуле (6.5) вычислить значения N_m. Например:

$$C_{12}^3 = \frac{12 \cdot 11 \cdot 10}{1 \cdot 2 \cdot 3} = 220; \quad N_3 = \frac{80 \cdot 220}{4096} \approx 4{,}3.$$

Значения C_{12}^m и N_m записаны соответственно в колонках 3 и 4 табл. 6.1; значения N_m получены путем умножения C_{12}^m на постоянный множитель:

$$\frac{N}{2^{12}} = \frac{80}{4096} = 0{,}0195.$$

Сравнивая теоретические и эмпирические частоты, видим, что хотя соответствие является довольно относительным, общий характер распределения передается все же правильно. Более объективно можно оценить совпадение при помощи критерия хи-квадрат (см. § 5.1).

ТАБЛИЦА 6.1

Число петушков	Фактическое число выводков	Биномиальные коэффициенты	Теоретическое число выводков
0	1	1	0,0
1	0	12	0,2
2	0	66	1,3
3	6	220	4,3
4	11	495	9,7
5	13	792	15,5
6	19	924	18,0
7	16	792	15,5
8	7	495	9,7
9	4	220	4,3
10	2	66	1,3
11	1	12	0,2
12	0	1	0,0
Сумма	80	4096	80,0

6.1.4. Пусть мы имеем некоторое биномиальное распределение с определенным значением параметра n, скажем $n = 10$. На

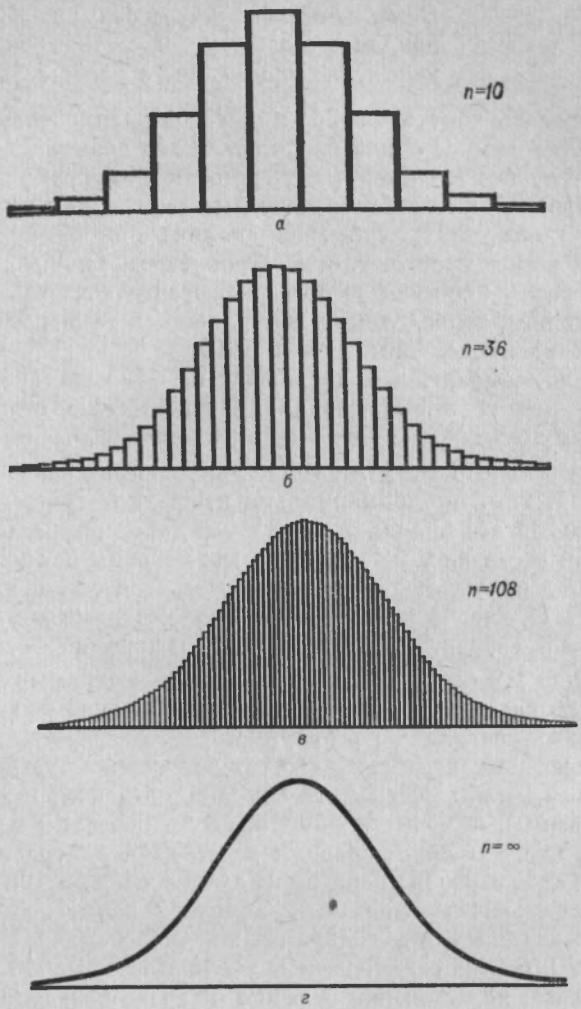


Рис. 6.1. Переход от биномиального распределения к нормальному.

рис. 6.1, *а* изображена диаграмма такого распределения при $\hat{p} = 1/2$; столбики этой диаграммы могут изображать, например, вероятности выпадения герба на x монетах при бросании горсти из $n = 10$ монет или вероятности рождения x самцов в пометах из 10 особей и т. д. За начало отсчета величин x примем значение $x_0 = n\hat{p} = 5$, при котором вероятность максимальна, т. е. будем отсчитывать x влево и вправо от x_0.

В дальнейших рассуждениях используются понятия дисперсии и стандартного отклонения, о которых см. в разделе **3.3.1.**

Если бросать горсть из 36 монет, то распределение вероятностей будет иным. С одной стороны, число возможных значений x увеличится, так что новая ступенчатая фигура будет шире; с другой стороны, она станет в общем ниже, так как площадь этой фигуры должна остаться без изменения: она ведь изображает сумму всех вероятностей, т. е. 1. Произведем теперь деформацию новой диаграммы: сожмем ее по горизонтали так, чтобы стандартное отклонение распределения изображалось таким же отрезком, как и для распределения при $n = 10$.

Расчет показывает, что дисперсия σ^2 биномиального распределения примерно пропорциональна показателю степени бинома n. Поэтому стандартное отклонение σ изменяется при изменении n приблизительно как $\sqrt{n}$. Поскольку ширина распределения равна $n + 1$, т. е. пропорциональна n, то гистограмма, получившаяся в результате произведенной нами деформации, будет шире, чем диаграмма на рис. 6.1, *а*. Если мы захотим сохранить прежнюю ширину рисунка, то на нем поместится лишь часть новой диаграммы. В нашем примере ширина диаграммы увеличилась в $37 : 11 \approx 3{,}4$ раза, а стандартное отклонение — примерно в $\sqrt{36} : \sqrt{10} \approx 1{,}9$ раза. Так как мы сжали диаграмму в 1,9 раза, то рисунок вместил только $1{,}9 : 3{,}4 \approx 0{,}58$ всей диаграммы; ее «хвосты» не поместились на рисунке.

После того как мы произведем также надлежащую деформацию по вертикали (чтобы обеспечить сохранение неизменной площади всей диаграммы), получится фигура, изображенная на рис. 6.1, *б*; повторяем, что это лишь средняя часть всей диаграммы.

На рис. 6.1, *в* изображена диаграмма для $n = 108$, преобразованная аналогичным образом. Точнее говоря, это опять-таки лишь центральная часть диаграммы: ведь по сравнению со случаем $n = 10$ ширина диаграммы увеличилась в $109 : 11 \approx 10$ раз, между тем как сужение графика было произведено во столько раз, во сколько увеличилось стандартное отклонение, т. е. лишь примерно в $\sqrt{108} : \sqrt{10} = 3{,}2$ раза. Следовательно, непоместившиеся на рисунке «хвосты» диаграммы здесь еще больше: поместившаяся на рисунке центральная часть составляет лишь $3{,}2 : 10 = 0{,}32$ всей диаграммы (но только по ширине; по площади же эта центральная часть содержит подавляющую долю всей вероятности). Сравнение рис. 6.1. *а*, *б* и *в* показывает, что общий характер распределения во всех трех случаях одинаков, но это распределение тем больше детализировано, чем больше значение n.

Если продолжить эту процедуру, беря все большие значения n, то в пределе при $n \to \infty$ верхние стороны столбиков сольются

в гладкую кривую (рис. 6.1, *г*). При этом «хвосты» графика распространяются неограниченно далеко по обе стороны от центра. Получившееся предельное распределение называется *нормальным*, или *гауссовым* (*распределение Гаусса*). Это распределение обычно получается при совместном воздействии ряда малых независимых (значит, случайно сочетающихся) факторов, число которых неограниченно велико. Такое условие (одновременное воздействие большого числа малых по сравнению с общей суммой факторов) выполняется в природе очень часто. Поэтому гауссово распределение и принято называть нормальным. Однако если какой-либо из факторов, не подчиняющийся сам нормальному распределению, играет преобладающую роль, то распределение не будет гауссовым; так как такой случай тоже может иметь место, то ясно, что нормальное распределение не следует считать универсальным.

Свойства нормального распределения рассмотрены в § 2.3.

§ 6.2. Доверительный интервал для доли (процента) вариант

Прежде чем читать этот параграф, надо обязательно познакомиться с понятиями дисперсии и стандартного отклонения (см. раздел **3.3.1***), стандартной ошибки (см. раздел* **3.5.1***) и доверительного интервала (см. раздел* **3.7.1***)*

6.2.1. Когда рассматриваемая совокупность является выборкой из некоторой бесконечной генеральной совокупности, то величина

$$p = \frac{n_1}{n_1 + n_2} = \frac{n_1}{n} \qquad (6.6)$$

будет выборочной оценкой генеральной доли $\hat{p}$. При этом естественно возникает вопрос об определении доверительного интервала для $\hat{p}$.

Строго эта задача решается с использованием биномиального распределения (6.2). Соответствующие расчеты очень громоздки, поэтому были составлены таблицы, в которых можно сразу найти 95%-ные и 99%-ные доверительные границы для $\hat{p}$ при заданных значениях n_1 и n_2: такие таблицы имеются, например, в сборнике таблиц Я. Янко (табл. 28, с. 181—194), Л. Н. Большева и Н. В. Смирнова (табл. 52, с. 348—359).

Ввиду большого объема этих таблиц они обычно приводятся лишь в специальных справочниках и не всегда имеются под рукой. Поэтому часто пользуются различными приближенными методами. Чаще всего применяется нормальное приближение (т. е.

замена биномиального распределения нормальным), при котором доверительные границы для $\hat{p}$ вычисляют по формулам:

$$p_н = p - u_p \sigma_p; \quad p_в = p + u_p \sigma_p. \tag{6.7}$$

Что касается стандартной ошибки σ_p, то ее можно найти так. Введем чисто формально некую условную шкалу, приписав одной из альтернатив значение $x_1 = 1$, а другой — значение $x_2 = 0$. Тогда заданное распределение запишется в виде:

x_i	1	0	Итого
n_i	n_1	n_2	n

так что чисто формально получится:

$$\bar{x} = \frac{1}{n}(n_1 x_1 + n_2 x_2) = \frac{1}{n}(n_1 \cdot 1 + n_2 \cdot 0) = \frac{n_1}{n}.$$

Но $n_1/n = p$; значит, в данной модели p имеет смысл $\bar{x}$, а поэтому σ_p можно найти, вычислив $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Стандартное отклонение σ является характеристикой генеральной совокупности. Продолжая наше формальное рассмотрение, при котором $x_1 = 1$ и $x_2 = 0$, но относя его на этот раз уже к генеральной совокупности, можно вычислить для нашего условного распределения:

$$\sigma^2 = \mathrm{M}\{[x - \mathrm{M}(x)]^2\} = \frac{1}{N}\left\{N_1\left(x_1 - \hat{p}\right)^2 + N_2\left(x_2 - \hat{p}\right)^2\right\} =$$

$$= \hat{p}\left(1 - \hat{p}\right)^2 + \left(1 - \hat{p}\right)\hat{p}^2 = \hat{p}\left(1 - \hat{p}\right),$$

так что:

$$\sigma = \sqrt{\hat{p}\left(1 - \hat{p}\right)}.$$

Это дает:

$$\sigma_p = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\hat{p}\left(1 - \hat{p}\right)}{n}}. \tag{6.8}$$

Генеральная доля $\hat{p}$ чаще всего неизвестна, поэтому, если располагают только данными о выборке, величину $\hat{p}$ заменяют ее выборочной оценкой p, считая приближенно:

$$\sigma_p \approx \sqrt{\frac{p(1-p)}{n}}; \tag{6.9}$$

если доля выражена в процентах, то:

$$\sigma_p \approx \sqrt{\frac{p(100-p)}{n}}. \qquad (6.9')$$

Пример 6.2. Из обследованных 430 случайно выбранных колосьев пшеницы 37 оказались пораженными головней. Каковы 95%-ные доверительные границы процента пораженности для данного поля?

Выборочный средний процент пораженности составляет:

$$p = \frac{37}{430} = 0{,}086 = 8{,}6\%.$$

Теперь по формуле (6.9′) находим:

$$\sigma_p = \sqrt{\frac{8{,}6 \cdot 91{,}4}{430}} \approx 1{,}35\%.$$

Тогда для 95%-ного доверительного уровня имеем доверительные границы:

$$p \pm 1{,}96\sigma_p = (8{,}6 \pm 2{,}6)\ \%,$$

т. е. 95%-ный доверительный интервал есть (6,0 ÷ 11,2)%.

Пример 6.3. При рентгеновском облучении 10 мышей дозой в 550 Р погибло 5 мышей. Каковы 99%-ные доверительные границы для доли мышей, погибающих под действием данной дозы облучения?

Имеем:

$$p = \frac{5}{10} = 0{,}5; \quad \sigma_p = \sqrt{\frac{0{,}5 \cdot 0{,}5}{10}} = 0{,}158;$$

поэтому при $P = 99\%$ (и $u_P = 2{,}58$) доверительные границы будут:

$$0{,}5 - 2{,}58 \cdot 0{,}158 = 0{,}5 - 0{,}408 = 0{,}092 = 9{,}2\%,$$

$$0{,}5 + 2{,}58 \cdot 0{,}158 = 0{,}5 + 0{,}408 = 0{,}908 = 90{,}8\%.$$

Так как найденный доверительный интервал перекрывает почти весь возможный диапазон расположения истинной доли погибающих мышей (от 0 до 100%), то следует заключить, что опыт вообще не дал почти никакого результата (кроме указания, что при данной дозе облучения выборка из 10 мышей недостаточно велика для нахождения ответа на поставленный вопрос).

6.2.2. При малых объемах выборок нормальное приближение дает слишком неточные результаты, особенно вследствие замены в формуле для $\sigma_{\hat{p}}$ неизвестной доли $\hat{p}$ ее оценкой p — ведь разли-

чие между p и p равно в среднем как раз величине σ_p, которая, согласно формуле (6.8), возрастает при уменьшении n.

Для исправления этого положения Р. Фишер предложил пользоваться вспомогательной величиной φ, связанной с p равенством:

$$p = \sin^2 \frac{\varphi}{2}, \tag{6.10}$$

откуда:

$$\varphi = 2 \, arc \, sin \sqrt{p}. \tag{6.11}$$

Эта величина, как показал Фишер, имеет распределение, близкое к нормальному; особенно удобно то, что ее стандартная ошибка зависит только от объема выборки n, причем очень простым образом:

$$\sigma_\varphi \approx \frac{1}{\sqrt{n}}, \tag{6.12}$$

(если φ измерять в радианах). Это приближение также предполагает не слишком малые n, но оно все же оказывается применимым при меньших n, чем нормальное приближение.

С целью получения доверительных границ для доли вариант, нужно найти φ по формуле (6.11) и вычислить:

$$\varphi \pm u_P \sigma_\varphi = \varphi \pm \frac{u_P}{\sqrt{n}},$$

где u_P — процентная точка нормального распределения для доверительной вероятности P, а затем по формуле (6.10) вычислить значения $p_н$ и $p_в$, соответствующие значениям:

$$\varphi_н = \varphi - \frac{u_P}{\sqrt{n}} \text{ и } \varphi_в = \varphi + \frac{u_P}{\sqrt{n}}.$$

Конечно, переход от p к φ и обратно по формулам (6.10) и (6.11) с применением обычной таблицы тригонометрических функций очень неудобен. Поэтому была составлена специальная таблица, которая непосредственно связывает значения p и φ (табл. VII Приложений).

Пример 6.4. В подвергнутом проверке ящике с консервами 3 банки из 64 оказались дефектными. Каков 95%-ный доверительный интервал для доли дефектных банок?

Имеем: $p = x/n = 3/64 = 0{,}047 = 4{,}7\%$, чему соответствует значение $\varphi = 0{,}437$ (по табл. VII Приложений). Поскольку

$$u_P \sigma_\varphi = \frac{1{,}96}{\sqrt{n}} = \frac{1{,}96}{8} = 0{,}245,$$

то границы доверительного интервала для φ будут:

$$\varphi_н = 0{,}437 - 0{,}245 = 0{,}192; \qquad \varphi_в = 0{,}437 + 0{,}245 = 0{,}682;$$

пользуясь опять табл. VII Приложений, получаем:

$$p_н = 0{,}9\%; \quad p_в = 11{,}2\%.$$

Расчет по формулам (6.9′) и (6.7) дал бы:

$$\sigma_p = \sqrt{\frac{4{,}7 \cdot 95{,}3}{64}} = 2{,}64\%;$$

$$p_н = 4{,}7 - 1{,}96 \cdot 2{,}64 = 0{,}5\%; \quad p_в = 4{,}7 + 1{,}96 \cdot 2{,}64 = 9{,}9\%.$$

Точные значения, найденные по таблицам для биномиального распределения, равны 1,0 и 13,3%. Следовательно, φ-преобразование дает в этом случае гораздо лучший результат, чем нормальное приближение.

Пример 6.5. Уточним результат примера 6.2, пользуясь φ-преобразованием: Здесь:

$$p = 8{,}6\%; \quad \varphi = 0{,}595; \quad u_p \sigma_\varphi = \frac{1.96}{\sqrt{430}} = 0{,}094,$$

так что:

$$\varphi_н = 0{,}595 - 0{,}094 = 0{,}501; \quad \varphi_в = 0{,}595 + 0{,}094 = 0{,}689;$$

поэтому по табл. VII Приложений:

$$p_н = 6{,}2\%; \quad p_в = 11{,}4\%.$$

Так как в этом случае n довольно велико, то различие между результатами, даваемыми φ-преобразованием и нормальным приближением, оказалось небольшим.

Пример 6.6. Пересчитаем, пользуясь φ-преобразованием, данные из примера 6.3:

$$p = 50\%; \quad \varphi = 1{,}571; \quad u_{0{,}99} \sigma_\varphi = \frac{2{,}58}{\sqrt{10}} = 0{,}816.$$

Поэтому:

$$\varphi_н = 1{,}571 - 0{,}816 = 0{,}755; \quad \varphi_в = 1{,}571 + 0{,}816 = 2{,}387,$$

что дает:

$$p_н = 13{,}6\%; \quad p_в = 86{,}4\%.$$

Конечно, этот доверительный интервал все еще слишком широк, чтобы иметь достаточное практическое значение.

При малых p можно получить хорошие результаты (т. е. близкие к тем, которые дает применение точного биномиального распределения) и притом весьма просто, при помощи распределения Пуассона. Об этом см. раздел **7.2.3.**

§ 6.3. Проверка гипотез о доле (проценте) вариант

Прежде чем читать этот параграф, обязательно прочтите § 4.1, в котором рассмотрены общие основы применения статистических критериев и разъясняются понятия, которые будут фигурировать в § 6.3: нулевая и альтернативная гипотезы, значимость различия, уровень значимости, критическое значение, односторонний и двусторонний критерии.

6.3.1. Во многих случаях доля (процент) вариант одного из двух типов в генеральной совокупности подсказывается существом изучаемого явления — например, соотношение 3 : 1 для доминантных и рецессивных форм при расщеплении гибридов с одним учитываемым признаком (т. е. $p_0 = 0{,}75$ для одной из форм — в данном случае доминантной), вероятность $p_0 = 0{,}5$ рождения животных каждого пола и т. д. Поскольку в действительности всегда изучается выборка, то нужно проверить, нельзя ли различие между наблюденной долей p и ожидаемой долей p_0 отнести за счет выборочного варьирования.

Используя биномиальное распределение, можно найти критические значения m_α (при $\alpha = 0{,}05$ или $\alpha = 0{,}01$) для заданных p_0 и n, после чего вычисляем критические значения $p_\alpha = m_\alpha/n$. В табл. 6.2 приведены критические значения m_α для разных n при $p_0 = 0{,}5$. Выбор именно этого значения p_0 определяется особой важностью данного случая для биологической практики. Табл. 6.2 часто называют таблицей *критерия знаков* (рассматривая знаки плюс и минус как две альтернативы).

Пример 6.7. В 10 выводках получились 51 курочка и 37 петушков. Противоречит ли это гипотезе $\hat{p} = 0{,}5$?

При $n = 51 + 37 = 88$ имеем по табл. 6.2 критические значения $m_{0,05} = 35$, $m_{0,01} = 32$. Так как фактическое значение $m = 37$ больше критического значения $m_{0,05}$, то нулевая гипотеза не отвергается.

В табл. 6.2 содержатся значения n до 100. При больших объемах выборки приходится пользоваться расчетным методом (с при-

менением табл. VII Приложений). Однако поскольку при $p_0 = 0{,}5$ распределение выборочных значений p симметрично, можно использовать также u-критерий. Более того, ввиду сравнительной простоты данного распределения можно построить упрощенный критерий:

$$m_\alpha = \frac{1}{2}\left(n - u_\alpha \sqrt{n}\right), \qquad (6.13)$$

причем нулевая гипотеза принимается при $m \geqslant m_\alpha$ и отвергается при $m < m_\alpha$. Как и ранее, m есть численность меньшей группы.

Пример 6.8. За 30 лет в области было зарегистрировано 1 359 814 рождений мальчиков и 1 285 047 рождений девочек. Можно ли на основании этих данных отвергнуть гипотезу $p = 0{,}5$ о вероятности рождений девочек?

ТАБЛИЦА 6.2

Критические значения m_α для менее часто встречающихся альтернатив (по книге Б. Л. ван дер Вардена, с. 416—417)

n	5%	1%	n	5%	1%	n	5%	1%	n	5%	1%
8	1	1	31	10	8	54	20	18	77	30	27
9	2	1	32	10	9	55	20	18	78	30	28
10	2	1	33	11	9	56	21	18	79	31	28
11	2	1	34	11	10	57	21	19	80	31	29
12	3	2	35	12	10	58	22	19	81	32	29
13	3	2	36	12	10	59	22	20	82	32	29
14	3	2	37	13	11	60	22	20	83	33	30
15	4	3	38	13	11	61	23	21	84	33	30
16	4	3	39	13	12	62	23	21	85	33	31
17	5	3	40	14	12	63	24	21	86	34	31
18	5	4	41	14	12	64	24	22	87	34	32
19	5	4	42	15	13	65	25	22	88	35	32
20	6	4	43	15	13	66	25	23	89	35	32
21	6	5	44	16	14	67	26	23	90	36	33
22	6	5	45	16	14	68	26	23	91	36	33
23	7	5	46	16	14	69	26	24	92	37	34
24	7	6	47	17	15	70	27	24	93	37	34
25	8	6	48	17	15	71	27	25	94	38	35
26	8	7	49	18	16	72	28	25	95	38	35
27	8	7	50	18	16	73	28	26	96	38	35
28	9	7	51	19	16	74	29	26	97	39	36
29	9	8	52	19	17	75	29	26	98	39	36
30	10	8	53	19	17	76	29	27	99	40	37
									100	40	37

Нулевая гипотеза принимается при $m \geqslant m_\alpha$ и отвергается при $m < m_\alpha$.

Применение u-критерия и формулы (6.9) приводит к результату:

$$p = \frac{1\,285\,047}{2\,644\,861} = 0{,}486;$$

$$\sigma_p = \sqrt{\frac{0{,}486 \cdot 0{,}514}{2\,644\,860}} \approx 0{,}000307; \quad u = \frac{0{,}500 - 0{,}486}{0{,}000307} \approx 45{,}5.$$

Поэтому гипотеза $\hat{p} = 0{,}5$ определенно опровергается. По формуле (6.13) получаем:

$$m_{0,01} = \frac{1}{2}\left(2\,644\,861 - 2{,}58\sqrt{2\,644\,861}\right) = 1\,320\,335,$$

в то время как фактическое значение $m = 1\,285\,047$ меньше. Следовательно, результат тот же — гипотеза $\hat{p} = 0{,}5$ опровергается.

6.3.2. Если $p_0 \neq 0{,}5$, то изложенные выше критерии неприменимы. В частности, u-критерий нельзя применять вследствие несимметричности биномиального распределения при $\hat{p} \neq 0{,}5$ (если только n не очень велико).

В этом случае используется то обстоятельство, что величина

$$\varphi = 2\arcsin\sqrt{p} \tag{6.11}$$

(см. **6.2.2**) имеет распределение, близкое к нормальному, при любом p. Поэтому для проверки гипотезы $p = p_0$ можно использовать критерий:

$$u = \frac{|\varphi - \varphi_0|}{\sigma_{\varphi-\varphi_0}} \geqslant u_\alpha. \tag{6.14}$$

Как указывалось в разделе **6.2.2**, $\sigma_\varphi \approx 1/\sqrt{n}$; поскольку φ_0 задано, то $\sigma_{\varphi_0} = 0$, так что:

$$\sigma_{\varphi-\varphi_0} = \sqrt{\sigma_\varphi^2 + \sigma_{\varphi_0}^2} = \sigma_\varphi \approx 1/\sqrt{n}.$$

Поэтому окончательно:

$$u = |\varphi - \varphi_0|\sqrt{n}. \tag{6.15}$$

Пример 6.9. При расщеплении гибридов *primula* получено 486 растений с гладкими листьями и 102 растения со сморщенными листьями. Соответствует ли это ожидаемому отношению 3 : 1 для доминантных и рецессивных форм?

В данном случае:

$$p = 102 : (486 + 102) = 0{,}179 = 17{,}9\%;$$

$$p_0 = 1 : (3 + 1) = 0{,}250 = 25{,}0\%.$$

По табл. VII Приложений находим: $\varphi = 0{,}874$; $\varphi_0 = 1{,}047$, так что:

$$u = (1{,}047 - 0{,}874)\sqrt{570} = 0{,}173 \cdot 23{,}9 = 4{,}14.$$

Это превышает $u_{0.01} = 2{,}58$ (и даже $u_{0{,}001} = 3{,}29$), поэтому нулевая гипотеза определенно отвергается. Причины отклонения от теоретического отношения требуют биологического анализа (например, одной из них может оказаться меньшая жизнеспособность растений со сморщенными листьями).

Конечно, этот критерий можно применять и при $p_0 = 0{,}5$. Но для этого частного случая в разделе **6.3.1** были описаны более простые критерии.

Если предполагаемая доля p_0 мала (например меньше 0,1), а объем выборки достаточно велик (100 и больше), то можно использовать совсем простой критерий, связанный с распределением Пуассона (см. раздел **7.2.3***).*

6.3.3. Выше нулевая и альтернативная гипотезы формулировались так:

$$H_0: p = p_0; \quad H_1: p \neq p_0,$$

и применялись двусторонние критерии. Можно также сформулировать односторонние задачи:

$$H_0: p \leqslant p_0, \quad H_1: p > p_0$$

или

$$H_0: p \geqslant p_0, \quad H_1: p < p_0;$$

в этом случае будут применяться односторонние критерии, т. е. уровень значимости при использовании тех же критических значений, что и выше, будет не α, а 2α (а для получения уровня значимости α надо будет брать другие критические значения).

Однако иногда задача формулируется еще и так:

$$H_0 . p \leqslant p_0, \quad H_1 . p \geqslant p_1.$$

Пример 6.10. В ядрах лейкоцитов циркулирующей крови человека наблюдаются определенной формы образования, причем у мужчин не более чем в 1,2% клеток и у женщин не менее чем в 5,6% клеток. Наблюдая мазок крови, мы хотим установить, принадлежит он мужчине или женщине.

Для решения таких задач удобнее всего пользоваться методом *последовательного анализа.*

Прежде чем читать дальше, прочтите раздел 4.5.1, в котором изложена идея этого метода.

Мы сейчас рассмотрим приложение этого метода к обсуждаемому в этой главе случаю альтернативного распределения, когда интересующим нас параметром является доля событий одного из двух возможных типов.

Итак, речь идет о независимых испытаниях, при которых может наступить или не наступить некоторое событие. Из формулы (6.2) следует, что если произведено n испытаний, то вероятность наступления событий m раз равна:

$$w_m = C_n^m \hat{p}^m \hat{q}^{n-m}.$$

Если $P_{0,n}$ есть вероятность m наступлений событий в n испытаниях при условии, что $\hat{p} = p_0$, а $P_{1,n}$ — соответствующая вероятность при условии, что $\hat{p} = p_1$, то:

$$P_{0,\,n} = C_n^m p_0^m q_0^{n-m}, \qquad P_{1,\,n} = C_n^m p_1^m q_1^{n-m}.$$

Тогда условие (***) из раздела 4.5.1 примет вид:

$$\frac{\alpha}{1-\beta} < \left(\frac{p_1}{p_0}\right)^m \left(\frac{q_1}{q_0}\right)^{n-m} < \frac{1-\alpha}{\beta}$$

или, после логарифмирования,

$$\lg \frac{\alpha}{1-\beta} < m \lg \frac{p_1}{p_0} - (n-m) \lg \frac{q_0}{q_1} < \lg \frac{1-\alpha}{\beta}$$

(здесь можно использовать десятичные логарифмы). Это можно переписать в виде:

$$\frac{\lg (q_0/q_1)}{\lg (p_1/p_0)}(n-m) - \frac{\lg [(1-\beta)/\alpha]}{\lg (p_1/p_0)} < m < \frac{\lg (q_0/q_1)}{\lg (p_1/p_0)}(n-m) + \frac{\lg [(1-\alpha)/\beta]}{\lg (p_1/p_0)}.$$

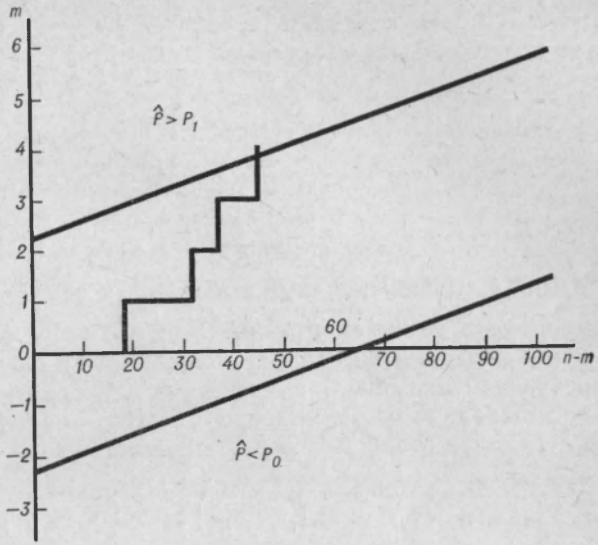
Поэтому если строить график, откладывая на оси ординат число наступлений события m, а на оси абсцисс число ненаступлений событий $n-m$, то полоса испытаний будет ограничена двумя параллельными прямыми с наклоном

$$a = \frac{\lg (q_0/q_1)}{\lg (p_1/p_0)}, \tag{6.16}$$

отсекающими на оси ординат соответственно

$$-b' = -\frac{\lg [(1-\beta)/\alpha]}{\lg (p_1/p_0)}, \quad b'' = \frac{\lg [(1-\alpha)/\beta]}{\lg (p_1/p_0)}. \tag{6.17}$$

Рис. 6.2. Последовательный анализ для доли вариант.



При пользовании этим графиком откладывают после каждого испытания единичный отрезок вверх (если событие наступило) или вправо (если событие не наступило). Когда получившаяся ступенчатая ломаная коснется одной из границ полосы испытаний (или пересечет ее), испытания заканчивают и принимают соответствующую гипотезу.

Применим изложенную методику к примеру 6.10, причем примем $\alpha = \beta = 0{,}05 = 5\%$. Тогда, поскольку $p_0 = 1{,}2\%$, $p_1 = 5{,}6\%$, имеем:

$$a = \frac{\lg (98{,}8/94{,}4)}{\lg (5{,}6/1{,}2)} = \frac{0{,}020}{0{,}569} = 0.035,$$

$$b' = b'' = \frac{\lg (95{,}0/5{,}0)}{\lg (5{,}6/1{,}2)} = \frac{1{,}279}{0{,}569} = 2.248.$$

Соответствующий график изображен на рис. 6.2; для удобства масштабы по осям различны. В исследовавшемся мазке образования встретились после 18-й, 32-й, 37-й и 45-й «пустых» клеток. Так как ломаная вышла в область $\hat{p} > p_1$, то кровь считаем женской.

§ 6.4. Сравнение двух выборочных долей вариант

Прежде чем читать этот параграф, обязательно прочтите § 4.1, в котором рассмотрены общие основы применения статистических критериев и разъясняются понятия, используемые в настоящем параграфе: нулевая гипотеза, уровень значимости, критическое значение.

6.4.1. Сравнение двух выборочных долей вариант имеет целью проверить гипотезу $\hat{p}_1 = \hat{p}_2$. Удобнее всего это делается при помощи критерия хи-квадрат, описанного в главе пятой. Но в интересующем нас теперь случае двух альтернативных распределений расчеты значительно упрощаются. В этом случае таблица содержит 2 строки и 2 столбца (поэтому ее обычно называют таблицей 2×2), причем:

$$n_1 - \hat{n}_1 = -\left(n_2 - \hat{n}_2\right) = -\left(n_3 - \hat{n}_3\right) = n_4 - \hat{n}_4,$$

так как сумма отклонений $n_i - \hat{n}_i$ в каждой строке и в каждом столбце должна равняться нулю. Совершенно элементарные, хотя несколько громоздкие, выкладки приводят выражение (5.1) к виду:

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)},$$

где n — объем выборки, а остальные обозначения ясны из табл. 6.3. При пользовании этой формулой отпадает необходимость в вычислении «теоретических» частот.

ТАБЛИЦА 6.3

	B_1	B_2	Сумма
A_1	a	b	$a+b$
A_2	c	d	$c+d$
Сумма	$a+c$	$b+d$	n

Специальный анализ показывает, что более правильный результат получается, если ввести в эту формулу поправку на группировку. В данном случае эта поправка особенно существенна

ввиду того, что группировка является предельно грубой — всего на два разряда. В исправленном виде:

$$\chi^2 = \frac{(|\,ad - bc\,| - n/2)^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)}\,n. \qquad (6.18)$$

Число степеней свободы равно:

$$(k_A - 1)\,(k_B - 1) = (2-1)(2-1) = 1,$$

поэтому в соответствии с табл. VII Приложений

$$\chi^2_{0,05} = 3{,}84; \quad \chi^2_{0,01} = 6{,}63.$$

Пример 6.11. Были проведены опыты иммунизации телят от туберкулеза: телятам сначала делали либо предохранительную прививку, либо прививку контрольных средств, а затем их заражали микобактериями туберкулеза. Результаты получились следующие: с прививкой заболели 6 из 20, без прививки заболели 16 из 19 (табл. 6.4). Для выяснения значимости действия иммунизации применим к этим данным χ^2-критерий.

ТАБЛИЦА 6.4

	Заболевшие	Незаболевшие	Сумма
С прививкой	6	14	20
Без прививки	16	3	19
Сумма	22	17	39

Расчет по формуле (6.18) дает

$$\chi^2 = \frac{(|\,6 \cdot 3 - 14 \cdot 16\,| - 39/2)^2}{20 \cdot 19 \cdot 22 \cdot 17} \cdot 39 = 9{,}55;$$

эта величина превышает $\chi^2_{0.01} = 6{,}63$, поэтому результат надо считать значимым.

6.4.2. Как было указано в разделе 5.1.2, следует избегать применения критерия χ^2, если какие-либо «теоретические» частоты меньше 3. Это особенно касается тех случаев, когда число степеней свободы невелико. Так как для таблиц 2×2 всегда $f = 1$, то к этим таблицам указанное ограничение относится в первую очередь. В таких случаях можно использовать формулу Фишера для вероятности того, что в двух выборках из одной генеральной совокупности

получится заданное четырехклеточное распределение. Формула Фишера имеет вид:

$$P = \frac{(a+b)!\,(c+d)!\,(a+c)!(b+d)!}{n!\,a!\,b!\,c!\,d!}. \qquad (6.19)$$

Знак «!» есть так называемый факториал (см. раздел 6.1.2).

Если окажется, что $P < \alpha$, то нулевая гипотеза отвергается; при $P \geqslant \alpha$ нулевая гипотеза принимается.

Пример 6.12. Из 10 больных, которых лечили способом A, у 3 состояние улучшилось, у 7 осталось без изменения. Из 7 больных, которых лечили способом $Б$, соответствующие числа составляют 5 и 2. Можно ли считать доказанным преимущество способа $Б$?

В данном случае таблица имеет следующий вид (табл. 6.5).

ТАБЛИЦА 6.5

Наличие улучшения	Способ лечения		Сумма
	A	$Б$	
+	3	5	8
—	7	2	9
Сумма	10	7	17

Подставляем данные в формулу (6.19):

$$P = \frac{10!\,7!\,8!\,9!}{17!\,3!\,5!\,7!\,2!}.$$

При недостатке опыта в вычислениях полезно выписать все множители (произведя очевидное сокращение на 7!):

$$P = \frac{(1\cdot2\cdot3\cdot4\cdot5\cdot6\cdot7\cdot8\cdot9\cdot10)\,(1\cdot2\cdot3\cdot4\cdot5\cdot6\cdot7\cdot8)\,(1\cdot2\cdot3\cdot4\cdot5\cdot6\cdot7\cdot8\cdot9)}{(1\cdot2\cdot3\cdot4\cdot5\cdot6\cdot7\cdot8\cdot9\cdot10\cdot11\cdot12\cdot13\cdot14\cdot15\cdot16\cdot17)\,(1\cdot2\cdot3)\,(1\cdot2\cdot3\cdot4\cdot5)\,(1\cdot2)}.$$

Теперь сразу видно, что большинство чисел сокращается, после чего остается:

$$P = \frac{4\cdot7\cdot9}{11\cdot13\cdot17} = 0{,}104.$$

Так как это больше, чем 0,05, то нулевая гипотеза не отвергается — преимущество способа $Б$ не доказано.

Вычисления можно упростить, используя то обстоятельство, что логарифм величины P равен сумме логарифмов чисел, стоящих в формуле (6.19) в числителе, минус сумма логарифмов чисел, стоящих в знаменателе. Так, для примера 6.12, имеем:

$$\lg P = (\lg 10! + \lg 7! + \lg 8! + \lg 9!) - (\lg 17! + \lg 3! + \lg 5! + \lg 7! + \lg 2!).$$

Логарифмы факториалов чисел до $n = 40$ равны:

n	$\lg n!$	n	$\lg n!$	n	$\lg n!$
2	0,301	15	12,116	28	29,484
3	0,778	16	13,321	29	30,947
4	1,380	17	14,551	30	32,424
5	2,079	18	15,806	31	33,915
6	2,857	19	17,085	32	35,420
7	3,701	20	18,386	33	36,939
8	4,606	21	19,708	34	38,470
9	5,560	22	21,051	35	40,014
10	6,560	23	22,412	36	41,571
11	7,601	24	23,793	37	43,139
12	8,680	25	25,191	38	44,719
13	9,794	26	26,606	39	46,310
14	10,940	27	28,037	40	47,912

Тогда для примера 6.12 имеем (после сокращения числителя и знаменателя на 7!):

Числитель	Знаменатель	Разность
$\lg 10! = 6{,}560$	$\lg 17! = 14{,}551$	$-17{,}709$
$\lg 8! = 4{,}606$	$\lg 3! = 0{,}778$	$+16{,}726$
$\lg 9! = 5{,}560$	$\lg 5! = 2{,}079$	$-0{,}983$
	$\lg 2! = 0{,}301$	
16,726	17,709	

Значит, $\lg P = -0{,}983 = \bar{1}{,}017$, что дает $P = 0{,}104$, как и ранее. Впрочем, можно и не переходить от $\lg P$ к P. Так как $\lg 0{,}01 = -2{,}0$ и $\lg 0{,}05 = \bar{2}{,}699 \approx -1{,}3$, то можно считать, что нулевая гипотеза отвергается при $|\lg P| > 2{,}0$ (для $\alpha = 0{,}01$) или при $|\lg P| > 1{,}3$ (для $\alpha = 0{,}05$). В примере 6.12 получилось $|\lg P| = 0{,}983 < 1{,}3$, поэтому нулевая гипотеза не отвергается.

Имеются специальные таблицы[1], позволяющие проверять значимость различия для четырехклеточных комплексов без всяких вычислений (при сравнительно малых численностях).

[1] См. «Таблицы» В. С. Генеса (1964) и «Statistical Tables...», Fisher and Yates (1957).

6.4.3. Значимость различия между двумя долями вариант можно также проверять, используя φ-преобразование (см. раздел 6.2.2). При сравнении двух выборок будем иметь:

$$\sigma_{\varphi_1-\varphi_2} = \sqrt{\sigma^2_{\varphi_1} + \sigma^2_{\varphi_2}} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{\frac{n_1 + n_2}{n_1 n_2}}. \quad (6.20)$$

Поскольку величины φ распределены нормально, то мы пользуемся u-критерием, т. е. вычисляем величину:

$$u = \frac{|\varphi_1 - \varphi_2|}{\sigma_{\varphi_1-\varphi_2}} = |\varphi_1 - \varphi_2| \sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \quad (6.21)$$

и сравниваем ее с критическими значениями $u_{0.05} = 1{,}96$ и $u_{0{,}01} = 2{,}58$ (понятно, что при вычислении u всегда из большей величины φ вычитается меньшая).

