АКАДЕМИЯ НАУК СССР
ИНСТИТУТ БИОЛОГИЧЕСКОЙ ФИЗИКИ

В. Ю. Урбах

БИОМЕТРИЧЕСКИЕ МЕТОДЫ

(СТАТИСТИЧЕСКАЯ ОБРАБОТКА ОПЫТНЫХ ДАННЫХ В БИОЛОГИИ, СЕЛЬСКОМ ХОЗЯЙСТВЕ И МЕДИЦИНЕ)

ИЗДАТЕЛЬСТВО «НАУКА»
Москва 1964

Настоящее руководство является существенно переработанным и значительно дополненным переизданием книги автора «Математическая статистика для биологов и медиков» (Изд-во АН СССР, 1963 г.)

Оно предназначено для лиц, ведущих исследования в различных областях биологии, медицины, сельского хозяйства и встречающихся с необходимостью статистической обработки опытных данных, а также обучающихся в высших учебных заведениях по этим специальностям.

В книге излагаются все основные методы современной биометрии, включая пробит-метод, последовательный анализ, непараметрические критерии и т. д.

Изложение построено так, что чтение книги не требует от читателя специальной математической подготовки, кроме знаний в объеме средней школы.

Теоретический материал иллюстрируется примерами из биологии и смежных дисциплин. Особое внимание обращено на подробный разбор техники расчетов, играющей во всех применениях биометрии первостепенную роль.

Книга содержит необходимые математико-статистические и другие вспомогательные таблицы.

ОТВЕТСТВЕННЫЙ РЕДАКТОР

доктор биологических наук Н. Н. ЛИВШИЦ

ПРЕДИСЛОВИЕ КО 2-МУ ИЗДАНИЮ

Книга автора «Математическая статистика для биологов и медиков» (Изд-во Академии наук СССР, 1963) разошлась сразу же по выходе и вызвала ряд положительных откликов в печати («Радиобиология», 1963, т. 3, № 4; «Вестник Академии медицинских наук», 1964, № 1; «Biometrics», 1964, т. 20, № 1). Учитывая пожелания, высказанные в рецензиях, а также то, что многие заявки на книгу остались неудовлетворенными, издательство «Наука» сочло целесообразным переиздать книгу.

Как в рецензиях, так и в письмах читателей и в беседах со специалистами отмечались наряду с определенными достоинствами книги и ее недостатки; высказывались также пожелания о внесении различных дополнений. Поэтому при подготовке повторного издания книга была существенно переработана и значительно дополнена (что дало основание изменить название книги). Изменено, прежде всего, расположение материала. Главы, посвященные двумерным совокупностям, перенесены в конец книги; это позволило сделать изложение соответствующих вопросов более компактным, не разрывая описание генеральных и выборочных характеристик. Далее, первые три главы прежнего издания объединены в одну главу о свойствах эмпирических совокупностей.

Из дополнений отметим следующие. В гл. 2 о теоретических распределениях введен параграф с изложением элементов теории вероятностей; рассмотрено общее биномиальное распределение (при $\hat{p} \neq 1/2$); дан вывод формулы распределения Пуассона; добавлен параграф о равномерном распределении и о композиции распределений. В гл. 3 (об оценке параметров по выборочным данным) более подробно описаны различные способы составления выборок; эта глава, а также следующая за ней гл. 4 о параметрических критериях различия сильно переработаны. В главе 5 (о дисперсионном анализе) изложены метод случайных (рэндомизированных) блоков и факторные схемы типа 2^k, а также обсужден вопрос о факторной доле разнообразия. В гл. 8 о корреляционном и регрессионном анализе введены разделы о корреляционных отношениях, о критерии линейности регрессии, о сравнении двух линий регрессии. Наконец, добавлено «Заключение», в котором излагается общая схема статистического анализа — своего рода «путеводитель» по книге. В то же время некоторые второстепенные разделы опущены. Значительно расширен список литературы

(более 50 названий вместо 19 в первом издании). Хронологическое расположение пособий заменено алфавитным (по авторам).

Изменены также и таблицы Приложений: добавлены таблицы случайных чисел; доверительных границ для стандартного отклонения при нормальном распределении; доверительных границ для параметра в распределении Пуассона; критерия для отбрасывания крайних вариант; преобразования вариант с пуассоновским распределением; квадратов чисел. Некоторые таблицы уточнены. В соответствии с новым расположением материала в книге изменено и расположение таблиц Приложения.

В ходе переработки книги были, естественно, устранены замеченные в ней неточности, неудачные формулировки, недостаточно обоснованные рекомендации. Существенную помощь в этом деле оказал автору научный редактор книги кандидат физико-математических наук Л. Н. Большев. Важные указания автор получил от академика АН УССР Б. В. Гнеденко и профессора Н. В. Терентьева. Автор выражает благодарность всем товарищам, принявшим в той или иной форме участие в обсуждении первого издания и рукописи второго издания книги, а также руководителю Центральной изотопной лаборатории Института биохимии им. А. Н. Баха Академии наук СССР доктору биологических наук И. Н. Верховской, чья постоянная поддержка и внимание сделали возможным появление обеих книг.

Автор

ИЗ ПРЕДИСЛОВИЯ К 1-МУ ИЗДАНИЮ

Биология давно уже перестала быть наукой только описательной. Современная биология — это в основном э к с п е р и м е н т а л ь н а я наука. Происходит постоянное совершенствование методов биологического эксперимента и измерительной техники. Биология все более становится т о ч н о й наукой.