Более правильные результаты получаются при введении так называемой поправки на группировку. Если $p_1 > p_2$, то производится замена:

$$p_1 \to p_1' = p_1 - \frac{1}{2n_1} \quad \text{или} \quad \frac{m_1}{n_1} \to \frac{m_1 - 0{,}5}{n_1},$$

$$p_2 \to p_2' = p_2 + \frac{1}{2n_2} \quad \text{или} \quad \frac{m_2}{n_2} \to \frac{m_2 + 0{,}5}{n_2},$$

после чего по табл. VII Приложений находят значения φ_1 и φ_2 соответствующие p_1' и p_2'.

Пример 6.13. В примере 6.11 разбирался опыт иммунизации телят от туберкулеза. Оценим значимость влияния прививки, используя φ-преобразование. Так как здесь проверяется $p_2 > p_1$, мы пишем:

$$p_1' = \frac{6 + 0{,}5}{20} = 0{,}325 = 32{,}5\%,$$

$$p_2' = \frac{16 - 0{,}5}{19} = 0{,}816 = 81{,}6\%.$$

Из табл. VII Приложений получаем:

$$\varphi_1 = 1{,}213; \quad \varphi_2 = 2{,}255.$$

Поэтому:

$$u = (2{,}255 - 1{,}213)\frac{20 \cdot 19}{20 + 19} = 1{,}042\frac{380}{39} = 3{,}25.$$

Так как $u > u_{0.01} = 2{,}58$, то разность $p_2 - p_1$ значима.

6.4.4. Анализ четырехклеточных таблиц может быть использован для сравнения двух порядковых совокупностей. Идея состоит в том, что отыскивают медиану суммарной совокупности (т. е. получившейся сведением обоих заданных рядов в один), после чего составляют табл. 6.6.

ТАБЛИЦА 6.6

	Ниже медианы	Выше медианы	Сумма
Первый ряд	a	b	$a + b$
Второй ряд	c	d	$c + d$
Сумма	$a + c = \frac{n}{2}$	$b + d = \frac{n}{2}$	n

О порядковых совокупностях и медиане см. разделы **2.2.1** *и* **3.1.5.**

Далее применяется φ-преобразование, критерий χ^2 или критерий Фишера (в зависимости от объема n). Ввиду того что в основу классификации здесь положена медиана, этот тест обычно называют *тестом медианы*.

Этот тест применяют и для количественных совокупностей, особенно когда теоретическое распределение неизвестно: ведь в этом случае крайние варианты, сильно отклоняющиеся от центра, нельзя отбрасывать, между тем они сильно влияют на положение среднего значения; положение же медианы почти не зависит от крайних вариант.

Пример 6.14. Две группы мышей из двух генетических линий получили одинаковую смертельную дозу облучения. Табл. 6.7 показывает число дней между облучением и гибелью. Можно ли считать, что мыши одной из линий более радиочувствительны?

ТАБЛИЦА 6.7

Первая группа	5	4	10	5	14	4		
Вторая группа	8	15	2	10	11	7	11	13

Составив общий ряд, имеем:

I 4 4 5 5 10 14

II 2 7 8 10 11 11 13 15

Так как здесь $Me = 9$, то таблица 2×2 будет иметь следующий вид (табл. 6.8).

ТАБЛИЦА 6.8

	Ниже медианы	Выше медианы	Сумма
Первый ряд	4	2	6
Второй ряд	3	5	8
Сумма	7	7	14

Применение критерия Фишера показывает, что значимой разницы нет ($P = 0{,}244 > 0{,}05$ или $|\lg P| = 0{,}611 < 1{,}3$).

Если суммарный объем выборок нечетный, то медиана приходится на одно из значений. Поскольку учитываются лишь значения «ниже медианы» и «выше медианы», то объем n в таблице типа табл. 6.8 оказывается на единицу меньше, чем сумма объемов выборок. Например, если в табл. 6.7 выбросить значение 10 из первой группы, то сводный ряд будет иметь вид:

I 4 4 5 5 14
II 2 7 8 10 11 11 13 15

с медианой $Me = 8$. Это даст табл. 6.9.

ТАБЛИЦА 6.9

	Ниже медианы	Выше медианы	Сумма
Первый ряд	4	1	5
Второй ряд	2	5	7
Сумма	6	6	12

К табл. 6.9 и нужно применять критерий.

6.4.5. Если разность долей вариант оказалась значимой, то вычисляется доверительный интервал для этой разности. Из формулы (6.20) очевидно, что ширина доверительного интервала определяется величиной:

$$u_P \, \sigma_{\varphi_1 - \varphi_2} = u_P \sqrt{\frac{n_1 + n_2}{n_1 n_2}}, \qquad (6.22)$$

так что границы этого интервала будут:

$$\Delta\varphi_н = (\varphi_1 - \varphi_2) - u_P \sqrt{\frac{n_1 + n_2}{n_1 n_2}},$$

$$\Delta\varphi_в = (\varphi_1 - \varphi_2) + u_P \sqrt{\frac{n_1 + n_2}{n_1 n_2}}.$$

В случае из примера 6.13 имеем при $P = 95\%$:

$$u_P \sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 1{,}96 \sqrt{\frac{39}{380}} = 0{,}628;$$

поэтому:

$$\Delta\varphi_н = 1{,}042 - 0{,}628 = 0{,}414; \quad \Delta\varphi_в = 1{,}042 + 0{,}628 = 1{,}670.$$

Этому соответствуют значения (см. табл. VII Приложений):

$$\Delta p_н = 4{,}2\%; \quad \Delta p_в = 55{,}0\%.$$

О понятиях доверительного интервала см. раздел 3.7.1. *О построении доверительного интервала для доли (процента) вариант подробно сказано в § 6.2.*

Глава седьмая

РАСПРЕДЕЛЕНИЕ РЕДКИХ СОБЫТИЙ

§ 7.1. Распределение Пуассона

7.1.1. Распределение Пуассона описывает реализации событий с малой вероятностью, но при большом числе испытаний.

Рассмотрим следующий пример. Мешок содержит 1000 каких-то мерок белых бобов, причем одна мерка вмещает 100 семян; таким образом, общее число семян в мешке составляет 100 000. Заменим 1000 штук белых семян таким же количеством черных. Тогда на каждые 100 семян будет приходиться одно черное семя. Если теперь после тщательного перемешивания всех семян зачерпывать из мешка по одной мерке семян, то на каждую порцию придется в среднем одно черное семя. Однако это совсем не означает, что на самом деле в каждой порции будет по одному такому семени. В некоторых порциях таких семян не окажется совсем, в других будет по одному семени, в некоторых по два, в каком-то числе порций — по три и т. д. Распределение числа порций, в которых окажется то или иное число черных семян, и дается законом Пуассона. Формула для этого распределения, которую мы даем без вывода[1], имеет вид:

$$n_m = n \frac{\lambda^m}{m!} e^{-\lambda}, \qquad (7.1)$$

Здесь λ есть среднее число интересующих нас событий (в нашем примере событием является попадание в мерку черного семени), приходящихся на одну пробу (в нашем примере на одну мерку семян), m есть фактическое число событий в одной пробе (это может быть также число ионизирующих частиц, попадающих в счетчик в одну минуту, или число рождений тройни за год, или число несчастных случаев за время t и т. д.— важно только, чтобы событие было достаточно редким, т. е. чтобы m было малым). Величина n_m есть число проб, в которых осуществилось m событий (в нашем случае — число мерок, в которых оказалось m черных семян).

---

[1] Эта формула получается как предельный случай биномиального распределения (см. раздел **6.1.2**) при малых p. Подробнее см. В. Ю. Урбах «Биометрические методы», с. 84—85.

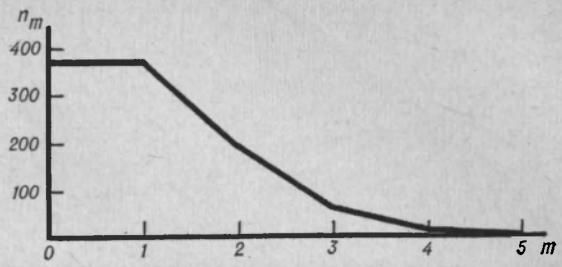


Рис. 7.1. Распределение Пуассона при $\lambda = 1$.

Знак «!» есть так называемый факториал: величина $k!$ есть произведение вида $1 \cdot 2 \cdot 3 \ldots (k-1)k$, причем, как было показано в разделе 6.1.2, $0! = 1$.

Величина $e = 2{,}718 \ldots$ есть основание натуральных логарифмов; она равна сумме бесконечного ряда:

$$\sum_{m=0}^{\infty} \frac{1}{m!} = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \ldots$$

Можно доказать, что

$$\sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = e^{\lambda}$$

при любом λ; отсюда следует:

$$\sum_{m=0}^{\infty} n_m = \sum_{m=0}^{\infty} n \frac{\lambda^m}{m!} e^{-\lambda} = n e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = n e^{-\lambda} e^{\lambda} = n,$$

как это и должно быть.

Если $p_m = n_m/n$ есть доля порций, содержащих m черных семян, то

$$p_m = \frac{\lambda^m}{m!} e^{-\lambda}. \qquad (7.2)$$

В рассмотренном примере $\lambda = 1$, так что $e^{-\lambda} \approx 1/2{,}718 \approx 0{,}368$. Поскольку $1^0/0! = 1$, $1^1/1! = 1$, $1^2/2! = 1/2$, $1^3/3! = 1/6$ и т. д., то число порций, в каждой из которых число черных семян равно 0; 1; 2; 3 и т. д., будет задаваться распределением, записанным в табл. 7.1 (учитывая также, что в данном случае $n = 1000$). Полигон частот этого распределения изображен на рис. 7.1.

Вероятность появления шести и больше черных семян в одной порции очень мала, так что вернее всего из 1000 порций таких не

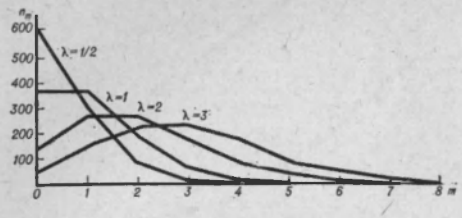


Рис. 7.2. Распределение Пуассона при разных значениях параметра.

ТАБЛИЦА 7.1

m	Число черных семян в одной порции	0	1	2	3	4	5	6
n_m	Число порций с данным числом черных семян	368	368	184	62	15	3	0

окажется ни одной (но в принципе распределение Пуассона распространяется вправо неограниченно).

7.1.2. Форма полигона частот зависит существенно от значения параметра λ. В нашем примере было $\lambda = 1$. Теперь рассмотрим для сравнения еще два случая, когда $\lambda < 1$ и $\lambda > 1$. В первом случае, очевидно, числитель в формуле (7.1) убывает при возрастании m, а так как знаменатель одновременно возрастает, то n_m все время уменьшается. Если же $\lambda > 1$, то в формуле (7.1) возрастает и числитель, и знаменатель. Но они возрастают по разному закону, так что сначала «перевешивает» числитель, а затем знаменатель. Распределения для нескольких значений λ записаны в табл. 7.2. Соответствующие полигоны частот изображены на рис. 7.2.

ТАБЛИЦА 7.2

m	0	1	2	3	4	5	6	7	8
$\lambda = \frac{1}{2}$	606	303	76	13	2				
$\lambda = 1$	368	368	184	62	15	3			
$\lambda = 2$	135	271	271	180	90	36	12	4	1
$\lambda = 3$	50	150	225	225	169	101	51	22	7

Как видно из табл. 7.2, при $\lambda < 1$ частоты монотонно убывают с возрастанием m, а при $\lambda > 1$ имеется максимум; значение $\lambda = 1$ является в этом отношении критическим. Чем больше λ, тем дальше отодвигается максимум, и асимметричность распределения становится все менее заметной. При достаточно больших λ распределение мало отличается от симметричного. Если $\lambda > 20$, то распределение Пуассона достаточно хорошо приближается законом Гаусса (т. е. нормальным распределением; об этом распределении см. § 2.3).

7.1.3. Единственный параметр λ распределения Пуассона полностью определяет все свойства этого распределения. Если бы мы захотели описать это распределение в привычных терминах математического ожидания, дисперсии, коэффициента асимметрии, то оказалось бы, что эти величины выражаются только через λ, точнее,

$$\mu = \sigma^2 = \langle \xi^3 \rangle = \lambda. \tag{7.3}$$

В этом нетрудно убедиться, произведя непосредственные вычисления; так, например:

$$\mu = \sum_{m=0}^{\infty} m \frac{\lambda^m}{m!} e^{-\lambda} = \sum_{m=1}^{\infty} m \frac{\lambda^m}{m!} e^{-\lambda} = e^{-\lambda} \sum_{m=1}^{\infty} \frac{\lambda \cdot \lambda^{m-1}}{(m-1)!} =$$

$$= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} e^{\lambda} = \lambda,$$

полагая, что $k = m - 1$. Аналогично доказывают, что $\sigma^2 = \lambda$ и $\langle \xi^3 \rangle = \lambda$, используя формулы (3.19) и (2.7), определяющие σ^2 и $\langle \xi^3 \rangle$. Очевидно, асимметрия распределения Пуассона всегда положительна.

О математическом ожидании, дисперсии и коэффициенте асимметрии см. соответственно разделы **3.1.1**, **3.3.1** *и* **2.4.1.**

7.1.4. Для проверки того, что выборка взята из генеральной совокупности с пуассоновым распределением, используют то обстоятельство, что для этого распределения $\sigma^2 = \mu$. Это значит, что статистики s^2 и $\bar{m}$ не должны значительно различаться. Точнее, используется критерий:

$$\chi^2 = f \frac{s^2}{\bar{m}}, \tag{7.4}$$

где f есть число значений m с ненулевыми частотами (например, в табл. 7.1 эти значения $m = 0, 1, 2, 3, 4, 5$, так что $f = 6$). Гипо-

теза о пуассоновом распределении отвергается, если значение χ^2 (хи-квадрат) превышает критическое значение $\chi^2_\alpha(f)$, взятое из табл. VI. Приложений для числа степеней свободы f.

О том, что такое хи-квадрат, подробнее сказано в главе пятой. Рекомендуем прочесть по крайней мере раздел 5.1.2. *О статистиках* s^2 *и* $\overline{m}$ *см разделы* 3.4.1 *и* 3.2.1.

Пример 7.1. Буква «ц» встречается в русском языке довольно редко, поэтому можно ожидать, что ее распределение будет удовлетворять закону Пуассона. В табл. 7.3 приведено распределение того, сколько раз эта буква встречается в отрывках из 100 слов (взятых из сочинений А. П. Чехова); таких отрывков было изучено 1000.

ТАБЛИЦА 7.3

Число букв «ц» в отрывке из 100 слов	0	1	2	3	4
Число отрывков с данным числом букв «ц»	752	207	38	3	0

Среднее число букв «ц» в 100 словах меньше единицы, поэтому частоты монотонно убывают. Расчет дает $\overline{m} = 0{,}292$, $s^2 = 0{,}301$. Тогда:

$$\chi^2 = 4\frac{0{,}301}{0{,}292} = 4{,}13,$$

в то время как $\chi^2_{0{,}05}(4) = 9{,}49$. Значит, данные из табл. 7.3 не противоречат предположению о пуассоновости распределения в генеральной совокупности.

Этот же результат можно было бы получить другим путем, используя критерий хи-квадрат для сравнения эмпирического распределения с теоретическим так, как это описано в § 5.1.

ТАБЛИЦА 7.4

m	$\bar{m}^m$	$m!$	$\frac{\bar{m}^m}{m!}$	$\hat{n}_m$
0	1,000	1	1,000	747,5
1	0,292	1	0,292	218,1
2	0,085	2	0,042	31,4
3	0,025	6	0,004	3,0
4	0,007	24	0,000	0,0
			1,338	1000,0

Вычисление теоретических частот $\hat{n}_m$ показано в табл. 7.4, причем учтено, что $\bar{m} = 0{,}292$ и что

$$\hat{n}_m = n \frac{\bar{m}^m}{m!} \Big/ \sum_{m=0}^{4} \frac{\bar{m}^m}{m!},$$

так как здесь

$$\sum_{m=0}^{4} \frac{\bar{m}^m}{m!} \approx e^{\bar{m}}.$$

Расчет величины χ^2 показан в табл. 7.5; эмпирические частоты взяты из табл. 7.3.

ТАБЛИЦА 7.5

m	n_m	$\hat{n}_m$	$n_m - \hat{n}_m$	(χ^2)
0	752	747,5	4,5	0,03
1	207	218,1	11,1	0,56
2	38	31,4	6,6	1,39
3	3	3,0	0,0	0,00
				1,98

В данном случае частоты связаны двумя условиями:

$$\sum n_m = n, \qquad \frac{1}{n} \sum n_m m = \bar{m},$$

так что $f = 4 - 2 = 2$. Из табл. VI Приложений находим $\chi^2_{0.05} = 5{,}99$. Так как $\chi^2 < \chi^2_{0,05}$, то различие между эмпирическим распределением и теоретическим незначимо — распределение из табл. 7.3 можно считать пуассоновским.

Сравнивая два способа решения задачи в последнем примере, мы убеждаемся в гораздо большей простоте первого способа. Это связано с тем, что он использует специальное свойство пуассоновского распределения — равенство математического ожидания и дисперсии. Общий метод сравнения эмпирического и теоретического распределений усложняется здесь необходимостью вычисления теоретических частот.

§ 7.2. Доверительный интервал для параметра распределения Пуассона

7.2.1. Рассмотрим конкретный случай распределения числа ионизирующих частиц, попадающих в счетчик в единицу времени (эта величина называется скоростью счета), при беспорядочном следовании их друг за другом. Скорость счета приходится измерять всегда, когда пользуются методом меченых атомов; поскольку этот метод широко применяется в биологических и медицинских исследованиях, разберем некоторые относящиеся сюда статистические вопросы.

Если бы мы не знали о том, что распределение числа импульсов в счетчике является пуассоновским, то измерения и статистическую обработку нужно было бы вести по обычной схеме: 1) сделать k измерений числа импульсов в минуту, что даст нам числа $x_i = x_1, x_2, \ldots, x_k$; 2) вычислить среднюю скорость счета $\bar{x} = \Sigma x_i/k$; 3) найти отклонения от среднего отдельных результатов: $\xi_i = x_i - \bar{x}$; 4) вычислить оценку стандартной ошибки среднего значения по формуле:

$$s_{\bar{x}} = \sqrt{\frac{\sum \xi_i^2}{k(k-1)}},$$

после чего, при достаточно большом k, получим окончательный результат в виде $\bar{x} + u_P s_{\bar{x}}$.

Понятия стандартной ошибки среднего значения и доверительного интервала для параметров разъясняются в разделах **3.5.1.** *и* **3.7.1.** *Прежде чем читать дальше, прочтите эти разделы.*

То обстоятельство, что распределение скорости счета является пуассоновским, значительно упрощает вычисления. Действитель-

но, это распределение имеет ту особенность, что для него $\sigma^2 = \lambda$ (см. раздел 7.1.3). Но тогда (если учесть, что при радиометрических измерениях n обычно достаточно велико) имеем приближенно:

$$\sigma = \sqrt{\bar{x}} = \sqrt{\frac{\sum x_i}{k}}; \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{k}} = \frac{\sqrt{\sum x_i}}{k},$$

так что результат имеет вид:

$$\bar{x} \pm u_P \, \sigma_{\bar{x}} = \frac{\sum x_i}{k} \pm u_P \frac{\sqrt{\sum x_i}}{k}.$$

Мы видим, что нет надобности знать отдельные x_i, а лишь их сумму Σx_i и число измерений k. Но в таком случае можно не делать k одноминутных измерений, а сделать одно k-минутное измерение. Если обозначить длительность измерения через τ (численно оно равно k), а общее число сосчитанных за это время импульсов через X, то окончательно результат запишется в виде:

$$\bar{x} \pm u_P \sigma_{\bar{x}} = \frac{X}{\tau} \pm u_P \frac{\sqrt{X}}{\tau}. \tag{7.5}$$

Пример 7.2. В табл. 7.6 приведены результаты десяти одноминутных измерений числа импульсов за счет фона. Мы видим, что действительно величины

$$\sqrt{\frac{\sum \xi_i^2}{k(k-1)}} = \sqrt{\frac{544}{10 \cdot 9}} = 2{,}46 \quad \text{и} \quad \frac{\sqrt{X}}{\tau} = \frac{\sqrt{550}}{10} = 2{,}34$$

достаточно близки.

ТАБЛИЦА 7.6

i	x_i	ξ_i	ξ_i^2	i	x_i	ξ_i	ξ_i^2
1	54	— 1	1	7	65	10	100
2	60	5	25	8	52	— 3	9
3	56	1	1	9	57	2	4
4	51	— 4	16	10	45	—10	100
5	67	12	144				
6	43	—12	144				
				Сумма	550	30+(—30)= = 0	544

Пример 7.3. Для определения фона торцового счетчика было произведено 20-минутное измерение; прибор насчитал 496 импульсов. Каковы 95%-ные доверительные границы скорости счета фона?

Имеем:

$$\bar{x} \pm u_P \sigma_{\bar{x}} \approx \frac{496}{20} \pm 1{,}96 \frac{\sqrt{496}}{20} = 24{,}8 \pm 1{,}96 \frac{22{,}3}{20} = 24{,}8 \pm 2{,}2,$$

так что получаем границы:

$$24{,}8 - 2{,}2 = 22{,}6 \approx 23; \quad 24{,}8 + 2{,}2 = 27 \text{ имп/мин.}$$

Пример 7.4. Определим 95% доверительный интервал для среднего числа букв «ц» в отрывках из 100 слов чеховских текстов (см. пример 7.1. из раздела 7.1.4).

В данном случае $X = 292$, $\tau = 1000$, так что:

$$\frac{X}{\tau} \pm u_{0,95} \frac{\sqrt{X}}{\tau} = \frac{292}{1000} \pm 1{,}96 \frac{\sqrt{292}}{1000} = 0{,}292 \pm 0{,}033,$$

т. е. границы доверительного интервала будут:

$$0{,}292 - 0{,}033 = 0{,}259; \quad 0{,}292 + 0{,}033 = 0{,}325.$$

Значение $s^2 = 0{,}301$ находится внутри этого интервала.

7.2.2. Формула (7.5), как уже указывалось, справедлива лишь при достаточно больших X, потому что она использует нормальное приближение. При не очень больших X надо вычислять доверительные границы, исходя из точного распределения Пуассона. Это довольно трудоемкая работа. Но поскольку распределение Пуассона имеет только один параметр, можно заранее табулировать доверительные границы при разных X (и при заданных доверительных уровнях P). В табл. 7.7 приводятся соответствующие значения при $P = 95\%$ и $P = 99\%$ для X от 0 до 50.

Пример 7.5. Найдем 99% доверительные границы при $X = 49$. По формуле (7.5) имеем (при $\tau = 1$):

$$49 \pm 2{,}58 \sqrt{49} = 49 \pm 18{,}06 = 30{,}94 \div 67{,}06.$$

Между тем в табл. 7.7 находим значения 32,85 и 70,08. Следовательно, даже при $X = 49$ формула (7.5) дает не очень точные значения доверительных границ. При меньших X различие еще заметнее.

ТАБЛИЦА 7.7

Доверительные границы для параметра распределения Пуассона
(по книге Л. Н. Большева и Н. В. Смирнова, с. 368 — 369)

X	$P=99\%$		$P=95\%$		X	$P=99\%$		$P=95\%$	
0	5,30	0,000	3,69	0,000	26	42,25	14,74	38,10	16,98
1	7,43	0,005	5,57	0,025	27	43,50	15,49	39,28	17,79
2	9,27	0,103	7,22	0,242	28	44,74	16,24	40,47	18,61
3	10,98	0,338	8,77	0,619	29	45,98	17,00	41,65	19,42
4	12,59	0,672	10,24	1,09	30	47,21	17,77	42,83	20,24
5	14,15	1,08	11,67	1,62	31	48,44	18,53	44,00	21,06
6	15,66	1,54	13,06	2,20	32	49,67	19,30	45,17	21,89
7	17,13	2,04	14,42	2,81	33	50,89	20,08	46,34	22,72
8	18,58	2,57	15,76	3,45	34	52,11	20,86	47,51	23,55
9	20,00	3,13	17,08	4,12	35	53,32	21,64	48,68	24,38
10	21,40	3,72	18,39	4,80	36	54,54	22,42	49,84	25,21
11	22,78	4,32	19,68	5,49	37	55,75	23,21	51,00	26,05
12	24,14	4,94	20,96	6,20	38	56,96	24,00	52,16	26,89
13	25,50	5,58	22,23	6,92	39	58,16	24,79	53,31	27,73
14	26,84	6,23	23,49	7,65	40	59,36	25,59	54,47	28,58
15	28,16	6,89	24,74	8,40	41	60,56	26,38	55,62	29,42
16	29,48	7,57	25,98	9,15	42	61,76	17,18	56,77	30,27
17	30,79	8,25	27,22	9,90	43	62,96	27,99	57,92	31,12
18	32,09	8,94	28,45	10,67	44	64,15	28,79	59,07	31,97
19	33,38	9,64	29,67	11,44	45	65,34	29,60	60,21	32,82
20	34,67	10,35	30,89	12,22	46	66,53	30,41	61,36	33,68
21	35,95	11,07	32,10	13,00	47	67,72	31,22	62,50	34,53
22	37,22	11,79	33,31	13,79	48	68,90	32,03	63,64	35,39
23	38,48	12,52	34,51	14,58	49	70,08	32,85	64,78	36,25
24	39,74	13,25	35,71	15,38	50	71,27	33,66	65,92	37,11
25	41,00	14,00	36,90	16,18					

7.2.3. Распределение Пуассона можно также использовать для решения задач альтернативного распределения, если доля p достаточно мала (см. первую сноску в разделе 7.1.1). В этом случае вычисляют нижнюю и верхнюю доверительные границы для $\hat{p}$ по формулам:

$$p_{\text{н}} = \frac{m_{\text{н}}}{n}, \quad p_{\text{в}} = \frac{m_{\text{в}}}{n}, \tag{7.6}$$

беря значения $m_{\text{н}}$ и $m_{\text{в}}$ непосредственно из табл. 7.7. Так, для примера 6.2 (из раздела **6.2.1**) находим в табл. 7.7 $x_{\text{н}} = 26{,}05$ и $x_{\text{в}} = 51{,}00$, так что:

$$p_{\text{н}} = \frac{26{,}05}{430} = 0{,}061 = 6{,}1\%; \quad p_{\text{в}} = \frac{51{,}00}{430} = 0{,}119 = 11{,}9\%.$$

Для примера 6.4 (из раздела **6.2.2**) $x_{\text{н}} = 0{,}619$, $x_{\text{в}} = 8{,}77$ и

$$p_н = \frac{0,619}{64} \approx 0,0097 \approx 1,0\%; \quad p_в = \frac{8,77}{64} = 0,137 = 13,7\%.$$

Эти результаты точнее, чем найденные другими приближенными методами, и в то же время получаются совсем просто.

Сравнение результатов разных методов расчета дано в табл. 7.8 (о φ-ireобразовании см. раздел **6.2.2**).

ТАБЛИЦА 7.8

Номер примера	Нормальное приближение	φ-преобразование	Приближение Пуассона	Точные значения
6.2	6,0 ÷ 11,2	6,2 ÷ 11,4	6,1 ÷ 11,9	6,1 ÷ 11,6
6.4	—0,5 ÷ 9,0	0,9 ÷ 11,2	1,0 ÷ 13,7	1,0 ÷ 13,3

§ 7.3. Проверка различия параметров двух распределений Пуассона

Прежде чем читать этот параграф, обязательно прочтите § 4.1, в котором рассмотрены общие основы применения статистических критериев и разъясняются понятия, используемые в настоящем параграфе: нулевая гипотеза, уровень значимости, критическое значение.

7.3.1 Если значения X достаточно велики, то для проверки различия параметров двух распределений Пуассона можно принять, что $\sigma_m \approx \sqrt{X}/\tau$ и распределение $\bar{m}$ нормально. Тогда получаем критерий:

$$u = \left(\frac{X_1}{\tau_1} - \frac{X_2}{\tau_2}\right) \Big/ \sqrt{\frac{X_1}{\tau_1^2} + \frac{X_2}{\tau_2^2}} > u_\alpha, \qquad (7.7)$$

при котором отвергается нулевая гипотеза $\lambda_1 = \lambda_2$ (об u-критерии см. раздел **4.2.2**).

Пример 7.6. При помещении под счетчик препарата, предполагаемого радиоактивным, за 5 мин зарегистрировано 187 импульсов, а фон составляет 496 импульсов за 20 мин (см. пример 7.3). Имеется ли превышение над фоном?

По формуле (7.7) получаем:

$$u = \left(\frac{187}{5} - \frac{496}{20}\right) \Big/ \sqrt{\frac{187}{25} + \frac{496}{400}} = (37,4 - 24,8)/2,95 = 4,27.$$

Это превышает $u_{0,001} = 3,29$, так что различие вполне значимо.

7.3.2. Когда $X < 50$, принятые приближения (т. е. $\sigma_m \approx \approx \sqrt{X/\tau}$ и нормальность распределения статистик $\bar{m} = X/\tau$) становятся недопустимо грубыми. В этом случае надо пользоваться табл. 7.9. Критические значения большего из двух X (для уровней значимости $\alpha = 0{,}05$ и $\alpha = 0{,}01$) ищутся в строке, отвечающей меньшему X, при этом оба $X = \Sigma m_i$ должны относиться к одному и тому же τ или k (т. е. времени счета, числу клеток цитометра, площади чашек и др.).

Пример 7. 7. Две бактериальные пробы были перенесены на две чашки Петри, и после инкубации в одной было обнаружено 26 колоний, а в другой 38 колоний. Можно ли считать различными значения параметров λ для этих двух проб?

В табл. 7.9 находим в строке $X_{\text{меньшее}} = 26$ два числа: 40 ($\alpha = 0{,}05$) и 46 ($\alpha = 0{,}01$). Фактическое $X_{\text{большее}} = 38$ не превышает числа 40 для $\alpha = 0{,}05$, так что нулевая гипотеза $\lambda_1 = \lambda_2$ не опровергается.

ТАБЛИЦА 7.9

Критерий различия параметров двух распределений Пуассона

Меньшее X	Уровень значимости		Меньшее X	Уровень значимости		Меньшее X	Уровень значимости	
	0,05	0,01		0,05	0,01		0,05	0,01
0	4	6	17	28	34	34	49	57
1	6	9	18	30	35	35	51	58
2	8	11	19	31	37	36	52	59
3	9	13	20	32	38	37	53	61
4	11	14	21	34	39	38	54	62
5	12	16	22	35	41	39	55	63
6	14	18	23	36	42	40	57	65
7	15	19	24	37	44	41	58	66
8	17	21	25	39	45	42	59	67
9	18	22	26	40	46	43	60	68
10	19	24	27	41	48	44	61	70
11	21	25	28	42	49	45	63	71
12	22	27	29	43	50	46	64	72
13	23	28	30	45	52	47	65	74
14	25	30	31	46	53	48	66	75
15	26	31	32	47	54	49	67	76
16	27	32	33	48	55	50	68	77

Гипотеза $\lambda_1 = \lambda_2$ отвергается, если при заданном меньшем X большее X превышает табличное значение.

Глава восьмая

РЕГРЕССИОННЫЙ И КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

§ 8.1. Связь между признаками. Метод наименьших квадратов

8.1.1. Отличительной чертой биологических объектов является многообразие признаков, характеризующих каждый из них. Так, животные можно характеризовать возрастом, размером, весом, различными физиологическими показателями и т. д. Имея однородную совокупность объектов, можно изучить распределение их по любому из их признаков.

Весьма часто можно усмотреть известную связь между вариациями по различным признакам. Например, чем больше размер животного, тем обычно больше его вес; известно также, что в однородном стаде те коровы, в молоке которых имеется больший процент жира, дают обычно меньший удой.

В простейшем случае связь между двумя переменными величинами строго однозначна. Например, вес образцов, сделанных из одного и того же материала, полностью определяется их объемом. Такую зависимость принято называть *функциональной*. Для биологических объектов связь обычно бывает менее «жесткой»: объекты с одинаковым значением одного признака имеют, как правило, разные значения по другим признакам. Такую связь между вариациями разных признаков называют *корреляцией* (дословный перевод: соотношение) между признаками.

Пример 8. 1. Измерение диаметров и высот 250 экземпляров сосны дало результаты:

Диаметр, *см* (x):	22	37	14	26	...
Высота, *м* (y):	19,3	24,1	20,6	19,0	...

Каждый из этих рядов может быть обработан отдельно. Но если желательно установить наличие и характер корреляции между обоими признаками (диаметром и высотой), то следует разнести обследованные экземпляры (вернее, соответствующие численные значения) в одну общую таблицу. Это будет, естественно, двумерная таблица. Поскольку число особей в данном случае велико, должна быть произведена группировка. Результат такой группировки приведен в табл. 8.1.

ТАБЛИЦА 8.1

Диаметр, *см* (x)	Высота. *м* (u)										n_x
	18	19	20	21	22	23	24	25	26	27	
15		1	6	4	3						14
20	1	3	15	29	20	8					76
25		1	8	18	49	20	6	1			103
30			1	4	5	12	8	5			35
35					1	3	6	4	1		13
40							1	3	2		6
45									1	1	2
n_y	1	5	30	55	78	43	21	13	4	1	250

Уже по виду таблицы (ее называют *корреляционной решеткой*) можно сделать заключение о наличии явной корреляции (связи) между диаметром дерева и его высотой. Действительно, толстые деревья, как правило, более высокие, чем тонкие. Однако мы видим, что однозначного соответствия между диаметром и толщиной все же нет — некоторые из тонких деревьев оказались выше, чем отдельные толстые деревья. Такая «размазанность» корреляции чрезвычайно характерна для биологических объектов, развитие которых определяется сложным переплетением многих факторов.

Важно отметить, что установление корреляции между признаками само по себе еще не дает оснований делать какие-либо заключения о причинно-следственных связях между ними. Так, в данном примере ни один из признаков не может считаться влияющим непосредственно на второй; вернее всего оба они обусловливаются в основном третьим признаком — возрастом дерева. В некоторых случаях корреляция вызывается тем, что один признак является следствием другого, например корреляция между числом зерен в колосе и урожаем на единицу площади. Задачей предстоящего анализа будет лишь установление самого факта корреляции и отыскание подходящих численных характеристик для выражения степени этой корреляции.

В случае несгруппированной совокупности может быть получено наглядное представление о наличии или отсутствии корреляции путем построения так называемого *корреляционного поля*.

Пример 8.2. Измерения длины головы (x) и длины грудного плавника (y) у 16 окуней дали результаты:

x	66	61	67	73	51	59	48	47	58	44	41	54	52	47	51	45
y	38	31	36	43	29	33	28	25	36	26	21	30	20	27	28	26

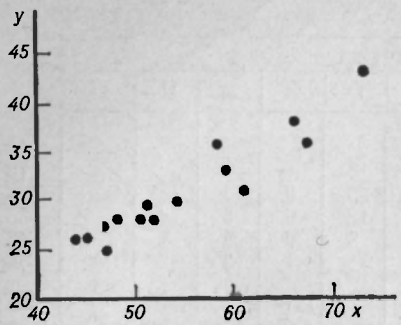


Рис. 8.1. Корреляционное поле.

Нанося точки на графике в выбранном масштабе, получаем картину, изображенную на рис. 8.1. (точечную диаграмму). Вытянутость корреляционного поля в диагональном направлении свидетельствует о несомненном наличии корреляции между обоими признаками.

Если число вариант велико, то корреляционное поле часто имеет вид более или менее правильного эллипса со сгущением точек в центре и сравнительно редким их расположением на периферии; отклонение осей эллипса от координатных направлений указывает на наличие корреляции. Вытянутость же эллипса не является объективным показателем, так как она зависит от принятых масштабов по осям координат.

8.1.2. Наряду с задачами, в которых случайно варьирующими являются обе связанные между собой величины, в практике биологических исследований часто встречаются случаи, когда одна из двух связанных между собой величин рассматривается именно как аргумент изучаемой функциональной зависимости, и процесс исследования состоит в том, что определяются значения некоторой варьирующей величины y при вполне определенных значениях аргумента x. Пусть, например, изучается зависимость числа хромосомных аберраций (y) от дозы облучения (x). Естественно поставить опыт так, чтобы дозы варьировали не случайно, а принимали определенные, заранее намеченные значения. Тогда при многократном повторении опытов будут случайно варьировать лишь значения y, относящиеся к одним и тем же дозам x; распределение же численностей по значениям x целиком зависит от экспериментатора.

8.1.3. Весьма часто исследуемая зависимость принадлежит к хорошо изученному типу, и ее аналитическое (алгебраическое) выражение точно известно; при этом целью исследования является определение численных параметров этой зависимости. Например, при радиометрическом исследовании образца почвы мы заранее знаем, что уменьшение активности образца происходит по закону радиоактивного распада:

$$A = A_0 e^{-kt}, \qquad (*)$$

но нам нужно определить значение константы распада k (знание ее позволит установить, какой радиоактивный изотоп обусловливает активность почвы).

Другой пример. Есть все основания считать, что привес животных вообще пропорционален количеству корма:

$$y = ax, \qquad (**)$$

но мы хотим найти коэффициент пропорциональности a для интересующего нас вида корма.

Каждому частному значению аргумента x отвечает ряд значений варьирующей величины y. Этот ряд, очевидно, составляет некоторое распределение, в котором может быть найдено свое математическое ожидание $\hat{y}_x$ или, в случае выборки, среднее значение $\overline{y}_x$. Эти средние значения $\overline{y}_x$ называют иногда *условными средними*: $\overline{y}_x$ есть среднее значение y при условии, что x имеет заданную величину.

О понятии математического ожидания см. раздел **3.1.1**, *а о среднем значении — раздел* **3.2.1**.

Рассматривая значения x и соответствующие им $\hat{y}_x$ как координаты точек на плоскости, мы можем по этим точкам провести более или менее плавную линию, изображающую зависимость $\hat{y}_x$ от x. Такую линию принято называть *линией регрессии*.

Разумеется, если мы имеем эмпирическую совокупность, то точки, изображающие зависимость $\overline{y}_x$ от x, никогда не ложатся на достаточно плавную линию. Поэтому речь может идти только о нахождении такой линии, которая проходила бы наиболее близко ко всем точкам. При этом, конечно, смысл этой «близости» можно понимать по-разному. Можно, например, считать наилучшей ту линию, при которой максимальное отклонение эмпирического значения от расчетного оказывается наименьшим; однако при этом отдельная, наиболее отклоняющаяся от общего хода точка оказывает несоразмерно большое влияние на определение расположения линии. Можно далее считать наилучшей ту линию, при которой площадь между этой линией и ломаной, соединяющей эмпирические точки $(x, \overline{y}_x)$, окажется наименьшей; этот критерий лишен недостатков предыдущего критерия, но он неудобен в вычислительном отношении.

В большинстве случаев наиболее целесообразным является критерий, исходящий из требования, чтобы наименьшей была сумма квадратов отклонений эмпирических точек от линии (так называемый *способ наименьших квадратов*). При этом наиболее отклоняющаяся точка не играет слишком большой и даже решающей роли; известным образом учитывается и требование второго из упомянутых критериев. Для того чтобы те $\overline{y}_x$, которые представляют меньшее число вариант, оказывали соответственно меньшее

влияние на расположение линии регрессии, следует каждую эмпирическую точку брать с надлежащим весом n_x/n. Это можно сделать иначе, используя для расчета непосредственно исходные данные корреляционной решетки, т. е. все точки корреляционного поля; это сильно упрощает вычисления.