Однако точность и достоверность результатов биологических экспериментов зависят не только от качеств экспериментальных методик. Дело в том, что свойства самих биологических объектов сильно варьируют в пределах популяций. Этим объясняется то исключительно большое значение, которое имеет использование статистических методов в биологии.

Настоящая книга предназначена специально для биологов (самого различного профиля). Этим и определяется характер изложения материала.

Прежде всего приходилось считаться с ограниченными математическими знаниями биологов и медиков, а также с тем, что они не склонны следить за цепью математических выкладок. В то же время необходимо учесть следующие обстоятельства.

Каждый из методов математической статистики имеет определенные границы применимости, а использование его вне этих границ может приводить к ошибочным выводам, которые особенно опасны потому, что имеют видимость математической строгости и точности. Для того чтобы избежать таких ошибок, нужно ясно представлять себе смысл и предпосылки применяемых формул. Поэтому мы старались не давать готовых рецептов, а по возможности обосновывать и разъяснять излагаемые методы и соотношения. Так как при этом приходилось ограничиваться самыми простыми математическими средствами, то все эти обоснования оставляют желать очень многого в смысле математической строгости. Но главным образом мы стремились к тому, чтобы они были максимально наглядными.

Назначение книги сказалось также на отборе материала. Сравнительно много внимания уделяется вопросам, имеющим второстепенное значение в математической статистике, но важным для практики биологического эксперимента. Зато некоторые разделы, не представляющие интереса для биологов, опущены, даже если они занимают важное место в логической структуре самой математической статистики.

Наконец, совершенно естественно, что в настоящем руководстве используются в основном примеры из биологии и смежных дисциплин. Оговоримся, что цифровые данные большей частью специально подобраны в соответствии с той целью, для 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; поэтому автор не счел себя вправе давать ссылки на источники, из которых взяты эти иллюстративные данные.

Несколько слов о терминологии и условных обозначениях. В пору расцвета у нас биометрических исследований (20—30-е годы) получили распространение терминология и обозначения, не совпадающие с принятыми в настоящее время в математической статистике (например, термины: варианса, альтернативная изменчивость и т. д.; обозначения: M — для среднего значения, m — для стандартной ошибки среднего и др.). Сохранение в дальнейшем традиционных биометрических обозначений и терминологии означало бы, что биологи и математики будут говорить на разных языках, что отнюдь не в интересах биологов. Поскольку одной из целей данного руководства является подготовить читателя к пользованию более полной статистической литературой, здесь применены общепринятые в математической статистике обозначения и терминология. При этом надо иметь в виду, что и в специальной математической литературе существует известный разнобой в этом отношении.

ПРИНЯТЫЕ УСЛОВНЫЕ ОБОЗНАЧЕНИЯ

A — коэффициент асимметрии
$b_{y/x}$ — оценка коэффициента регрессии
$b_{(y/x)z}$ — оценка частного коэффициента регрессии
D — максимальная разность частостей
d — разность рангов
E — коэффициент эксцесса
F — отношение оценок дисперсии
f — число степеней свободы
G — критерий Кохрена
H_0 — нулевая гипотеза
H_1 — альтернативная гипотеза
h — порядок момента
K — коэффициент взаимной сопряженности
k — число разрядов группировки
M_h — моменты h-го порядка
m_h — начальные моменты h-го порядка
Me — медиана
Mo — мода
N — объем совокупности
n — объем выборки
n_i — частоты разрядов
P — вероятность
p — доля вариант при альтернативном распределении
R — размах варьирования
r — оценка коэффициента корреляции
$r_{xy(z)}$ — оценка частного коэффициента корреляции
S — число серий
s — оценка стандартного отклонения
$s_{\bar{x}}$ — оценка стандартной ошибки среднего значения
s_i — накопленные частоты
T — сумма рангов
T^{Δ} — критерий Вилкоксона для сопряженных пар
t — аргумент распределения Стьюдента
u — аргумент нормального распределения
v — коэффициент вариации
X — критерий ван дер Вардена
x_i — значения вариант
$\hat{x}$ — среднее значение
$\bar{x}$ — оценка среднего значения
y' — пробит
Z — число знаков
z — преобразованный коэффициент корреляции
z_i — накопленные частости
α — уровень значимости
β — вероятность ошибки II рода
$\beta_{y/x}$ — коэффициент регрессии
$\beta_{(y/x)z}$ — частный коэффициент регрессии
Δ — разности частот
ε — относительная неточность
$\eta_{y/x}$ — корреляционное отношение
$\theta(u)$ — площадь под кривой нормального распределения
$\theta^*(t)$ — площадь под кривой распределения Стьюдента
λ — критерий Колмогорова — Смирнова
μ_h — центральные моменты h-го порядка
ν — параметр биномиального распределения
ν_i — частости
ξ_i — отклонения
ρ — коэффициент корреляции
ρ_h — основные моменты h-го порядка
ρ^S — показатель корреляции рангов
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 Пирсона
$\Psi(p)$ — функция, обратная интегралу вероятностей
$!$ — факториал
$|\ |$ — знак модуля (абсолютного значения)

ЗАДАЧИ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ НАБЛЮДЕНИЙ В БИОЛОГИИ

Для биологических объектов характерно то, что они в подавляющем большинстве составляют однородные популяции (виды, породы, сорта и т. д.), 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 полученных из этого изучения данных составить себе представление о свойствах популяции в целом.