Условие минимума суммы квадратов отклонений позволяет свести задачу к системе уравнений, в которой неизвестными являются разыскиваемые параметры (число таких уравнений равно, конечно, числу неизвестных параметров). В общем случае эти уравнения очень сложны. Но если зависимость может быть выражена степенным полиномом (многочленом), т. е. в виде:

$$\hat{y}_x = a_0 + a_1 x + a_2 x^2 + \ldots + a_h x^h, \tag{8.1}$$

то дело упрощается. Это связано с тем, что в выражении (8.1) величина $\hat{y}_x$ зависит от параметров a_0, a_1, a_2, ..., a_h линейно, а поэтому упомянутая выше система уравнений получается также л и н е й н о й. Система же линейных уравнений может быть решена совершенно элементарными приемами, например, способом подстановки или способом исключения неизвестных. В соответствии с формулой (8.1) отклонения $y - \hat{y}_x$ равны:

$$y - \hat{y}_x = y - a_0 - a_1 x - a_2 x^2 - \ldots - a_h x^h.$$

Можно показать, что сумма $\Sigma (y - \hat{y}_x)^2$ имеет минимальное значение, если одновременно выполняются равенства:

$$\left.\begin{aligned}
&a_0 n + a_1 \Sigma x + a_2 \Sigma x^2 + \ldots + a_h \Sigma x^h = \Sigma y;\\
&a_0 \Sigma x + a_1 \Sigma x^2 + a_2 \Sigma x^3 + \ldots + a_h \Sigma x^{h+1} = \Sigma yx;\\
&a_0 \Sigma x^2 + a_1 \Sigma x^3 + a_2 \Sigma x^4 + \ldots + a_h \Sigma x^{h+2} = \Sigma yx^2;\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_0 \Sigma x^h + a_1 \Sigma x^{h+1} + a_2 \Sigma x^{h+2} + \ldots + a_h \Sigma x^{2h} = \Sigma yx^h.
\end{aligned}\right\} \tag{8.2}$$

Эту систему равенств называют *системой нормальных уравнений*. Величины a_0, a_1, a_2, ... a_h являются искомыми неизвестными, а суммы Σx, Σx^2, ... Σx^{2h}, вычисляемые по эмпирическим данным, являются коэффициентами при этих неизвестных. Число уравнений системы (8.2) равно $h + 1$, как и число неизвестных параметров a_0 ..., a_h.

§ 8.2. Линейная регрессия

8.2.1. Пусть предполагаемая зависимость между x и y является линейной, т. е. имеет вид:

$$\hat{y}_x = a_0 + a_1 x. \tag{8.3}$$

Здесь $h = 1$, так что нормальных уравнений будет $h + 1 = 2$. В первом уравнении наивысшая степень x будет $h = 1$, а во втором уравнении $h + 1 = 2$. Таким образом, имеем систему:

$$\left.\begin{aligned} a_0 n + a_1 \sum x &= \sum y; \\ a_0 \sum x + a_1 \sum x^2 &= \sum yx. \end{aligned}\right\} \tag{8.4}$$

Решая эту систему уравнений, находим:

$$a_1 = \frac{\sum xy - \sum x \sum y/n}{\sum x^2 - (\sum x)^2/n}, \tag{8.5}$$

$$a_0 = (\sum y - a_1 \sum x)/n. \tag{8.6}$$

Величину a_1 называют *коэффициентом регрессии* (точнее, коэффициентом линейной регрессии) и обычно обозначают $\beta_{y/x}$. Если подставить в уравнение (8.3) значение a_0 из уравнения (8.5), то получим:

$$\hat{y}_x - \hat{y} = \beta_{y/x}\left(x - \hat{x}\right). \tag{8.7}$$

В таком виде чаще всего записывают уравнение линейной регрессии.

Разделив числитель и знаменатель в уравнении (8.5) на n и пользуясь равенством (3.19), перепишем a_1 в виде:

$$\beta_{y/x} = a_1 = \frac{\sum (x - \hat{x})\ (y - \hat{y})/n}{\sum (x - \hat{x})^2/n} = \frac{cov\{x,\ y\}}{\sigma^2_x}, \tag{8.8}$$

где $\sigma^2{}_x$ есть дисперсия значений x (см. раздел **3.3.1**), а величина

$$cov\{x, y\} = \frac{1}{n} \sum \left(x - \hat{x}\right)\left(y - \hat{y}\right) \tag{8.9}$$

называется *ковариацией* варьирующих величин x и y.

В случае группированной совокупности

$$\beta_{y/x} = \frac{\sum n_{xy}\ (x - \hat{x})\ (y - \hat{y})}{\sum n_x\ (x - \hat{x})^2}. \tag{8.10}$$

Если рассматриваемая эмпирическая совокупность является выборкой из какой-то генеральной совокупности, то центрами рассеяния считаются выборочные средние $\bar{x}$ и $\bar{y}$. Тогда величина

$$b_{y/x} = \frac{\sum n_{xy}\,(x - \bar{x})\,(y - \bar{y})}{\sum n_x\,(x - \bar{x})^2} \tag{8.11}$$

будет выборочной оценкой коэффициента регрессии $\beta_{y/x}$.

Разумеется, при использовании выборок в регрессионном анализе возникают многие дополнительные задачи: нахождение доверительного интервала для коэффициента регрессии, проверка гипотез об этом коэффициенте, оценка значимости различия между двумя линиями регрессии, проверка линейности регрессии и др. По указанным вопросам см. §§ 8.5, 8.6 и 8.7. Сами понятия доверительного интервала, статистической проверки гипотез, значимости различия обсуждаются в разделах **3.7.1** *и* **4.1.1.**

Поскольку для эмпирических совокупностей условные средние $\bar{y}_x$ не лежат на прямых регрессии, то для них не выполняются уравнения (8.7). Поэтому подстановка в выражение

$$y_x = \bar{y} + b_{y|x}(x - \bar{x}) \qquad (8.12)$$

опытных значений x и y будет давать не $\bar{y}_x$, а какие-то другие величины, которые обозначим $\tilde{y}_x$. Эти величины являются ординатами точек, лежащих на прямой регрессии; мы будем называть их *выравненными* условными средними.

Пример 8.3. В табл. 8.2 приведены данные о систолическом (верхнем) давлении крови (y) у 20 женщин разного возраста (x). Построим уравнение регрессии для этих данных.

ТАБЛИЦА 8.2

Возраст (лет) x	71	33	31	55	63	49	58	38	36	64
Давление крови (мм) y	173	118	125	155	153	160	148	142	110	142

Продолжение

Возраст (лет) x	45	68	42	76	34	75	78	62	66	46
Давление крови (мм) y	128	160	136	150	121	166	154	135	146	127

В данном случае удобно для упрощения вычислений ввести вместо y переменную $y' = y - 100$. После этого расчет дает:

$$\sum x = 1090, \quad \sum x^2 = 64056, \quad \sum y' = 851, \quad \sum xy' = 50\,392,$$

так что по формулам (8.5) и (8.6) получаем:

$$a_1 = \frac{50392 - 1090 \cdot 851/20}{64056 - (1090)^2/20} = \frac{4012,5}{4651,0} = 0,863,$$

$$a_0 = (851 - 0,863 \cdot 1090)/20 = -4,5.$$

Следовательно, уравнение регрессии будет:

$$y_x = 95,5 + 0,863x,$$

если от величин y' вернуться к исходным y.

8.2.2. В некоторых случаях предполагаемая зависимость между y и x не является линейной, но может быть приведена к таковой надлежащим преобразованием координат.

Пример 8.4. При облучении гамма-лучами фермента наблюдается падение его активности. В табл. 8.3 приведены значения активности A (в процентах к начальной) при разных дозах

ТАБЛИЦА 8.3

x — D, тыс. P	A	$y = \lg A$	x^2	xy
0	100,0	2,000	0	0,0
3	83,5	1,922	9	5,8
7,5	77,0	1,886	56	14,1
15	39,9	1,600	225	24,0
30	21,8	1,338	900	40,1
45	10,7	1,030	2025	46,4
60	4,43	0,646	3600	38,8
160,5		10,422	6815	169,2
$\sum x$		$\sum y$	$\sum x^2$	$\sum xy$

облучения D. Известно, что активность убывает при увеличении дозы облучения по показательному закону:

$$A = A_0 e^{-\gamma D}$$

(e — основание натуральных логарифмов). Требуется найти коэффициент γ.

Данную нелинейную зависимость можно привести к линейному виду, если выполнить преобразование:

$$\lg A - \lg A_0 = -\gamma D \lg e.$$

Теперь мы ищем регрессию между $D = x$ и $\lg A = y$. Проведя

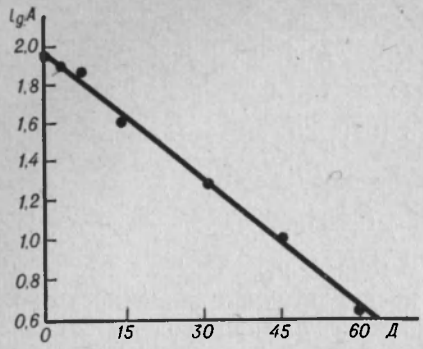


Рис. 8.2. Зависимость активности фермента от дозы облучения.

обычным образом вычисления (они показаны в табл. 8.3), находим:

$$\sum (x-\bar{x})(y-\bar{y}) = \sum xy - \frac{\sum x \sum y}{n} =$$

$$= 169{,}2 - \frac{160{,}5 \cdot 10{,}42}{7} = -69{,}8;$$

$$\sum (x-\bar{x})^2 = \sum x^2 - \frac{\sum x^2}{n} = 6815 - \frac{160{,}5^2}{7} = 3135;$$

$$b_{y|x} = \frac{-69{,}8}{3135} = -0{,}0223.$$

Мы получили величину $-\gamma \lg e$; так как $\lg e = 0{,}4343$, то

$$\gamma = \frac{0{,}0223}{0{,}4343} = 0{,}0514.$$

Зависимость между $\lg A$ и D изображена на рпе. 8.2. Там же показана линия регрессии, проведенная по найденному значению b.

Аналогичным способом поступаем, если предполагаемая функциональная зависимость имеет другой вид,— стараемся привести ее к линейной подходящим преобразованием. Несколько примеров таких преобразований[1] дает табл. 8.4.

При пользовании всеми этими преобразованиями следует помнить, что оценки параметров исходных уравнений, получаемые в результате таких вычислений, являются наилучшими в том смысле, что они удовлетворяют принципу наименьших квадратов относительно преобразованных уравнений, а не исходных.

[1] Более подробно этот вопрос изложен в книге А. Хальда (гл. XVII, § 7).

ТАБЛИЦА 8.4

Исходная функция	Искомые параметры	Преобразованное уравнение	Вспомогательные величины
$y = ax^s$	$a,\ s$	$\lg y = \lg a + s \lg x$	$y' = \lg y;\ a' = \lg a$ $x' = \lg x$
$y = b \lg ax$	$a,\ b$	$y = b \lg a + b \lg x$	$a' = b \lg a,\quad x' = \lg x$
$y = a + \frac{b}{x}$	$a,\ b$	$y = a + b\frac{1}{x}$	$x' = \frac{1}{x}$
$y = ax + bx^3$	$a,\ b$	$\frac{y}{x} = a + bx^2$	$y' = \frac{y}{x},\quad x' = x^2$
$y = ae^{-b/x}$	$a.\ b$	$\lg y = \lg a - \frac{b \lg e}{x}$	$y' = \lg y,\quad x' = \frac{1}{x}$ $a' = \lg a,\quad b' = b \lg e$

§ 8.3. Коэффициент корреляции

8.3.1. В § 8.2 (раздел 8.2.1) было получено выражение

$$\operatorname{cov}\{x, y\} = \frac{1}{n} \sum \left(x - \hat{x}\right)\left(y - \hat{y}\right), \tag{8.13}$$

названное ковариацией признаков x и y. Нетрудно убедиться в том, что эта величина может характеризовать количественно степень (или «силу») корреляции (связи) между двумя признаками. Действительно, при сильной корреляции положительные отклонения $x - \hat{x}$ будут чаще всего сочетаться с положительными же отклонениями $y - \hat{y}$, а отрицательные $x - \hat{x}$ — с отрицательными $y - \hat{y}$. Поэтому произведения $(x - \hat{x})\ (y - \hat{y})$ будут, как правило, положительными и при суммировании будут складываться, так что сумма в формуле (8.13) будет иметь почти максимальное значение. В случае же слабой корреляции положительные $x - \hat{x}$ будут примерно одинаково часто сочетаться как с положительными, так и с отрицательными $y - \hat{y}$; то же можно сказать и об отрицательных $x - \hat{x}$. В результате сумма в формуле (8.13) будет содержать примерно равное число положительных и отрицательных произведений $(x - \hat{x})\ (y - \hat{y})$, поэтому при суммировании будет происходить почти полная компенсация, и сумма будет близка к нулю.

Мера корреляции не должна, однако, меняться при переходе от одних единиц к другим в величинах x и y. Поэтому ясно, что вместо самих отклонений $x - \hat{x}$ и $y - \hat{y}$ следует в данном случае подставлять приведенные безразмерные величины $(x - \hat{x})/\sigma\{x\}$ $(y - \hat{y})/\sigma\{y\}$. Таким образом, в качестве меры корреляции мы можем принять величину:

$$\rho = \frac{1}{n} \sum \frac{x - \hat{x}}{\sigma_x} \cdot \frac{y - \hat{y}}{\sigma_y} = \frac{\frac{1}{n} \sum (x - \hat{x})(y - \hat{y})}{\sigma_x \sigma_y}, \qquad (8.14)$$

т. е.

$$\rho = \frac{\operatorname{cov}\{x, y\}}{\sigma_x \sigma_y}, \qquad (8.15)$$

где $\sigma_x = \sigma\{x\}$ и $\sigma_y = \sigma\{y\}$ — стандартные отклонения значений x и y (см. раздел **3.3.1**). Эта величина называется *коэффициентом корреляции* между x и y. Для группированных совокупностей можно вычислять ρ по формуле:

$$\rho = \frac{\sum n_{xy} \left(x - \hat{x}\right)\left(y - \hat{y}\right)}{\sqrt{\sum n_x \left(x - \hat{x}\right)^2 \sum n_y \left(y - \hat{y}\right)^2}}. \qquad (8.16)$$

Несложный анализ (который мы здесь, однако, опускаем) показывает, что коэффициент корреляции не может превосходить по абсолютной величине единицу, т. е. он может принимать значения лишь в пределах от -1 до $+1$. Если $\rho = +1$, то имеет место полная положительная корреляция, а если $\rho = -1$, то полная отрицательная корреляция. Если же $\rho = 0$, то корреляция отсутствует; при этом линия регрессии представляет собой прямую, параллельную оси абсцисс, т. е. $\hat{y}_x$ не зависит от x. При вычислении знак коэффициента регрессии получается автоматически — он совпадает со знаком $\operatorname{cov}\{x, y\}$. Выборочной оценкой коэффициента корреляции будет:

$$r = \frac{\sum \left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\sqrt{\sum \left(x - \bar{x}\right)^2 \sum \left(y - \bar{y}\right)^2}}, \qquad (8.17)$$

причем для удобства вычислений следует учесть, что

$$\sum (x - \bar{x})(y - \bar{y}) = \sum xy - \sum x \sum y / n,$$

$$\sum (x - \bar{x})^2 = \sum x^2 - (\sum x)^2 / n,$$

$$\sum (y - \bar{y})^2 = \sum y^2 - (\sum y)^2 / n.$$

Однако эта оценка является смещенной. Почти несмещенную оценку коэффициента корреляции можно получить по формуле:

$$r^* = r\left\{1 + \frac{1-r^2}{2(n-4)}\right\}. \qquad (8.18)$$

Пример 8.5. Найдем оценку коэффициента корреляции между систолическим давлением крови у женщин и возрастом (см. пример 8.3 в 8.1.1).

Для вычисления r нужно, в дополнение к полученным выше величинам Σx, Σx^2, $\Sigma y'$, $\Sigma xy'$, вычислить еще $\Sigma(y')^2 = 41795$. После этого имеем:

$$\sum (y' - \bar{y'})^2 = 41795 - 851^2/20 = 5585,$$

так что:

$$r = \frac{4012{,}5}{\sqrt{4651 \cdot 5585}} = 0{,}787;$$

значения $\Sigma(x - \bar{x})^2 = 4651$ и $\Sigma(x - \bar{x})(y' - \bar{y'}) = 4012{,}5$ взяты из примера § 8.2. Теперь по формуле (8.18) получаем:

$$r^* = 0{,}787\left(1 + \frac{1 - 0{,}787^2}{2 \cdot 16}\right) = 0{,}796.$$

Далее здесь говорится о построении доверительного интервала для коэффициента корреляции и о сравнении двух коэффициентов корреляции. Прежде чем читать об этом, обязательно просмотрите разделы **3.7.1** *и* **4.4.1**, *в которых разъясняются понятия доверительного интервала, статистической проверки гипотез, значимости различия. О стандартной ошибке и о распределении Стьюдента говорится в разделах* **3.5.1** *и* **3.7.2**.

8.3.2. Чтобы построить доверительный интервал для коэффициента корреляции, нужно знать распределение величины $(r - \rho)/s_r$, где s_r — оценка стандартной ошибки коэффициента корреляции. Напомним, что при построении доверительного интервала для среднего значения исходят из того, что величина $(\bar{x} - \hat{x})/s_{\bar{x}}$ имеет распределение Стьюдента. Последнее же условие выполняется тогда, когда выборочные средние распределены нормально (для чего в свою очередь требуется, чтобы распределение вариант x в совокупности не слишком отличалось от нормального). Ясно, что в случае выборочного коэффициента корреляции распределение заведомо отличается от нормального, так как r вообще может принимать значения только от -1 до $+1$. Это отклонение от нормального распределения тем заметнее, чем ближе генеральный коэф-

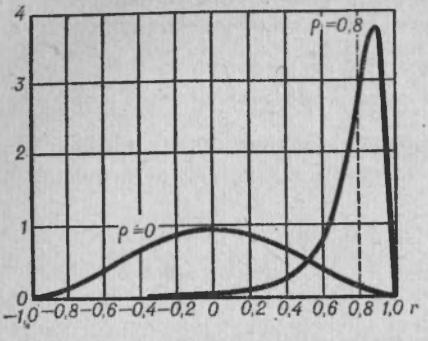


Рис. 8.3. Распределение выборочного коэффициента корреляции.

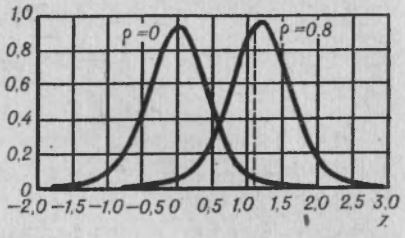


Рис. 8.4. Распределение преобразованного выборочного коэффициента корреляции.

фициент ρ к единице (рис. 8.3); в то же время именно последний случай представляет наибольший интерес.

Чтобы обойти это затруднение, было предложено (Р. Фишер) ввести вспомогательную величину:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}, \tag{8.19}$$

связанную взаимно однозначно с r; при изменении r от -1 до $+1$ величина z меняется от $-\infty$ до $+\infty$, как и всякая нормально распределенная величина. Математический анализ показывает, что распределение величины z мало отклоняется от нормального даже при близких к 1 значениях ρ (на рис. 8.4 изображен график плотности распределения z при $\rho = 0$ и $\rho = 0{,}8$). Доказывается также, что стандартная ошибка величины z равна:

$$\sigma_z = \frac{1}{\sqrt{n-3}}, \tag{8.20}$$

где n — объем выборки.

Для перехода от r к z и обратно составлены таблицы, облегчающие вычисления (табл. 8.5 и 8.6).

Пример 8.6. Найдем 95%-ный доверительный интервал для ρ по данным из табл. 8.2 (см. пример 8.5 в этом параграфе, $r^* = 0{,}796$).

В нашем случае $n = 20$, так что по формуле (8.20) получаем:

$$\sigma_z = \frac{1}{\sqrt{20-3}} = \frac{1}{\sqrt{17}} = 0{,}2425.$$

ТАБЛИЦА 8.5

Значения r для z от 0,00 до 2,99. Ноль целых и запятая опущены (по книге А. М. Длина, с. 432)

z	0	1	2	3	4	5	6	7	8	9
0,0	0000	0100	0200	0300	0400	0500	0599	0699	0798	0898
1	0997	1096	1194	1293	1391	1489	1586	1684	1781	1877
2	1974	2070	2165	2260	2355	2449	2543	2636	2729	2821
3	2913	3004	3095	3185	3275	3364	3452	3540	3627	3714
4	3800	3885	3969	4053	4136	4219	4301	4382	4462	4542
5	4621	4699	4777	4854	4930	5005	5080	5154	5227	5299
6	5370	5441	5511	5580	5649	5717	5784	5850	5915	5980
7	6044	6107	6169	6231	6291	6351	6411	6469	6427	6584
8	6640	6696	6751	6805	6858	6911	6963	7014	7064	7114
9	7163	7211	7259	7306	7352	7398	7443	7487	7531	7574
1,0	7616	7658	7699	7739	7779	7818	7857	7895	7932	7969
1	8005	8041	8076	8110	8144	8178	8210	8243	8275	8306
2	8337	8367	8397	8426	8455	8483	8511	8538	8565	8591
3	8617	8643	8668	8692	8717	8741	8764	8787	8810	8832
4	8854	8875	8896	8917	8937	8957	8977	8996	9015	9033
5	9051	9069	9087	9104	9121	9138	9154	9170	9186	9201
6	9217	9232	9246	9261	9275	9289	9302	9316	9329	9341
7	9354	9366	9379	9391	9402	9414	9425	9436	9447	9458
8	9468	9478	9488	9498	9508	9518	9527	9536	9545	9554
9	9562	9571	9579	9587	9595	9603	9611	9618	9626	9633
2,0	9640	9647	9654	9661	9668	9674	9680	9686	9693	9699
1	9704	9716	9716	9722	9727	9732	9738	9743	9748	9753
2	9757	9762	9767	9771	9776	9780	9785	9789	9793	9797
3	9801	9805	9809	9812	9816	9820	9823	9827	9830	9834
4	9837	9840	9843	9846	9849	9852	9855	9858	9861	9864
5	9866	9869	9871	9874	9876	9879	9881	9884	9886	9888
6	9890	9892	9894	9897	9899	9901	9903	9904	9906	9908
7	9910	9912	9914	9915	9917	9919	9920	9922	9923	9925
8	9926	9928	9929	9931	9932	9933	9935	9936	9937	9938
9	9940	9941	9942	9943	9944	9945	9946	9947	9948	9949

Поскольку значению $r^* = 0{,}796$ соответствует $z = 1{,}0877$ (по табл. 8.6 с применением интерполяции), то

$$z_н;\ z_в = z \mp u_{0,95}\sigma_z = 1{,}0877 \mp 1{,}96 \cdot 0{,}2425 = 0{,}6124;\ 1{,}5630.$$

Этим значениям z соответствуют (по табл. 8.5) значения $r^*_н = 0{,}546$, $r^*_в = 0{,}916$. Это и будут 95%-ные доверительные границы генерального коэффициента корреляции.

Пример 8.7. Для двух групп больных дифтерией ($n' = 186$, $n'' = 219$) найдены оценки коэффициентов корреляции, характеризующих связь между сроками введения противодифтерийной сыворотки и летальностью заболевания: $r' = 0{,}962$, $r'' =$

ТАБЛИЦА 8.6

Значения величины $z(r) = \frac{1}{2} \ln \frac{1+r}{1-r}$

(по книге А. М. Длина, с. 431).

r	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0500	0,0601	0,0701	0,0802	0,0902
0,1	0,1003	0,1104	0,1206	0,1307	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
0,2	0,2027	0,2132	0,2237	0,2342	0,2448	0,2554	0,2661	0,2769	0,2877	0,2986
0,3	0,3095	0,3205	0,3316	0,3428	0,3541	0,3654	0,3769	0,3884	0,4001	0,4118
0,4	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
0,5	0,5493	0,5627	0,5763	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
0,6	0,6931	0,7089	0,7250	0,7414	0,7582	0,7753	0,7928	0,8107	0,8291	0,8480
0,7	0,8673	0,8872	0,9076	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
0,8	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
0,9	1,4722	1,5275	1,5890	1,6584	1,7380	1,8318	1,9459	2,0923	2,2976	2,6467
0,99	2,6466	2,6996	2,7587	2,8257	2,9031	2,9945	3,1063	3,2504	3,4534	3,8002

$= 0{,}971$. Найдем общую по обеим группам оценку коэффициента корреляции и 99%-ные доверительные границы для ρ.

Пользуясь табл. 8.6, находим:

$$z' = 1{,}9752, \quad z'' = 2{,}1128.$$

Производим усреднение, взвешивая по обратным дисперсиям:

$$1/\sigma_z^2 = n' - 3 = 183, \qquad 1/\sigma_z^2 = n'' - 3 = 216,$$

$$z = \frac{183 \cdot 1{,}9752 + 216 \cdot 2{,}1128}{183 + 216} = \frac{817{,}83}{399} = 2{,}0497.$$

Далее вычисляем:

$$\sigma_z = \sqrt{\sigma_{z'}^2 + \sigma_{z''}^2} = \sqrt{\frac{1}{n' - 3} + \frac{1}{n'' - 3}} =$$

$$= \sqrt{0{,}005464 + 0{,}004629} = 0{,}10046.$$

Значит,

$$z_н;\ z_в = 2{,}0497 \mp 2{,}58 \cdot 0{,}10046 = 1{,}7905;\ 2{,}3089.$$

Теперь по табл. 8.5 получаем:

$$r_н = 0{,}946, \quad r_в = 0{,}980.$$

8.3.3. Переход от r к z используется также при сравнении выборочных коэффициентов корреляции, т. е. при проверке гипотезы $\rho' = \rho''$.

Поскольку распределение выборочных z может считаться нормальным, то критерием различия будет:

$$u_{z'-z''} = \frac{|z'-z''|}{\sigma_{z'-z''}} > u_{\alpha}, \tag{8.21}$$

где, согласно формуле (8.20) и (3.22),

$$\sigma_{z'-z''} = \sqrt{\frac{1}{n'-3} + \frac{1}{n''-3}}. \tag{8.22}$$

Пример 8.8. В примере 8.7 расчет средней для двух выборок ($n' = 186$, $n'' = 219$) оценки коэффициента корреляции исходил из допущения, что различие между двумя выборочными коэффициентами корреляции ($r' = 0{,}962$, $r'' = 0{,}971$) незначимо, так что они могут считаться оценками одного и того же параметра ρ. Верно ли это допущение?

Как было найдено в примере 8.7, $z' = 1{,}9752$, $z'' = 2{,}1128$ и

$$\sigma_{z''-z'} = \sqrt{\sigma^2_{z'} + \sigma^2_{z''}} = 0{,}10046.$$

значит,

$$u_{z''-z'} = \frac{2{,}1128 - 1{,}9752}{0{,}10046} = 1{,}37,$$

в то время как $u_{0{,}05} = 1{,}96$. Следовательно, различие незначимо.

8.3.4. Критерий (8.21) пригоден для сравнения не только двух выборочных коэффициентов корреляции, но и для сравнения выборочного и теоретического коэффициентов. В частности, особый интерес представляет случай, когда теоретическое значение коэффициента корреляции равно нулю, т. е. когда нулевая гипотеза состоит в том, что корреляция отсутствует. Поскольку в теоретической совокупности предполагается $n = \infty$, то $\sigma_{z(\text{выб})-z(\text{теор})}$ дается просто формулой (8.20). Таким образом, коэффициент корреляции может считаться значимо отличным от нуля, если:

$$z\sqrt{n-3} > u_{\alpha} \tag{8.23}$$

или, иными словами, если

$$z > \frac{u_{\alpha}}{\sqrt{n-3}} \equiv z_{\alpha}(n), \tag{8.24}$$

причем здесь применяется двусторонний критерий. Так как r и z связаны взаимно однозначно равенством (8.19), то можно вычислить значения r_{α}, соответствующие каждому из значений z_{α}. Критические значения r_{α} даны в табл. 8.7.

ТАБЛИЦА 8.7

Критические значения r_α выборочного коэффициента корреляции. r незначим при $r \leq r_\alpha$ и значим при $r > r_\alpha$ (по книге Д. Б. Оуэна, с. 510)

n	$\alpha = 5\%$	$\alpha = 1\%$	n	$\alpha = 5\%$	$\alpha = 1\%$
4	0,950	0,990	26	0,388	0,496
5	0,878	0,959	27	0,381	0,487
6	0,811	0,917	28	0,374	0,478
7	0,754	0,874	29	0,367	0,470
8	0,707	0,834	30	0,361	0,463
9	0,666	0,798	35	0,332	0,435
10	0,632	0,765	40	0,310	0,407
11	0,602	0,735	45	0,292	0,384
12	0,576	0,708	50	0,277	0,364
13	0,553	0,684	60	0,253	0,333
14	0,532	0,661	70	0,234	0,308
15	0,514	0,641	80	0,219	0,288
16	0,497	0,623	90	0,206	0,272
17	0,482	0,606	100	0,196	0,258
18	0,468	0,590	125	0,175	0,230
19	0,456	0,575	150	0,160	0,210
20	0,444	0,561	200	0,138	0,182
21	0,433	0,549	250	0,124	0,163
22	0,423	0,537	300	0,113	0,148
23	0,413	0,526	400	0,098	0,128
24	0,404	0,515	500	0,088	0,115
25	0,396	0,505	1000	0,062	0,081

§ 8.4. Корреляция рангов

О порядковых признаках и рангах см. раздел **2.2.1.** *Для понимания второй части этого параграфа требуется знать понятия доверительной вероятности, статистической проверки гипотез, критериев значимости; об этом см. разделы* **3.7.1** *и* **4.4.1.**

8.4.1. Изложенный в предыдущем параграфе способ, позволяющий определить степень связанности между двумя признаками, может быть в известной мере применен и к порядковым совокупностям, где каждая варианта характеризуется не численным значением, а лишь своим рангом.

Пример 8.9. Бегуны, ранги которых при построении по росту были 1, 2, ... 10, заняли на состязании места:

6, 5, 1, 4, 2, 7, 8, 10, 3, 9.

Как велика корреляция между ростом и быстротой бега?

Если условно считать, что ранг варианты есть некоторая единица измерения, то корреляцию между ранжированными признаками можно характеризовать обычным выражением:

$$\rho = \frac{\sum\left(x-\hat{x}\right)\left(y-\hat{y}\right)}{\sqrt{\sum\left(x-\hat{x}\right)^2 \sum\left(y-\hat{y}\right)^2}}. \qquad (8.16)$$

Однако последовательным рангам, как правило, не соответствуют равноотстоящие значения. Это приводит к тому, что найденной таким способом величине (в данном случае ее называют *показателем корреляции рангов* и обозначают через ρ^S) нельзя приписывать такую же количественную определенность, как коэффициенту корреляции для признаков с количественной градацией вариант (отсюда, между прочим, следует, что вычисление величины ρ^S с точностью более двух значащих цифр не имеет смысла).

Для примера 8.9 расчет по формуле (8.16) дает $\rho^S \approx 0{,}45$. Корреляция между рангами оказалась сравнительно небольшой.

Поскольку варианты ранжированной совокупности составляют последовательный ряд рангов от 1 до n (где n — объем совокупности), в большинстве случаев совпадающий с рядом натуральных чисел, то вычисление показателя корреляции рангов может быть упрощено. Действительно, пусть d есть разность между рангом признака y и соответствующим ему рангом признака x: $d = y - x$; в нашем примере $d_1 = y_1 - x_1 = 6-1 = 5$, $d_2 = y_2 - x_2 = 5-2 = 3$ и т. д.

Если бы корреляция была полной, так что ранги вариант по обоим признакам были всегда одинаковы (т. е. варианта с рангом 1 по одному признаку имела бы также ранг 1 и по другому признаку, и так для всех вариант), то все d были бы равны нулю. Всякое отступление от полной корреляции должно приводить к появлению отличных от нуля значений d. Поэтому совокупность значений d может служить мерой корреляции. Эту совокупность можно характеризовать одним числом, например, средней разностью $d_{ср}$. Однако это не должна быть средняя арифметическая $\bar{d} = \Sigma d/n$, так как сумма Σd всегда равна нулю (в самом деле, $\Sigma d = \Sigma(y - x) = \Sigma y - \Sigma x = 0$); как обычно в таких случаях, для средней разности можно принять среднюю квадратическую, связанную с суммой квадратов Σd^2. Очевидно, формула должна быть такова, чтобы ρ^S было тем меньше, чем больше Σd^2.

Для получения этой формулы будем исходить из формулы (8.16). В данном случае

$$\Sigma x = \Sigma y = 1 + 2 + \ldots + n;$$

$$\Sigma x^2 = \Sigma y^2 = 1^2 + 2^2 + \ldots + n^2.$$

Из алгебры известно, что

$$1+2+\ldots+n=\frac{n(n+1)}{2},$$

$$1^2+2^2+\ldots+n^2=\frac{n(n+1)(2n+1)}{6},$$

так что при помощи элементарных преобразований получаем:

$$\sum\left(x-\hat{x}\right)^2=\sum\left(y-\hat{y}\right)^2=\frac{n(n^2-1)}{12},$$

$$\sum\left(x-\hat{x}\right)\left(y-\hat{y}\right)=\frac{n(n^2-1)}{12}-\frac{1}{2}\sum d^2.$$

Поэтому окончательно (формула Спирмена):

$$\rho^S=\frac{\sum\left(x-\hat{x}\right)\left(y-\hat{y}\right)}{\sqrt{\sum\left(x-\hat{x}\right)^2\sum\left(y-\hat{y}\right)^2}}=1-\frac{6\sum d^2}{n(n^2-1)}. \qquad (8.25)$$

Выборочную оценку этой величины обозначают r^S.
Применяя эту формулу к нашему примеру (см. табл. 8.8), имеем:

$$\sum d^2=25+9+4+0+9+\ldots+36+1=90,$$

$$n(n^2-1)=10(100-1)=990, \qquad r^S=1-\frac{6\cdot 90}{990}\approx 0{,}45,$$

как и ранее.

ТАБЛИЦА 8.8

x	1	2	3	4	5	6	7	8	9	10
y	6	5	1	4	2	7	8	10	3	9
d	5	3	—2	0	—3	1	1	2	—6	—1
d^2	25	9	4	0	9	1	1	4	36	1

Приведем еще один пример нахождения показателя корреляции рангов.

Пример 8.10. Цветные диски, имевшие порядок оттенков 1, 2, ..., 15, были расположены испытуемым в следующем порядке:

7 4 2 3 1 10 6 8 9 5 11 15 14 12 13.

Очевидно, показатель корреляции между действительными и наблюденными рангами будет характеризовать способность испытуемого различать оттенки цветов. Составив табл. 8.9, имеем:

$$\sum d^2 = 118, \quad n(n^2-1) = 15(225-1) = 3360;$$

$$r^S = 1 - \frac{6 \cdot 118}{3360} \approx 1 - 0{,}21 = 0{,}79.$$

ТАБЛИЦА 8.9

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y	7	4	2	3	1	10	6	8	9	5	11	15	14	12	13
d	6	2	−1	−1	−4	4	−1	0	0	−5	0	3	1	−2	−2
d^2	36	4	1	1	16	16	1	0	0	25	0	9	1	4	4

В заключение отметим, что вычисление показателя корреляции рангов может быть использовано для грубой, но быстрой оценки коэффициента корреляции при небольшом числе вариант.

Пример 8.11. Произведем приближенную оценку коэффициента корреляции между систолическим давлением крови у женщин и возрастом по данным табл. 8.2.

Составив табл. 8.10, имеем $\Sigma d^2 = 292{,}5$, поэтому

$$r^S = 1 - \frac{6 \cdot 292{,}5}{20 \cdot 399} = 1 - 0{,}22 = 0{,}78,$$

что весьма близко к найденной в примере 8.5. оценке коэффициента корреляции 0,787.

Конечно, не всегда совпадение бывает таким хорошим, но в большинстве случаев значения r^S и r получаются довольно близкими. Теоретический анализ показывает, что максимальное расхождение между ρ^S и ρ не превышает 3%; оно достигается при $\rho \approx 0{,}6$, а при $\rho \to 0$ и $\rho \to 1$ это расхождение стремится к нулю. Что касается большей простоты вычисления r^S, то она очевидна из сравнения примеров 8.11 и 8.5.

ТАБЛИЦА 8.10

Возраст (лет) x	71	33	31	55	63	49	58	38	36	64
Ранг x	17	2	1	10	13	9	11	5	4	14
Давление крови (мм рт. ст.) y	173	118	125	155	153	160	148	142	110	142
Ранг y	20	2	4	16	14	17,5	11,5	9,5	1	9,5
d	3	0	3	6	1	8,5	0,5	4,5	−3	−4,5
d^2	3	0	9	36	1	72,25	0,25	20,25	9	20,25

Продолжение

Возраст (лет) x	45	68	42	76	34	75	78	62	66	46
Ранг x	7	16	6	19	3	18	20	12	15	8
Давление крови (мм рт. ст.) y	128	160	136	150	121	166	154	135	148	127
Ранг y	6	17,5	8	13	3	19	15	7	11,5	5
d	−1	1,5	2	−6	0	1	−5	−5	−3,5	−3
d^2	1	2,25	4	36	0	1	25	25	12,5	9

8.4.2. В разделе 8.3.4 было показано, что коэффициент корреляции генеральной совокупности может считаться отличным от нуля (с определенной вероятностью) только тогда, когда выборочный коэффициент корреляции превышает некоторое минимальное значение, зависящее от выбранного уровня значимости и от объема выборки.

Аналогично можно указать минимальные доверительные значения выборочного показателя корреляции рангов, r^S. Значения $r^S_{0,05}(n)$ и $r^S_{0,01}(n)$ приведены в табл. 8.11.

ТАБЛИЦА 8.11

Критические значения r_a^S выборочного показателя корреляции рангов

n	5%	1%	n	5%	1%	n	5%	1%
5	0,94		17	0,48	0,62	29	0,37	0,48
6	0,85		18	0,47	0,60	30	0,36	0,47
7	0,78	0,94	19	0,46	0,58	31	0,36	0,46
8	0,72	0,88	20	0,45	0,57	32	0,36	0,45
9	0,68	0,83	21	0,44	0,56	33	0,34	0,45
10	0,64	0,79	22	0,43	0,54	34	0,34	0,44
11	0,61	0,76	23	0,42	0,53	35	0,33	0,43
12	0,58	0,73	24	0,41	0,52	36	0,33	0,43
13	0,56	0,70	25	0,40	0,51	37	0,33	0,42
14	0,54	0,68	26	0,39	0,50	38	0,32	0,41
15	0,52	0,66	27	0,38	0,49	39	0,32	0,41
16	0,50	0,64	28	0,38	0,48	40	0,31	0,40

r^S незначим при $r^S \ll r_a^S$ и значим при $r^S > r_a^S$.

§ 8.5. Доверительная зона регрессии

Чтобы понять содержание настоящего параграфа и правильно пользоваться приводимыми в нем рекомендациями, необходимо иметь четкое представление о ряде таких статистических понятий, как дисперсия и стандартное отклонение (см. раздел **3.3.1***), их выборочные оценки (см. раздел* **3.4.1***), стандартная ошибка (см. раздел* **3.5.1***), доверительный интервал (см. раздел* **3.7.1***), распределение Стьюдента (см. раздел* **3.7.2***), условное среднее (см. раздел* **8.1.3***), линия регрессии (см. раздел* **8.1.3***), выравненное условное среднее (см. раздел* **8.2.1***), коэффициент регрессии (см. раздел* **8.2.1***), коэффициент корреляции (см. раздел* **8.3.1***).*

8.5.1. Выборочное значение коэффициента регрессии[1] b является оценкой соответствующего генерального коэффициента β и варьирует около него с дисперсией σ_b^2. Это значит, что «истинная» линия регрессии заключена (при больших объемах выборки с вероятностью 68,3%) внутри пары вертикальных углов, образованных пересечением в точке (x, y) двух прямых PP и QQ с наклонами

[1] Для упрощения обозначений мы будем в дальнейшем писать b и β вместо $b_{y/x}$ и $\beta_{y/x}$.

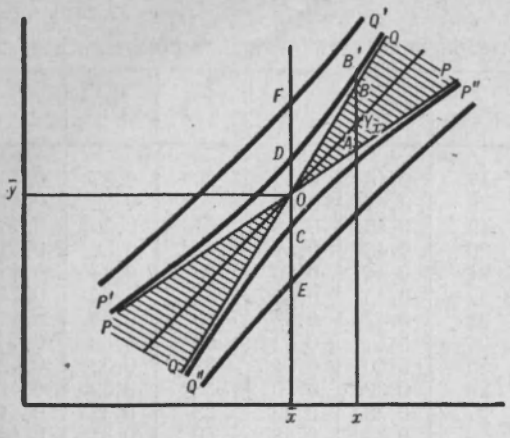


Рис. 8.5. Доверительные зоны регрессии.

$b - \sigma_b$ и $b + \sigma_b$ (заштрихованная часть на рис. 8.5). Поэтому для каждого значения x «истинное» значение выравненного условного среднего заключено (с вероятностью 68,3%) в интервале $\bar{y}_x \pm \pm \sigma_b(x - \bar{x})$; на рис. 8.5 этот интервал изображен отрезком AB.