Однако разные особи одной и той же популяции всегда в какой-то мере различаются между собой; так как развитие живого организма определяется очень многими и весьма разнообразными условиями внутреннего и внешнего порядка, то практически исключена возможность того, чтобы все эти условия оказались совершенно одинаковыми для каких-нибудь двух особей. Поэтому в результате изучения у ряда особей какого-либо качественного или количественного признака будет получаться не одно какое-нибудь значение, а целый ряд значений, обычно не совпадающих между собой. Каждое из этих значений на первый взгляд имеет одинаковые основания считаться истинным значением признака изучаемой популяции. В какой степени верно это заключение, будет видно из дальнейшего. Пока же у нас нет оснований отдавать предпочтение какому-либо из полученных значений, и поэтому будет правильным считать, что данную популяцию следует характеризовать в отношении изучаемого свойства всей совокупностью полученных значений.

Однако неудобно иметь дело с таким множеством характеристик. Желательно «уплотнить информацию», т. е. получить сравнительно небольшой набор каких-то *сводных* показателей, который был бы легко обозримым. Одной из важнейших задач статистической обработки и является выявление таких немногих показателей — *параметров*, которые в компактной форме, но в то же время достаточно полно характеризуют свойства совокупности. Этому посвящена гл. 1.

Вторая важная задача, которую решает математическая статистика в биологии, возникает в связи с тем, что хотя целью исследования является определение свойств той или иной биологической популяции, исследователь почти никогда не имеет возможности изучить в с е особи этой популяции. Это связано, с одной стороны, с тем, что, как правило, биологические популяции чрезвычайно многочисленны, что совершенно исключает возможность изучения всех членов популяции. С другой стороны, в результате исследования опытный материал часто приводится в негодность или даже полностью уничтожается (например при химических анализах, при вскрытиях и т. д.). Поэтому обычно изучается лишь часть популяции, которую принято называть *выборкой* из *генеральной совокупности* — совокупности всех экземпляров или особей, или членов данной совокупности, которые вообще в принципе могут относиться к этой совокупности.

Заметим, что при экономических исследованиях весьма часто имеют дело с генеральными совокупностями. Например, для того чтобы выяснить, насколько успешно ведется хозяйство на некоторой ферме, можно сравнить значения тех или иных обобщенных показателей за ряд лет. При этом стадо на данной ферме интересует исследователя не как представитель всего вида (или породы) скота, а само по себе, как самостоятельный объект изучения. Напротив, в экспериментальной биологии почти всегда имеют дело с выборками — генеральной совокупностью здесь обычно является бесконечное множество однотипных экспериментов, которые в принципе можно было бы провести.

В связи с заменой изучения генеральной совокупности изучением выборки из нее возникает целый ряд вопросов: в какой степени свойства выборки отражают свойства генеральной совокупности; какую информацию о значениях параметров генеральной совокупности может дать определение характеристик выборки; насколько точна эта информация и т. д. Эти вопросы рассматриваются в гл. 3.

Третья не менее важная задача статистического метода — получение достаточно надежных заключений о различии между генеральными совокупностями на основании информации о выборках из них. Дело в том, что из-за случайностей, сопровождающих составление выборок, разные выборки из одной и той же генеральной совокупности всегда имеют неодинаковые значения обобщенных характеристик. Поэтому сам по себе факт несовпадения значений этих характеристик у двух выборочных совокупностей еще не может считаться доказательством того, что эти выборки взяты из разных генеральных совокупностей. Нужны какие-то критерии, которые позволяли бы каждый раз выяснить, является ли различие между двумя эмпирическими сово-

купностями просто различием между двумя выборками из одной и той же генеральной совокупности или же оно отражает отличие генеральных совокупностей. Эти и сходные вопросы рассматриваются в гл. 4—7.

Важной областью применения математической статистики в биологии является также изучение связей между различными признаками биологических объектов, например между размерами и весом животных, количеством удобрения и урожайностью, числом рождений в данном году и числом браков в предыдущем году и т. д. Вариабильность биологических объектов по каждому из таких признаков делает неизбежным применение статистических методов. Для таких, как принято говорить, *двумерных* (или *многомерных*) совокупностей возникают те же задачи (о которых говорилось выше), что и для одномерных совокупностей. Кроме того, имеются, конечно, и специфические задачи изучения связей между признаками. Эти вопросы рассмотрены в гл. 8—10.

Разумеется, указанные проблемы далеко не исчерпывают всей области применения статистических методов в биологии. Это лишь наиболее общие, фундаментальные вопросы математической статистики; но с ними биологу-исследователю приходится сталкиваться чаще всего. Кроме того, без достаточного знакомства с ними нельзя освоить и другие, более специальные методы.

Для того чтобы научиться грамотно применять методы статистики, более или менее свободно ориентироваться в их многообразии и уметь при решении каждой конкретной биологической задачи подобрать наиболее подходящий показатель, способ расчета, критерий и т. д., необходимо систематическое ознакомление с предметом. Уместно привести по этому поводу следующие строки из превосходной книги Д. Финни (1957, стр. 11)[1]:

«Мне хотелось бы очень серьезно предупредить читателей, которые относятся к статистической науке, как к поваренной книге. Попытки ознакомиться лишь с одним или двумя методами применительно к узкому полю научной деятельности или поиски в учебнике метода для решения одного, только что возникающего вопроса без настоящего понимания логических основ метода часто приводят к совершенно неправильной обработке данных». И еще: «Книга рассчитана на систематическое чтение, а не на случайные справки».

Однако систематическое изучение математической статистики вовсе не означает, что исследователь-биолог должен постоянно держать в голове содержание всей книги. Поэтому, чтобы облегчить работу с книгой не только как с учебным, но и как со справочным руководством, в «Заключении» изложена общая схема

[1] См. список литературы в конце книги.

статистического анализа с подробными ссылками на соответствующие места книги. Надо надеяться, что это будет полезным дополнением к обычному справочному аппарату (оглавление, предметный указатель) и к перечню важнейших формул, приведенному в конце книги.

Многие биологи имеют обыкновение сетовать на то, что применение статистических методов требует большой вычислительной работы. Конечно, в некоторых случаях расчеты могут оказаться весьма громоздкими. Но это относится главным образом к отдельным специальным задачам (гармонический анализ случайных процессов, нахождение дискриминантных функций, многофакторный и многоступенчатый дисперсионный анализ и т. д.). Методы, излагаемые в этой книге, предназначены для решения более простых (но наиболее часто встречающихся) задач. Связанные с ними расчеты редко превышают несколько часов, что, конечно, не идет ни в какое сравнение с продолжительностью непосредственно биологических исследований, длящихся обычно недели и месяцы.

Длительность расчетов может быть сильно сокращена применением всякого рода вычислительных устройств. Для излагаемых здесь методов целесообразно использование логарифмической линейки. Время, затраченное на освоение линейки, с лихвой окупается; к тому же она обеспечивает точность (три значащих цифры), которая для многих задач вполне достаточна. Однако когда требуется большая точность, нужно пользоваться арифмометром или простой счетной машиной [1]. В некоторых случаях оказывается полезной таблица квадратов чисел и квадратных корней, которую мы даем в конце книги (табл. XXIX Приложений).

В этой книге широко употребляются буквы греческого алфавита. Учитывая, что в биологической литературе эти буквы применяются сравнительно мало, даем здесь греческий алфавит (только строчные буквы):

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

---

[1] Типы таких машин, распространенных в СССР, описаны в книге Л. С. Хренова «Малые вычислительные машины», М., ГИФМЛ, 1963. О некоторых особенностях статистических расчетов на подобных машинах см. в книге Бейли (1962, гл. 15).

СВОЙСТВА ЭМПИРИЧЕСКИХ СТАТИСТИЧЕСКИХ СОВОКУПНОСТЕЙ

§ 1. Классификация и группировка вариант

Подлежащий статистической обработке «сырой» материал представляет собой ряд значений, относящихся к одному и тому же признаку. В силу тех или иных причин эти значения, вообще говоря, не совпадают между собой (как принято говорить, они варьируют). Такой ряд значений называют *статистической совокупностью*, а каждый член этой совокупности — *вариантой*. Число вариант в совокупности называется *объемом* совокупности; это число будем обозначать 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 семян, процент жира в молоке, число деревьев на делянке и т. д. В этом случае первоначальное упорядочение совокупности состоит в том, что варианты располагаются в порядке возрастания или убывания их численных значений.

Пусть, например, измерялась длина зерна пшеницы и были получены следующие значения (в *мм*):

5,39 5,42 5,38 5,47 5,51 5,30 5,40 5,40 5,28 5,43.

Тогда упорядоченный ряд будет иметь вид

5,28 5,30 5,38 5,39 5,40 5,40 5,42 5,43 5,47 5,51.

Другой пример: регистрировалось число деревьев определенного вида на каждой из 10 выбранных делянок; получены числа:

15 13 15 14 16 14 12 14 14 15.

Здесь упорядоченный ряд будет

12 13 14 14 14 14 15 15 15 16.

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.

Прежде чем излагать способы группировки, отметим известное различие между двумя только что приведенными совокупностями. Именно, число деревьев не может быть дробным, т. е. одно число в совокупности может отличаться от другого не меньше, чем на единицу; в то же время одно зерно может отличаться по длине от другого в принципе на любую малую величину, так что число градаций в каком-либо выбранном промежутке зависит только от принятой точности измерения. Совокупности первого типа (как, например, по числу деревьев на делянке) называют *дискретными*, а совокупности второго типа (например по длине зерен) — *непрерывными*.

Заметим, однако, что если интервал между наименьшим и наибольшим значениями велик по сравнению с наименьшей градацией (т. е. разностью соседних значений), то число градаций очень велико, и тогда отличие дискретной совокупности от непрерывной становится практически несущественным. С другой стороны, значения непрерывной совокупности записываются всегда с конечным числом десятичных знаков, так что одно значение может отличаться от другого не меньше, чем на единицу последней значащей цифры; поэтому данная совокупность может часто рассматриваться как дискретная.

По этим причинам практически имеет смысл говорить не столько о различии между дискретными и непрерывными совокупностями, сколько о различии между совокупностями с малым или большим числом градаций. Поэтому, употребляя в дальнейшем термины «дискретная» и «непрерывная» (в соответствии с общепринятой терминологией), мы будем вкладывать в них тот условный смысл, который вытекает из приведенных выше соображений.

Из сказанного ясно, что способ группировки должен быть различен для совокупностей с малым и большим числом градаций. В первом случае подсчитывается, 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, имеющих соответствующее значение.

Пример 1. В табл. 1 записаны результаты подсчета числа деревьев на 60 делянках (приведенные на стр. 11 данные пред-

ставляют собой первые 10 чисел из этой таблицы). Следует произвести группировку этих вариант.

Таблица 1

15	13	15	14	16	14	12	14	14	15	13	11	13	15	14
13	15	14	12	14	14	15	15	12	12	13	15	16	14	13
14	13	15	14	13	14	15	14	15	14	14	13	14	15	13
14	13	13	11	12	14	13	12	11	15	15	13	13	13	14

Практически это делается следующим образом. Заготавливается таблица, разделенная на три вертикальные колонки (табл. 2). В первой колонке выписываются сверху вниз все возможные значения вариант (обычно в порядке возрастания). Просматривая подряд заданную совокупность вариант, заполняют вторую более широкую колонку. Так, в табл. 1 первым встречается число 15, поэтому во второй колонке нашей «разносной таблицы» мы ставим черточку против значения «15», вторым является число 13, и ставится черточка против значения «13»; третье число — опять 15, так что против значения «15» ставится вторая черточка, и т. д. Для упрощения последующего подсчета удобней эти пометки группировать по пять, как это сделано в табл. 2[1]. Результат подсчета по каждому из несовпадающих значений записывается в третьей колонке таблицы.