Однако имеется некоторая неопределенность не только в наклоне «истинной» линии регрессии $\bar{y}_x$, но и в ее положении по высоте (т. е. в направлении y). Очевидно, среднее отклонение вариант от линии регрессии в направлении y будет характеризоваться величиной:

$$\sqrt{\frac{1}{n}\sum_x\sum_y\left(y-\hat{y}_x\right)^2} = \sigma\left\{y-\hat{y}_x\right\}, \qquad (*)$$

а поэтому стандартная ошибка в расположении линии регрессии (в смысле смещения по вертикали) будет:

$$\sigma_{y_x}^{(\text{верт.})} = \frac{\sigma\left\{y-\hat{y}_x\right\}}{\sqrt{n}}; \qquad (**)$$

интервал $\bar{y} + \sigma_{y_x}{}^{(\text{верт.})}$ изображен отрезком CD на рис. 8.5.

Величина $\sigma_{y_x}{}^{(\text{верт.})}$ имеет размерность y; например, при изучении регрессии веса мышей на их возраст $\sigma_{y_x}{}^{(\text{верт.})}$ имеет размерность веса. Что касается величины σ_b, то она должна иметь ту же размерность, что и b, т. е. размерность y/x. Величина с такой размерностью получится, если мы разделим $\sigma_{y_x}{}^{(\text{верт.})}$ на σ_x, т. е. примем:

$$\sigma_b = \frac{\sigma_{y_x}^{(\text{верт.})}}{\sigma_x}. \qquad (8.26)$$

В первом приближении полная стандартная ошибка истинного условного среднего будет:

$$\sigma_{y_x} = \sqrt{\left[\sigma_b\left(x-\bar{x}\right)\right]^2 + \left[\sigma_{y_x}^{(\text{верт.})}\right]^2} \qquad (***)$$

или, подставив значение σ_b из (8.25),

$$\sigma_{y_x} = \sigma_{y_x}^{(\text{верт.})} \sqrt{\left(\frac{x-\bar{x}}{\sigma_x}\right)^2 + 1}. \qquad (8.27)$$

При $x = \bar{x}$ первое слагаемое под корнем равно нулю, и мы получаем $\sigma_{y_x} = \sigma_{y_x}^{(\text{верт.})}$; при больших же значениях $x - \bar{x}$ второе слагаемое (единица) под корнем может быть отброшено, и тогда границами доверительного интервала для $\hat{y}_x$ можно считать прямые PP и QQ на рис.8.5. Вообще же границы доверительного интервала для $\hat{y}$ представляют собой пару кривых $P'DQ'$ и $Q''CP''$ — две ветви гиперболы. Область, заключенная между этими кривыми, называется *доверительной зоной регрессии* (в данном случае это будет 68,3%-ная зона).

Практически построение P%-ной доверительной зоны производится следующим образом[1]. Вычислив по формуле (**) величину $\sigma_{y_x}^{(\text{верт.})}$, откладывают значения $\bar{y} - t_P \sigma_{y_x}^{(\text{верт.})}$ и $\bar{y} + t_P \sigma_{y_x}^{(\text{верт.})}$ на вертикали $x = \bar{x}$ (точки C и D на рис. 8.5). Затем по формуле (8.26) находят σ_b и проводят через точку $(\bar{x}, \bar{y})$ две прямые с наклонами $b - t_P\sigma_b$ и $b + t_P\sigma_b$ (прямые PP и QQ на рис. 8.5). После этого с помощью лекала проводят кривые, проходящие через точки C и D и приближающиеся асимптотически к прямым PP и QQ. При этом надо учитывать, что кривые становятся достаточно близкими к своим асимптотам при $x - \bar{x} \approx 10\,\sigma_x$ (тогда $BB' \approx \approx \frac{1}{20} OD$).

Наибольшую трудность представляет вычисление величины $\sigma\{y - \hat{y}_x\}$ по формуле (*). Однако это вычисление можно упростить. Именно, если подставить в (*) значения $\hat{y}_x$ из уравнения регрессии (8.7), то после ряда тождественных преобразований можно получить:

$$\sigma^2\left\{y - \hat{y}_x\right\} = \sigma_y^2 - \frac{[cov\,\{x, y\}]^2}{\sigma_x^2} = \sigma_y^2(1-\rho^2), \qquad (8.28)$$

[1] Описываемая процедура дает приближенный результат. Точный способ построения доверительной зоны регрессии приведен в сборнике таблиц Л. Н. Большева и Н. В. Смирнова, с. 81—84.

так что

$$\sigma_{y_x}^{(\text{верт.})} = \sqrt{\sigma_y^2 \frac{1-\rho^2}{n}}.$$

Выборочная оценка этой величины будет:

$$s_{y_x}^{(\text{верт.})} = \sqrt{\frac{(1-r^2)\sum(y-\bar{y})^2}{n(n-2)}}, \qquad (8.29)$$

и тогда в соответствии с формулой (8.26)

$$s_b = \sqrt{\frac{\sum(y-\bar{y})^2}{\sum(x-\bar{x})^2} \cdot \frac{1-r^2}{n-2}}. \qquad (8.30)$$

При помощи величины s_b можно проверять значимость регрессии, т. е. правильность гипотезы $\beta = 0$. Эта проверка основывается на том, что величина

$$t = \frac{b-\beta}{s_b},$$

где s_b дается формулой (8.30), имеет распределение Стьюдента с $n-2$ степенями свободы.

Пользуясь формулами (8.27), (8.26) и (8.30), запишем доверительную зону линейной регрессии в виде:

$$y_x = \bar{y} + b(x-\bar{x}) \pm$$

$$t_P \sqrt{\frac{1-r^2}{n-2}\sum(y-\bar{y})^2\left[\frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2} + \frac{1}{n}\right]}.$$

8.5.2. В некоторых случаях может представлять интерес доверительная зона не для условных средних $\hat{y}_x$, а для результатов отдельных измерений $y(x)$. Рассмотрим такой пример. Диагностическим признаком определения нарушений сократительной способности сердечной мышцы может служить изменение продолжительности фазы изгнания. Так как эта величина (обозначим ее y) связана корреляционно с общей продолжительностью сердечного цикла (которую обозначим x), а последняя сама варьирует даже в норме, то совокупность значений y, соответствующих норме (конечно, при определенном уровне доверительной вероятности), занимает некоторую зону около линии регрессии y на x, размер которой зависит от выбранного доверительного уровня. Построим по результатам обследования здоровых людей такую $P\%$-ную доверительную зону; тогда, если точка, изображающая значения x и y для какого-нибудь обследуемого, окажется вне указанной зоны, то это будет свидетельствовать с вероятностью P о наличии

у него нарушения сократительной способности сердечной мышцы.

При нахождении доверительной зоны для $y(x)$ нужно в выражении (***) этого параграфа учесть также варьирование отдельных значений y около условных средних $\overline{y}_x$. Это варьирование отражает величина $\sigma\{y - \overline{y}_x\}$, равная, согласно (**), $\sqrt{n}\,\sigma_{\overline{y}_x}^{\text{верт.}}$. Тогда вместо (8.27) будем иметь:

$$\sigma_{y(x)} = \sigma_{\overline{y}_x}^{(\text{верт.})} \sqrt{\left(\frac{x-\overline{x}}{\sigma_x}\right)^2 + 1 + n}\,. \qquad (8.31)$$

так что границы доверительной зоны для $y(x)$ будут пересекать вертикаль $x = \overline{x}$ в точках (E и F на рис. 8.5), отстоящих от $\overline{y}$ на $\pm t_P\,\sigma_{\overline{y}_x}^{(\text{верт.})}\sqrt{n+1}$; используя оценку $\sigma_{\overline{y}_x}^{(\text{верт.})}$ из (8.29), имеем:

$$\left[t_P\,\hat{\sigma}_{y(x)}\right]_{x=\overline{x}} = t_P\sqrt{\frac{n+1}{n(n-2)}(1-r^2)\sum(y-\overline{y})^2}\,. \qquad (8.32)$$

Пример 8.12. В табл. 8.12 записаны результаты серии

ТАБЛИЦА 8.12

Доза, $10^3 P$ (x)	Остаточный уровень размножения (y), %									
1	94	96	97	92	95	93	96	94	95	
2	87	91	86	88	88	90	89	89	95	87
3	83	85	82	84	81	81	85	83		
4	77	71	77	79	76	78	75	79	75	
5	71	68	70	69	69	68	72			
6	63	66	64	64	63	65	67	65	68	62
7	62	58	60	59	64	61	63			

опытов, в которых определялось уменьшение темпа размножения одного вида бактерий под действием рентгеновского облучения; результаты выражены в процентах к уровню размножения необлученных бактерий. Проведем регрессионный анализ этих данных.

Здесь численности, отвечающие отдельным значениям аргумента x, полностью определяются волей экспериментатора, который повторял опыт с каждой дозой облучения то или иное число раз, руководствуясь собственными соображениями. Следовательно, распределение этих численностей не является статистическим, так что здесь применима изложенная выше методика построения доверительной зоны для $y(x)$.

Для получения корреляционной решетки обычного вида надо произвести группировку вариант. Результат этой группировки представлен в табл. 8.13, там же даны все промежуточные расчеты.

ТАБЛИЦА 8.13

x/y	58	63	68	73	78	83	88	93	98	n_x	x	$n_x x$	$n_x x^2$	Y_x	$\bar{y}_x$	$Y_x x$
1								6	3	9	−3	−27	81	30	3,33	−90
2							8	2		10	−2	−20	40	22	2,20	−44
3						8				8	−1	−8	8	8	1,00	− 8
4				3	6					9	0	0	0	−3	−0,33	0
5			5	2						7	1	7	7	−12	−1,71	−12
6		7	3							10	2	20	40	−27	−2,70	−54
7	4	4								8	3	24	72	−28	−3,50	−84
n_y	4	11	8	5	6	8	8	8	8	61		−4	248			−292
y	− 4	−3	−2	−1	0	1	2	3	4			$X_{(1)}$	$X_{(2)}$			(XY)
$n_y y$	−16	33	−16	−5	0	8	16	24	12	−10	$Y_{(1)}$					
$n_y y^2$	64	99	32	5	0	8	32	72	48	360	$Y_{(2)}$					

Для удобства записи использованы обозначения:

$$\sum_x n_x x = X_{(1)}; \quad \sum_x n_x x^2 = X_{(2)};$$

$$\sum_y n_y y = Y_{(1)}, \quad \sum_y n_y y^2 = Y_{(2)};$$

$$\sum_y n_{xy} y = Y_x; \quad \sum_x Y_x x = \sum_x \sum_y n_{xy} xy = (XY).$$

Тогда:

$$\sum_x n_x (x - \bar{x})^2 = X_{(2)} - \frac{X_{(1)}^2}{n}; \quad \sum_y n_y (y - \bar{y})^2 = Y_{(2)} - \frac{Y_{(1)}^2}{n},$$

$$\sum_x \sum_y n_{xy} (x - \bar{x})(y - \bar{y}) = (XY) - \frac{X_{(1)} Y_{(1)}}{n},$$

поскольку в этих обозначениях:

$$\bar{x} = \frac{X_{(1)}}{n} \qquad \bar{y} = \frac{Y_{(1)}}{n}.$$

Вычисление промежуточных величин $X_{(1)}$, $X_{(2)}$, $Y_{(1)}$, $Y_{(2)}$, (XY) производится непосредственно в корреляционной таблице; разумеется, все расчеты ведутся в единицах условной шкалы.

Теперь находим, как обычно, по формулам § 8.3:

$$\sum_x \sum_y n_{xy} (x - \bar{x})(y - \bar{y}) = -292 - \frac{(-4)(-10)}{61} =$$

$$= -292 - 0{,}7 = -292{,}7;$$

$$\sum_x n_x (x - \bar{x})^2 = 248 - \frac{(-4)^2}{61} = 248 - 0{,}3 = 247{,}7;$$

$$\sum_y n_y (y - \bar{y})^2 = 360 - \frac{(-10)^2}{61} = 360 - 1{,}6 = 358{,}4.$$

Тогда:

$$r^2 = \frac{(-292{,}7)^2}{247{,}7 \cdot 358{,}4} = 0{,}965.$$

Согласно формуле (8.29), в выражение для $s_{y_x}^{(\text{верт.})}$, определяющее ширину доверительных зон, входит $1 - r^2$. Ввиду того что в данном случае коэффициент корреляции весьма близок к единице, величина $1 - r^2$ оказывается очень малой (0,035). Это предъявляет высокие требования к точности определения r^2. Очевидно, следует ожидать, что в проделанном выше расчете точ-

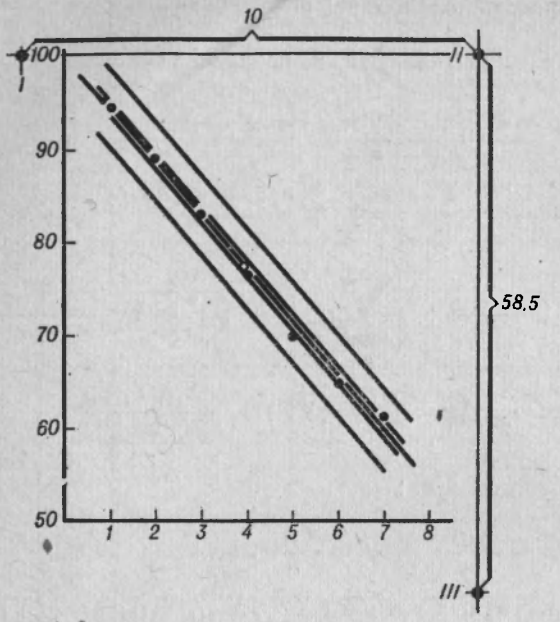


Рис. 8.6. Регрессионный анализ уменьшения темпа размножения бактерий при облучении.

ность могла пострадать из-за группировки вариант — в данном случае довольно грубой. Поэтому лучше провести расчет, не делая группировки. Такой расчет показан в табл. 8.14. Вследствие сравнительной малочисленности вариант вычисления по несгруппированным данным оказываются не слишком громоздкими; чтобы не иметь дела с большими числами, начало отсчета сдвинуто — все варианты уменьшены на 80.

Используя полученные в табл. 8.14 значения, находим:

$$\sum_{x,\ y}(x-\bar{x})(y-\bar{y}) = -2102 - \frac{(-166)\,240}{61} = -2102 + 653{,}1 = -1448{,}9.$$

$$\sum_{x}(x-x)^2 = 1192 - \frac{240^2}{61} = 1192 - 944{,}3 = 247{,}7;$$

$$s_x^2 = \frac{247{,}7}{60} = 4{,}13;$$

$$\sum_{y}(y-\bar{y})^2 = 9190 - \frac{(-166)^2}{61} = 9190 - 451{,}7 = 8738{,}3.$$

Поэтому

$$b = \frac{-1448,9}{247,7} = -5,85; \qquad r^2 = \frac{(-1148,9)^2}{247,7 \cdot 8738,3} = 0,9699.$$

Кроме того,

$$\bar{x} = \frac{240}{61} = 3,94; \qquad \bar{y} = 80 + \frac{(-166)}{61} = 77,28.$$

Через точку с координатами $\bar{x}$, $\bar{y}$ должна проходить прямая регрессии. Построение ее проводим следующим образом. Прежде всего, выбрав подходящий масштаб, откладываем на графике точку с координатами $\bar{x} = 3,94$; $\bar{y} = 77,28$ (на рис. 8.6 эта точка указана светлым кружком). Затем от какой-нибудь точки в левом верхнем углу графика (на рис. 8.6 это точка I) откладываем вправо 10 единиц в масштабе x и от найденной точки II откладываем вниз (так как $b < 0$) отрезок длиной $|b| \cdot 10 = 5,85 \cdot 10 = 58,5$ единицы в масштабе y (точка III). Наконец, проводим через точку $(\bar{x}, \bar{y})$ прямую, параллельную прямой, соединяющей точки I и III. Сплошными кружками на рис. 8.6 изображены эмпирические условные средние $\bar{y}_x$, взятые из табл. 8.14.

Теперь построим доверительные зоны для $\hat{y}_x$ и $y(x)$. По формуле (8.29) находим:

$$s_{\hat{y}_x}^{(\text{верт.})} = \sqrt{\frac{0,0301 \cdot 8738,3}{61 \cdot 59}} = \sqrt{0,0731} = 0,270.$$

Для максимального в данной задаче значения $|x - \bar{x}| \approx 3$ имеем:

$$\frac{n(x-\bar{x})^2}{\sum (x-\bar{x})^2} = \frac{61 \cdot 9}{247,7} = 2,216.$$

Если выбрать $P = 95\%$, то $t_P = 2,00$ (при $n - 2 \approx 60$). Поэтому:

$$\left[t_P \, s_{\bar{y}_x} \right]_{x-\bar{x}=0} = 2,00 \cdot 0,270 = 0,54;$$

$$\left[t_P \, s_{\bar{y}_x} \right]_{x-\bar{x}=3} = 2,00 \cdot 0,270 \sqrt{2,216 + 1} = 0,97;$$

$$\left[t_P \, s_{y(x)} \right]_{x-\bar{x}=0} = 2,00 \cdot 0,270 \sqrt{61 + 1} = 4,25;$$

$$\left[t_P \, s_{y(x)} \right]_{x-\bar{x}=3} = 2,00 \cdot 0,270 \sqrt{61 + 1 + 2,216} = 4,33.$$

На вертикали $x = 3,94$ откладываем точки с ординатами $77,28 - 0,54 = 76,74$ и $77,28 + 0,54 = 77,82$ для зоны $\hat{y}_x$ и точки с ординатами $77,28 - 4,25 = 73,03$ и $77,28 + 4,25 = 81,53$ для зоны $y(x)$. На вертикалях $x = 1$ и $x = 7$ надо отложить вверх и вниз от линии регрессии по 0,96 для $\hat{y}_x$ и по 4,33 для $y(x)$. По всем этим точкам и проводим границы доверительных

ТАБЛИЦА 8.14

Доза 10^3 Р (x)	Остаточный уровень размножения (y)	Y_x	$Y_x x$	n_x	$\bar{y}_x$	$n_x x$	$n_x x^2$
1	14 16 17 12 15 13 16 14 15	132	132	9	14,7	9	9
2	7 11 6 8 8 10 9 9 15	90	180	10	9,0	20	40
3	3 5 2 4 1 1 5 3	24	72	8	3,0	24	72
4	—3 —9 —3 —1 —4 —2 —5 —1 —5	—33	—132	9	—3,7	36	144
5	—9 —12 —10 —11 —11 —12 —8	—73	—365	7	—10,4	35	175
6	—17 —14 —16 —16 —17 —15 —13 —15 —12 —18	—153	—918	10	—15,3	60	360
7	—18 —22 —20 —21 —16 —20 —19 —17	—153	—1071	8	—19,1	56	392
	$Y_{(2)} = 9190$	—166 $Y_{(1)}$	—2102 (XY)	61 n		240 $X_{(1)}$	1192 $X_{(2)}$

ТАБЛИЦА 8.16

Доза, $10^3 P$ (x)	Остаточный уровень размножения, % (y)	Y_x	$Y_x x$	n_x	$\bar{y}_x$	$n_x x$	$n_x x^2$
1	13 16 14 16 15 15	89	89	6	14,8	6	6
2	8 7 12 9 10 9 11 9	75	150	8	9,4	16	32
3	5 2 2 4 6 3 6 4	32	96	8	4,0	24	72
4	—7 —2 —4 0 —3 —1 —1	—18	—72	7	—2,6	28	112
5	—7 —10 —10 —8 —6 —6 —9 —7	—63	—315	8	—7,9	40	200
6	—13 —11 —15 —14 —9	—62	—372	5	—12,4	30	180
	$Y_{(2)} = \Sigma y^2 = 3581$	53 $Y_{(1)}$	—424 (XY)	42 n		144 $X_{(1)}$	602 $X_{(2)}$

зон (см. рис. 8.6). В данном случае это почти прямые, параллельные линии регрессии.

8.5.3. Когда величина x имеет такие же статистические свойства, как и величина y, точки корреляционного поля заполняют область, близкую по форме к эллипсу[1]. Построение такого *корреляционного эллипса* требует обычно довольно громоздких вычислений, поэтому мы не будем его описывать[2].

8.5.4. Сделаем следующее знамечание. Согласно формуле (8.26), величина σ_b^2 обратно пропорциональна дисперсии σ_x^2. Поэтому наклон прямой регрессии будет найден с тем меньшей неопределенностью, чем шире будет интервал значений x. Это обстоятельство может быть использовано для снижения σ_b^2, когда x не случайная, а задаваемая в эксперименте величина.

Если линейность регрессии не вызывает сомнения, то величина σ_x^2 может быть увеличена и без расширения интервала используемых значений x. Пусть, например, условия биохимического эксперимента позволяют задать значения температуры от 21 до 36° С. Если для изучения зависимости активности фермента от температуры намечено провести измерения для шести разных температур, то выгоднее (в смысле получения наименьшего значения σ_b выбрать не точки 21; 24; 27; 30; 33; 36°, а точки 21 и 36°, произведя по три измерения при каждой из этих температур. Однако если линейность регрессии нуждается в проверке, то следует использовать равноотстоящие точки по всему интервалу температур.

§ 8.6. Сравнение двух линий регрессии

В этом параграфе используются те же понятия, которые были перечислены (вместе со ссылками на соответствующие разделы) в начале § 8.5. Кроме того, используются еще понятия стандартной ошибки коэффициента регрессии (см. раздел 8.5.1*) и F-критерия (см. раздел* 4.6.1*).*

8.6.1. На практике иногда возникает задача сравнения двух линий регрессии — например, зависимость урожайности от количества удобрений для двух сортов или ход размножения бактерий в двух средах и т. д.

---

[1] Но доверительные зоны условных средних $\hat{y}_x$ и $\hat{x}_y$, как и в предыдущем случае, ограничены парами гипербол, которые строятся по формуле (8.26).

[2] См. Лукомский Я. И. Теория корреляции и ее применение к анализу производства. М., «Госстатиздат», 1961 и Линник Ю. В. Метод наименьших квадратов и основы математико-статистической обработки наблюдений. М., «Физматгиз», 1962.

Уравнения сравниваемых линий регрессии запишем в обычном виде:

$$\tilde{y}_{x\,(1)} - \bar{y}_{(1)} = b_{(1)}\left(x - \bar{x}_{(1)}\right);$$

$$\tilde{y}_{x\,(2)} - \bar{y}_{(2)} = b_{(2)}\left(x - \bar{x}_{(2)}\right);$$

цифра в скобках обозначает номер прямой.

Чтобы различие между двумя линиями регрессии было незначимо, нужно прежде всего чтобы не различались значимо их угловые коэффициенты $b_{(1)}$ и $b_{(2)}$. Проверка нулевой гипотезы $\beta_{(1)} = \beta_{(2)}$ основана на том, что величина

$$t = \frac{|b_{(1)} - b_{(2)}|}{s_{b_{(1)} - b_{(2)}}} \qquad (8.33)$$

имеет распределение Стьюдента[1]. Однако последнее справедливо только в том случае, когда обе линии регрессии характеризуются одной и той же случайной дисперсией $\sigma^2\{y - \hat{y}_x\}$ (которую будем обозначать в дальнейшем просто σ^2). Поэтому анализ должен начинаться с проверки этого предположения. Проверка производится при помощи F-критерия, причем оценки для σ_1^2 и σ_2^2 находят по формуле:

$$s^2 = s_y^2(1 - r^2) = \frac{\sum n_y (y - \bar{y})^2}{n - 2}(1 - r^2), \qquad (8.34)$$

которая получается из формулы (8.27) путем замены величин σ_y^2 и ρ^2 их выборочными оценками s_y^2 и r^2; числа степеней свободы равны соответственно $n_{(1)} - 2$ и $n_{(2)} - 2$.

Примем обозначения:

$$\Sigma_{xx} = \sum_x n_x (x - \bar{x}) = \sum_x n_x x^2 - n\bar{x}^2;$$

$$\Sigma_{yy} = \sum_y n_y (y - \bar{y})^2 = \sum_y n_y y^2 - n\bar{y}^2;$$

$$\Sigma_{xy} = \sum_x \sum_y n_{xy} (x - \bar{x})(y - \bar{y}) = \sum_x \sum_y n_{xy} xy - n\bar{x}\bar{y}.$$

Если различие между $s_{(1)}^2 = (1 - r_{(1)}^2)\Sigma_{yy(1)}/(n_{(1)} - 2)$ и $s_{(2)}^2 = (1 - r_{(2)}^2)\Sigma_{yy(2)}/(n_{(2)} - 2)$ оказалось незначимым, можно приступить к вычислению t по формуле (8.33).

---

[1] Если распределение исходных величин не очень отклоняется от нормального.

Так как, согласно формулам (8.26) и (**), $\sigma_t^2 = \sigma^2 \{y - \bar{y}_x\}/\sum' xx$, то оценкой $\sigma^2_{b(1)-b(2)}$ будет:

$$s^2_{b(1)-b(2)} = s^2_{b(1)} + s^2_{b(2)} = s^2\left(\frac{1}{\sum xx(1)} + \frac{1}{\sum xx(2)}\right)$$

или

$$s^2_{b(1)-b(2)} = s^2 \frac{\sum xx(1) + \sum xx(2)}{\sum xx(1)\ \sum xx(2)}, \tag{8.35}$$

где s^2 получается объединением соответствующих оценок для обеих прямых (коль скоро различие между ними незначимо). Объединение оценок дисперсии производится как обычно — путем взвешивания по числу степеней свободы:

$$s^2 = \frac{(n_{(1)} - 2)\, s^2_{(1)} + (n_{(2)} - 2)\, s^2_{(2)}}{n_{(1)} + n_{(2)} - 4}. \tag{8.36}$$

Последовательное применение формул (8.36), (8.35) и (8.33) дает значение t, которое сравнивают с критическим значением t_α для $f = n_{(1)} + n_{(2)} - 4$ степеней свободы.

Если отношение $s^2_{(1)}/s^2_{(2)}$ окажется значимым, так что дисперсии для обеих прямых регрессии надо считать различными, то проверка гипотезы $\beta_{(1)} = \beta_{(2)}$ должна производиться по приближенному t-критерию, который был описан в разделе **4.2.4.**

8.6.2. Пусть значение t, вычисленное тем или иным способом, оказалось незначимым. Тогда обе линии регрессии можно считать параллельными, и общая оценка для углового коэффициента β найдется как среднее взвешенное из $b_{(1)}$ и $b_{(2)}$, причем весами будут служить обратные значения дисперсий $s^2_{b(1)}$ и $s^2_{b(2)}$; после сокращения на общую величину $\sigma\{y - \hat{y}_x\}$ получается:

$$\bar{b} = \frac{\sum xx(1)\ b_{(1)} + \sum xx(2)\ b_{(2)}}{\sum xx(1) + \sum xx(2)}. \tag{8.37}$$

Оценка дисперсии этой величины (которая нужна для построения доверительного интервала для β) равна:

$$s^2_{\bar{b}} = \frac{s^2}{\sum xx(1) + \sum xx(2)}, \tag{8.38}$$

причем s^2 берут из формулы (8.36).

8.6.3. Если линии регрессии оказались параллельными (т. е. имеют общий угловой коэффициент b), то их уравнения примут вид:

$$\tilde{y}_{x\,(1)} - \bar{y}_{(1)} = b\left(x - \bar{x}_{(1)}\right),$$

$$\tilde{y}_{x(2)} - \bar{y}_{(2)} = \bar{b}\left(x - \bar{x}_{(2)}\right)$$

или, после некоторой перестановки,

$$\tilde{y}_{x(1)} = \left(\bar{y}_{(1)} - \bar{b}\,\bar{x}_{(1)}\right) + \bar{b}\,x,$$

$$\tilde{y}_{x(2)} = \left(\bar{y}_{(2)} - \bar{b}\,\bar{x}_{(2)}\right) + \bar{b}\,x.$$

Чтобы эти линии были не только параллельными, но и совпадали, должно выполняться условие:

$$\bar{y}_{(1)} - \bar{b}\,\bar{x}_{(1)} = \bar{y}_{(2)} - \bar{b}\,\bar{x}_{(2)},$$

которое можно записать так:

$$\frac{\bar{y}_{(1)} - \bar{y}_{(2)}}{\bar{x}_{(1)} - \bar{x}_{(2)}} = \bar{b}.$$

Если же величина

$$\frac{\bar{y}_{(1)} - \bar{y}_{(2)}}{\bar{x}_{(1)} - \bar{x}_{(2)}} = b^*, \tag{8.39}$$

полученная подстановкой эмпирических чисел $\bar{x}_{(1)}$, $\bar{x}_{(2)}$, $\bar{y}_{(1)}$, $\bar{y}_{(2)}$, не равна величине $\bar{b}$, полученной из формулы (8.37), то прямые не совпадают. Поэтому вопрос о совпадении двух параллельных прямых регрессии можно решать при помощи t-критерия, сравнивая с табличным значением t_α величину

$$t = \frac{b^* - \bar{b}}{s_{b^* - \bar{b}}}. \tag{8.40}$$

Величину s_{b^*-b} находим из равенства:

$$s^2_{b^*-\bar{b}} = s^2_{b^*} + s^2_{\bar{b}}, \tag{8.41}$$

причем:

$$s^2_{b^*} = \frac{s^2}{\left(\bar{x}_{(1)} - \bar{x}_{(2)}\right)^2}\left(\frac{1}{n_{(1)}} + \frac{1}{n_{(2)}}\right), \tag{8.42}$$

а $s^2_{\bar{b}}$ вычисляют по формуле (8.38).

Если различие окажется значимым, то можно вычислить разность:

$$d_{(1)-(2)} = \left[\bar{y}_{(1)} - \bar{b}\left(x - \bar{x}_{(1)}\right)\right] - \left[\bar{y}_{(2)} - \bar{b}\left(x - \bar{x}_{(2)}\right)\right] =$$
$$= \left(\bar{y}_{(1)} - \bar{y}_{(2)}\right) - \bar{b}\left(\bar{x}_{(1)} - \bar{x}_{(2)}\right),$$

которую удобней представить в эквивалентной форме:

$$d_{(1)-(2)} = \left(\bar{x}_{(1)} - \bar{x}_{(2)}\right)\left(b^* - \bar{b}\right). \qquad (8.43)$$

Оценка дисперсии этой величины равна:

$$s_d^2 = s^2\left[\frac{1}{n_{(1)}} + \frac{1}{n_{(2)}} + \frac{\left(\bar{x}_{(1)} - \bar{x}_{(2)}\right)^2}{s_{x(1)}^2 + s_{x(2)}^2}\right]. \qquad (8.44)$$

Пример 8.13. В примере 8.12 из § 8.5 описывалось уменьшение темпа размножения бактерий под действием рентгеновского облучения. Эти опыты были повторены с бактериями другого штамма. Совпадение результатов позволило бы непосредственно сопоставлять данные всех дальнейших экспериментов (например, по изысканию защитных веществ), проведенных на обоих штаммах. Дозовая зависимость остаточного уровня размножения для второго штамма приведена в табл. 8.15.

ТАБЛИЦА 8.15

Доза, $10^3 P$ (x)	Остаточный уровень размножения, % (y)							
1	93	96	94	96	95	95		
2	88	87	92	89	90	89	91	89
3	85	82	84	82	86	83	86	84
4	73	78	76	80	77	79	79	
5	73	70	70	72	74	74	71	73
6	67	69	65	66	71			

Непосредственное сопоставление данных табл. 8.15 и 8.12 не дает оснований считать, что штаммы различны. Однако окончательный вывод можно будет сделать лишь после количественного регрессионного анализа.

Основные расчеты для этого анализа выполнены в табл. 8.16 (см. стр. 212), которая совершенно аналогична табл. 8.14. По полученным данным находим:

$$\sum_{xy} = -424 - \frac{53 \cdot 144}{42} = -424 - 181{,}7 = -605{,}7;$$

$$\sum_{xx} = 602 - \frac{144^2}{42} = 602 - 493{,}7 = 108{,}3,$$

$$\sum_{yy} = 3581 - \frac{53^2}{42} = 3581 - 66{,}9 = 3514{,}1,$$

так что:

$$b = \frac{-605{,}7}{108{,}3} = -5{,}593; \qquad r^2 = \frac{(-605{,}7)^2}{108{,}3 \cdot 3514{,}1} = 0{,}9640.$$

Теперь найдем:

$$s^2 = \frac{\sum_{yy}(1-r^2)}{n-2} = \frac{3514{,}1 \cdot 0{,}0360}{40} = 3{,}163.$$

Наконец,

$$\bar{x} = \frac{144}{42} = 3{,}429, \qquad \bar{y} = 80 + \frac{53}{42} = 81{,}262.$$

Таким образом, мы имеем для двух линий регрессии два набора эмпирических оценок:

$n_{(1)} = 61;$	$n_{(2)} = 42;$
$\bar{x}_{(1)} = 3{,}934;$	$\bar{x}_{(2)} = 3{,}429;$
$\bar{y}_{(1)} = 77{,}279;$	$\bar{y}_{(2)} = 81{,}262;$
$\sum_{xx(1)} = 247{,}7;$	$\sum_{xx(2)} = 108{,}3;$
$b_{(1)} = -5{,}849;$	$b_{(2)} = -5{,}593;$
$s^2_{(1)} = 4{,}458;$	$s^2_{(2)} = 3{,}163;$

Так как в дальнейшем нам понадобятся разности $\bar{x}_{(1)} - \bar{x}_{(2)}$, $\bar{y}_{(1)} - \bar{y}_{(2)}$, $b_{(1)} - b_{(2)}$ и $b^* - b$, то значения $\bar{x}_{(1)}$, $\bar{y}_{(1)}$ и b из § 8.5 на всякий случай вычислены заново более точно — с сохранением еще одного десятичного знака; величина $s^2_{(1)}$ получена по формуле (8.34).

Анализ начинаем со сравнения $s^2_{(1)}$ и $s^2_{(2)}$. Так как

$$F = \frac{s^2_{(1)}}{s^2_{(2)}} = \frac{4{,}458}{3{,}163} = 1{,}41 < F_{0{,}05}(40;\ 59) = 1{,}64,$$

то различие между дисперсиями $s^2_{(1)}$ и $s^2_{(2)}$ незначимо, что позволяет пользоваться t-критерием для сравнения $b_{(1)}$ и $b_{(2)}$. По-

скольку $s^2_{(1)}$ и $s^2_{(2)}$ не различаются значимо, то мы считаем их оценками одной и той же общей дисперсии σ^2, наилучшую оценку которой находим по формуле (8.36):

$$s^2 = \frac{59 \cdot 4{,}458 + 40 \cdot 3{,}163}{59 + 40} = \frac{389{,}5}{99} = 3{.}934.$$

Теперь по формуле (8.35) находим:

$$s^2_{b_{(1)} - b_{(2)}} = 3{,}934\,\frac{247{,}7 + 108{,}3}{247{,}7 \cdot 108{,}3} = 0{.}0522,$$

$$s_{b_{(1)} - b_{(2)}} = \sqrt{0{,}0522} = 0{,}228.$$

Поэтому, согласно формуле (8.33),

$$t_{b_{(1)} - b_{(2)}} = \frac{5{,}849 - 5{,}593}{0{,}228} = \frac{0{,}256}{0{,}228} = 1{.}12.$$

Это меньше, чем $t_{0{,}05}(99) = 1{,}98$, так что значимого различия между $b_{(1)}$ и $b_{(2)}$ нет. Следовательно, обе прямые линии регрессии могут считаться параллельными с общим угловым коэффициентом, оцениваемым, согласно формуле (8.37), величиной

$$\bar{b} = \frac{247{,}7\,(-5{,}849) + 108{,}3\,(-5{,}593)}{247{,}7 + 108{,}3} = -\frac{2054{,}5}{356{,}0} = -5{,}771.$$

Для оценки дисперсии σ^2_b получаем по формуле (8.38):

$$s^2_{\bar{b}} = \frac{3{,}934}{356{,}0} = 0{.}01105.$$

Проверим теперь, не совпадают ли обе линии регрессии. Для этого по формуле (8.39) вычисляем:

$$b^* = \frac{77{,}279 - 81{,}262}{3{,}934 - 3{,}429} = -\frac{3{,}983}{0{,}505} = -7{.}887.$$

Далее по формулам (8.42) и (8.41) находим:

$$s^2_{b^*} = \frac{3{,}934}{0{,}505^2}\,\frac{61 + 42}{61 \cdot 42} = 0{.}6202,$$

$$s^2_{b^* - \bar{b}} = 0{,}6202 + 0{,}0110 = 0{,}6312,$$

$$s_{b^* - \bar{b}} = \sqrt{0{,}6312} = 0{,}795.$$

Поэтому

$$t_{b^* - \bar{b}} = \frac{7{,}89 - 5{,}75}{0{,}795} = \frac{2{,}14}{0{,}795} = 2{,}69.$$

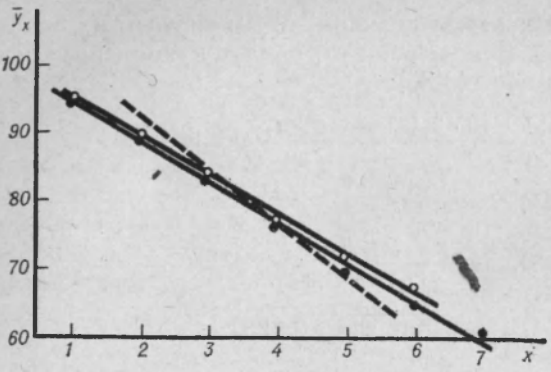


Рис. 8.7. Сравнение двух линий регрессии.

Это превышает $t_{0,01}(99) = 2{,}63$, так что приходится отвергнуть гипотезу о совпадении обеих прямых. На рис. 8.7 изображены сплошными и светлыми кружками точки $(x, \bar{y}_x)$ соответственно для первой и второй регрессии, а сплошными линиями — их прямые регрессии. Пунктирная линия изображает прямую с угловым коэффициентом b^*, проходящую через центры обеих регрессий (жирные точки).

Поскольку две прямые регрессии значимо не совпадают, мы вычислим по формуле (8.43) разность $d_{(2)-(1)}$ (из более высокой вычитаем более низкую). Значения $\bar{x}_{(2)} - \bar{x}_{(1)} = 0{,}505$ и $b^* - \bar{b} = -2{,}14$ были получены ранее, так что

$$d_{(2)-(1)} = (-0{,}505)(-2{,}14) = 1{,}08.$$

Следовательно, при всех дозах облучения остаточный уровень размножения у второго штамма на $\sim 1\%$ выше, чем у первого.

Поскольку

$$s_d^2 = 3{,}934\,(0{,}0164 + 0{,}0238 + 0{,}007) = 3{,}934 \cdot 0{,}0409 = 0{,}1609,$$

то

$$s_d = \sqrt{0{,}1609} = 0{,}401 \approx 0{,}40.$$

Значит, 95%-ные доверительные пределы для $d_{(2)-(1)}$ будут:

$$1{,}08 - 1{,}98 \cdot 0{,}40 = 0{,}29; \quad 1{,}08 + 1{,}98 \cdot 0{,}40 = 1{,}87.$$

Об одновременном сравнении нескольких линий регрессии см. в книгах А. Хальда, с. 495 и Дж. У. Снедекора, гл. 13.

§ 8.7. Нелинейная регрессия

8.7.1. Во многих случаях теоретические соображения или вид корреляционного поля (см. раздел **8.1.1.**) могут подсказывать нелинейный характер зависимости между двумя признаками. Конечно, такое предположение должно быть прежде всего проверено статистически, т. е. должна быть при помощи соответствующих критериев опровергнута нулевая гипотеза о том, что зависимость в данной эмпирической совокупности (если она есть выборка из некоторой генеральной совокупности) является линейной.

Ниже будет использовано понятие распределения Стьюдента и его критического значения (см. раздел **3.7.2***).*

В общем случае такая проверка требует довольно громоздких расчетов (см. книгу А. К. Митропольского, гл. VII, § 1). Мы разберем здесь простой частный случай, когда получены три ряда значений y для трех значений x, таких, что $x_3 - x_2 = x_2 - x_1$, т. е. значение x_2 находится посередине интервала $x_1 \div x_3$. В этом случае критерием линейности регрессии может служить отношение:

$$t = \frac{\bar{y}_1 + \bar{y}_3 - 2\bar{y}_2}{\sqrt{2\sum d^2/n\,(n-1)}}, \qquad (8.45)$$

где $\bar{y}_1$, $\bar{y}_2$, $\bar{y}_3$ — средние значения трех рядов измерений y, соответствующих трем значениям x_1, x_2, x_3; n есть общее число всех измерений, а

$$\sum d^2 = \sum (y_1 - \bar{y}_1)^2 + \sum (y_2 - \bar{y}_2)^2 + \sum (y_3 - \bar{y}_3)^2 .$$

Величину t, вычисленную по формуле (8.45), сравнивают с критическим значением распределения Стьюдента $t_\alpha(f)$ для числа степеней свободы $f = 3(n-1)$. Если число измеренных значений y для разных x неодинаково ($n_1 \neq n_2 \neq n_3$), то в (8.45) производится замена:

$$\frac{\sum d^2}{n\,(n-1)} \to \frac{\sum d^2}{n_1 + n_2 + n_3 - 3}\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3}\right),$$

а $t_\alpha(f)$ берут для числа степеней свободы $f = n_1 + n_2 + n_3 - 3$.