Таблица 2

Значения	Пометки	Число значений
11	𝍸	3
12	𝍸 |	6
13	𝍸 𝍸 𝍸 |	16
14	𝍸 𝍸 𝍸 ||||	19
15	𝍸 𝍸 ||||	14
16	||	2

Во избежание ошибок нужно повторить разноску вариант, просматривая их в обратном порядке — от конца к началу.

[1] Часто при разноске используются так называемые «конвертики»: сначала ставятся четыре точки по углам (: :), затем они соединяются четырьмя черточками по сторонам квадрата (|‾|) и, наконец, проводятся две черточки по диагоналям (|X|); такой «конвертик» означает группу из 10 вариант.

Кроме того, сумма чисел третьей колонки должна совпасть с объемом совокупности, подсчитанным до разноски. В нашем случае сгруппированный ряд имеет вид табл. 3.

Таблица 3

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* (ибо 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:

$$\sum_{i=1}^{k} n_i = N \tag{1.1}$$

(читается: «сумма n_i от 1 до k равна N»). Знак $\sum_{i=1}^{k}$, стоящий перед n_i, указывает, что производится суммирование разных n_i со всеми значениями i от 1 до k (в нашем случае $k = 6$); иными словами, выражение $\sum_{i=1}^{6}$ есть сокращенная запись суммы $n_1 + n_2 + n_3 + n_4 + n_5 + n_6$. Часто пишут упрощенно $\sum n_i$, 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

$$\nu_i = \frac{n_i}{N}. \tag{1.2}$$

В нашем примере

поэтому

$$\nu_1 = \frac{3}{60} = 0{,}05; \quad \nu_2 = \frac{6}{60} = 0{,}10; \quad \nu_3 = \frac{16}{60} = 0.27;$$

$$\nu_4 = \frac{19}{60} = 0{,}32; \quad \nu_5 = \frac{14}{60} = 0{,}23; \quad \nu_6 = \frac{2}{60} = 0{,}03.$$

Очевидно,

$$\sum_i{}' \nu_i = \sum \frac{n_i}{N} = \frac{\Sigma n_i}{N} = \frac{N}{N} = 1. \tag{1.3}$$

Если совокупность имеет много градаций, то группировка заключается в том, что диапазон вариаций делится на определенное число частей (*разрядов*), а затем подсчитывается 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 (скажем, 9—12), а при малом объеме — меньше (6—9).

Следующий вопрос, возникающий при группировке непрерывного распределения, — вопрос о ширине разрядов Δx и о расположении их границ. Решение этого вопроса рассмотрим на примере совокупности, представленной в табл. 4 (длина зерен пшеницы в *мм*; первые десять вариант уже приводились выше).

Самым простым было бы разделить разность между наибольшим ($x_{max} = 5{,}69$) и наименьшим ($x_{min} = 5{,}18$) значениями на принятое число частей (скажем на $k = 11$), после чего границы разрядов находятся сразу.

В нашем примере $5{,}69 - 5{,}18 = 0{,}51$ и $\Delta x = 0{,}51 : 11 = 0{,}04636\ldots$ Мы видим, что здесь $x_{max} - x_{min}$ не делится нацело на k (в пределах принятой точности); надо сказать, что это имеет место почти всегда. В таких случаях производят округление ширины разряда, разумеется, в сторону увеличения, ибо в противном случае общая ширина интервала вариации уменьшилась бы, так что крайние значения вариант не попали бы в него. При таком округлении весь интервал несколько расширяется, причем расширение можно произвести как в сторону меньших, так и в сторону больших значений.

Таблица 4

5,39	5,43	5,49	5,42	5,45
5,42	5,52	5,35	5,45	5,37
5,38	5,45	5,48	5,32	5,48
5,47	5,26	5,26	5,44	5,46
5,51	5,33	5,55	5,58	5,51
5,30	5,43	5,46	5,50	5,29
5,40	5,50	5,41	5,36	5,42
5,40	5,44	5,55	5,44	5,69
5,28	5,47	5,37	5,50	5,60
5,43	5,52	5,45	5,37	5,45
5,46	5,48	5,54	5,47	5,38
5,53	5,34	5,32	5,50	5,46
5,55	5,36	5,52	5,44	5,52
5,47	5,59	5,39	5,28	5,43
5,24	5,45	5,62	5,31	5,41
5,44	5,44	5,40	5,64	5,18
5,54	5,34	5,23	5,46	5,61
5,66	5,33	5,45	5,47	5,36
5,43	5,41	5,47	5,57	5,39
5,42	5,54	5,40	5,58	5,44

Если какое-либо значение попадает на границу разрядов, то его относят к левому разряду. Так, если границы разрядов суть

$$.;\ 5{,}30;\ 5{,}35;\ 5{,}40;\ 5{,}45;\ \ldots,$$

то значение 5,35 относят к разряду 5,30 — 5,35. Иными словами, этот разряд содержит значения x, удовлетворяющие условию $5{,}30 < x \leqslant 5{,}35$.