8.7.2. В дальнейшем изложении мы будем исходить из того, что нелинейный характер зависимости установлен статистически или не подлежит сомнению по каким-либо другим соображениям, и вопрос заключается лишь в оценке соответствующих параметров этой зависимости. При этом будем рассматривать только такие

нелинейные зависимости, которые выражаются полиномами, т. е. в виде:

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_h x^h$$

(см. раздел 8.1.3).

О зависимостях, которые могут быть сведены к линейным, см. раздел 8.2.2. *Так называемым S-образным зависимостям типа доза—эффект посвящена глава девятая. Для выравнивания зависимостей (произвольного вида) может также применяться способ скользящего среднего (см. раздел* 8.7.4).

Если предполагаемая зависимость между y и x квадратична:

$$y = a_0 + a_1 x + a_2 x^2, \tag{8.46}$$

то система нормальных уравнений (см. раздел 8.1.3) имеет вид:

$$\left.\begin{aligned} a_0 n + a_1 \sum x + a_2 \sum x^2 &= \sum y, \\ a_0 \sum x + a_1 \sum x^2 + a_2 \sum x^3 &= \sum yx, \\ a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4 &= \sum yx^2. \end{aligned}\right\} \tag{8.47}$$

Как и в случае линейной зависимости, можно вывести формулы для коэффициентов a_0, a_1 и a_2. Формулы получаются несколько менее громоздкими, если изобразить зависимость в виде:

$$\tilde{y} = b_0 + b_1 (x - \bar{x}) + b_2 [(x - \bar{x})^2 - \gamma], \tag{8.48}$$

где

$$\bar{x} = \frac{1}{n} \sum x, \quad \gamma = \frac{1}{n} \sum (x - \bar{x})^2. \tag{8.49}$$

Тогда:

$$\left.\begin{aligned} b_0 = \bar{y} &= \frac{1}{n} \sum y, \quad b_1 = \frac{\sum (x - \bar{x}) y}{\sum (x - \bar{x})^2}, \\ b_2 &= \frac{\sum (x - \bar{x})^2 y - n\gamma \bar{y}}{\sum (x - \bar{x})^4 - n\gamma^2}. \end{aligned}\right\} \tag{8.50}$$

Пример 8.14. В табл. 8.17 приведены данные об удоях коров за первую, вторую и т. д. лактации. Значения y_i представляют собой усредненные данные для группы коров.

ТАБЛИЦА 8.17

x (номер лактации)	1	2	3		5	6	7	8	9
y (удой в ц за лактацию)	30,7	32,9	35,8	38,2	37,2	39,2	39,9	38,4	37,1

Вид графика на рис. 8.8 (ломаная линия) подсказывает, что подходящим аналитическим выражением для этой зависимости может служить квадратичная парабола типа (8.46).

ТАБЛИЦА 8.18

x	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^4$	y'	$(x-\bar{x})\,y'$	$(x-\bar{x})^2 y'$	$\tilde{y}$
1	—4	16	256	0,7	—2,8	11,2	30,34
2	—3	9	81	2,9	—8,7	26,1	33,45
3	—2	4	16	5,8	—11,6	23,2	35,92
4	—1	1	1	8,2	—8,2	8,2	37,74
5	0	0	0	8,7	0	0	38,92
6	1	1	1	9,2	8,2	9,2	39,45
7	2	4	16	9,9	19,8	39,6	39,34
8	3	9	81	8,4	25,2	75,6	38,58
9	4	16	256	7,1	28,4	113,6	37,18
45		60	708	60,9	51,3	306,7	

Заменив для сокращения вычислений y на $y' = y - 30$, составляем вспомогательную табл. 8.18; в ней мы производим все промежуточные расчеты, необходимые для получения сумм, входящих в формулы (8.50).

Используя табличные данные, имеем:

$$\bar{x} = \frac{45}{9} = 5, \quad \bar{y} = \frac{60}{9} = 6{,}667, \quad b_0 = \frac{60{,}9}{9} = 6{,}767,$$

$$b_1 = \frac{51{,}3}{60} = 0{,}855, \quad b_1 = \frac{306{,}7 - 9 \cdot \frac{60}{9} \cdot \frac{60{,}9}{9}}{708 - 9 \cdot \frac{400}{9}} =$$

$$= \frac{-99{,}3}{308} = -0.3224.$$

Поэтому:

$$\tilde{y} = 30 + 6{,}77 + 0{,}855\,(x - 5) - 0{,}3224\,[(x - 5)^2 - 6{,}667]. \quad (*)$$

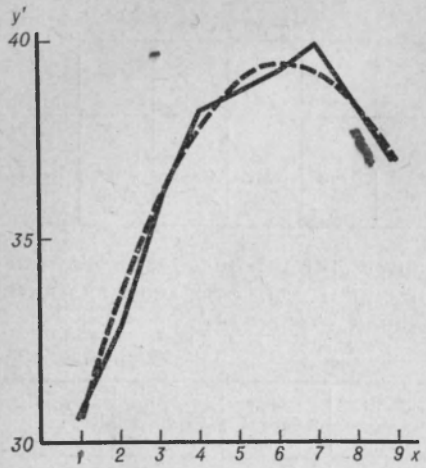


Рис. 8.8. Квадратичная регрессия.

Раскрыв скобки в (*), придем к форме (8.46). Значения $\tilde{y}$, вычисленные по формуле (*), проставлены в последнем столбце табл. 8.18. Кривая, изображающая функцию (*), нанесена пунктиром на рис. 8.8.

8.7.3. Во многих случаях отсутствуют какие-либо теоретические соображения о виде уравнения регрессии, и выравнивание при помощи полинома имеет просто целью получение достаточно хорошей интерполяционной формулы. При этом всегда возникает вопрос о том, какова должна быть наивысшая степень аргумента x в полиноме.

Приступая к анализу характера зависимости, нужно прежде всего составить графическое изображение ряда; вид графика может многое подсказать о виде регрессии. Однако имеется и более точный способ решения поставленной задачи. Этот способ основан на следующем. Если функция $y = f(x)$ линейна, то разности соседних значений y одинаковы (при равноотстоящих значениях x). Так, функция

$$y = 3 + 2x$$

дает ряд:

x: 1 2 3 4 5 и т. д.,
y: 5 7 9 11 13 и т. д.

Разности соседних значений y (или приращения Δy) все равны 2. Если функция $y = f(x)$ квадратична, то приращения Δy неодинаковы, то зато одинаковы приращения этих приращений (обозначим их $\Delta^2 y$). Например, если

$$y = x^2,$$

то ряд будет:

x : 1 2 3 4 5 ...
y : 1 4 9 16 25 ...

Приращения здесь равны:

$$\Delta y_1 = y_2 - y_1 = 4 - 1 = 3;$$
$$\Delta y_2 = y_3 - y_2 = 9 - 4 = 5;$$
$$\Delta y_3 = y_4 - y_3 = 16 - 9 = 7 \text{ и т. д.}$$

т. е. они различны, но приращения второго порядка

$$\Delta^2 y_1 - \Delta y_2 - \Delta y_1 = 5 - 3 = 2;$$
$$\Delta^2 y_2 = \Delta y_3 - \Delta y_2 = 7 - 5 = 2 \text{ и т. д.},$$

одинаковы. Аналогично можно показать, что если в полиноме наивысшая степень аргумента равна h, то одинаковыми окажутся приращения h-го порядка. Поэтому если мы имеем ряд, являющийся табличной записью некоторого полинома, порядок которого нам неизвестен, то нужно составить сначала ряд приращений первого порядка, затем ряд приращений второго порядка и т. д. до тех пор, пока не получится ряд одинаковых приращений; порядок этого ряда и укажет степень полинома.

Пусть, например, имеется ряд:

x :	1	2	3	4	5	6	7	8
y :	−2	−4	−2	10	38	88	166	278

Для анализа приращений составим табл. 8.19.

ТАБЛИЦА 8.19

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1	−2	−2	4	6
2	−4	2	10	6
3	−2	12	16	6
4	10	28	22	6
5	38	50	28	6
6	88	78	34	
7	166	112		
8	278			

Так как одинаковыми оказались разности третьего порядка, то мы заключаем, что ряд описывается полиномом третьей степени (в данном случае $y = -2 + 3x - 4x^2 + x^3$).

Когда мы имеем эмпирический ряд, то не приходится ожидать, чтобы какие-либо приращения были строго одинаковыми. Мы можем только требовать, чтобы изменения этих приращений не носили систематический характер, а имели бы вид случайных вариаций. Составим, например, таблицу приращений (табл. 8.20) для ряда из табл. 18.17.

ТАБЛИЦА 8.20

x	y	Δy	$\Delta^2 y$
1	30,7	2,2	0,7
2	32,9	2,9	—0,5
3	35,8	2,4	—1,9
4	38,2	0,5	0,0
5	38,7	0,5	0,2
6	39,2	0,7	—2,2
7	39,9	—1,5	0,2
8	38,4	—1,3	—1,1
9	37,1	—2,4	—
10	34,7	—	—

Здесь изменения Δy при переходе от низких значений x к высшим являются несомненно систематическими (значения Δy в общем убывают), но изменения $\Delta^2 y$ не обнаруживают какой-либо определенной тенденции; поэтому мы заключаем, что заданный эмпирический ряд может быть удовлетворительно выравнен квадратичной функцией вида:

$$\tilde{y} = a_0 + a_1 x + a_2 x^2.$$

Если значения x не равноотстоящие, то для проверки линейности полинома нужно составить ряд о т н о с и т е л ь н ы х приращений $\Delta y/\Delta x$.

8.7.4. Выравнивание эмпирического ряда при помощи полинома

$$\tilde{y} = a_0 + a_1 x + a_2 x^2 + \ldots + a_h x^h \qquad (8.1)$$

оправдано, строго говоря, лишь в тех случаях, когда имеются какие-либо теоретические основания считать, что зависимость между y и x именно такова (с учетом также преобразований, описанных в разделе 8.2.2). Если же таких оснований нет, то полином (8.1) будет играть роль просто «подгоночной» интерполяционной формулы. Однако использование такой формулы может быть целесообразным только тогда, когда эта формула достаточно проста, т. е. когда степень полинома не очень высока. Но если ни линейная, ни квадратичная функции не оказываются подходящими (это выясняется из анализа приращений), то выравнивать при помощи полинома третьей, четвертой и более высокой степени не имеет особого смысла (за исключением случаев, когда имеются теоретические соображения в пользу именно такой зависимости). Гораздо удобнее оказывается в этих случаях оставить зависимость в табличной форме, ограничивая задачу выравнивания только исключением влияния случайных вариаций.

Это достигается применением так называемого *способа скользящего среднего*. Идея этого способа состоит в том, что если для всей эмпирической кривой нельзя подобрать сравнительно простое алгебраическое выражение, то это можно сделать для отдельных отрезков данной кривой; разбив кривую на достаточно малые отрезки, можно каждый из них описывать даже линейной функцией (как известно, на этом основано линейное интерполирование при пользовании, например, таблицами логарифмов, тригонометрических функций и т. п.).

Может показаться, что практическое выполнение этой программы должно потребовать многочисленных вычислений, связанных с нахождением большого числа уравнений регрессии (по числу отрезков, на которые мы разбили весь интервал). К счастью, это не так. Действительно, рассмотрим отрезок, включающий три соседние точки, например:

$$x_4,\ x_5,\ x_6$$
$$y_4,\ y_5,\ y_6.$$

Прямая линия регрессии, которую можно провести по этим трем точкам, проходит, как мы знаем, через точку с координатами

$$\bar{x} = \frac{x_4 + x_5 + x_6}{3}, \qquad \bar{y} = \frac{y_4 + y_5 + y_6}{3}.$$

Но по принятому способу выравнивания через эту точку проходит и выравненная кривая. Поэтому мы можем считать, что эти равенства дают нам координаты одной из выравненных точек, причем естественно приписать ей в данном случае номер 5 — номер средней точки «триады». Если бы мы вычислили еще коэффициент регрессии b по указанным выше трем парам значений, то мы бы знали также направление выравненной кривой в точке $(\tilde{x}_5, \tilde{y}_5)$. Но эта дополнительная информация не оправдывает вычислительную работу, связанную с определением b. Поэтому обычно ограничиваются нахождением выравненных точек $(\tilde{x}_i, \tilde{y}_i)$.

Выравненные точки $(\tilde{x}_6, \tilde{y}_6)$, $(\tilde{x}_7, \tilde{y}_7)$ и т. д. найдем из равенств:

$$\tilde{x}_6 = \frac{x_5 + x_6 + x_7}{3}, \qquad \tilde{y}_6 = \frac{y_5 + y_6 + y_7}{3},$$

$$\tilde{x}_7 = \frac{x_6 + x_7 + x_8}{3}, \qquad \tilde{y}_7 = \frac{y_6 + y_7 + y_8}{3}$$

и т. д. Так как тройки, на которые мы разбиваем весь интервал, не следуют одна за другой, а перекрываются (т. е. мы пользуемся не системой 123; 456; 789 и т. д., а системой 123, 234; 345 и т. д.),

то средние $\tilde{x}_i$, $\tilde{y}_i$ называют *скользящими*; отсюда и название разбираемого метода выравнивания.

Процедура выравнивания значительно упрощается, если значения x_i являются равноотстоящими. Тогда $\tilde{x}_5 = x_5$, $\tilde{x}_6 = x_6$ и т. д. (т. е. вообще $\tilde{x}_i = x_i$), так что остается находить значения:

$$\tilde{y}_i = (y_{i-1} + y_i + y_{i+1}) : 3. \qquad (8.51)$$

При выравнивании не обязательно пользоваться именно тройками чисел. Однако удобнее, чтобы число точек, по которым производится усреднение, было нечетным, так как иначе не будет выполняться условие $\tilde{x}_i = x_i$. Кроме того, должно быть учтено следующее обстоятельство. Если кривизна ожидаемой выравненной кривой велика, то ломаная может служить достаточно удовлетворительным приближением к кривой только тогда, когда она состоит из коротких отрезков; иначе говоря, усреднение должно производиться по малому числу точек. В то же время чем больше число точек, по которому производится усреднение, тем меньше будет сказываться влияние случайных вариаций. Поэтому если эмпирический ряд имеет большую кривизну и малые флуктуации, то лучше усреднять по трем точкам; если же ряд имеет малую кривизну и сильные флуктуации, то усреднение следует производить по 7—9 точкам. В большинстве случаев оптимальным может считаться усреднение по пяти точкам.

$$\tilde{y}_i = (y_{i-2} + y_{i-1} + y_i + y_{i+1} + y_{i+2}) : 5. \qquad (8.52)$$

Более подробно о способе скользящего среднего, в том числе о так называемом *взвешенном скользящем среднем*, см. в книге автора «Биометрические методы» (гл. 9, § 6 и 8).

§ 8.8. Множественная линейная регрессия

Чтобы понять содержание этого параграфа, необходимо предварительно прочесть § 8.1 и 8.2.

8.8 1. Любая величина y, характеризующая биологический объект, связана корреляционно не с одной какой-нибудь переменной x, а со многими переменными (обозначим их x_1, x_2, ...). Если изучается регрессионная зависимость y от одной переменной (как это было описано во всех предыдущих параграфах этой главы), то это просто означает, что в данном конкретном случае влия-

ние других факторов вносит достаточно малый вклад в изменение значений y. Однако во многих случаях изменения одного аргумента могут объяснить лишь небольшую часть общего изменения y (количественным выражением этого факта будет малое значение коэффициента корреляции ρ_{xy}, точнее — его квадрата ρ^2_{xy}; о коэффициенте корреляции см. § 8.3). Тогда имеет смысл рассмотреть регрессионную зависимость y также от других переменных, что приводит к так называемой *множественной регрессии* вида:

$$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_h x_h \qquad (8.53)$$

(если ограничиться только изучением линейных связей).

Так, в примере 8.3 (§ 8.2) рассматривалась зависимость систолического (верхнего) давления крови у женщин от возраста; было получено уравнение линейной регрессии:

$$\tilde{y} = 95{,}5 + 0{,}863\,x,$$

а в примере 8.5 из § 8.3 — оценка коэффициента корреляции для этих данных $r = 0{,}787$. Поскольку $r^2 = 0{,}62$, то можно считать, что лишь 62% изменения признака y связано с учтенным фактором x — возрастом; остальные 38% приходится отнести за счет неучтенных или не поддающихся контролю факторов. Это в свою очередь приведет к значительной ширине доверительной зоны регрессии, которая, согласно § 8.5, как раз и определяется величиной $1 - \rho^2$.

В связи с этим можно попытаться выделить еще какой-нибудь фактор, могущий оказывать систематическое влияние на изучаемый признак. В данном случае таким фактором может быть, например, вес. Если учесть этот дополнительный признак, то мы придем к табл. 8.21, причем прежнее x (возраст) обозначено x_1, а вес — x_2.

ТАБЛИЦА 8.21

Возраст (лет) x_1	71	33	31	55	63	49	58	38	36	64
Вес (кг) x_2	79	64	66	81	77	80	76	76	54	65
Давление крови (мм рт. ст.) y	173	118	125	155	153	160	148	142	110	142

Продолжение

Возраст (лет) x_1	45	68	42	76	34	75	78	62	66	46
Вес (кг) x_2	67	81	70	65	62	82	74	69	71	59
Давление крови (мм рт. ст.) y	128	160	136	150	121	166	154	135	148	127

Для того чтобы дальнейшие выкладки имели более симметричный вид, удобнее записать уравнение регрессии в форме:

$$\tilde{y} = \beta_0 + \beta_1 (x_1 - \mu_1) + \beta_2 (x_2 - \mu_2) + \ldots + \beta_h (x_h - \mu_h), \quad (8.54)$$

где μ обозначает математическое ожидание (см. раздел **3.1.1**). При работе с выборкой мы, естественно, запишем:

$$\tilde{y} = b_0 + b_1 (x_1 - \bar{x}_1) + b_2 (x_2 - \bar{x}_2) + \ldots + b_h (x_h - \bar{x}_h), \quad (8.55)$$

где $\bar{x}$ есть среднее значение (см. раздел **3.2.1**).

Можно показать, что если искать коэффициенты $b_0, b_1, \ldots, b_h$ по методу наименьших квадратов (см. § 8.1), то мы придем к системе нормальных уравнений:

$$\left.\begin{aligned} \sum \xi_1 \xi_1 b_1 + \sum \xi_1 \xi_2 b_2 + \ldots + \sum \xi_1 \xi_h b_h &= \sum \xi_1 \eta, \\ \sum \xi_2 \xi_1 b_1 + \sum \xi_2 \xi_2 b_2 + \ldots + \sum \xi_2 \xi_h b_h &= \sum \xi_2 \eta, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ \sum \xi_h \xi_1 b_1 + \sum \xi_h \xi_2 b_2 + \ldots + \sum \xi_h \xi_h b_h &= \sum \xi_h \eta, \end{aligned}\right\} \quad (8.56)$$

где $\xi = x - \bar{x}$, $\eta = y - \bar{y}$.

Если переменных много, то система (8.56) решается на вычислительной машине; для решения таких задач машины всех типов имеют стандартные программы. В случае двух аргументов, как в разобранном выше примере, система уравнений будет:

$$\left.\begin{aligned} \sum \xi_1^2 b_1 + \sum \xi_1 \xi_2 b_2 &= \sum \xi_1 \eta, \\ \sum \xi_2 \xi_1 b_1 + \sum \xi_2^2 b_2 &= \sum \xi_2 \eta \end{aligned}\right\} \quad (8.57)$$

(причем $\sum \xi_2 \xi_1 = \sum \xi_1 \xi_2$), откуда:

$$\left.\begin{aligned} b_1 &= \left[\left(\sum \xi_1 \eta\right)\left(\sum \xi_2^2\right) - \left(\sum \xi_2 \eta\right)\left(\sum \xi_1 \xi_2\right)\right] / D, \\ b_2 &= \left[\left(\sum \xi_2 \eta\right)\left(\sum \xi_1^2\right) - \left(\sum \xi_1 \eta\right)\left(\sum \xi_1 \xi_2\right)\right] / D, \\ D &= \left(\sum \xi_1^2\right)\left(\sum \xi_2^2\right) - \left(\sum \xi_1 \xi_2\right)^2. \end{aligned}\right\} \quad (8.58)$$

Применим эти формулы к приведенным выше численным данным. В примере 8.3 были получены значения $\Sigma\,\xi_1^2 = 4651{,}0$. $\Sigma\xi_1\eta = 4012{,}5$. Пользуясь данными из табл. 8.21, дополнительно находим $\Sigma\xi_2^2 = 1241{,}8$, $\Sigma\xi_1\xi_2 = 1223{,}0$, $\Sigma\xi_2\eta = 2311{,}1$. Тогда:

$$b_1 = \frac{4012{,}5 \cdot 1241{,}8 - 2311{,}1 \cdot 1223{,}0}{4651{,}0 \cdot 1241{,}8 - (1223{,}0)^2} = 0{,}504,$$

$$b_2 = \frac{2311{,}1 \cdot 14651{,}0 - 4012{,}5 \cdot 1223{,}0}{4651{,}0 \cdot 1241{,}8 - (1223{,}0)^2} = 1{,}365.$$

Учитывая также, что $\overline{x}_1 = 54{,}5$, $\overline{x}_2 = 70{,}9$, $\overline{y} = 142{,}55$, получаем уравнение регрессии:

$$\widetilde{y} = 142{,}55 + 0{,}504\,(x_1 - 54{,}5) + 1{,}365\,(x_2 - 70{,}9) =$$
$$= 18{,}30 + 0{,}504\,x_1 + 1{,}365\,x_2.$$

Теперь для коэффициента регрессии давления крови на возраст[1] получилось другое значение по сравнению с примером 8.5 из § 8.3, а именно 0,504 вместо 0,863. В множественной регрессии это происходит всегда — исключение какой-либо переменной или прибавление повой изменяет все коэффициенты регрессии. Причиной является взаимная коррелированность всех факторов.

8.8.2. Подставив в уравнение регрессии значения x_1 и x_2 из табл. 8.21, получим регрессионные (выравненные) значения $\widetilde{y}$, после чего можно будет вычислить величину $\Sigma(y - \widetilde{y})^2$, т. е. варьирование переменной y за счет отклонения от регрессии. Однако это вычисление можно произвести проще, по формуле:

$$\Sigma\left(y - \widetilde{y}\right)^2 = \Sigma\,\eta^2 - \left(b_1 \Sigma\,\xi_1\eta + b_2 \Sigma\,\xi_2\eta\right); \qquad (8.59)$$

вычитаемое в правой части есть варьирование за счет регрессии. В нашем случае получаем:

$$\Sigma\left(y - \widetilde{y}\right)^2 = 5585 - (0{,}504 \cdot 4012{,}5 + 1{,}365 \cdot 2311{,}1) = 5585 - 5177 = 408,$$

что составляет лишь 7% от полной суммы $\Sigma(y - \overline{y})^2 = 5585$ вместо прежних 38%. Таким образом, оба учтенных фактора — возраст и вес, вместе объясняют 93% всего варьирования переменной y.

Не следует, однако, думать, что на фактор веса (x_2) приходится 93% — 62% = 31% всего варьирования, т. е. вдвое меньше, чем на фактор возраста. Дело в том, что сами переменные x_1 и

[1] Напоминаем, что, как и вообще в этой книге, рассматриваемый пример чисто условный.

x_2 находятся между собой в корреляционной связи (очевидно, $r_{x_1x_2} = 1223{,}0/\sqrt{4651{,}0 \cdot 1241{,}8} = 0{,}509$). Поэтому в $r^2_{y,x_1x_2} = 0{,}62$ уже вошла частично связь между y и x_2. Отсюда видно, что при наличии нескольких факторов оценка роли каждого из них требует большой осторожности.

Далее используются понятия дисперсии (см. раздел **3.3.1***), стандартной ошибки (см. раздел* **3.5.1***), доверительного интервала (см. раздел* **3.7.1***), t-критерия значимости (см. разделы* **4.1.1** *и* **4.1.2***), доверительной зоны регрессии (см. раздел* **8.5.1***).*

Оценка дисперсии, связанной с отклонением от регрессии (дисперсия «случайного варьирования»), равна, очевидно:

$$s^2 = \frac{\sum (y - \tilde{y})^2}{n - m}, \tag{8.60}$$

где n — объем выборки, а m — число случайных переменных (в рассматриваемом случае это x_1, x_2 и y, так что $m = 3$). Она позволяет оценить стандартную ошибку «свободного члена» b_0:

$$s^2_{b_0} = s^2/n = \sum \left(y - \tilde{y}\right)^2 / n(n - m), \tag{8.61}$$

причем $\Sigma(y - \tilde{y})^2$ вычисляется по формуле (8.59). В нашем примере:

$$s_{b_0} = \sqrt{408/20 \cdot 17} = 1{,}095.$$

Стандартные ошибки коэффициентов регрессии b_1 и b_2 можно оценить по формулам:

$$s^2_{b_i} = s^2_{y_0} \sum \xi^2_i / D, \qquad i = 1,\ 2, \tag{8.62}$$

где D дается в формуле (8.58). Для нашего примера получается:

$$s_{b_1} = 0{,}0361, \qquad s_{b_2} = 0{,}0187.$$

Выражение (8.62) можно использовать для построения доверительных интервалов и для проверки значимости коэффициентов регрессии b_1 и b_2. В рассматриваемом примере:

$$t_1 = b_1 / s_{b_1} = 0{,}504/0{,}0361 = 14{,}0, \quad f = 17,$$

$$t_2 = b_2 / s_{b_2} = 1{,}365/0{,}0187 = 73{,}0, \quad f = 17,$$

т. е. оба коэффициента значимы.

8.8.3. Доверительная зона регрессии, т. е. доверительные интервалы для величин $\tilde{y}$, соответствующих заданным парам значений (x_1, x_2), будет:

$$\tilde{y}(x_1, x_2) \pm t_P \, s_{\tilde{y}},$$

где

$$s_{\tilde{y}} = \sqrt{\frac{\sum (y - \tilde{y})^2}{n - m}\left[\frac{1}{n} + \frac{\sum \xi_1^2}{D}(x_1 - \bar{x}_1)^2 + \frac{\sum \xi_2^2}{D}(x_2 - \bar{x}_2)^2 + \right.}$$

$$\left. + \frac{\sum \xi_1 \xi_2}{D}(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)\right],$$

а $\tilde{y}(x_1, x_2)$ вычисляется по уравнению регрессии.

В рассматриваемом примере

$$n = 20, \quad \bar{x}_1 = 54{,}5, \quad \bar{x}_2 = 70{,}9, \quad \frac{\sum (y - \tilde{y})^2}{n - m} = \frac{408}{17} = 24{,}0,$$

$$\frac{\sum \xi_1^2}{D} = 0{,}001086, \quad \frac{\sum \xi_2^2}{D} = 0{,}000290, \quad \frac{\sum \xi_1 \xi_2}{D} = 0{,}001304.$$

Поэтому если принять $P = 0{,}95 = 95\%$, так что $t_{0,95}(17) = 2{,}11$, то, например, при $x_1 = 60$ лет и $x_2 = 73$ кг получим доверительный интервал: $129{,}88 \pm 2{,}72 = 127{,}16 \div 132{,}60$.

Если число учитываемых факторов больше двух, то нахождение стандартных ошибок коэффициентов регрессии становится громоздким; при большом числе переменных такая задача должна решаться при помощи вычислительных машин. Соответствующая методика изложена достаточно подробно в книгах Дж. У. Снедекора (гл. 14, п. 10) и Р. Фишера (§ 29).

Глава девятая

ОБРАБОТКА КРИВЫХ «ДОЗА — ЭФФЕКТ»

§ 9.1. Задачи статистической обработки кривых «доза — эффект»

9.1.1. В главе восьмой рассматриваются зависимости между двумя варьирующими величинами, x и y, каждая из которых может принимать различные количественные значения. Между тем в биологических и медицинских исследованиях часто сталкиваются с задачами, в которых градации либо аргумента (x), либо функции (y) характеризуются качественно. В первом случае (качественные градации аргумента, причем y может зависеть одновременно от нескольких аргументов x_A, x_B, ...) мы приходим к так называемому дисперсионному анализу. Нас будет сейчас интересовать второй случай, когда качественными являются градации «ответов» y, а еще точнее — когда таких «ответов» может быть только два: наличие или отсутствие какого-либо эффекта (гибель, судороги, излечение и др.). Аргументом при этом является переменная, уровни которой выражаются количественно; обычно это доза какого-либо фармакологического препарата, токсина, ионизирующего излучения и т. п.

Как правило, опыт ставят так, чтобы дозы варьировали не случайно, а принимали бы определенные значения. Что касается «ответов» y, то они обычно определяются не только интенсивностью воздействия (т. е. значением аргумента x), но и множеством неконтролируемых факторов. Это приводит к тому, что если воздействовать одной и той же дозой на группу тест-объектов (животные, чашки Петри с клеточной культурой и др.), то часть этих тест-объектов даст положительный ответ, а другая часть — отрицательный, хотя группу тест-объектов подбирали максимально однородной.

Доля p тест-объектов, дающих положительный ответ, может рассматриваться как оценка вероятности реагирования $\tilde{p}$ при данной дозе (интенсивности воздействия). Эта доля p может быть предметом изучения сама по себе: пользуясь методами, описанными в главе шестой, можно строить доверительные интервалы для $\tilde{p}$, выяснять значимость различия между двумя значениями p_1 и p_2, соответствующими двум определенным дозам $Д_1$ и $Д_2$ и др. Однако, с другой стороны, предметом внимания исследова-

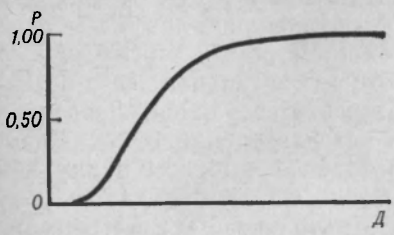


Рис. 9.1. Кривая «доза—эффект».

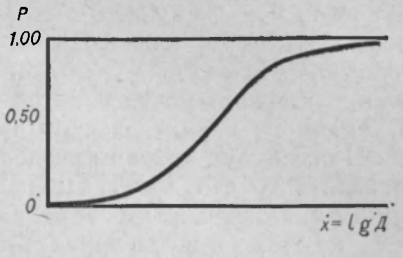


Рис. 9.2. Логарифмическая кривая «доза—эффект».

теля может быть связь между вероятностью реагирования $\hat{p}$, оцениваемой долей положительных ответов p, и интенсивностью воздействия (доза $Д$). Именно этот вопрос и будет рассматриваться в настоящей главе.

Обычно доля p возрастает с увеличением дозы $Д$. Зависимость p от $Д$ изображается S-образной кривой (см. рис. 9.1), называемой *кривой эффекта* или *кривой «доза—эффект»*[1]. Как правило, значения p изменяются от 0 до 1, т. е. при очень малых дозах ни один тест-объект не реагирует, а при достаточно больших дозах реагируют все тест-объекты. Иногда, впрочем, некоторые тест-объекты оказываются совсем нечувствительными к изучаемому действующему фактору и не реагируют даже на очень большие дозы. Тогда кривая «доза—эффект» возрастает только до некоторого значения, меньшего единицы. Такие случаи мы в дальнейшем не будем рассматривать.

Чаще всего кривая эффекта несимметрична, но при замене доз их логарифмами она становится симметричной (рис. 9.2). Поэтому мы будем далее считать, что аргументом являются значения $l = lg\ Д$.

Для краткости будем иногда называть величину l просто дозой.

9.1.2. Хотя зависимость между вероятностью реагирования $\hat{p}$ и дозой $Д$ описывается всей совокупностью экспериментальных данных, графическим изображением которой является кривая «доза—эффект», обычно стараются извлечь из этой совокупности данных некоторые немногие численные характеристики, отражающие наиболее существенные стороны явления. Одной из таких харак-

[1] Разумеется, было бы вполне законно применять это название и в тех случаях, когда эффект выражается численно не в виде доли положительных ответов, а в виде результатов каких-то измерений или счета (артериальное давление, число лейкоцитов и т. п.). Но по установившейся традиции название «кривая «доза — эффект» применяется обычно только к регрессиям первого типа.

теристик может служить та доза, которая вызывает эффект у 50% тест-объектов; ее называют 50%-ной эффективной дозой и обозначают $ЭД_{50}$ (в частности, для токсинов и некоторых других поражающих агентов употребляется 50%-ная летальная доза $ЛД_{50}$). Именно эту характеристику чаще всего и стараются определить.

Ниже будут описаны несколько методов оценки $ЭД_{50}$. Может показаться, что было бы целесообразнее выбрать наилучший из этих методов и им и ограничиться. Однако, в данном случае, как и во многих других, понятия «лучше» и «хуже» неприменимы.

Каждый из методов имеет свои преимущества и недостатки, а также свою область применения. Так, в одних обстоятельствах лучшим оказывается один метод, в других — другой. Поэтому можно настоятельно рекомендовать читателю сначала прочесть всю главу и выбрать метод, наиболее адекватный решаемой биологической задаче, сообразуясь с ограничениями, преимуществами и недостатками каждого из методов, указанных в соответствующих местах, а затем уже более детально разобраться в обосновании и вычислительной процедуре выбранного метода.

При любом методе расчета величина $ЭД_{50}$ оценивается по выборочным данным, поэтому, естественно, должны строиться доверительные интервалы для нее.

Для понимания дальнейшего потребуется знакомство с понятиями среднего значения (см. раздел 3.2.1), стандартной ошибки (см. раздел 3.5.1), доверительного интервала (см. раздел 3.7.1), критериев различия (см. раздел 4.1.1). Если Вы недостаточно хорошо знакомы с этими понятиями, обязательно просмотрите указанные разделы.

§ 9.2. Метод Рида и Менча

9.2.1. Метод Рида и Менча — один из простейших и наиболее часто употребляемых методов определения $ЭД_{50}$. Сущность этого метода поясним на численном примере.

Пример 9.1. В табл. 9.1 даны результаты некоторого исследования, причем в графе 2 указано общее число тест-объектов, получивших ту или иную дозу, а в графах 3 и 4 — соответственно числа этих объектов, давших положительный или отрицательный ответ.

Как видно из чисел графы 5, процент положительных ответов нарастает с увеличением дозы, причем при $l = 3{,}2$ этот процент (42,9) ниже 50, а при $l = 3{,}6$ он выше 50 (83,3).

ТАБЛИЦА 9.1

Логарифм дозы	Число тест-объектов в опыте	Частота эффекта		Положительные ответы, %	Накопленная частота эффекта				Пробит
		есть	нет		есть	нет	сумма	процент	
1	2	3	4	5	6	7	8		
2,4	6	0	6	0,0	0	17	17	0,0	3,16
2,8	7	1	6	14,3	1	11	12	8,2	3,93
3,2	7	3	4	42,9	4	5	9	44,5	4,82
3,6	6	5	1	83,3	9	1	10	90,0	5,97
4,0	6	6	0	100,0	15	0	15	100,0	6,84

Считая, что в средней части графика эффект—доза (рис. 9.3) зависимость может рассматриваться как линейная, естественно применить для интервала доз от $l = 3{,}2$ до $l = 3{,}6$ линейную интерполяцию и записать:

$$\lg ЭД_{50} = 3{,}2 + (3{,}6 - 3{,}2)\frac{50{,}0 - 42{,}9}{83{,}3 - 42{,}9} = 3{,}27,$$

так что $ЭД_{50} = 1{,}86 \cdot 10^3$.

Однако при таком подсчете используются только два опыта—для двух доз, близких к $ЭД_{50}$. Ясно, что данные опытов, относящихся к другим дозам, могли бы уточнить результат, но при описанной выше методике нет возможности включить эти данные в расчет.

Указанный недостаток устранен в методе Рида и Менча. Этот метод исходит из того естественного допущения, что если некоторый тест-объект дал положительный ответ при какой-либо дозе, то он дал бы такой же ответ и при более высоких дозах; и наоборот, если тест-объект дал отрицательный ответ при определенной дозе, то он дал бы также отрицательный ответ и при всех меньших дозах. Поэтому, например, можно считать, что при дозе $l = 3{,}2$ дают положительный ответ не только те три тест-объекта, которые получили эту дозу и дали положительный ответ, но и тот один тест-объект, который дал положительный ответ при меньшей дозе $l = 2{,}8$. В то же время при указанной дозе $l = 3{,}2$ дают отрицательный ответ как те четыре тест-объекта, которые получили эту дозу и дали отрицательный ответ, так и тот один тест-объект, который дал отрицательный ответ при еще большей дозе $l = 3{,}6$. В соответствии со сказанным эффективность

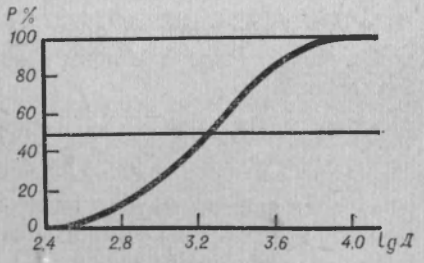


Рис. 9.3. Определение $ЭД_{50}$.

каждой дозы должна характеризоваться не соотношением числа положительных и отрицательных ответов в опыте с этой только дозой, а соотношением н а к о п л е н н ы х частот, включающих надлежащим образом результаты опытов с другими дозами.

Указанные накопленные частоты для рассматриваемого примера записаны в табл. 9.1 в графах 6 и 7, причем накопление числа положительных ответов происходит, естественно, с увеличением дозы (т. е. включает результаты меньших доз), а накопление числа отрицательных ответов — с уменьшением дозы (включает результаты более высоких доз). Числа в графе 8 представляют как бы численности групп, отвечающих тем или иным дозам и обладающих тем свойством, что одна часть этой группы дает положительный ответ при данной и еще меньших дозах, а другая часть дает отрицательный ответ при данной и еще больших дозах. Тогда скорректированный, с учетом результатов опытов с другими дозами, процент положительных ответов при данной дозе дается числами в графе 9. Эти-то проценты и употребляются теперь для нахождения $ЭД_{50}$, опять-таки с применением линейной интерполяции. Если ввести обозначения: A — скорректированный процент положительных ответов, ближайший к 50% снизу (в нашем примере $A = 44{,}5\%$), B — скорректированный процент положительных ответов, ближайший к 50% сверху (в нашем примере $B = 90{,}0\%$), а l_A и l_B — соответствующие дозы (в нашем примере $l_A = 3{,}2$, $l_B = 3{,}6$), то

$$\lg ЭД_{50} = l_{50} = l_A + (l_B - l_A)\frac{50 - A}{B - A}. \qquad (9.1)$$

Используя числа из нашего примера, получаем:

$$\lg ЭД_{50} = 3{,}2 + (3{,}6 - 3{,}2)\frac{50{,}0 - 44{,}5}{90{,}0 - 44{,}5} = 3{,}25$$

и

$$ЭД_{50} = 1{,}78 \cdot 10^3.$$

9.2.2. Стандартную ошибку величины $l_{50} = lg\ ЭД_{50}$, найденной методом Рида и Менча, оценивают по формуле (которую даем без вывода):

$$s_{l_{50}} = \sqrt{\frac{0{,}79\,kh\,(l_{75} - l_{25})}{N}}, \qquad (9.2)$$

где k есть число доз, h — интервал между двумя соседними дозами (в этом методе дозы должны быть равноотстоящими), N — общее число тест-объектов, а l_{25} и l_{75} находятся по тому же правилу, что и l_{50}, т. е. линейной интерполяцией.