Иногда сдвигают границы разрядов на половину последнего знака, т. е. берут для границ разрядов значения

$$.;\ 5{,}305;\ 5{,}355;\ 5{,}405;\ 5{,}455;\ .$$

Тогда значение 5,35 попадет в разряд 5,305 — 5,355.

С учетом этого обстоятельства мы теперь можем перейти к вопросу о положении границ всего интервала. Если принять $\Delta x = 0{,}05$ и $k = 11$, то весь диапазон вариации будет иметь ширину $0{,}05 \cdot 11 = 0{,}55$, а поэтому в качестве его границ можно принять либо 5,175—5,725 (расширив диапазон в сторону больших значений), либо 5,145—5,695 (расширив его в сторону меньших значений), либо какие-нибудь промежуточные значения (например 5,165—5,715). Чтобы пояснить, как отразится на виде

распределения тот или иной выбор границ диапазона вариации, приведем результаты группировки для двух крайних случаев *а* и *б* (табл. 5). В первом случае мы получаем одну наибольшую

Таблица 5

Границы разрядов	Пометки	Частоты
	Случай *а*	
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
	Случай *б*	
5,145—5,195	|	1
5,195—5,245	||	2
5,245—5,295	||||	5
5,295—5,345	|||| |||	8
5,345—5,395	|||| |||| ||	12
5,395—5,445	|||| |||| |||| |||| |||	23
5,445—5,495	|||| |||| |||| |||| ||	22
5,495—5,545	|||| |||| ||||	14
5,545—5,595	|||| ||	7
5,595—5,645	||||	4
5,645—5,695	||	2
Всего		100

частоту примерно в середине ряда (в данном случае получилось точно в середине), а при уходе от значения с максимальной частотой вправо и влево частоты убывают. Во втором случае мы имеем в средней части ряда две близкие наибольшие частоты.

Несмотря на такую видимую несхожесть этих рядов, они отражают одно и то же фактическое распределение. Как будет показано ниже, те обобщенные параметры, которыми мы будем в дальнейшем описывать свойства распределения, получатся в обоих случаях одинаковыми (что в свою очередь будет свидетельством достаточно объективного характера этих параметров). Поэтому в общем безразлично, в какую сторону расширены границы интервала вариации.

При составлении таблицы распределения для каждого разряда указывается либо середина разряда (например числа 5,20; 5,25; 5,30 и т. д.), либо границы разрядов (например 5,175—5,225; 5,225—5,275 и т. д.).

Перейдем теперь ко второму типу признаков — порядковому. К этому типу 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, довольно 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*.

Следует отметить, что при углубленном анализе всегда могут быть найдены более или менее рациональные количественные основания для ранжирования — оценка числа и тяжести ошибок, подсчет очков по той или иной системе и т. д. При этом оказывается возможным характеризовать каждую варианту не ее рангом, а соответствующим числовым значением. Если это не всегда делается, то только потому, что часто информация о рангах оказывается достаточной для целей исследования. Например, не всегда требуется знание и, следовательно, измерение точного роста людей — в некоторых случаях достаточно и проще построить их по росту и произвести таким образом ранжирование.

Необходимо подчеркнуть важную особенность порядковых совокупностей: численное значение, отражаемое рангом k, не есть среднее из численных значений, отражаемых рангами $k-1$ и $k+1$. Например, в совокупности, записанной в табл. 6, рангу 2, среднему между рангами 1 и 3, соответствует рост 185 *см*, в то время как середина интервала между 186 и 180 *см* приходится на число 183 *см*; таким образом, рост человека с рангом 2 не является в данном случае средним между ростом людей с рангами 1 и 3. Это относится в той или иной степени и к другим рангам. Данная особенность приводит к тому, что информация, даваемая порядковой градацией, менее полна, чем в случае количественной градации.

Таблица 6

Рост, *см*	186	185	180	177	173	171	164	160	158	155
Ранг	1	2	3	4	5	6	7	8	9	10

Во многих случаях одинаковый балл приписывается нескольким вариантам. Так как полное число рангов должно равняться числу вариант, то применяется следующий прием: ранг совпадающих вариант считается равным среднему из их порядковых номеров. Например, в табл. 7 три варианты имеют баллы 2.

Таблица 7

Балл	1	2	2	2	3	4	4	5	6	8	11	15
Порядковый номер	1	2	3	4	5	6	7	8	9	10	11	12
Ранг	1	3	3	3	5	6,5	6,5	8	9	10	11	12

Так как они имеют одинаковые баллы, то им нужно приписать одинаковый ранг; с другой стороны, они занимают места со второго по четвертое. Поэтому им всем приписывается ранг

$$\frac{2+3+4}{3}=3,$$

а следующей варианте с баллом 3, стоящей на пятом месте, ранг 5. Аналогично двум вариантам с баллом 4, стоящим на шестом

и седьмом местах, приписывается ранг

$$\frac{6+7}{2} = 6{,}5,$$

а следующей за ними восьмой по порядку варианте с баллом 5 — ранг 8.

Наконец, рассмотрим 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 т. д. Конечно, при известных условиях и здесь может оказаться целесообразным ранжирование: по степени пигментации в первом случае, по летальности или длительности — во втором, по экономической выгодности — в третьем. Но нас сейчас будет интересовать именно тот случай, когда ранжирование не производится.

Группировка вариант, отобранных по качественному признаку, состоит в классификации их по градациям этого признака. Не останавливаясь на элементарном вопросе о процедуре классификации, приведем пример сгруппированной совокупности (доход от разных видов сельскохозяйственной продукции в тыс. руб. — табл. 8).