Если число тест-объектов при всех дозах одно и то же ($n = N/k$), то, очевидно,

$$s_{l_{50}} = \sqrt{\frac{0{,}79\,h\,(l_{75} - l_{25})}{n}}. \qquad (9.3)$$

В нашем примере $h = 0{,}4$ и

$$l_{25} = 2{,}8 + 0{,}4\,\frac{25{,}0 - 8{,}2}{44{,}5 - 8{,}2} = 2{,}99,$$

$$l_{75} = 3{,}2 + 0{,}4\,\frac{75{,}0 - 44{,}5}{90{,}0 - 44{,}5} = 3{,}47,$$

так что

$$s_{l_{50}} = \sqrt{\frac{0{,}79 \cdot 5 \cdot 0{,}4\,(3{,}47 - 2{,}99)}{32}} = 0{,}154.$$

9.2.3. Доверительный интервал для $ЭД_{50}$ (см. раздел 3.7.1) должен, конечно, строиться с учетом распределения выборочных l_{50}. Как правило, это распределение неизвестно или очень сложно и поэтому оно аппроксимируется нормальным распределением. Тогда для границ доверительного интервала получаем:

$$l_{50} \pm u_P\, s_{l_{50}},$$

где u_P — аргумент интеграла вероятностей для доверительной вероятности P (см. раздел 2.3.2):

$$u_{0{,}95} = 1{,}96, \quad u_{0{,}99} = 2{,}58.$$

В разбираемом примере 95%-ный доверительный интервал для l_{50} будет:

$$3{,}25 - 1{,}96 \cdot 0{,}154 = 2{,}95,$$

$$3{,}25 + 1{,}96 \cdot 0{,}154 = 3{,}55.$$

что дает для $ЭД_{50}$ доверительный интервал $0{,}89 \cdot 10^3 \div 3{,}54 \cdot 10^3$.

9.2.4. Значимость различия в эффективности двух препаратов, инфекционных агентов, видов излучений и т. д. есть значимость сдвига соответствующих кривых доза — эффект вдоль оси доз (точнее — оси логарифмов доз). Ее можно оценить как значимость различия соответствующих $ЭД_{50}$, т. е. по критерию:

$$u = \frac{l_{50}(1) - l_{50}(2)}{\sqrt{s^2_{l_{50}(1)} + s^2_{l_{50}(2)}}}. \qquad (9.4)$$

Как и при нахождении доверительных интервалов, распределение выборочных l_{50} рассматривается как приближенно нормаль-

ное, так что u сравнивается с u_P — аргументом интеграла вероятностей.

В заключение укажем, что метод Рида и Менча можно применять, если значения $l = lg$ Д являются равноотстоящими, а численности групп для разных доз одинаковы (или почти одинаковы). Кроме того, крайне желательно, чтобы эти численности не были меньше 4.

§ 9.3. Метод Кербера

9.3.1. Описанный в § 9.2 метод определения $ЭД_{50}$ Рида и Менча имеет ряд ограничений: значения $l = lg$ Д должны быть равноотстоящими, численности групп для всех доз должны быть одинаковы. Кроме того, применение линейной интерполяции в решающем пункте вычисления $ЭД_{50}$ — использование формулы (9.1) — может внести значительную ошибку в результат, если дозы l_A и l_B не очень близки к l_{50}, так что на участке от l_A до l_B зависимость доза — эффект заметно криволинейна.

Метод Кербера, который мы сейчас рассмотрим, свободен от этих недостатков и сохраняет в то же время основную здоровую идею метода Рида и Менча: если тест-объект дал положительный ответ при дозе l, то он дал бы также положительный ответ и при более высокой дозе, а если он дал отрицательный ответ при этой дозе, то такой же отрицательный ответ он дал бы и при меньшей дозе. В таком случае l_{50} есть такая доза, что половина тест-объектов дает положительный ответ при этой и всех меньших (именно меньших) дозах, а другая половина тест-объектов дает отрицательный ответ при этой и всех больших дозах. Графически это можно выразить как равенство заштрихованных площадей на рис. 9.4.

Пусть l_0 обозначает дозу, при которой вероятность положительного ответа есть $p_0 \approx 0$, а l_{100} обозначает дозу, при которой эта вероятность есть $p_{100} \approx 100\% = 1$. Обозначим через S площадь под кривой доза — эффект в пределах от l_0 до l_{100}, т. е. площадь фигуры $l_0QPl_{100}l_0$. Очевидно, можно также записать:

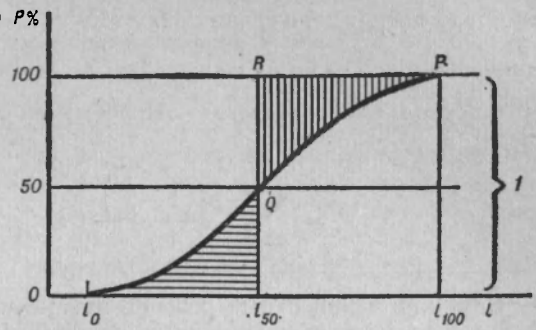
$$S = \text{пл. } l_0 Ql_{50} l_0 + \text{пл. } l_{50} QPl_{100} l_{50}.$$

Но доза l_{50} была определена так, чтобы выполнялось равенство пл. $l_0Ql_{50}l_0 =$ пл. $QRPQ$. Используя это равенство, имеем:

$$S = \text{пл. } QRPQ + \text{пл. } l_{50} QPl_{100} l_{50} = \text{пл. } l_{50} RPl_{100} l_{50}.$$

Высота последнего прямоугольника равна 1 (максимальная вероятность положительного ответа), поэтому его площадь численно равна длине его основания, т. е. $S = l_{100} - l_{50}$, откуда $l_{50} = l_{100} - S$. Задача сводится, таким образом, к нахождению площади фигуры под кривой доза—эффект.

Рис. 9.4. Нахождение $ЭД_{50}$ по методу Кербера.



Пусть теперь $l_{(0)}, l_{(1)}, l_{(2)}, \ldots, l_{(k)}, l_{(k+1)}$ обозначают дозы, употреблявшиеся в опыте, причем индексы в скобках обозначают просто номера доз, а не вероятности положительных ответов, как у l_{50} или l_{100}. Примем, далее, что $l_{(0)} = l_0$, $l_{(k+1)} = l_{100}$. Очевидно, площадь S под кривой можно аппроксимировать суммой трапеций, площадь каждой из которых равна:

$$\frac{\hat{p}_i + \hat{p}_{i+1}}{2}\left(l_{(i+1)} - l_{(i)}\right).$$

Тогда:

$$S \approx \frac{1}{2}\left(\hat{p}_0 + \hat{p}_1\right)\left(l_{(1)} - l_{(0)}\right) + \left(\hat{p}_1 + \hat{p}_2\right)\left(l_{(2)} - l_{(1)}\right) + \ldots +$$
$$+ \left(\hat{p}_k + \hat{p}_{k+1}\right)\left(l_{(k+1)} - l_{(k)}\right).$$

Произведя элементарное преобразование с приведением подобных членов и учитывая, что $\hat{p} \approx 0$ и $\hat{p}_{k+1} \approx 1$, а также заменяя неизвестные вероятности $\hat{p}$ положительных ответов соответствующими долями p, получаем:

$$l_{50} \approx \frac{1}{2}\left(l_{(k)} + l_{(k+1)}\right) - \frac{1}{2}\sum_{i=1}^{k} p_i\left(l_{(i+1)} - l_{(i-1)}\right). \qquad (9.5)$$

Здесь дозы могут быть не равноотстоящими, а численности n_i групп тест-объектов при разных дозах могут быть различны. Даже лучше, если эти численности для средних доз больше, чем для крайних.

9.3.2. При оценке стандартной ошибки величины l_{50} исходят из того, что выборочная l_{50}, согласно формуле (9.5), есть сумма одного постоянного слагаемого $\frac{1}{2}(l_{(k)} + l_{(k+1)})$, дисперсия которо-

го равна нулю, и k случайных слагаемых вида $\frac{1}{2}\ p_i(l_{(i+1)} - l_{(i-1)})$, причем в каждом из них случайным является p_i, а $\frac{1}{2}(l_{(i+1)} - l_{(i-1)})$ есть постоянный множитель. Поэтому на основании формул (3.21), (3.20) и (6.8) имеем:

$$\sigma^2_{l_{50}} = \sum_{i=1}^{k} \frac{\hat{p}_i\left(1-\hat{p}_i\right)}{4n_i}\left(l_{(i+1)} - l_{(i-1)}\right)^2.$$

Заменяя неизвестные $\hat{p}$ эмпирическими p, а числа n_i величинами $n_i - 1$, получаем оценку:

$$s^2_{l_{50}} = \sum_{i=1}^{k} \frac{p_i(1-p_i)}{4(n_i-1)}\left(l_{(i+1)} - l_{(i-1)}\right)^2. \qquad (9.6)$$

9.3.3. Разумеется, формулы (9.5) и (9.6) значительно упрощаются, если дозы взяты равноотстоящими: $l_{(i+1)} - l_{(i)} = h$ для всех i. Тогда:

$$\frac{1}{2}\left(l_{(k)} + l_{(k+1)}\right) = \frac{1}{2}\left(l_{(k+1)} - h + l_{(k+1)}\right) = l_{(k+1)} - \frac{h}{2} =$$

$$= l_{100} - \frac{h}{2}.$$

$$\frac{1}{2}\left(l_{(i+1)} - l_{(i-1)}\right) = h,$$

так что:

$$l_{50} = l_{100} - h\left(\sum_{i=1}^{k} p_i + \frac{1}{2}\right), \qquad (9.7)$$

$$s^2_{l_{50}} = h^2 \sum_{i=1}^{k} \frac{p_i(1-p_i)}{n_i - 1}. \qquad (9.8)$$

Для данных из табл. 9.1 получаем:

$$l_{50} = 4{,}00 - 0{,}4\,(0{,}143 + 0{,}429 + 0{,}833 + 0{,}500) = 3{,}24,$$

$$s_{l_{50}} = 0{,}4\sqrt{\frac{0{,}143 \cdot 0{,}857}{6} + \frac{0{,}429 \cdot 0{,}571}{6} + \frac{0{,}833 \cdot 0{,}167}{5}} = 0{,}119.$$

Это дает 95%-ный доверительный интервал для $\hat{l}_{50}$:

$$3{,}24 - 1{,}96 \cdot 0{,}119 = 3{,}01,$$

$$3{,}24 + 1{,}96 \cdot 0{,}119 = 3{,}47$$

или доверительный интервал для $ЭД_{50}$: $1{,}02 \cdot 10^3 \div 2{,}95 \cdot 10^3$.

9.3.4. Мы видим, что вычисления по методу Кербера более просты, а доверительный интервал более узкий, чем в методе Рида и Менча. Кроме того, здесь нет такой грубой аппроксимации, как линейная интерполяция в не совсем линейной части кривой «доза—эффект». И, наконец, метод Кербера не требует ни равноотстоящих доз в серии опытов, ни одинаковой численности тест-объектов для каждой дозы.

Но с другой стороны и метод Кербера имеет определенные ограничения и недостатки. Прежде всего диапазон доз должен быть настолько широк, чтобы включать дозы с $p = 0$ и $p = 1$. Далее, замена площади, ограниченной кривой, на сумму трапеций не может не сказаться на точности результата. Но значительно снижает точность результата то обстоятельство, что получаемое значение $ЭД_{50}$ в решающей степени зависит от значения l_{100}, которое определяется экспериментально не слишком точно.

§ 9.4. Пробит-метод

Прежде чем читать этот параграф, необходимо ознакомиться с некоторыми свойствами нормального распределения (см. раздел **2.3.2***), а также с понятиями математического ожидания (см. раздел* **3.1.1***) и стандартного отклонения (см. раздел* **3.5.1***).*

9.4.1. От ограничений и недостатков метода Рида—Менча и метода Кербера можно избавиться, если принять, что в генеральной совокупности кривая «доза—эффект» (после преобразования $l = lg$ Д, см. рис. 9.2) совпадает с графиком функции нормального распределения (рис. 2.7). Тогда доли p положительных ответов можно приравнять к накопленным частостям z нормального распределения, для которых имеет место равенство:

$$z = \Phi\left(\frac{x - \mu}{\sigma}\right),$$

где Ф— интеграл вероятностей, μ и σ — математическое ожидание и стандартное отклонение распределения. Заменяя z на p, x на l и μ на l_{50}, имеем:

$$p = \Phi\left(\frac{l - l_{50}}{\sigma}\right). \qquad (*)$$

Построим теперь график, откладывая по оси ординат (т. е. по вертикальной оси) не значения p, а значения:

$$y' = \Psi(p),$$

где Ψ— функция, обратная к интегралу вероятностей Ф. Последнее означает, что

$$p = \Phi(y'),$$

а тогда сравнение с (*) дает:

$$y' = \frac{l - l_{50}}{\sigma},$$

что можно переписать в виде:

$$y' = \frac{1}{\sigma} l - \frac{l_{50}}{\sigma},$$

т. е. в виде уравнения прямой линии. Следовательно, если откладывать по оси абсцисс значения l, а по оси ординат — значения $y' = \Psi(l)$, то точки расположатся вдоль прямой линии (имеющей наклон $1/\sigma$).

При $p < 0{,}5$ значения y' отрицательны, что представляет собой известное неудобство. Во избежание этого заменяют величины y' величинами $y = y' + a$, где a — некоторое положительное число. Его надо выбрать так, чтобы оно превышало по абсолютной величине все отрицательные значения $\Psi(p)$, которые могут встретиться на практике. Это будет выполнено, если принять $a = 5$, так как значению $\Psi(p) = +5$ соответствует очень малое значение $p \approx 0{,}0000003$. Таким образом, по оси ординат мы будем откладывать величины:

$$y = \Psi(p) + 5. \qquad (9.9)$$

Эти величины называют *пробитами* (от английского probability unit — вероятностная единица), в связи с чем излагаемый метод анализа кривых «доза—эффект», использующий в качестве модели график интеграла вероятностей, называется *пробит-методом*, или *пробит-анализом* (Ч. Блисс).

Если численности тест-объектов в отдельных группах невелики, то можно обойтись без вычисления доли положительных ответов и пользования таблицей интеграла вероятностей (табл. III Приложений): имеется специально разработанная, очень удобная таблица, в которой пробит для данного опыта находится непосредственно по общему числу тест-объектов в опыте и числу положительных ответов (табл. 9.2). Например, если из 9 тест-объектов 6 дали положительный ответ, то пробит будет 5,43. Значения пробитов для $p = 0$ и $p = 1$ нельзя найти из равенства $y' = \Psi(p)$,

ТАБЛИЦА 9.2

Значения пробитов (по книге А. Н. Кудрина и Г. Т. Пономаревой, с. 218)

Число тест-объектов в группе	Число тест-объектов, у которых наблюдается изучаемый эффект															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3	3,50	4,57	5,43	6,50	—	—	—	—	—	—	—	—	—	—	—	—
4	3,36	4,33	5,00	5,67	6,64	—	—	—	—	—	—	—	—	—	—	—
5	3,25	4,16	4,75	5,25	5,84	6,75	—	—	—	—	—	—	—	—	—	—
6	3,16	4,03	4,57	5,00	5,43	5,97	6,84	—	—	—	—	—	—	—	—	—
7	3,10	3,93	4,43	4,82	5,18	5,57	6,07	6,90	—	—	—	—	—	—	—	—
8	3,04	3,85	4,33	4,68	5,00	5,32	5,67	6,15	6,96	—	—	—	—	—	—	—
9	2,99	3,78	4,23	4,57	4,86	5,14	5,43	5,77	6,22	7,01	—	—	—	—	—	—
10	2,95	3,72	4,16	4,48	4,75	5,00	5,25	5,52	5,84	6,28	7,05	—	—	—	—	—
11	2,90	3,67	4,09	4,40	4,65	4,89	5,11	5,35	5,60	5,91	6,33	7,10	—	—	—	—
12	2,88	3,61	4,03	4,33	4,57	4,79	5,00	6,21	5,43	5,67	5,97	6,39	7,12	—	—	—
13	2,84	3,57	3,98	4,26	4,50	4,71	4,90	5,10	5,29	5,50	5,74	6,02	6,43	7,16	—	—
14	2,81	3,53	3,93	4,21	4,43	4,63	4,82	5,00	5,18	5,37	5,57	5,79	6,07	6,47	7,19	—
15	2,78	3,50	3,89	4,16	4,38	4,57	4,75	4,92	5,08	5,25	5,43	5,62	5,84	6,11	6,50	7,22

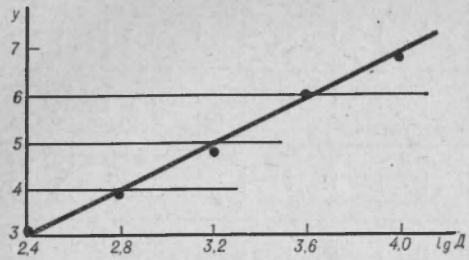


Рис. 9.5. Пробит-метод анализа кривых «доза — эффект».

так как $\Psi(0) = -\infty$ и $\Psi(1) = \infty$. В таких случаях можно рекомендовать производить замену $p = 0 \to p^* = \frac{1}{5n_i}$, $p = 1 \to p^* = 1 - \frac{1}{5n_i}$, после чего уже можно вычислять пробиты из равенства $y' = \Psi(p^*)$. Соответствующие значения записаны в табл. 9.2.

9.4.2. Чтобы определить l_{50} (а затем и $ЭД_{50}$), нужно найти абсциссу той точки прямой, ордината которой равна $y = 5$ (так как это соответствует $y' = 0$ и $p = 0{,}5$). Поэтому нужно прежде всего провести прямую, используя экспериментальные точки l, y. В первом приближении это можно сделать графически, пользуясь прозрачной линейкой. При этом надо придавать отдельным точкам тем больший «вес», чем ближе к 5 соответствующий пробит.

Существуют расчетные методы, которые позволяют более точно провести прямую, лучше всего отвечающую экспериментальным точкам (см. монографии М. Л. Беленького, 1963; Б. С. Бессмертного, 1967; А. Н. Кудрина и Г. И. Пономаревой, 1967). Однако в пробит-анализе главной причиной неточностей является обычно то, что логарифмическая кривая эффекта, даже для генеральной совокупности, не имеет в точности нормальную форму. Поэтому использование этих более сложных вычислительных методов чаще всего приводит к кажущемуся повышению точности.

Пример 9.2. Для данных из табл. 9.1 применение пробит-метода дает (см. числа в графе 10 табл. 9.1 и рис. 9.5) $l_{50} = 3{,}21$ и $ЭД_{50} = 1{,}62 \cdot 10^3$.

9.4.3. Для оценки стандартной ошибки $\sigma_{l_{50}}$ были предложены разные методы[1]. По-видимому, достаточно простой и в то же время достаточно точной является формула

$$s_{l_{50}} = \frac{\sigma}{\sqrt{N'/2}}, \qquad (9.10)$$

где N' есть число тест-объектов, для которых значения пробитов находятся в пределах от 3,5 до 6,5 (в нашем примере $N' = 20$).

[1] См., например, книгу C.I. Bliss.

Величину σ оценивают по графику: точки прямой, имеющие ординаты 4,0 и 6,0, соответствуют значениям l, равным $l_{50} - \sigma$ и $l_{50} + \sigma$, так что

$$\sigma = l_{y=5,0} - l_{y=4,0} = l_{y=6,0} - l_{y=5,0};$$

более точную оценку σ дает выражение:

$$\sigma = \frac{1}{2}\left(l_{y=6,0} - l_{y=4,0}\right). \qquad (9.11)$$

Для разбираемого примера $l_{y=4,0} = 2{,}79$, $l_{y=6,0} = 3{,}63$, так что $\sigma = \frac{1}{2}(3{,}63 - 2{,}79) = 0{,}42$. Пользуясь формулой (9.10), находим:

$$s_{l_{50}} = 0{,}42 / \sqrt{10} = 0{,}133,$$

поэтому 95%-ные доверительные интервалы для $\hat{l}_{50}$ и $\hat{\text{ЭД}}_{50}$ будут:

$$\hat{l}_{50} : 3{,}21 \pm 1{,}96 \cdot 0{,}133 = 2{,}95 \div 3{,}47,$$

$$\text{ЭД}_{50} : 0{,}89 \cdot 10^3 \div 2{,}96 \cdot 10^3.$$

В заключение отметим еще раз, что применение пробит-метода не требует ни равноотстоящих доз, ни одинаковой численности тест-объектов в группах. Далее, дозы с $p = 0$ и $p = 1$ могут как присутствовать, так и отсутствовать. И точность этого метода, по сравнению с другими описанными выше методами, достаточно высока. Вместе с тем слабым пунктом пробит-анализа является использование допущения о нормальности кривой «доза—эффект».

Глава десятая

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

§ 10.1. Задачи дискриминантного анализа

10.1.1. В биологии и медицине часто возникает задача об отнесении индивида к одному из известных классов. Это могут быть разные биологические виды, к одному из которых принадлежит изучаемый экземпляр, или разные заболевания, одним из которых страдает обследуемый больной.

Вопрос об отнесении данного индивида к одному из классов решается всегда на основе изучения тех или иных признаков, характеризующих этот индивид, и сравнения их с признаками, характерными для каждого из «конкурирующих» классов.

К сожалению, чаще всего интервалы варьирования численных значений признаков у объектов, принадлежащих к разным классам, перекрываются. В этом случае отнесение к определенному классу индивида, у которого значение рассматриваемого признака лежит в области перекрытия, требует специального анализа. Например, у цветов *Enotera muricata* число лепестков варьирует от 10 до 16, а у *Enotera biennis* — от 14 до 25. Поэтому, если у какого-либо растения число лепестков оказалось $x < 14$, то его можно отнести к *E. muricata*, а если $x > 16$, то его можно отнести к *E. biennis*[1]; но если подсчет числа лепестков даст для x значение, лежащее между 13 и 17, то сделать определенное заключение о принадлежности растения к тому или другому из двух видов будет нельзя.

Самым простым было бы исключить этот сомнительный интервал из таксономии, отказываясь при $14 \leqslant x \leqslant 16$ от каких-либо заключений. Но обычно область перекрытия настолько велика, что исключение ее сделало бы практически непригодными для отнесения почти все признаки. При нормальном же распределении, которое встречается довольно часто, «хвосты» распределений в обеих группах тянутся в обе стороны неограниченно далеко, так что в принципе вообще нет таких участков, где не было бы перекрытия. Поэтому представляется целесообразным выбрать какое-либо одно значение x_0 внутри упомянутого интервала пере-

---

[1] Здесь и в дальнейшем имеются в виду выводы, которые можно сделать на основании изучения лишь тех признаков, которые в данном месте обсуждаются.

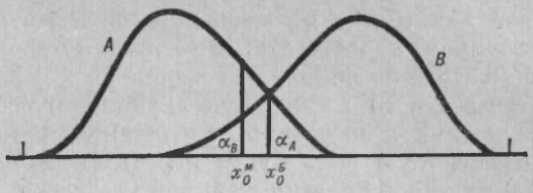


Рис. 10.1. Решающие правила при разграничении одномерных совокупностей.

крытия в качестве граничного, относя индивиды с $x < x_0$ к одной из групп (назовем ее группой A), а индивиды с $x > x_0$ к другой группе (группа B).

Очевидно, что при этом те индивиды с $x < x_0$, которые на самом деле принадлежат к группе A будут отнесены правильно, а те индивиды с $x < x_0$, которые на самом деле принадлежат к группе B, будут отнесены неправильно; наоборот, индивиды с $x > x_0$, принадлежащие к группе B, будут отнесены правильно, а индивиды с $x > x_0$, принадлежащие к группе A, будут отнесены неправильно. Соответствующие доли неправильных отнесений обозначим через α_A, и α_B.

Что касается выбора граничной точки x_0, то проще всего было бы выбрать x_0 в середине интервала перекрытия. Однако, как мы сейчас увидим, можно предложить другие способы выбора x_0, которые, уступая первому в простоте, имеют зато более важные преимущества. Кроме того, и это самое важное, чаще всего (например, при нормальном или близком к нему распределении) вообще невозможно указать определенные границы интервала перекрытия.

На рис. 10.1 видно, что имеет смысл выбрать в качестве x_0 абсциссу точки пересечения обеих кривых плотности распределения. В самом деле, ордината любой точки кривой плотности распределения пропорциональна вероятности того, что индивид имеет заданное значение x. Поэтому значение $x_0^Б$, выбранное из условия $p_A(x_0^Б) = p_B(x_0^Б)$, имеет следующее важное свойство: если $x < x_0^Б$, то более вероятно, что индивид принадлежит к группе A, а если $x > x_0^Б$, то более вероятно, что индивид принадлежит к группе B.

О плотности распределения см. раздел **2.2.5.** *Кроме того, обязательно прочтите весь § 2.3, в котором описывается нормальное распределение: в настоящей главе постоянно будут использоваться свойства этого распределения.*

10.1.2. Задача отнесения несколько видоизменяется, если имеется та или иная дополнительная информация о двух рассматриваемых группах. Прежде всего исследователь может распола-

гать данными о соотношении численностей обеих групп. Если, например, подлежащий отнесению к одному из двух видов экземпляр скворца обнаружен в районе, где значительно преобладают птицы одного из видов, то знание этого факта не может, конечно, не повлиять на результат отнесения. В медицинской диагностике аналогичную роль будут играть сведения о распространенности той или иной болезни в данной местности, о роде занятий больного в случае профессиональных заболеваний, о наличии в данный момент эпидемии и т. д.

Эти сведения при прочих равных условиях повышают вероятность того, что индивид принадлежит к одной из двух групп и соответственно понижают вероятность принадлежать к другой группе. Тем самым они нарушают равенство $p_A = p_B$, которое было положено ранее в основу нахождения граничного значения x_0.

Указанную дополнительную информацию можно выразить в виде *априорных вероятностей* q_A и q_B, приписываемых каждой группе. С учетом этих априорных вероятностей условие $p_A = p_B$, определяющее граничное значение x_0, заменится условием:

$$q_A p_A = q_B p_B.$$

И еще одно обстоятельство должно быть принято во внимание: ошибки при отнесении могут иметь неодинаковую «цену». Пусть, например, производится диспансерное обследование какой-либо группы населения. Тогда ошибка, состоящая в отнесении здорового человека к больным, имеет неприятным последствием некоторую моральную травму, необходимость проведения дополнительных анализов и др. Но зато противоположная ошибка, когда больной человек отнесен к здоровым, может привести к тому, что не будет начато своевременно лечение, а это может иметь более опасные последствия.

Если «цена» одного ошибочного отнесения индивида не в свою группу составляет соответственно C_A и C_B, то за все α_A ошибок придется «уплатить» $C_A \alpha_A$, а за все α_B ошибки придется «уплатить» $C_B \alpha_B$. Общая «плата» за ошибки будет $C_A \alpha_A + C_B \alpha_B$. Естественно выбрать граничное значение x_0 так, чтобы указанная общая «плата» за ошибки была минимальной. Этот критерий оптимальности носит название *критерия Байеса*. Можно показать, что использование критерия Байеса приводит к следующему уравнению для нахождения $x_0^Б$.

$$C_A q_A p_A(x_0^Б) = C_B q_B p_B(x_0^Б).$$

Наряду с критерием Байеса можно использовать другой критерий, не требующий знания априорных вероятностей q_A и q_B. На рис. 10.1 видно, что при перемещении граничной точки x_0 вправо ошибка отнесения α_A уменьшается, а α_B увеличивается;

при перемещении точки x_0 влево α_A и α_B меняются в обратном направлении. Очевидно, можно найти такую точку x_0^{II}, при которой $\alpha_A = \alpha_B$. При этом условии максимальная ошибка (т. е. наибольшая из двух ошибок) имеет минимальное значение, поэтому данный критерий называют *минимаксным критерием*. Если распределения в группах A и B симметричны и имеют одинаковые стандартные отклонения σ, то x_0^{II} и x_0^{E} совпадают.

В случае нормального распределения имеем:

$$\alpha_A(x) = \Phi\left(-\frac{x-\mu^A}{\sigma_A}\right), \quad \alpha_B(x) = \Phi\left(\frac{x-\mu^B}{\sigma_B}\right), \quad (10.1)$$

где Ф— интеграл вероятностей (см. раздел **2.3.2**), а μ и σ — математическое ожидание и стандартное отклонение распределения (см. разделы **3.1.1** и **3.3.1**). Тогда условие минимакса гласит:

$$\Phi\left(-\frac{x_0-\mu^A}{\sigma_A}\right) = \Phi\left(\frac{x_0-\mu^B}{\sigma_B}\right).$$

Так как функция Ф — однозначная, то равенство значений функции означает равенство значений аргумента, так что получается:

$$x_0 = \frac{\sigma_A\,\mu^B + \sigma_B\,\mu^A}{\sigma_A + \sigma_B}. \quad (10.2)$$

Если совокупности A и B представлены выборками, то математические ожидания μ^A, μ^B и стандартные отклонения σ_A, σ_B заменяются их выборочными оценками $\bar{x}^A$ и $\bar{x}^B$ и s_A, s_B (см. разделы **3.2.1** и **3.4.1**). Это относится и ко всем последующим формулам.

10.1.3. Хотя, согласно идее дискриминантного анализа, все индивиды со значениями $x < x_0$ относятся к группе A, представляется естественным полагать, что при значениях x, далеких от x_0, отнесение индивида к этой группе более бесспорно, чем при значениях x, близких к x_0. Это обстоятельство можно выразить количественно. Из соображений, относящихся к рис. 10.1, следует, что вероятности P_A и P_B того, что индивид со значением x принадлежит к группе A или B, пропорциональны плотностям вероятностей распределений $p_A(x)$ и $p_B(x)$:

$$\frac{P_A}{P_B} = \frac{p_A(x)}{p_B(x)}.$$

Учитывая, что $P_A + P_B = 1$, получим:

$$P_A = \frac{1}{1 + p_B(x)/p_A(x)}, \quad P_B = \frac{1}{1 + p_A(x)/p_B(x)}.$$

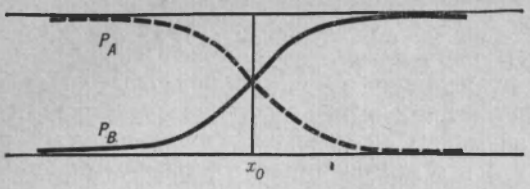


Рис. 10.2. Логистическая функция.

Пусть распределения нормальны и для простоты дисперсии одинаковы. В этом случае x_0 определяется формулой (10.2), и если выбрать начало координат в точке x_0, то μ^A и μ^B будут иметь координаты: $-\frac{d}{2}$ и $+\frac{d}{2}$, где $d = \mu^A - \mu^B$.

Тогда:

$$p_A(x) / p_B(x) = exp\left\{-\left(x + \frac{d}{2}\right)^2 / 2\sigma^2 + \left(x - \frac{d}{2}\right)^2 / 2\sigma^2\right\} =$$

$$= exp\left\{-\frac{d}{\sigma^2} x\right\},$$

так что:

$$P_B = \frac{1}{1 + e^{-(d/\sigma^2)x}}. \qquad (10.3)$$

Это так называемая *логистическая функция*. График ее изображен на рис. 10.2. Если $x < x_0$, то $P_B < \frac{1}{2}$ и $P_A > \frac{1}{2}$, а если $x > x_0$, то $P_B > \frac{1}{2}$ и $P_A < \frac{1}{2}$. Значения функции (10.3) при разных $l = \frac{d}{\sigma^2} x$ приведены в табл. 10.1. Например, если $d = 8$ и $\sigma = 4$, то при $x = 3$ имеем:

$$P_B = 0{,}818, \quad P_A = 1 - P_B = 0{,}182.$$

ТАБЛИЦА 10.1

Значения функции $P(l) = 1/(1 + e^{-l})$. Ноль целых и запятая опущены

	0	1	2	3	4	5	6	7	8	9
0,	500	525	550	574	599	622	646	668	690	711
1,	731	750	768	786	802	818	832	846	858	870
2,	881	891	900	909	917	924	931	937	943	948
3,	953	957	961	964	968	971	973	976	978	980
4,	982	984	985	987	988	989	990	991	992	993
5,	9933	9939	9945	9950	9955	9959	9963	9967	9970	9973

Если вероятность P_B (или P_A) оказалась недостаточно высокой, то можно предполагать, что данный индивид принадлежит

к той части (численно определяемой величиной α) индивидов, которые относятся неправильно формальным дискриминантным анализом. Тогда можно в некоторых случаях изменить решение на противоположное, если имеется какая-либо дополнительная неформальная информация. Можно показать, что выбор в качестве «достаточной высокой» вероятности 0,8 обладает некоторыми свойствами оптимальности.

10.1.4. Важнейшее значение в проблеме дискриминации имеет размер доли неправильных отнесений. Признак x пригоден для решения задачи дискриминации только тогда, когда эта доля достаточно мала.

В случае нормального распределения доля неправильных отнесений равна согласно формуле (10.1):

$$\alpha_A = \alpha_B = \Phi\left(-\frac{x_0 - \mu^A}{\sigma_A}\right) = \Phi\left(\frac{x_0 - \mu^B}{\sigma_B}\right).$$

Подставляя сюда x_0 из (10.2), получаем:

$$\alpha_A = \alpha_B = \Phi\left(-\frac{\mu_B - \mu_A}{\sigma_A + \sigma_B}\right) = \Phi\left(-\frac{d}{\sigma_A + \sigma_B}\right), \qquad (10.4)$$

где $d = \mu^B - \mu^A$.

§ 10.2. Линейная дискриминантная функция

10.2.1. Почти всегда оказывается, что ни один из признаков не обеспечивает достаточно удовлетворительного разграничения двух совокупностей, т. е. достаточной малости неверных отнесений. В этом случае можно надеяться улучшить положение, используя одновременно два или большее число признаков, самих по себе недостаточных. Надежда основана на том, что весьма часто индивиды, попадающие в область перекрытия по одному признаку, находятся вне области перекрытия по другому признаку, в то время как индивиды, попадающие в область перекрытия по второму признаку, находятся вне области перекрытия по первому признаку.

Разберем этот вопрос подробнее. Если индивиды распределены одновременно по двум признакам, то плотность распределения изображается не кривой на плоскости, а некоторой поверхностью в трехмерном пространстве. Если теперь провести плоскость, параллельную плоскости координат xOy, то она пересечет поверхность распределения по некоторой кривой. При нормальном распределении эта кривая будет эллипсом, который называется *корреляционным эллипсом*. Используя этот эллипс в качестве направляющей, построим цилиндрическую поверхность, перпендикулярную к координатной плоскости. Если внутри этого цилиндра

находится $P\%$ всех индивидов, то корреляционный эллипс назовем P-процентным.

Далее используются понятия дисперсии (см. раздел **3.3.1***) и ее выборочной оценки (см. раздел* **3.4.1***), а также ковариации (см. раздел* **8.2.1***).*

При заданном P величина и форма корреляционного эллипса, а также направление его осей зависит от дисперсий по признакам x и y, т. е. от величин σ^2_{xx} и σ^2_{yy}, а также от ковариации σ^2_{xy}. В практике дискриминантного анализа используют конечные выборки, так что σ^2_{xx}, σ^2_{yy} и σ^2_{yx} заменяются их выборочными оценками:

$$\left.\begin{aligned} s^2_{xx} &= \frac{1}{n-1}\sum (x-\bar{x})^2, \qquad s^2_{yy} = \frac{1}{n-1}\sum (y-\bar{y})^2, \\ s^2_{xy} &= \frac{1}{n-1}\sum (x-\bar{x})(y-\bar{y}). \end{aligned}\right\} \quad (*)$$

Эти величины можно записать в виде таблицы

$$S = \begin{pmatrix} s^2_{xx} & s^2_{xy} \\ s^2_{yx} & s^2_{yy} \end{pmatrix},$$

которую называют *ковариационной матрицей*; из (*) видно, что $s^2_{yx} = s^2_{xy}$.

Изобразим на чертеже (рис. 10.3) два $P\%$-ных корреляционных эллипса с разными положениями центров $(\bar{x}^A, \bar{y}^A)$ и $(\bar{x}^B, \bar{y}^B)$ и разными ковариационными матрицами $S^A \neq S^B$. Если бы учитывался только признак x, то все индивиды, попадающие в вертикально заштрихованные части эллипсов, оказались бы в области перекрытия и не могли бы быть однозначно отнесены. Учет одного лишь признака y также был бы недостаточен, так как в этом случае в области перекрытия оказались бы индивиды, попадающие в горизонтально заштрихованные части эллипсов. Только совместный учет обоих признаков дает возможность произвести полную однозначную классификацию (точнее — с ошибкой, не превышающей $100 - P$). При этом можно произвести достаточно полное однозначное отнесение и в тех случаях, когда часть индивидов попадает в область перекрытия по каждому из признаков; на рис. 10.3 это части эллипсов с двойной штриховкой.

Отнесение каждого нового индивида производится в соответствии с тем, в какой из эллипсов он попадает — точнее, по какую сторону от граничной линии (на рис. 10.3 она проведена пункти-

ром). Поскольку P%-ные эллипсы в этом примере не перекрываются, ошибка отнесения будет меньше $100 - P$.

Из сказанного ясно, что задача отнесения сводится по существу к нахождению граничной линии на плоскости xOy. В общем случае эта граничная линия представляет собой некоторую кривую. Но в первом приближении эту кривую можно аппроксимировать прямой линией. Такое приближенное решение задачи мы сейчас и рассмотрим.

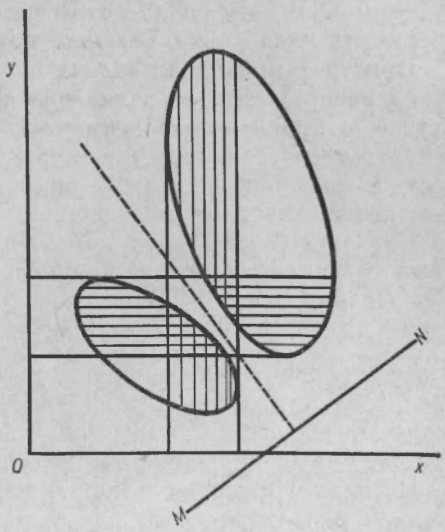


Рис. 10.3. Линейное разграничение двумерных совокупностей.

Если два двумерных распределения, корреляционные эллипсы которых изображены на рис. 10.3, спроектировать на ось Ox, то получатся два одномерных распределения, которые довольно сильно перекрываются. То же получится, если проектировать на ось Oy. Но если бы мы стали проектировать, например, на прямую MN (рис. 10.3), то перекрывание было бы гораздо меньше. Очевидно, надо выбрать направление MN так, чтобы перекрывание было наименьшим. Если затем провести прямую, перпендикулярную к MN и пересекающую ее в подходящем месте, то эту прямую и можно будет считать граничной линией. Такой метод разграничения был предложен Р. Фишером.

Задачу о выборе направления MN можно заменить задачей о повороте системы координат, что в свою очередь сводится к некоторому линейному преобразованию координат:

$$X = a_1 x + a_2 y,$$
$$Y = b_1 x + b_2 y,$$

где коэффициенты a_1, a_2, b_1, b_2 определяются углом поворота системы координат. Метод Фишера состоит в проектировании обоих распределений на одну из новых осей. Если условиться проектировать всегда на ось OX, то значения Y вообще не понадобятся, поскольку все частоты, относящиеся к каждому значению X, просто суммируются, независимо от их значений Y. Таким образом, требуется найти только a_1 и a_2.

10.2.2. Чаще всего использование одновременно двух признаков все еще не обеспечивает достаточной малости доли неверных

отнесений. В этом случае отнесение может быть улучшено, если увеличить число используемых признаков.

При трех признаках задача будет состоять в отыскании некоторой поверхности, разделяющей два эллипсоида. В более общем случае m признаков геометрическая интерпретация задачи будет такова: требуется найти такую $(m-1)$-мерную гиперповерхность в пространстве m измерений, которая лучше всего разделяет два m-мерных гиперэллипсоида.

Как и в случае двух измерений, эта геометрическая задача может быть заменена алгебраической задачей о нахождении преобразования

$$X = a_1 x_1 + a_2 x_2 + \ldots + a_m x_m, \tag{10.5}$$

превращающего два m-мерных распределения в два одномерных распределения, причем коэффициенты $a_1, a_2, \ldots, a_m$ должны быть выбраны так, чтобы получившиеся два одномерных распределения перекрывались как можно меньше.

Последнее требование можно выразить количественно так: расстояние между центрами новых распределений, т. е. $X^B - \bar{X}^A$, отнесенное к их среднему стандартному отклонению[1] $\langle s\{X\}\rangle$, должно быть максимальным. Чтобы не иметь дела с квадратными корнями, можно сформулировать условие наилучшего разделения в виде:

$$\frac{(\bar{X}^B - X^A)^2}{\langle s^2\{X\}\rangle} = \text{maximum}. \tag{10.6}$$

Можно показать, что требование (10.6) приводит к следующей системе уравнения[2] для неизвестных $a_1, a_2, \ldots, a_m$:

$$\left.\begin{aligned} S_{11} a_1 + S_{12} a_2 + \ldots + S_{1m} a_m &= d_1, \\ S_{21} a_1 + S_{22} a_2 + \ldots + S_{2m} a_m &= d_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ S_{m1} a_1 + S_{m2} a_2 + \ldots + S_{mm} a_m &= d_m \end{aligned}\right\}, \tag{10.7}$$

где

$$\left.\begin{aligned} d_j &= \bar{x}_j^B - \bar{x}_j^A, \\ S_{ij} &= \frac{1}{n_A + n_B - 2}\left[\sum\left(x_i^A - \bar{x}_i^A\right)\left(x_j^A - \bar{x}_j^A\right) + \right. \\ &\quad \left. + \sum\left(x_i^B - \bar{x}_i^B\right)\left(x_j^B - \bar{x}_j^B\right)\right] \end{aligned}\right\}. \tag{10.8}$$

[1] О стандартном отклонении см. раздел 3.3.1.

[2] Тем, кто знаком с дифференциальным исчислением, напоминаем, что если функция от многих переменных имеет максимум в какой-либо точке, то все частные производные от функции по этим переменным в этой точке равны нулю. Это и приводит к системе уравнений (10.7), причем в качестве аргументов функции рассматриваются в данном случае величины $a_1, a_2 \ldots, a_m$.

Если ввести величины

$$S_{ij}^A = \frac{1}{n_A - 1} \sum \left(x_i^A - \bar{x}_i^A \right) \left(x_j^A - \bar{x}_j^A \right),$$

$$S_{ij}^B = \frac{1}{n_B - 1} \sum \left(x_i^B - \bar{x}_i^B \right) \left(x_j^B - \bar{x}_j^B \right),$$

то S_{ij} в (10.8) можно будет записать в виде:

$$\tilde{S}_{ij} = \frac{(n_A - 1) S_{ij}^A + (n_B - 1) S_{ij}^B}{n_A + n_B - 2}.$$

При сравнительно небольшом числе признаков m систему можно решать методами, известными из элементарной алгебры, но если m велико, то приходится прибегать к помощи вычислительных машин; для решения такой задачи все типы машин имеют готовые стандартные программы.

После того как найдены коэффициенты $a_1, a_2, \ldots, a_m$, можно по формуле (10.5) вычислять значения:

$$\left.\begin{aligned} \bar{X}^A &= a_1 \bar{x}_1^A + a_2 \bar{x}_2^A + \ldots + a_m \bar{x}_m^A, \\ \bar{X}^B &= a_1 \bar{x}_1^B + a_2 \bar{x}_2^B + \ldots + a_m \bar{x}_m^B, \end{aligned}\right\} \qquad (10.9)$$

которые понадобятся для того, чтобы из формулы (10.2) найти граничное значение X_0. Для нахождения X_0 требуется также знать величины $\sigma\{X^A\}$ и $\sigma\{X^B\}$ или их выборочные оценки $s\{X^A\}$ и $s\{X^B\}$. Очевидно (ниже K есть A или B),

$$s^2\{X^K\} = \frac{1}{n_K - 1} \sum \left(X^K - \bar{X}^K \right)^2 =$$

$$= \frac{1}{n_K - 1} \sum \left[\left(a_1 x_1^K + a_2 x_2^K + \ldots + a_m x_m^K \right) - \right.$$

$$\left. - \left(a_1 \bar{x}_1^K + a_2 \bar{x}_2^K + \ldots + a_m \bar{x}_m^K \right) \right]^2 = \frac{1}{n_K - 1} \sum \left[a_1 \left(x_1^K - \right.\right.$$

$$\left.\left. - \bar{x}_1^K \right) + a_2 \left(x_2^K - \bar{x}_2^K \right) + \ldots + a_m \left(x_m^K - \bar{x}_m^K \right) \right]^2.$$

Используя формулу квадрата суммы и затем произведя перегруппировку слагаемых, можно привести последнее выражение к более удобному для вычислений виду:

$$s^2\{X^K\} = \sum_{i,j} a_i a_j S_{ij}^K. \qquad (10.10)$$

Теперь, вычислив значения $\bar{X}^A$, $\bar{X}^B$, а также

$$s\{X^A\} = \sqrt{s^2\{X^A\}}, \qquad s\{X^B\} = \sqrt{s^2\{X^B\}}$$

находим из формулы (10.2) граничное значение X_0. После этого можно составить функцию:

$$D(x_1, x_2, \ldots, x_m) = X(x_1, x_2, \ldots, x_m) - X_0 = \\ = a_1 x_1 + a_2 x_2 + \ldots + a_m x_m - X_0. \qquad (10.11)$$

Она называется *дискриминантной функцией*. Если при подстановке значений $x_1, x_2, \ldots, x_m$ какого-либо индивида получится $X \leqslant X_0$, т. е. $D \leqslant 0$, то индивида следует отнести к группе A, а если получится $X > X_0$, т. е. $D > 0$, то его следует отнести к группе B.

Когда априорные вероятности q_A и q_B обеих совокупностей неодинаковы, к X_0 надо добавить величину $2{,}3 \lg (q_A/q_B)$.

10.2.3. Чтобы определить вероятность неправильного отнесения, воспользуемся формулой (10.4). В данном случае $d\{X\} = X^B - X^A$, а величину $\sigma_A + \sigma_B = \sigma\{X^A\} + \sigma\{X^B\}$ заменим усредненной оценкой $2\sigma\{X\}$, где

$$\sigma^2\{X\} \approx s^2\{X\} = \sum_{i,j} a_i a_j S_{ij}.$$

Очевидно:

$$d\{X\} = \bar{X}^B - \bar{X}^A = \sum_i a_i \bar{x}_i^B - \sum_i a_i \bar{x}_i^A = \sum_i a_i (\bar{x}_i^B - \bar{x}_i^A) = \sum_i a_i d_i.$$

Но, согласно формуле (10.7), $d_i = \Sigma S_{ij} a_j$. Поэтому:

$$d\{X\} = \sum_i a_i \sum_j a_j S_{ij} = \sum_{i,j} a_i a_j S_{ij} = s^2\{X\}.$$

Тогда для аргумента интеграла вероятностей в формуле (10.4) получаем:

$$u_\alpha = \frac{d\{X\}}{2s\{X\}} = \frac{d\{X\}}{2\sqrt{d\{X\}}} = \frac{1}{2}\sqrt{d\{X\}} = \\ = \frac{1}{2}\sqrt{a_1 d_1 + a_2 d_2 + \ldots + a_m d_m}. \qquad (10.12)$$

Равенство $s^2\{X\} = d\{X\}$ используем также для вычисления вероятностей того, что индивид с некоторым набором значений $x_1, x_2, \ldots, x_m$ принадлежит к той или другой группе. Учитывая, что дискриминантная функция (10.11) превращает многомерное распределение в одномерное, можно воспользоваться формулой (10.3) для соответствующих вероятностей. В данном случае следует произвести замену:

$$d \to d\{X\}, \quad \sigma^2 \to \sigma^2\{X\} \approx s^2\{X\}, \quad x \to X - X_0,$$

и тогда аргумент $l = \left(\frac{d}{\sigma^2}\right) x$ логистической функции примет простой

вид: $l = a_1 x_1 + a_2 x_2 + \ldots + a_m x_m - X_0 = D(x_1, x_2, \ldots, x_m)$.

Пример 10.1. В табл. 10.2 приведены распределения для двух чистых линий пшеницы ($n_A = 35$ и $n_B = 56$) по двум признакам — индексу плотности колоса и числу зерен в колосе; частоты, относящиеся к разным линиям, набраны различным шрифтом. В нижней строке и крайнем правом столбце выписаны частные суммы частот.

ТАБЛИЦА 10.2

y	$\tilde{y}$ \ $\tilde{x}$	15	16	17	18	19	20	21	22	
—4	18					1				1
—3	21			1	1		1	1	1	4/1
—2	24		1	3	6	3	1	3	2	13/6
—1	27	1	2	4	3	1	4	4	3	10/12
0	30	1	1	2	1	5	7	4	2	5/18
1	33		1	1	1	3	3	1		2/8
2	36				1	2	2	1		6
3	39					1	1	1		3
4	42				1					1
5	45						1			1
		2	5	11	11	4	1	1	8	\ 35
					3	12	19	14	8	56 \

Если изобразить распределения для обеих линий отдельно по каждому из двух признаков (рис. 10.4 *а* и *б*), то окажется, что эти распределения значительно перекрываются. Поэтому ни один из этих признаков в отдельности не пригоден для отнесения. Между тем совместное использование обоих признаков позволяет произвести отнесение с довольно малой ошибкой.

Даже из простого обозрения табл. 10.2 видно, что разделение будет достаточно хорошим, если провести граничную линию приблизительно по диагонали, отделяющей левую верхнюю часть от правой нижней части табл. 10.2. Однако, пользуясь изложенным выше методом, можно получить более точное решение задачи.

Для упрощения расчетов полезно применить кодирование координат; в данном случае удобно принять:

$$x = \tilde{x} - 19, \quad y = \left(\tilde{y} - 30\right) : 3.$$

Теперь вычисляем величины, входящие в систему уравнений (10.7). Сначала найдем d_x и d_y:

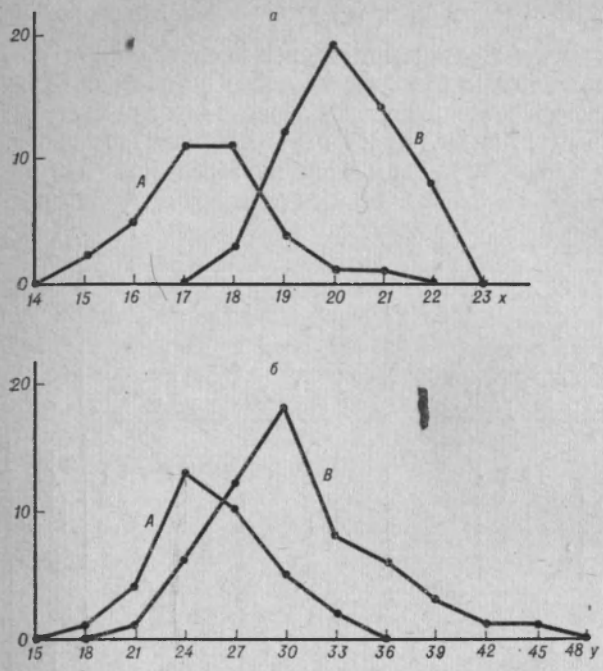


Рис. 10.4. Перекрывание двух двумерных совокупностей по отдельным признакам.

$$\bar{x}^A = [2(-4) + 5(-3) + 11(-2) + 11(-1) + 4 \cdot 0 + 1 \cdot 1 + 1 \cdot 2] : 35 = -53/35 = -1{,}5143,$$

$$\bar{x}^B = [3(-1) + 12 \cdot 0 + 19 \cdot 1 + 14 \cdot 2 + 8 \cdot 3] : 56 = 68/56 = 1{,}2143,$$

$$\bar{y}^A = [1(-4) + 4(-3) + 13(-2) + 10(-1) + 5 \cdot 0 + 2 \cdot 1] : 35 = -50/35 = -1{,}4286,$$

$$\bar{y}^B = [1(-3) + 6(-2) + 12(-1) + 18 \cdot 0 + 8 \cdot 1 + 6 \cdot 2 + 3 \cdot 3 + 1 \cdot 4 + 1 \cdot 5] : 56 = 11/56 = 0{,}1964,$$

так что

$$d_x = -1{,}5143 - 1{,}2143 = -2{,}7286,$$

$$d_y = -1{,}4286 - 0{,}1964 = -1{,}6250.$$

Далее находим:

$$\sum (x^A)^2 = 137, \quad \sum (y^A)^2 = 116, \quad \sum x^A y^A = 46;$$

$$\sum (x^B)^2 = 150; \quad \sum (y^B)^2 = 145, \quad \sum x^B y^B = -36.$$

Поскольку

$$(n_K - 1) S_{ij}^K = \sum (x_i^K - \bar{x}_i^K)(x_j^K - \bar{x}_j^K) = \sum x_i^K x_j^K - \frac{\sum x_i^K \sum x_j^K}{n_K},$$

то

$$34\, S_{xx}^A = 137 - \frac{53^2}{35} = 56{,}7, \qquad 55\, S_{xx}^B = 150 - \frac{68^2}{56} = 67{,}5,$$

$$34\, S_{yy}^A = 116 - \frac{50^2}{35} = 44{,}5, \qquad 55\, S_{yy}^B = 145 - \frac{11^2}{56} = 142{,}2$$

$$34\, S_{xy}^A = 46 - \frac{53 \cdot 50}{35} = -29{,}7, \qquad 55\, S_{xy}^B = -36 - \frac{68 \cdot 11}{56} = -48{,}1.$$

Поэтому

$$S_{xx} = \frac{56{,}7 + 67{,}5}{35 + 56 - 2} = 1{,}3954, \qquad S_{yy} = \frac{44{,}5 + 142{,}2}{35 + 56 - 2} = 2{,}0976,$$

$$S_{xy} = \frac{-29{,}7 - 48{,}1}{35 + 56 - 2} = -0{,}8741.$$

Следовательно, система уравнений (10.7) имеет вид:

$$1{,}3954\, a_x - 0{,}8741\, a_y = -2{,}7286,$$
$$-0{,}8741\, a_x + 2{,}0976\, a_y = -1{,}6250,$$

откуда: $a_x = -3{,}3029$, $a_y = -2{,}1511$. Таким образом, дискриминантная функция будет:

$$D(x, y) = -3{,}3029\, x - 2{,}1511\, y - X_0,$$

где граничное значение X_0 еще предстоит найти.

Поскольку результат применения дискриминантной функции не зависит от изменения масштаба, можно в данном случае упростить ее, разделив все члены на $-3{,}3029$. Тогда дискриминантная функция примет вид:

$$D^*(x, y) = x + 0{,}6513y - X_0^*.$$

В рассмотренном двумерном случае можно сразу записать общее решение для коэффициента $a = a_y$ (при нормировке $a_x = 1$) в виде простой формулы:

$$a = \frac{d_y S_{xx} - d_x S_{xy}}{d_x S_{yy} - d_y S_{xy}}. \qquad (10.13)$$

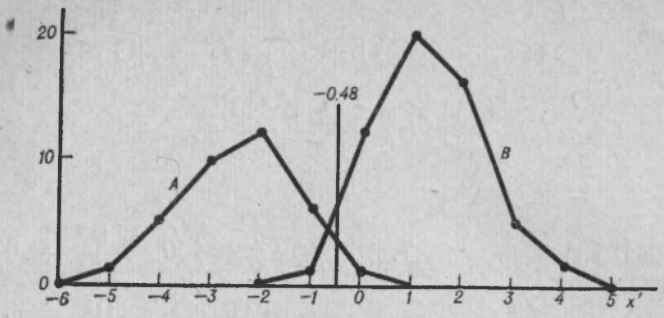


Рис. 10.5. Перекрывание двух двумерных совокупностей после применения дискриминантной функции.

Подставляя наши значения, находим сразу:

$$a = \frac{-1{,}6250 \cdot 1{,}3954 - (-2{,}7286)(-0{,}8741)}{-2{,}7286 \cdot 2{,}0976 - (-1{,}6250)(-0{,}8741)} = 0{,}6513.$$

Это значение a мы пока используем для вычисления новых координат $x' = x + ay$ отдельных точек. Например, для точки с координатами $\widetilde{x}^A = 20$, $\widetilde{y}^A = 21$, т. е. $x^A - 1$, $y^A = -3$, имеем:

$$x'_A = 1 + 0{,}6513(-3) = -0{,}954;$$

аналогично вычисляют новые координаты x' для всех остальных клеток табл. 10.2. Полученные значения x' записаны в табл. 10.3 в порядке их расположения в табл. 10.2 (по строкам, как при чтении); данные для разных генетических линий записаны отдельно. Они сгруппированы в табл. 10.4, причем в строку (−5) записаны значения, лежащие между −5,5 и −4,5, в строку (−4) — значения, лежащие между −4,5 и −3,5, и т. д.

Распределения из табл. 10.4 изображены на рис. 10.5. Как видно, перекрывание здесь гораздо меньше.

Граничное значение x_0' можно найти по формуле (10.2), для чего надо вычислить $\bar{x}'_A$, $\bar{x}'_B$, s'_A и s'_B. Эти значения можно было бы получить, исходя из распределений, записанных в табл. 10.4. Однако в этом нет надобности. Величины $\bar{x}'_A$ и $\bar{x}'_B$ можно получить просто по формуле $\bar{x}' = \bar{x} + a\bar{y}$:

$$\bar{x}'_A = -1{,}5143 + 0{,}6513(-1{,}4286) = -2{,}4447,$$

$$\bar{x}'_B = 1{,}2143 + 0{,}6513 \cdot 0{,}1964 = 1{,}3422.$$

Далее, по формуле (10.10) имеем:

$$(s^2)'_K = S^K_{xx} + 2aS^K_{xy} + a^2 S^K_{yy},$$

ТАБЛИЦА 10.3

A		B	
x'	n	x'	n
−2,605	1	1,046	1
−3,954	1	−0,303	1
−2,954	1	0,697	3
−0,954	1	1,697	2
+0,046	1	−0,651	1
−4,303	1	0,349	4
−3,303	3	1,349	4
−2,303	6	2,349	3
−1,303	3	0,000	5
−4,651	1	1,000	7
−3,651	2	2,000	4
−2,651	4	3,000	2
−1,651	3	−0,349	1
−4,000	1	0,651	3
−3,000	1	1,651	3
−2,000	2	2,651	1
−1,000	1	0,303	1
−2,349	1	1,303	2
−1,349	1	2,303	2
		2,303	1
		1,954	1
	35	2,954	1
		3,954	1
		1,605	1
		4,256	1
			56

ТАБЛИЦА 10.4

x'	A	B
−5	1	
−4	5	
−3	10	
−2	12	
−1	6	1
0	1	12
1		20
2		16
3		5
4		2
	35	56

так что

$$\left(s_A^2\right)' = [56{,}7 + 2 \cdot 0{,}6513\,(-29{,}7) + 0{,}6513^2 \cdot 44{,}5] : 34 = 1{,}0850,$$

$$\left(s_B^2\right)' = [67{,}5 + 2 \cdot 0{,}6513\,(-48{,}1) + 0{,}6513^2 \cdot 142{,}2] : 55 = 1{,}1848$$

$$s_A' = \sqrt{1{,}0850} = 1{,}0416, \quad s_B' = \sqrt{1{,}1848} = 1{,}0885.$$

Теперь по формуле (10.2) находим граничное значение на координатной оси x':

$$x_0' = \frac{1{,}041 \cdot 1{,}3422 + 1{,}0885\,(-2{,}4447)}{1{,}0416 + 1{,}0885} = -0{,}5929.$$

Так как

$$x_0' = x_0 + ay_0 = \tilde{x}_0 - 19 + 0{,}6513\,(\tilde{y}_0 - 30) : 3 =$$

$$= \tilde{x}_0 + 0{,}2171\,\tilde{y}_0 - 25{,}513,$$

то равенство $x_0' = -0{,}5929$ есть по существу уравнение граничной линии в старых координатах $\tilde{x}$, $\tilde{y}$; а именно:

$$\tilde{x}_0 + 0{,}2171\,\tilde{y}_0 - 25{,}513 = -0{,}5929,$$

т. е. уравнение граничной линии будет:

$$\tilde{x}_0 + 0{,}2171\,\tilde{y}_0 - 24{,}920 = 0.$$

Следовательно, дискриминантная функция имеет вид:

$$D^*(\tilde{x},\ \tilde{y}) = \tilde{x} + 0{,}2171\,\tilde{y} - 24{,}920.$$

Пользуясь формулой (10.3), мы можем определить вероятность неправильного отнесения:

$$\alpha_A = \alpha_B = \Phi\left(-\frac{\bar{x}_B' - \bar{x}_A'}{s_A' + s_B'}\right) = \Phi\left(-\frac{1{,}3422 - (-2{,}4447)}{1{,}0416 + 1{,}0885}\right) =$$

$$= \Phi(-1{,}7778) \approx 0{,}038 = 3{,}8\,\%$$

(последнее число найдено по табл. III Приложений); это можно считать вполне удовлетворительным.

Заметим, что во всех расчетах нам не понадобилось ни распределение значений x' из табл. 10.4, ни сами эти значения, записанные в табл. 10.3. Мы их использовали для чисто иллюстративных целей — для построения графика на рис. 10.5. Поэтому при практическом решении задачи классификации новые значения x' можно не вычислять.

ТАБЛИЦА 10.5

	1	2	3	4	5	6	$\bar{x}$
1	193	—108	94	—49	—19	—927	—2,57
2		322	—172	90	223	1392	4,93
3			271	—121	—136	—1214	—4,07
4	$(n_A - 1)\,S_{ij}^A$			149	—11	816	3,54
5	$(n_A = 100)$				2119	588	7,39
6						9646	33,71
1	276	—253	148	—98	—59	—1449	—4,85
2		624	—324	298	—36	2558	9,97
3			452	—230	45	—1908	—7,87
4	$(n_B - 1)\,S_{ij}^B$			336	66	1531	6,09
5	$(n_B = 93)$				5081	1688	12,59
6						15544	60,05

Пример 10.2. В табл. 10.5 даны значения $(n_K - 1)\, S_{ij}^K$ и $\bar{x}_i^K$ для признаков, используемых в диагностике заболеваний сердца по баллистокардиограммам; смысл отдельных признаков нас сейчас не интересует. Найдем дискриминантную функцию по этим данным.

ТАБЛИЦА 10.6

	1	2	3	4	5	6	d_i	$a_i \cdot 10^2$
1	469	−361	242	−147	−78	−2376	−2,28	−0,081557
2		946	−496	388	187	3950	5,04	57,713
3			723	−351	−91	−3122	−3,80	−34,957
4				485	55	2347	2,55	0,61922
5					7200	2276	5,20	10,073
6	$(n_A + n_B - 2)\, S_{ij}$					25190	26,34	5,6143

В табл. 10.6 мы записали значения $(n_A + n_B - 2)S_{ij} = (n_A - 1)S_{ij}^A + (n_B - 1)S_{ij}^B$ и $d_i = \bar{x}_i^B - \bar{x}_i^A$. Эти значения используем для нахождения величин a_i из системы уравнений (10.7); они также выписаны в табл. 10.6. Теперь по формулам (10.9) и (10.10) вычисляем:

$$\bar{X}^A = 6{,}9290, \qquad \bar{X}^B = 13{,}1863,$$

$$s^2\{X^A\} = 3{,}7671, \qquad s^2\{X^B\} = 0{,}75947,$$

$$s\{X^A\} = 1{,}9409, \qquad s\{X^B\} = 0{,}87147,$$

так что формула (10.2) дает:

$$X_0 = \frac{1{,}9409 \cdot 13{,}1863 + 0{,}87147 \cdot 6{,}9290}{1{,}9490 + 0{,}87147} = 11{,}247.$$

Следовательно, дискриминантная функция для этой задачи имеет вид:

$$D(x_1, x_2, x_3, x_4, x_5, x_6) = -0{,}081557x_1 + 57{,}713x_2 - 34{,}957x_3 + \\ + 0{,}61922x_4 + 10{,}073x_5 + 5{,}6143x_6 - 1124{,}7$$

(все коэффициенты и X_0 увеличены для удобства в 100 раз).

Теперь оценим вероятность ошибочных отнесений, пользуясь формулой (10.1). В данном случае:

$$d\{X\} = a_1 d_1 + \ldots + a_6 d_6 = 6{,}257,$$

так что по формуле (10.1) получаем:

$$u_\alpha = -\frac{1}{2}\sqrt{6{,}257} \approx -1{,}25$$

и

$$\alpha = \Phi(u_\alpha) \approx 0{,}106 = 10{,}6\%.$$

Найдем еще, к какой группе и с какой вероятностью следует отнести какого-либо индивида — например, имеющего значения признаков:

$$x_1^* = -3{,}01; \quad x_2^* = 5{,}72; \quad x_3^* = -6{,}11; \quad x_4^* = 4{,}20;$$

$$x_5^* = 10{,}24; \quad x_6^* = 48{,}31.$$

Подставляя эти значения в дискриминантную функцию, получаем (конечно, без множителя 100) значение $-2{,}038$. Так как $D(X) < 0$, то индивида следует отнести к группе A. Далее, пользуясь табл. 10.1, находим: $P_B = 0{,}115$, $P_A = 0{,}885$.

§ 10.3. Отнесение индивидов к одной из нескольких групп

10.3.1. Весьма часто число групп, к одной из которых может в принципе принадлежать индивид, больше двух. Примером является отнесение особи насекомого к одному из нескольких видов, или дифференциальная диагностика нескольких типов шизофрении, или выбор наиболее подходящего (для данного больного) из нескольких способов лечения и т. д.

Очевидно, если имеется g групп в многомерном пространстве, то между ними можно провести $h = g(g-1)/2$ разделяющих «плоскостей» (многообразия, разделяющие многомерные области, называют гиперповерхностями, а в линейном случае — гиперплоскостями). При увеличении числа групп число разделяющих гиперплоскостей быстро возрастает. Поэтому, если бы нужно было для каждой гиперповерхности находить заново дискриминантную функцию, то потребовалось бы решить соответствующее число систем уравнений (например, в случае $g = 6$ групп пришлось бы решить $h = 6(6-1)/2 = 15$ систем уравнений). К счастью, такой необходимости нет: можно доказать простыми преобразованиями, что если найдены дискриминантные функции для различения групп A и B и групп A и C, то дискриминантную функцию различения групп B и C находят просто:

$$D_{BC}(x) = D_{AC}(x) - D_{AB}(x).$$

Значит, в случае трех групп надо найти непосредственно (т. е. решая систему уравнений) только две дискриминантные функции, третью же можно выразить просто в виде разности первых двух. Аналогичным образом можно показать, что в случае g групп требуется решить только $g-1$ систем уравнений. При отнесении индивида к одной из групп нужно подставлять его значения x тоже только в $g-1$ дискриминантных функций.

Если имеется g групп A, B, C, ..., M, N, то в качестве набора «нужных» $g-1$ дискриминантных функций можно принять D_{AN}, D_{BN}, ..., D_{MN}, т. е. сравнивая все время только с одной из групп (в данном случае с группой N) все остальные группы. Индивида со значением признаков x_1', x_2'..., x_m' относят к той группе K, для которой значение $D_{KN}(x_1', \ldots, x_m')$ оказалось наибольшим. Заметим при этом, что всегда $D_{NN}(x_1, \ldots, x_m) = 0$, так как равны нулю все $d_i = \bar{x}_i^N - \bar{x}_i^N$; поэтому D_{NN} может оказаться наибольшим из всех D_{KN} только тогда, когда все остальные D_{KN} (при $K \neq N$) отрицательны. В этом случае индивида относят к группе N.

Пример 10.3. В табл. 10.7 и 10.8 приведены данные, касающиеся разделения пяти подвидов насекомых *Colicoides variipennis* по 4 признакам, а именно матрица S_{ij} и значения $\bar{x}_i^K$.

ТАБЛИЦА 10.7

S_{ij}

0,03748	—0,00067	0,00844	0,00474
	0,00778	0,00333	0,00385
		1,06044	—0,01541
			0,59385

ТАБЛИЦА 10.8

K	$\bar{x}_1^K$	$\bar{x}_2^K$	$\bar{x}_3^K$	$\bar{x}_4^K$	n_K
I	3,0583	1,7639	14,4444	0,1111	36
II	2,5286	1,5357	11,9286	2,7500	28
III	2,2345	1,2552	12,1724	0,8276	29
IV	2,4037	1,6037	12,2222	1,7778	27
V	2,2550	1,6300	14,0000	0,6500	20

По этим данным были получены коэффициенты четырех дискриминантных функций $D_{KV}(x)$, для чего были решены четыре системы уравнений типа (10.7); эти коэффициенты записаны в табл. 10.9.

ТАБЛИЦА 10.9

K	a_1^K	a_2^K	a_3^K	a_4^K	x_K^0	$D_{K,V}(x^*)$
I	21,9	19,6	0,2	—1,2	93,3	—12,4
II	7,1	—12,4	—1,4	3,5	—19,0	4,9
III	—1,1	—47,9	—1,5	0,6	—91,6	6,1
IV	4,1	— 3,3	—1,6	1,8	—15,4	1,1

Чтобы отнести к одной из этих групп индивида со значениями $x_1^* = 2{,}46$, $x_2^* = 1{,}31$, $x_3^* = 13{,}82$, $x_4^* = 1{,}15$, мы вычисляем по формуле (10.11) четыре значения $D_{KV}(x)$; они записаны в последнем столбце табл. 10.9. Наибольшим оказалось $D_{III\,V}(x) = 6{,}1$, что позволяет отнести данного индивида к группе III.

10.3.2. Как уже указывалось, предполагается, что ковариационные матрицы всех групп одинаковы. Поскольку на практике используются обычно выборки, то ковариационные матрицы для разных групп всегда оказываются различными. Если считать, что эти различия обусловлены только выборочным варьированием, то естественно принять все S_{ij}^K за разные выборочные оценки одной и той же матрицы Σ, наилучшей оценкой которой будет тогда

$$S_{ij} = \frac{1}{n-g}\left[(n_A - 1)\,S_{ij}^A + \ldots + (n_G - 1)\,S_{ij}^G\right]. \qquad (10.14)$$

Эта матрица затем используется для нахождения дискриминантных функций из системы уравнений (10.7). Разумеется, значимость различий между ковариационными матрицами можно проверить при помощи специального статистического критерия, но применение его оказывается чаще всего очень громоздким.

10.3.3. Если число групп велико (скажем, больше 10), то разделение их при помощи $g - 1$ дискриминантных функций оказывается весьма громоздким. Поэтому иногда применяют «*метод наилучших дискриминаторов*», несколько упрощающий решение задачи и, главное, приводящий к очень наглядному представлению семейства совокупностей. Идея этого метода состоит в следующем. В задаче с двумя совокупностями дискриминантная функция получается из условия:

$$Q_2 = \frac{(\bar{X}^A - \bar{X}^B)^2}{\langle s^2\{X\}\rangle} = \text{maximum}, \qquad (10.6)$$

т. е. из условия, чтобы нормированное «расстояние» между двумя преобразованными совокупностями оказалось наибольшим. В случае многих совокупностей это условие можно обобщить следующим образом:

$$Q_g = \frac{\sum_{K=1}^{G}(\bar{X}^K - \bar{X})^2}{\langle s^2\{X^K\}\rangle} = \text{maximum}, \qquad (10.15)$$

где $\bar{X}$ есть среднее по всем группам, а g обозначает число групп. Числитель характеризует рассеяние м е ж д у совокупностями, а знаменатель — рассеяние в н у т р и совокупностей.

Можно показать, что если размерность пространства (т. е. в данном случае число признаков) достаточно велика (точнее, долж-

но быть $m \geqslant g-1$), то величина Q_g, как функция от координат $x_1, x_2, \ldots, x_m$, имеет не один, а $g-1$ относительных максимумов, т. е. существует $g-1$ направлений, при проектировании на которые разделение всего набора из g групп получается лучше, чем при проектировании на соседние с ними направления.

Обозначим $max\ Q_g = \psi^{(l)}$ ($l = 1, 2, \ldots, g-1$). Каждому значению $\psi^{(l)}$ соответствует определенное направление проектирования, т. е. определенный набор $a_i^{(l)}$ коэффициентов в преобразовании:

$$X^{(l)} = \sum_{i=1}^{m} a_i^{(l)} x_i. \tag{10.16}$$

Величины $\bar{X}^{(l)}$, задаваемые наборами $a_i^{(l)}$, называют *дискриминаторами*. Нахождение наборов $a_i^{(l)}$ в общем сводится к следующему. Сначала вычисляют так называемые матрицы межгруппового рассеяния B_{ij} и внутригруппового рассеяния W_{ij}, равные:

$$B_{ij} = \frac{1}{n} \sum_{K=1}^{G} n_K \left(\bar{x}_i^K - \bar{x}_i\right)\left(\bar{x}_j^K - \bar{x}_j\right), \tag{10.17}$$

$$W_{ij} = \frac{1}{n} \sum_{K=1}^{G} \sum_{\gamma} \left(x_{i\gamma}^K - \bar{x}_i^K\right)\left(x_{j\gamma}^K - \bar{x}_j^K\right), \tag{10.18}$$

где γ — номера индивидов. Затем находят $g-1$ значений $\psi^{(l)}$ как корни уравнения:

$$det \parallel B_{ij} - \psi W_{ij} \parallel = 0 \tag{10.19}$$

($det \parallel M \parallel$ есть знак так называемого определителя матрицы $\|M\|$). И, наконец, для каждого $\psi^{(l)}$ находят набор $a_i^{(l)}$ из системы уравнений

$$\sum_{i=1}^{m} \left(B_{ij} - \varphi^{(l)} W_{ij}\right) a_i^{(l)} = 0, \quad j = 1, 2, \ldots, m. \tag{10.20}$$

Решение уравнений (10.19) и (10.20) требует применения ЭВМ. Для всех типов ЭВМ имеются стандартные программы, решающие эти задачи («нахождение собственных значений и собственных векторов матриц»).

Найдя все $g-1$ дискриминаторов, можно достаточно хорошо разделить все g групп, поскольку различные дискриминаторы в известном смысле дополняют друг друга: группы, плохо разделяемые каким-либо одним дискриминатором, зато хорошо разделяются каким-либо другим.

Впрочем, в общем разделение получается все же хуже, чем при использовании $g - 1$ попарных дискриминантных функций, поэтому может показаться, что метод дискриминаторов вообще мало полезен. Но это не так. Если в m-мерном пространстве группы сравнительно далеки друг от друга, то уже 2—3 дискриминаторов достаточно, чтобы разделить все группы — при условии, что использованы наилучшие дискриминаторы. Наилучшими дискриминаторами будут, очевидно, те $a_i^{(l)}$, которые отвечают наибольшим $\psi^{(l)}$. Так как значения $\psi^{(l)}$ получаются как корни уравнения (10.19), то всегда можно выбрать наибольшие из них (обозначим их ψ^* и ψ^{**}, а соответствующие им дискриминаторы a_i^*, a_i^{**} и X^*, X^{**}). Теперь поступаем следующим образом. Используя найденные a_i^* и a_i^{**}, вычисляем по формуле (10.16) для каждого индивида два значения X^* и X^{**} и считаем эти значения координатами на двумерной плоскости. Таким образом, каждый индивид изобразится точкой на плоскости, причем индивиды из одной и той же группы будут располагаться вокруг общего центра этой группы, и разные группы будут, вообще говоря, «разнесены» одна от другой. По существу мы получаем как бы проекцию семейства m-мерных эллипсоидов на двумерную плоскость, причем наилучшую проекцию (в отношении разделения групп), поскольку использованы наилучшие дискриминаторы. Очевидно, проецирование производится на ту плоскость, которая задается двумя направлениями a_i^* и a_i^{**} в m-мерном пространстве. Однако диаграмма (X^*, X^{**}) не совпадает в точности с указанной двумерной проекцией: дело в том, что на нашей диаграмме координатные оси X^* и X^{**} составляют привычный прямой угол (т. е. используется наиболее удобная прямоугольная декартова система координат), в то время как направления a_i^* и a_i^{**} пересекаются не обязательно под прямым углом.

Ясно, что использование двух наилучших дискриминаторов дает лучшее разделение семейства групп, чем использование каких-либо двух дискриминантных функций: ведь каждая дискриминантная функция хороша для своей пары групп и в общем непригодна для остальных групп, в то время как каждый дискриминатор строится сразу для всего семейства групп.

Таким образом, метод наилучших дискриминаторов позволяет получить очень наглядное изображение сразу для большого семейства совокупностей, причем с наилучшим разделением. Он особенно удобен, когда имеется сравнительно большое число не слишком похожих групп. Чаще всего это бывает в таксономических задачах. При использовании описанного метода отнесение новых индивидов к одной из групп производится графически: вычисляют для него значения X^* и X^{**}, наносят изображающую точку этого индивида на диаграмму (X^*, X^*) и смотрят, в какой из «ареалов» эта точка попадает.

Пример 10.4. Применим метод наилучших дискриминаторов к данным, описанным в примере 10.3. Матрицей внутригруппового рассеяния W_{ij} здесь будет матрица S_{ij} из табл. 10.7. Далее используя значения $\overline{x}^K$ из табл. 10.8, получаем по формуле (10 17) матрицу межгруппового рассеяния B_{ij} (табл. 10.10).

ТАБЛИЦА 10.10

0,1036	0,0424	0,2219	—0,1103
	0,0311	0,1367	—0,0417
		1,1669	—0,8382
B_{ij}			0,9214

Теперь из уравнения (10.19) получаются корни $\psi^{(i)}$, записанные здесь в порядке убывания:

$$\psi^{(1)} = 7{,}061, \quad \psi^{(2)} = 1{,}621, \quad \psi^{(3)} = 0{,}801, \quad \psi^{(4)} = 0{,}048.$$

Им отвечают в соответствии с уравнением (10.20) коэффициенты дискриминаторов $a_i^{(l)}$:

$a_i^{(1)}$	$a_i^{(2)}$	$a_i^{(3)}$	$a_i^{(4)}$
1,227	—0,248	0,946	0,052
3,369	—1,936	1,327	—0,942
0,109	0,175	—0,053	0,259
—0,149	—0,495	0,011	0,205

(все эти числа получены, разумеется, при помощи ЭВМ).

Используя первые два дискриминатора $a_i^{(1)} = a_i^*$ и $a_i^{(2)} = a_i^{**}$, строим график в осях X^*, X^{**} (рис. 10.6). Все группы разделились довольно хорошо, за исключением групп II и IV. В соответствии со сказанным выше можно было бы надеяться, что эти две группы разделятся какой-нибудь другой парой дискриминаторов. Однако в данном случае это не так, поскольку указанные две группы сильно перекрываются и в четырехмерном пространстве: даже специально для этой пары групп построенная дискриминантная функция $D_{II,\,IV}(x_1, x_2, x_3, x_4)$ обеспечивает отнесение индивидов в одну из этих групп с вероятностью ошибки более 31%.

Для рассматривавшегося выше индивида $x_1' = 2{,}46$, $x_2' = 1{,}31$, $x_3' = 13{,}82$, $x_4' = 1{,}15$ получаем $X^* = 3{,}77$, $X^{**} = -1{,}30$. Нанеся эту точку на график (на рис. 10.6 она отмечена треугольником), мы видим, что индивид должен быть отнесен к группе III (этот же результат был получен раньше при помощи набора четырех дискриминантных функций).

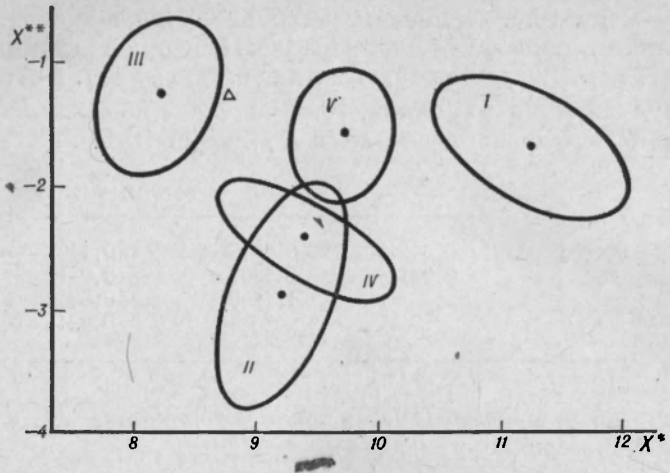


Рис. 10.6. Разграничение нескольких многомерных совокупностей методом наилучших дискриминаторов.