Таблица 8

Зерно	391
Молоко	27
Мясо	138
Овощи	64
Подсолнечник .	33
Сахарная свекла	162
Фрукты .	46
Всего .	861

Группировка состояла в том, что объединялись данные по разным видам зерна, мяса, фруктов, а также данные по подразделениям хозяйства. Естественно, что в зависимости от целей исследования группировка могла бы быть более мелкой (отдельно по разным видам зерна, мяса) или, наоборот, более крупной (продукты полеводства, продукты животноводства и т. д.).

Особенно важным и часто встречающимся в биологических исследованиях является частный случай, когда имеются только

два возможных класса группировки (две альтернативы). Например, если изучается 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 или непоявление какого-либо рефлекса, рождение особи того или иного пола, расщепление гибридов (по одному признаку) на две формы и т. д.

§ 2. Графическое представление распределения

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.

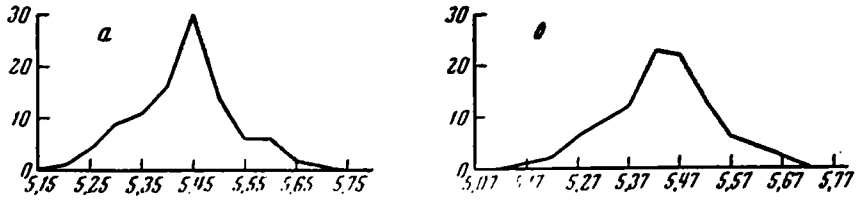

Рис. 1

Среди многих способов графического изображения распределений чаще всего применяются два способа: построение *полигона* (т. е. многоугольника) *частот* и построение *гистограммы* (столбчатой диаграммы).

В первом случае все значения, лежащие в данном разряде, «стягиваются» к середине этого разряда. Например, в разряд 5,375—5,425 (см. табл. 5) 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. 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 1 : 2.

На рис. 1, *а* и *б* представлены полигоны частот распределений, записанных в табл. 5. От крайних разрядов отложено в обе

стороны еще по одному разряду с нулевыми частотами для придания полигону частот завершенного вида.

На гистограмме каждый разряд изображается прямоугольником с шириной, пропорциональной ширине разряда, и с высотой, пропорциональной частоте данного разряда. Для распределений, приведенных в табл. 5, тогда получается картина, представленная на рис. 2, *а* и *б*.

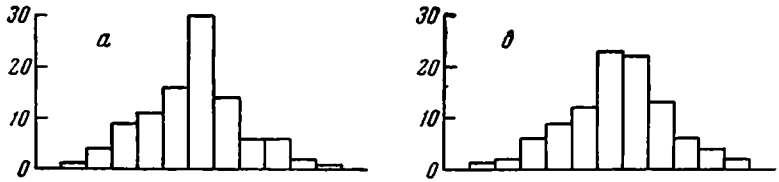

Рис. 2

Изображение распределения при помощи гистрограммы 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. С другой стороны, если совокупность существенно дискретна, то естественно изображать ее полигоном частот.

Рассмотрим теперь тот предельный случай, когда число градаций очень велико; соответственно будем считать очень большим объем совокупности. В этом случае можно в принципе взять 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. (Вопросу о возможных формах кривой распределения будет посвящена гл. 2.)

В интервал абсцисс от x до $x + \Delta x$ попадет примерно $\Delta n = f(x)\Delta x$ вариант (рис. 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 в целях упрощения дальнейших вычислений, а возникающая из-за такой замены погрешность может быть исправлена введением в окончательные результаты надлежащих поправок, о чем будет сказано в § 6 этой главы; впрочем, в большинстве случаев, при не очень грубой группировке, эти поправки можно не учитывать.

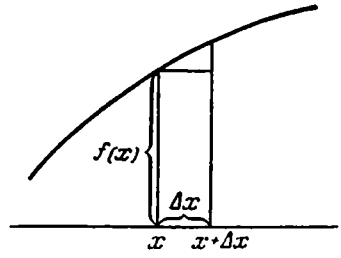

Рис. 3

Иногда приходится (или удобней) 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 интервала). Очевидно, при одинаковой ширине всех интервалов плотности относятся между собой так же, как частоты, поэтому переходить от частот к плотностям нет надобности.

При порядковых градациях графическое изображение излишне, так как в принципе каждому рангу отвечает одна варианта; случаи совпадающих рангов, о которых упоминалось выше, не составляют ту характерную особенность совокупности, которую следует специально подчеркивать.

При качественной классификации употребляется, как правило, изображение с помощью гистограммы. На рис. 4 показана гистограмма для совокупности, приведенной в табл. 8. Важно подчеркнуть, что, в отличие от количественной и порядковой группировки, при качественной классификации можно произвольно переставлять местами разряды группировки, располагая их в любом порядке (в данном случае они расположены просто по алфавиту). Поэтому здесь точная форма гистограммы не имеет никакого значения. Иногда оказывается удобным расположить

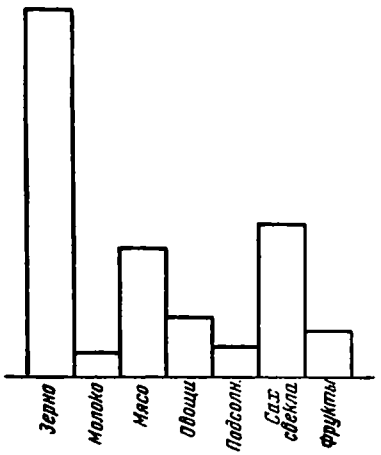

Рис. 4

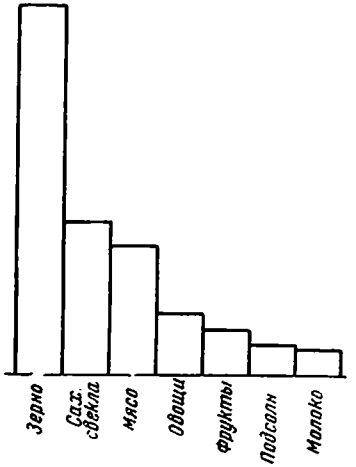

Рис. 5

разряды по убывающим или возрастающим частотам, если нет особых оснований для какого-либо другого расположения. Пример такого расположения (для той же совокупности) показан на рис. 5.