§ 10.4. Учет влияния выборочных вариаций

10.4.1. При решении практических задач дискриминантного анализа почти всегда имеют дело не с генеральными совокупностями, а с выборками из них. Но, как известно, средние значения, дисперсии и ковариации в отдельных выборках всегда в той или иной мере отличаются от соответствующих параметров генеральных совокупностей. Поскольку при нахождении коэффициентов дискриминантной функции использовались выборочные значения (оценки) параметров, то ясно, что эта дискриминантная функция будет как бы приспособлена к «своим» выборкам, т. е. она будет лучше всего разделять именно эти выборки. Применение же дискриминантной функции $D(x)$, полученной по выборкам, к новым индивидам из соответствующих генеральных совокупностей, т. е. к индивидам, не вошедшим в использовавшиеся для построения $D(x)$ выборки, неизбежно будет сопровождаться большим числом неверных отнесений. Если учесть, что, согласно формулам (10.1) и (10.12),

$$\alpha = \Phi(-u_\alpha), \quad u_\alpha = \frac{1}{2}\sqrt{d\{X\}},$$

то можно сказать, что выборочная оценка $d\{X\}$ («расстояния» между совокупностями) является как бы завышенной. Чтобы устранить это кажущееся завышение, применяют формулу:

$$\tilde{d}\{X\} = \frac{n_A + n_B - m - 3}{n_A + n_B - 2}\left[d\{X\} - m\left(\frac{1}{n_A} + \frac{1}{n_B}\right)\right]. \qquad (10.21)$$

Так, в примере 10.2 (§ 10.2) мы получили $d\{X\} = 6{,}257$ и $\alpha = 10{,}6\%$ при $m = 6$, $n_A = 100$, $n_B = 93$. Тогда:

$$\tilde{d}\{X\} = \frac{184}{191}\left[6{,}257 - 6\left(\frac{1}{100} + \frac{1}{93}\right)\right] = 5{,}908$$

и $\alpha = 11{,}2\%$.

В следующем далее разделе используются понятия критерия Стьюдента (см. раздел 4.2.1) и F-критерия Фишера (см. раздел 4.6.1).

10.4.2. Использование в дискриминантном анализе выборок вместо генеральных совокупностей выдвигает еще такую проблему: нахождение границы между двумя совокупностями, выполняемое на основании изучения выборок, может иметь смысл только тогда, когда различия между двумя выборками статистически значимы. Последнее может быть проверено при помощи критерия Стьюдента:

$$t = \frac{d}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{d}{s}\sqrt{\frac{n_A\, n_B}{n_A + n_B}}, \qquad (10.22)$$

причем в нашем случае $d = d\{X\}$ и $s = s\{X\} = \sqrt{s^2\{X\}}$. Но (см. раздел 10.2.3) $s^2\{X\} = d\{X\}$, поэтому

$$t = \sqrt{d\{X\}\frac{n_A\, n_B}{n_A + n_B}}. \qquad (10.23)$$

Согласно формуле (10.12), $\sqrt{d\{X\}} = 2u_\alpha$, и тогда

$$t = 2u_\alpha\sqrt{\frac{n_A\, n_B}{n_A + n_B}}.$$

Если объемы выборок из A и B не очень различаются, так что в первом приближении можно считать $n_A \approx n_B \approx N/2$ (где N — суммарный объем обеих выборок), то

$$t \approx u_\alpha\sqrt{N}.$$

Поскольку в подобных задачах N никогда не бывает очень малым, то t всегда по крайней мере в несколько раз больше, чем u_α. Но, как уже отмечалось выше, данный метод отнесения имеет практическое значение только при достаточно малых значениях α. Это значит, что в практически интересных случаях u_α не меньше единицы, так что t равно по крайней мере нескольким единицам. Но тогда значимость различия между выборками весьма высока.

Существует более точный и строгий критерий значимости различий между группами: величина

$$F = \frac{n_A n_B}{n_A + n_B} \cdot \frac{n_A + n_B - m - 1}{n_A + n_B - 2} \cdot \frac{d\{X\}}{m} \qquad (10.24)$$

должна превышать табличное значение $F_\beta(f_1, f_2)$, где через β здесь обозначен уровень значимости, а

$$f_1 = m; \quad f_2 = n_A + n_B - m - 1.$$

Так, для данных из примера 10.2 имеем:

$$F = \frac{186}{191} \cdot \frac{100 \cdot 93}{193} \cdot \frac{6{,}257}{6} = 48{,}94,$$

$$f_1 = 6, \quad f_2 = 186,$$

в то время как $F_{0,01}(6; 186) \approx 2{,}91$. По формуле (10.23) получаем $t \approx 17{,}4$, что также указывает на вполне значимое различие между группами.

10.4.3. Нетрудно показать, что при добавлении новых признаков $d\{X\}$ всегда увеличивается. Но для вероятности неверных отнесений реальное значение имеет не $d\{X\}$, а величина $\tilde{d}\{X\}$, определяемая формулой (10.21). Между тем из этой формулы следует, что увеличение m само по себе уменьшает $\tilde{d}\{X\}$. Поэтому если возрастание $d\{X\}$ недостаточно велико, то $\tilde{d}\{X\}$ может даже уменьшиться при добавлении новых признаков. И, наоборот, исключение некоторых признаков, дающих очень малый вклад в $d\{X\}$, может даже увеличить $\tilde{d}\{X\}$— за счет уменьшения m в формуле (10.21). Причину этого петрудно понять: при использовании выборочных данных каждый признак вносит в диагностическую систему не только дополнительную возможность различения двух совокупностей, но и дополнительный статистический «шум». Если «разрешающая способность» какого-либо признака невелика, то выигрыш от включения его в систему перекрывается вредом от добавляемого им «шума».

Введем обозначения:

$$\omega = d_m\{X\} - d_{m-1}\{X\}, \quad \tilde{\omega} = \tilde{d}_m\{X\} - \tilde{d}_{m-1}\{X\},$$

где индексы m и $m-1$ показывают число использованных признаков. Из формулы (10.21) имеем после несложных преобразований:

$$(n_A + n_B - 2)\,\tilde{\omega} = (n_A + n_B - m - 2)\,\omega - d_m\{X\} -$$

$$- (n_A + n_B - 2m - 2)(n_A + n_B)/n_A n_B .$$

Мы ищем условия, при которых исключение одного признака не увеличивает $\widetilde{d}\{X\}$, т. е. при которых $\widetilde{\omega} \leqslant 0$, Очевидно, это будет при

$$\omega \leqslant \omega^* = \frac{d_m\{X\} \dotplus (n_A \dotplus n_B - 2m - 2)/(n_A \dotplus n_B)\ n_A\ n_B}{n_A + n_B - m - 2}. \quad (10.25)$$

В то же время вклад каждого i-го признака, т. е. величину ω_i можно оценить по приближенной формуле[1]:

$$\omega_i \approx a_i^2 S_{ii}, \quad (10.26)$$

где a_i и S_{ii} — отвечающие данному признаку коэффициент дискриминантной функции и диагональный элемент ковариационной матрицы. Следовательно, i-й признак можно исключать, если ω_i из формулы (10.26) меньше, чем ω^* из (10.25). Это условие $\omega_i \leqslant \omega^*$, т. е. $a_i^2 S_{ii} \leqslant \omega^*$, можно переписать в таком виде:

$$|a_i| \leqslant \sqrt{\frac{\omega^*}{S_{ii}}} \equiv a_i^*. \quad (10.27)$$

Пример 10.5. В примере 10.2. (§ 10.2) мы имели: $m = 6$, $n_A = 100, n_B = 93$, $d\{X\} = 6{,}257$. Поэтому:

$$\omega^* = \frac{6{,}257 + \dfrac{(100 + 93 - 12 - 2)(100 + 93)}{100 \cdot 93}}{100 + 93 - 2} = 0{,}052.$$

Тогда для использованных признаков получаем по формуле (10.27) значения a_i^*, записанные в табл. 10.11 (значения a_i и S_{ii} взяты из табл.10.6).

ТАБЛИЦА 10.11

i	1	2	3	4	5	6
S_{ii}	2,435	4,953	3,785	2,539	37,696	131,885
$a_i^* \cdot 10^2$	14,581	10,267	11,744	14,339	3,722	1,990
$\lvert a_i \rvert \cdot 10^2$	0,0816	57,713	34,957	0,619	10,073	5,614

Сравнивая теперь значения $|a_i|$ и a_i^*, заключаем, что признаки 1 и 4 могут быть (и даже должны быть) отброшены.

[1]Вывод этой формулы см. V. Yu. U r b a k h. Biometrics, 1971, v. 27, p. 531.

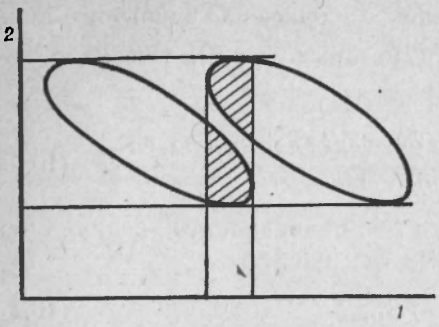


Рис. 10.7. Роль отдельных признаков при разграничении совокупностей с коррелированными признаками.

10.4.4. По изложенной выше методике вопрос о возможности исключения того или иного признака решается лишь после того, как вся работа по нахождению коэффициентов дискриминантной функции уже произведена. Конечно, хотелось бы иметь способ судить о «полезности» каждого из признаков до начала всех расчетов. К сожалению, эта задача пока не решена. В частности, из-за коррелированности признаков нельзя основываться на рассмотрении свойств отдельных признаков. Например, кажется вполне резонным, что признак бесполезен, если распределения двух совокупностей по этому признаку сильно перекрываются. Однако такое заключение совершенно ошибочно: достаточно посмотреть на рис. 10.7. По признаку *1* обе совокупности перекрываются довольно сильно, так что учет одного только этого признака не обеспечивает хорошего разделения. По признаку *2* имеет место полное перекрывание и поэтому, казалось бы, он совершенно бесполезен для диагностики. Между тем на рис. 10.7 видно, что именно дополнительный учет этого признака позволяет произвести достаточно полное разделение совокупностей. Таким образом, нельзя решать вопрос о диагностической полезности признака, исходя только из значимости различия между распределениями в двух совокупностях по этому признаку.

10.4.5. В заключение рассмотрим еще один вопрос, возникающий часто в тех случаях, когда число признаков велико (обычно это бывает в медицинских приложениях дискриминантного анализа). Представим себе, что в выборках, по которым мы хотим построить дискриминантную функцию, некоторые индивиды определены не полностью, т. е. по тем или иным причинам для них отсутствуют значения отдельных признаков. При методике, излагавшейся выше, такие индивиды должны быть исключены из обучающей выборки. Но если каждый индивид описывается, например, 28 признаками, то исключение из анализа индивида из-за одного пробела сопровождается потерей заметного количества информации. Поэтому целесообразно «восполнить» этот пробел, используя какую-либо «разумную» оценку вместо недостающего значения. Ра-

зумеется, это может привести к некоторому искажению результатов. Но если оценка выбрана не слишком плохо, то это искажение будет намного меньше, чем искажение, вызванное уменьшением объема выборки при исключении из анализа данного индивида.

В простейшем случае в качестве оценки недостающего значения $x_{i\gamma}^{\kappa}$ можно принять величину $\overline{x_i}^{K}$, т. е. среднее значение, вычисленное по имеющимся значениям. Однако при этом происходит некоторое занижение величин S_{ij}^{K}, так как вклад от индивида γ в эти величины всегда будет равен нулю. Чтобы устранить это занижение, можно умножить S_{ij}^{K} на $\frac{n}{n-1}$ (или на $\frac{n}{n-f}$, если недостающих значений было f). Эта методика требует, чтобы пробелы были распределены в таблице значений более или менее беспорядочно.

Приложения

Вспомогательные математико-статистические таблицы (в квадратных скобках указаны разделы, в которых описываются таблицы)

Ряд вспомогательных таблиц математической статистики имеется также в тексте:

Таблица I. Случайные числа [1.2.1]

489	583	156	835	988	912	038	460	869	420	522	935	877	665	020	555	379	124	878	544
755	579	550	487	477	864	349	012	250	633	759	554	080	074	001	249	224	368	102	672
303	895	371	196	231	918	380	438	547	644	351	634	323	623	803	374	191	464	696	529
068	803	832	119	350	120	026	684	657	304	613	428	796	447	503	654	254	336	536	944
148	534	105	368	890	473	240	652	435	422	815	144	649	638	137	070	345	865	456	708
780	277	316	013	867	938	930	203	696	769	187	951	991	245	700	564	352	891	249	568
184	179	554	088	254	435	965	154	209	069	916	972	885	275	144	034	122	213	666	230
524	341	860	565	981	842	171	284	707	008	146	291	354	694	377	336	460	585	415	358
920	826	238	402	937	993	332	327	875	230	978	947	380	425	267	285	130	722	164	573
453	653	645	497	969	682	191	976	361	334	473	938	899	348	641	652	852	296	538	456
162	797	002	707	880	660	446	883	768	881	645	219	807	301	279	168	305	937	120	547
042	192	175	851	432	635	757	656	660	389	470	702	958	080	925	519	127	233	452	341
045	730	005	704	345	275	738	862	556	333	880	257	163	439	276	353	912	731	033	294
083	261	277	998	298	204	965	028	936	148	762	713	189	090	989	273	213	935	321	820
965	855	765	703	678	841	543	308	732	289	023	589	740	424	924	005	969	636	237	227
690	408	098	629	819	219	241	128	853	921	292	426	573	903	916	576	368	270	641	033
867	656	016	220	533	345	227	904	138	537	505	127	255	276	233	956	118	199	380	340
295	795	112	761	575	837	336	232	403	345	323	615	410	365	117	417	176	434	240	455
672	536	966	773	412	114	930	697	919	569	422	507	670	013	351	886	268	469	584	596
653	472	113	735	469	545	331	303	814	394	438	376	328	649	327	110	548	955	275	890

Таблица II. Значения $\theta(u)$ — площади под нормальной кривой в пределах от $\mu - u\sigma$ до $\mu + u\sigma$. Ноль целых и запятая опущены [2.3.2]

u	0	1	2	3	4	5	6	7	8	9
0,0	0000	0080	0160	0239	0319	0399	0478	0558	0638	0717
0,1	0797	0876	0955	1034	1113	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2585	2661	2737	2812	2886	2961	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6680	6729	6778
1,0	6827	6875	6923	6970	7017	7063	7109	7154	7199	7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7995	8029
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8335	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8789	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9090
1,7	9109	9127	9146	9164	9181	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9426	9439	9451	9464	9488	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9586	9596	9606	9616	9625	9634
2,1	9643	9651	9660	9668	9676	9684	9692	9700	9707	9714
2,2	9722	9729	9736	9743	9749	9756	9762	9768	9774	9780
2,3	9786	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9840	9845	9849	9853	9857	9861	9865	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9909	9912	9915	9917	9920	9922	9924	9926	9929
2,7	9931	9933	9935	9937	9939	9940	9942	9944	9946	9947
2,8	9949	9950	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972
3,0	9973									

Таблица III. Значения $\Phi(u)$ — площадь под нормальной кривой в пределах от $-\infty$ до u (интеграл вероятностей). Ноль целых и запятая опущены [2.3.2]

u	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
0,0	5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0,1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0,2	5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0,3	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0,4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0,5	6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0,6	7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0,7	7580	7611	7642	7673	7703	7734	7764	7794	7823	7852
0,8	7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0,9	8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1,0	8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1,0	8643	8665	8686	8708	8729	8749	8770	8790	8810	8830
1,2	8849	8869	8888	8907	8925	8944	8962	8980	8997	9015
1,3	9032	9049	9066	9082	9099	9115	9131	9147	9162	9177
1,4	9192	9207	9222	9236	9251	9265	9278	9292	9306	9319
1,5	9332	9345	9357	9370	9382	9394	9406	9118	9430	9441
1,6	9452	9463	9474	9484	9495	9505	9515	9525	9535	9545
1,7	9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
1,8	9641	9648	9656	9664	9671	9778	9686	9693	9700	9706
1,9	9713	9719	9726	9732	9738	9744	9750	9756	9762	9767
2,0	9772	9778	9783	9788	9793	9798	9803	9808	9812	9817
2,1	9821	9826	9830	9834	9838	9842	9846	9850	9854	9857
2,2	9861	9865	9868	9871	9875	9878	9881	9884	9887	9890
2,3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
2,4	9918	9920	9922	9925	9927	9929	9931	9932	9934	9936
2,5	9938	9940	9941	9943	9945	9946	9948	9949	9951	9952
2,6	9953	9955	9956	9957	9959	9960	9961	9962	9963	9964
2,7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974
2,8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981
2,9	9981	9982	9982	9983	9984	9984	9985	9985	9986	9986

	,0	,1	,2	,3	,4	,5	,6	,7	,8	,9
3	$9^{2}86$	$9^{3}03$	$9^{3}31$	$9^{3}52$	$9^{3}66$	$9^{3}77$	$9^{3}84$	$9^{3}89$	$9^{4}28$	$9^{4}52$
4	$9^{4}68$	$9^{4}79$	$9^{4}87$	$9^{5}15$	$9^{5}46$	$9^{5}66$	$9^{5}79$	$9^{5}87$	$9^{6}21$	$9^{6}52$

Примеры: $\Phi(2{,}36) = 0{,}9909$, $\Phi(3{,}8) = 0{,}999928$.

Таблица IV. Значения $t_P = t_\alpha$ (критерия Стьюдента). Нулевая гипотеза принимается при $t \leq t_\alpha$ и отвергается при $t > t_\alpha$ [3.7.2]

f	Доверительные уровни P			f	Доверительные уровни P		
	95%	99%	99,9%		95%	99%	99,9%
2	4,30	9,93	31,60	21	2,08	2,83	3,82
3	3,18	5,84	12,94	22	2,07	2,82	3,79
4	2,78	4,60	8,61	23	2,07	2,81	3,77
5	2,57	4,03	6,86	24	2,06	2,80	3,75
6	2,45	3,71	5,96	25	2,06	2,79	3,73
7	2,37	3,50	5,41	26	2,06	2,78	3,71
8	2,31	3,36	5,04	27	2,05	2,77	3,69
9	2,26	3,25	4,78	28	2,05	2,76	3,67
10	2,23	3,17	4,59	29	2,04	2,76	3,66
11	2,20	3,11	4,44	30	2,04	2,75	3,65
12	2,18	3,06	4,32	40	2,02	2,70	3,55
13	2,16	3,01	4,22	50	2,01	2,68	3,50
14	2,15	2,98	4,14	60	2,00	2,66	3,46
15	2,13	2,95	4,07	80	1,99	2,64	3,42
16	2,12	2,92	4,02	100	1,98	2,63	3,39
17	2,11	2,90	3,97	120	1,97	2,60	3,37
18	2,10	2,88	3,92	200	1,97	2,60	3,34
19	2,09	2,86	3,88	500	1,96	2,59	3,31
20	2,09	2,85	3,85	∞	1,96	2,58	3,29
f	5%	1%	0,1%	f	5%	1%	0,1%
	Уровни значимости α				Уровни значимости α		

Таблица V. **Критические значения F_α (критерия Фишера). Верхнее число — $F_{0,05}$, нижнее число — $F_{0,01}$. Нулевая гипотеза принимается при $F \ll F_\alpha$ и отвергается при $F \geqslant F_\alpha$ [4.6.1]**

f_2 — число степеней свободы знаменателя	f_1 — число степеней свободы числителя										
	1	2	3	4	5	6	7	8	9	10	∞
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,50
	98,49	99,01	99,17	99,25	99,30	99,33	99,34	99,36	99,38	99,40	99,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,53
	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	26,12
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,63
	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,54	13,46
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,36
	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,02
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,67
	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	6,88
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,23
	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62	5,65
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	2,93
	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	4,86
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	2,71
	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26	4,31
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,54
	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	3,91
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,30
	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	3,36
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,13
	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,00
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,01
	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	2,75
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	1,92
	8,28	6,01	5,09	4,58	4,25	4,01	3,85	3,71	3,60	3,51	2,57
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	1,84
	8,10	5,85	4,94	4,43	4,10	3,87	3,71	3,56	3,45	3,37	2,42
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	1,73
	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,25	3,17	2,21
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	1,62
	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,06	2,98	2,01
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	1,51
	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,88	2,80	1,81
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,04	1,99	1,39
	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	1,60
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92	1,28
	6,90	4,82	3,98	3,51	3,20	2,99	2,82	2,69	2,59	2,51	1,43
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,00
	6,63	4,60	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	1,09

Таблица VI. **Критические значения χ^2_α. Нулевая гипотеза принимается при $\chi^2 \leqslant \chi^2_\alpha$ и отвергается при $\chi^2 > \chi^2_\alpha$ [5.1.2]**

f	5%	1%	f	5%	1%	f	5%	1%
1	3,84	6,63	18	28,9	34,8	35	49,8	57,3
2	5,99	9,21	19	30,1	36,2	36	51,0	58,6
3	7,81	11,3	20	31,4	37,6	37	52,2	59,9
4	9,49	13,3	21	32,7	38,9	38	53,4	61,2
5	11,1	15,1	22	33,9	40,3	39	54,6	62,4
6	12,6	16,8	23	35,2	41,6	40	55,8	63,7
7	14,1	18,5	24	36,4	43,0	41	56,9	65,0
8	15,5	20,1	25	37,7	44,3	42	58,1	66,2
9	16,9	21,7	26	38,9	45,6	43	59,3	67,5
10	18,3	23,2	27	40,1	47,0	44	60,5	68,7
11	19,7	24,7	28	41,3	48,3	45	61,7	70,0
12	21,0	26,2	29	42,6	49,6	46	62,8	71,2
13	22,4	27,7	30	43,8	50,9	47	64,0	72,4
14	23,7	29,1	31	45,0	52,2	48	65,2	73,7
15	25,0	30,6	32	46,2	53,5	49	66,3	74,9
16	26,3	32,0	33	47,4	54,8	50	67,5	76,2
17	27,6	33,4	34	48,6	56,1			

Таблица VII. Значения $\varphi = 2 \arcsin \sqrt{p}$ [6.2.2]

%	0	1	2	3	4	5	6	7	8	9
0,0	0,000	0,020	0,028	0,035	0,040	0,045	0,049	0,053	0,057	0,060
0,1	0,063	0,066	0,069	0,072	0,075	0,077	0,080	0,082	0,085	0,087
0,2	0,089	0,092	0,094	0,096	0,098	0,100	0,102	0,104	0,106	0,108
0,3	0,110	0,111	0,113	0,115	0,117	0,118	0,120	0,122	0,123	0,125
0,4	0,127	0,128	0,130	0,131	0,133	0,134	0,136	0,137	0,139	0,140
0,5	0,142	0,143	0,144	0,146	0,147	0,148	0,150	0,151	0,153	0,154
0,6	0,155	0,156	0,158	0,159	0,160	0,161	0,163	0,164	0,165	0,166
0,7	0,168	0,169	0,170	0,171	0,172	0,173	0,175	0,176	0,177	0,178
0,8	0,179	0,180	0,182	0,183	0,184	0,185	0,186	0,187	0,188	0,189
0,9	0,190	0,191	0,192	0,193	0,194	0,195	0,196	0,197	0,198	0,199
1	0,200	0,210	0,220	0,229	0,237	0,246	0,254	0,262	0,269	0,277
2	0,284	0,291	0,298	0,304	0,311	0,318	0,324	0,330	0,336	0,342
3	0,348	0,354	0,360	0,365	0,371	0,376	0,382	0,387	0,392	0,398
4	0,403	0,408	0,413	0,418	0,423	0,428	0,432	0,437	0,442	0,446
5	0,451	0,456	0,460	0,465	0,469	0,473	0,478	0,482	0,486	0,491
6	0,495	0,499	0,503	0,507	0,512	0,516	0,520	0,524	0,528	0,532
7	0,536	0,539	0,543	0,547	0,551	0,555	0,559	0,562	0,566	0,570
8	0,574	0,577	0,581	0,584	0,588	0,592	0,595	0,599	0,602	0,606
9	0,609	0,613	0,616	0,620	0,623	0,627	0,630	0,633	0,637	0,640
10	0,644	0,647	0,650	0,653	0,657	0,660	0,663	0,666	0,670	0,673
11	0,676	0,679	0,682	0,686	0,689	0,692	0,695	0,698	0,701	0,704
12	0,707	0,711	0,714	0,717	0,720	0,723	0,726	0,729	0,732	0,735
13	0,738	0,741	0,744	0,747	0,750	0,752	0,755	0,758	0,761	0,764
14	0,767	0,770	0,773	0,776	0,778	0,781	0,784	0,787	0,790	0,793
15	0,795	0,798	0,801	0,804	0,807	0,809	0,812	0,815	0,818	0,820
16	0,823	0,826	0,828	0,831	0,834	0,837	0,839	0,842	0,845	0,847
17	0,850	0,853	0,855	0,858	0,861	0,863	0,866	0,868	0,871	0,874
18	0,876	0,879	0,881	0,884	0,887	0,889	0,892	0,894	0,897	0,900
19	0,902	0,905	0,907	0,910	0,912	0,915	0,917	0,920	0,922	0,925
20	0,927	0,930	0,932	0,935	0,937	0,940	0,942	0,945	0,947	0,950
21	0,952	0,955	0,957	0,959	0,962	0,964	0,967	0,969	0,972	0,974
22	0,976	0,979	0,981	0,984	0,986	0,988	0,991	0,993	0,996	0,998
23	1,000	1,003	1,005	1,007	1,010	1,012	1,015	1,017	1,019	1,022
24	1,024	1,026	1,029	1,031	1,033	1,036	1,038	1,040	1,043	1,045
25	1,047	1,050	1,052	1,054	1,056	1,059	1,061	1,063	1,066	1,068
26	1,070	1,072	1,075	1,077	1,079	1,082	1,084	1,086	1,088	1,091
27	1,093	1,095	1,097	1,100	1,102	1,104	1,106	1,109	1,111	1,113
28	1,115	1,117	1,120	1,122	1,124	1,126	1,129	1,131	1,133	1,135
29	1,137	1,140	1,142	1,144	1,146	1,148	1,151	1,153	1,155	1,157
30	1,159	1,161	1,164	1,166	1,168	1,170	1,172	1,174	1,177	1,179
31	1,181	1,183	1,185	1,187	1,190	1,192	1,194	1,196	1,198	1,200
32	1,203	1,205	1,207	1,209	1,211	1,213	1,215	1,217	1,220	1,222
33	1,224	1,226	1,228	1,230	1,232	1,234	1,237	1,239	1,241	1,243
34	1,245	1,247	1,249	1,251	1,254	1,256	1,258	1,260	1,262	1,264

Продолжение

%	0	1	2	3	4	5	6	7	8	9
35	1,266	1,268	1,270	1,272	1,274	1,277	1,279	1,281	1,283	1,285
36	1,287	1,289	1,291	1,293	1,295	1,297	1,299	1,302	1,304	1,306
37	1,308	1,310	1,312	1,314	1,316	1,318	1,320	1,322	1,324	1,326
38	1,328	1,330	1,333	1,335	1,337	1,339	1,341	1,343	1,345	1,347
39	1,349	1,351	1,353	1,355	1,357	1,359	1,361	1,363	1,365	1,367
40	1,369	1,371	1,374	1,376	1,378	1,380	1,382	1,384	1,386	1,388
41	1,390	1,392	1,394	1,396	1,398	1,400	1,402	1,404	1,406	1,408
42	1,410	1,412	1,414	1,416	1,418	1,420	1,422	1,424	1,426	1,428
43	1,430	1,432	1,434	1,436	1,438	1,440	1,442	1,444	1,446	1,448
44	1,451	1,453	1,455	1,457	1,459	1,461	1,463	1,465	1,467	1,469
45	1,471	1,473	1,475	1,477	1,479	1,481	1,483	1,485	1,487	1,489
46	1,491	1,493	1,495	1,497	1,499	1,501	1,503	1,505	1,507	1,509
47	1,511	1,513	1,515	1,517	1,519	1,521	1,523	1,525	1,527	1,529
48	1,531	1,533	1,535	1,537	1,539	1,541	1,543	1,545	1,547	1,549
49	1,551	1,553	1,555	1,557	1,559	1,561	1,563	1,565	1,567	1,569
50	1,571	1,573	1,575	1,577	1,579	1,581	1,583	1,585	1,587	1,589
51	1,591	1,593	1,595	1,597	1,599	1,601	1,603	1,605	1,607	1,609
52	1,611	1,613	1,615	1,617	1,619	1,621	1,623	1,625	1,627	1,629
53	1,631	1,633	1,635	1,637	1,639	1,641	1,643	1,645	1,647	1,649
54	1,651	1,653	1,655	1,657	1,659	1,661	1,663	1,665	1,667	1,669
55	1,671	1,673	1,675	1,677	1,679	1,681	1,683	1,685	1,687	1,689
56	1,691	1,693	1,695	1,697	1,699	1,701	1,703	1,705	1,707	1,709
57	1,711	1,713	1,715	1,717	1,719	1,721	1,723	1,725	1,727	1,729
58	1,731	1,734	1,736	1,738	1,740	1,742	1,744	1,746	1,748	1,750
59	1,752	1,754	1,756	1,758	1,760	1,762	1,764	1,766	1,768	1,770
60	1,772	1,774	1,776	1,778	1,780	1,782	1,784	1,786	1,789	1,791
61	1,793	1,795	1,797	1,799	1,801	1,803	1,805	1,807	1,809	1,811
62	1,813	1,815	1,817	1,819	1,821	1,823	1,826	1,828	1,830	1,832
63	1,834	1,836	1,838	1,840	1,842	1,844	1,846	1,848	1,850	1,853
64	1,855	1,857	1,859	1,861	1,863	1,865	1,867	1,869	1,871	1,873
65	1,875	1,878	1,880	1,882	1,884	1,886	1,888	1,890	1,892	1,894
66	1,897	1,899	1,901	1,903	1,905	1,907	1,909	1,911	1,913	1,916
67	1,918	1,920	1,922	1,924	1,926	1,928	1,930	1,933	1,935	1,937
68	1,939	1,941	1,943	1,946	1,948	1,950	1,952	1,954	1,956	1,958
69	1,961	1,965	1,965	1,967	1,969	1,921	1,974	1,976	1,978	1,980
70	1,982	1,984	1,987	1,989	1,991	1,993	1,995	1,998	2,000	2,002
71	2,004	2,006	2,009	2,011	2,013	2,015	2,018	2,020	2,022	2,024
72	2,026	2,029	2,031	2,033	2,035	2,038	2,040	2,042	2,044	2,047
73	2,049	2,051	2,053	2,056	2,058	2,060	2,062	2,065	2,067	2,069
74	2,071	2,074	2,076	2,078	2,081	2,083	2,085	2,087	2,090	2,092
75	2,094	2,097	2,099	2,101	2,104	2,106	2,108	2,110	2,113	2,115
76	2,118	2,120	2,122	2,125	2,127	2,129	2,132	2,134	2,136	2,139
77	2,141	2,144	2,146	2,148	2,151	2,153	2,156	2,158	2,160	2,163
78	2,165	2,168	2,170	2,172	2,175	2,177	2,180	2,182	2,185	2,187
79	2,190	2,192	2,194	2,197	2,199	2,202	2,204	2,207	2,209	2,212

Продолжение

%	0	1	2	3	4	5	6	7	8	9
80	2,214	2,217	2,219	2,222	2,224	2,227	2,229	2,232	2,234	2,237
81	2,240	2,242	2,245	2,247	2,250	2,252	2,255	2,258	2,260	2,263
82	2,265	2,268	2,271	2,273	2,276	2,278	2,281	2,284	2,286	2,289
83	2,292	2,284	2,297	2,300	2,302	2,305	2,308	2,310	2,313	2,316
84	2,319	2,321	2,324	2,327	2,330	2,332	2,335	2,338	2,341	2,343
85	2,346	2,349	2,352	2,355	2,357	2,360	2,363	2,366	2,369	2,372
86	2,375	2,377	2,380	2,383	2,386	2,389	2,392	2,395	2,398	2,401
87	2,404	2,407	2,410	2,413	2,416	2,419	2,422	2,425	2,428	2,431
88	2,434	2,437	2,440	2,443	2,447	2,450	2,453	2,456	2,459	2,462
89	2,465	2,469	2,472	2,475	2,478	2,482	2,485	2,488	2,491	2,495
90	2,498	2,501	2,505	2,508	2,512	2,515	2,518	2,522	2,525	2,529
91	2,532	2,536	2,539	2,543	2,546	2,550	2,554	2,557	2,561	2,564
92	2,568	2,572	2,575	2,579	2,583	2,587	2,591	2,594	2,598	2,602
93	2,606	2,610	2,614	2,618	2,622	2,626	2,630	2,634	2,638	2,642
94	2,647	2,651	2,655	2,659	2,663	2,669	2,673	2,677	2,681	2,686
95	2,691	2,695	2,700	2,705	2,709	2,714	2,719	2,724	2,729	2,734
96	2,739	2,744	2,749	2,754	2,760	2,765	2,771	2,776	2,782	2,788
97	2,793	2,799	2,805	2,811	2,818	2,824	2,830	2,837	2,844	2,851
98	2,858	2,865	2,872	2,880	2,888	2,896	2,904	2,913	2,922	2,931
99,0	2,941	2,942	2,943	2,944	2,945	2,946	2,948	2,949	2,950	2,951
99,1	2,952	2,953	2,954	2,955	2,956	2,957	2,958	2,959	2,960	2,961
99,2	2,963	2,964	2,965	2,966	2,967	2,968	2,969	2,971	2,972	2,973
99,3	2,974	2,975	2,976	2,978	2,979	2,980	2,981	2,983	2,984	2,985
99,4	2,987	2,988	2,989	2,990	2,992	2,993	2,995	2,996	2,997	2,999
99,5	3,000	3,002	3,003	3,004	3,006	3,007	3,009	3,010	3,012	3,013
99,6	3,015	3,017	3,018	3,020	3,022	3,023	3,025	3,027	3,028	3,030
99,7	3,032	3,034	3,036	3,038	3,040	3,041	3,044	3,046	3,048	3,050
99,8	3,052	3,054	3,057	3,059	3,062	3,064	3,067	3,069	3,072	3,075
99,9	3,078	3,082	3,085	3,089	3,093	3,097	3,101	3,107	3,113	3,122
100	3,142									

ЛИТЕРАТУРА

Адлер Ю. П. Введение в планирование эксперимента. М., «Металлургия». 1969.

Аксютина З. М. Элементы математической оценки результатов наблюдений в биологических и рыбохозяйственных исследованиях. М., «Пищевая промышленность», 1968, 288 с.

Ашмарин И. П., Васильев Н. Н., Амбросов В. А. Быстрые методы статистической обработки и планирование экспериментов. Л., Изд-во ЛГУ, 1971, 78 с.

Ашмарин И. П., Воробьев А. А. Статистические методы в микробиологических исследованиях. Л., «Медгиз», 1962.

Беленький М. Л. Элементы количественной оценки фармакологического эффекта. Л., «Медгиз», 1963, 152 с.

Бернстейн А. Справочник статистических решений. М., «Статистика», 1968.

Бессмертный Б. С. Математическая статистика в клинической, профилактической и экспериментальной медицине. М., «Медицина», 1967, 303 с.

Бессмертный Б. С., Ткачева М. Н. Статистические методы в эпидемиологии. М., «Медгиз», 1961, 203 с.

Бирюкова Р. Н. Статистика в клинических исследованиях. М., «Медицина», 1964, 192 с.

Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. М., «Наука», 1965, 464 с.

Боярский А. Я. Статистические методы в экспериментальных медицинских исследованиях. М., «Медгиз», 1955, 263 с.

Боярский Э. А. Порядковые статистики. М., «Статистика», 1972.

Генес В. С. Таблицы достоверных различий между группами наблюдений по качественным показателям. М., «Медицина», 1964, 80 с.

Гублер Е. В., Генкин А. А. Применение непараметрических критериев статистики в медико-биологических исследованиях. Л., «Медицина», 1973, 141 с.

Каминский Л. С. Обработка клинических и лабораторных данных. Применение статистики в научной и практической работе врача. Л., «Медгиз», 1959, 196 с.

Кувшинников П. А. Статистический метод в клинических исследованиях. М., «Медгиз», 1955, 188 с.

Кудрин А. Н., Пономарева Г. Т. Применение математики в экспериментальной и клинической медицине. М., «Медицина», 1967, 356 с.

Максимов В. Н., Федоров В. Д. Применение методов математического планирования эксперимента при отыскании оптимальных условий культивирования микроорганизмов. М., МГУ, 1969, 128 с.

Меркое А. М. Общая теория и методика санитарно-статистического исследования. М., «Медгиз», 1963, 288 с.

Митропольский А. К. Техника статистических вычислений. М., «Наука», 1971, 576 с.

Налимов В. В. Применение математической статистики при анализе вещества. М., «Физматгиз», 1960, 431 с.

Налимов В. В., Чернова Н. А. Статистические методы планирования экстремальных экспериментов. М., «Наука», 1965, 340 с.

Новые идеи в планировании эксперимента. М., «Наука», 1969, 334 с. Авт.: Адлер Ю. П., Андрукович П. Ф., Бродский В. З. и др.

Оуэн Д. Б. Сборник статистических таблиц. Вычислительный центр АН СССР, М., 1966.
Плохинский Н. А. Дисперсионный анализ. Новосибирск. Изд-во Сиб. отд. АН СССР, 1960, 124 с.
Плохинский Н. А. Биометрия. М., МГУ, 1970, 366 с.
Поляков Л. Е. (ред.). Статистические методы исследования в медицине и здравоохранении. Л., «Медицина», 1971, 200 с.
Рокицкий П. Ф. Биологическая статистика. Минск, «Вышэйш. школа», 1967, 327 с.
Сепетлиев Д. А. Статистические методы в научных медицинских исследованиях. М., «Медицина», 1968, 419 с.
Смирнов Н. В., Дунин-Барковский И. В. Краткий курс математической статистики для технических приложений. М., «Физматгиз», 1959.
Урбах В. Ю. Математическая статистика для биологов и медиков. М. Изд-во АН СССР, 1963, 323 с.
Урбах В. Ю. Биометрические методы. М., «Наука», 1964, 415 с.
Янко Я. Математико-статистические таблицы. М., Госстатиздат, 1961.
Бейли Н. Статистические методы в биологии. Пер. с англ. М., «Мир», 1963, 271 с.
Bancroft H. Introduction to biostatistics. New York, 1957, 210 p.
Bliss C. I. Statistics in biology. New York, 1967.
Cochran W. G., Cox G. M. Experimental designs. New York, 1957, 611 p.
Финни Д. Введение в теорию планирования экспериментов. Пер. с англ. М., «Наука», 1970, 287 с.
Fisher R. A. The design of experiments. London, 1951.
Fisher R. A. Statistical methods and scientific inference. Edinburgh, 1956, 175 p.
Fisher R. A., Yates F. Statistical tables for biological, agricultural and medical research. London, 1957, 138 p.
Фишер Р. А. Статистические методы для исследователей. М., Госстатиздат, 1958.
Хальд А. Математическая статистика с техническими приложениями. Пер. с англ. М., «Изд-во иностр. лит-ры», 1956, 664 с.
Хикс Ч. Основные принципы планирования эксперимента. Пер. с англ. М., «Мир», 1967, 406 с.
Хилл А. Б. Основы медицинской статистики. Пер. с англ. М., «Медгиз», 1958, 306 с.
Юл. Дж. Э., Кендалл М. Дж. Теория статистики. Пер. с англ. М., Госстатиздат, 1960, 779 с.
Lienert G. A. Ferteilungsfreie Methoden in der Biostatistik. A. Meisenheim am Glan, 1962.
Linder A. Statistische Methoden für Naturwissenschaftler, Mediziner und Ingenieure. Basel, 1960.
Otto E. Biometrie. Einführung und Anleitung zur Auswertung tierzüchtlicher Ergebnisse. Berlin, 1958.
Pearson E. S., Hartley H. O. Biometrika Tables for Statisticians. Cambridge, 1956.
Снедекор Дж. У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. М., Сельхозиздат, 1961.
Уишарт Дж., Сандерс Г. Основы методики полевого опыта. М. Изд-во «Иностранной лит-ры», 1958.
Ван-дер-Варден Б. Л. Математическая статистика. Пер. с нем. М. Изд-во «Иностранной литературы», 1960, 434 с.
Weber E. Grundriss der biologischen Statistik fur Naturwissenschaftler und Mediziner. Jena, 1948.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

УРБАХ ВИКТОР ЮЛЬЕВИЧ

Статистический анализ в биологических и медицинских исследованиях

Редактор *А. Н. Лисенков*
Техн. редактор *Н. И. Людковская*
Корректор *Е. С. Беляева*
Художественный редактор *Л. С. Бирюкова*
Переплет художника *А. Е. Григорьева*

Сдано в набор 31/XII 1974 г. Подписано к печати 22/V 1975 г. Формат бумаги 60×90¹/₁₆. 18,5 печ. л. (условных 18,5 л.) 16,23 уч.-изд. л. Бум. тип. № 2. Тираж 3000 экз. Т-09713 МН-71.

Издательство «Медицина».
Москва, Петроверигский пер., 6/8

Заказ 803. Ярославский полиграфкомбинат «Союзполиграфпрома» при Государственном комитете Совета Министров СССР по делам издательств, полиграфии и книжной торговли. 150014, Ярославль, ул. Свободы, 97.
Цена 1 р. 83 к.