§ 3. Положение статистического ряда. Среднее значение

Как уже говорилось выше, одной из основных задач статистической обработки наблюдений является нахождение небольшого набора показателей, представляющих в обобщенном виде свойства данной статистической совокупности. Очевидно, существенной

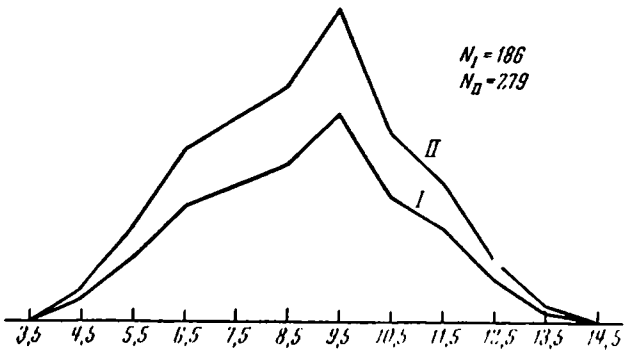

Рис. 6

характеристикой статистической совокупности является объем N, т. е. число исследованных особей или произведенных экспериментов. Совокупность, представленная в табл. 1, имеет объем $N = 60$; совокупность из табл. 4 имеет объем $N = 100$.

Изменение объема совокупности при неизменном характере ее статистического распределения означает пропорциональное увеличение или уменьшение всех частот. Геометрически это эквивалентно растяжению или сжатию графика (полигона частот или гистограммы) по вертикали (рис. 6).

Следующим важным свойством совокупности является положение ряда распределения. В табл. 9 представлены распределения по росту для 1000 мужчин и для 1000 женщин, а на рис. 7 — соответствующие полигоны частот. Мы видим, что графики выглядят в общем одинаково, но они сдвинуты один относительно другого.

Таблица 9

Рост, см	Число мужчин	Число женщин
134	—	1
137	—	6
140	—	19
143	1	61
146	2	127
149	8	186
152	26	209
155	65	179
158	120	121
161	179	59
164	201	23
167	172	7
170	120	2
173	64	—
176	28	—
179	10	—
182	3	—
185	1	—
Итого .	1000	1000

Для того чтобы оценить величину этого сдвига, нужно иметь какой-то количественный параметр, который бы характеризовал положение каждого из сравниваемых статистических рядов. Очевидно, правильней всего было бы характеризовать положение ряда положением его середины. Но, как будет видно из дальней-

шего, можно указать несколько различных параметров, могущих рассматриваться как «середина» ряда.

В зависимости от характера задачи, целей конкретного исследования, соображений удобства и т. д. оказывается более предпочтительным выбор того или другого из этих параметров.

«Естественным» определением середины ряда можно было бы считать полусумму крайних значений. Однако, как это выяснится в дальнейшем, такой параметр недостаточно «характерен» для эмпирических статистических совокупностей.

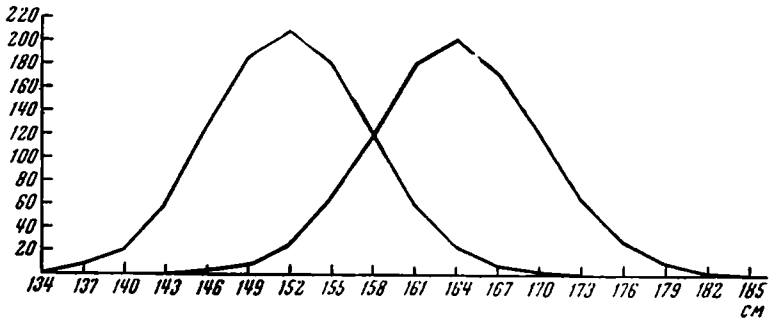

Рис. 7

Наиболее употребительной характеристикой положения статистического ряда является среднее арифметическое значение, называемое просто *средним значением* (обозначим его $\hat{x}$). По определению,

$$\hat{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{j=1}^{N} x_j, \tag{1.4}$$

где x_j ($j = 1, 2, \quad , N$) — значения вариант.

Разность

$$\xi_j = x_j - \hat{x}$$

назовем *отклонением* варианты x_j от среднего значения. Можно показать следующее:

1. Сумма всех отклонений равна нулю

$$\sum_j \xi_j = \sum_j (x_j - \hat{x}) = 0; \tag{1.5}$$

действительно,

$$\sum_j (x_j - \hat{x}) = \sum_j x_j - N\frac{1}{N}\sum_j x_j = 0.$$

2. Сумма квадратов отклонений вариант от $\hat{x}$ меньше, чем сумма квадратов отклонений от любого другого значения x; иначе

говоря, $\hat{x}$ есть то значение x, для которого сумма $\sum_j (x_j - x)^2$ имеет минимальное значение:

$$\sum_j (x_j - \hat{x})^2 < \sum_j (x_j - x)^2$$

при всех $x \neq \hat{x}$. Чтобы доказать справедливость этого неравенства, надо записать $\sum_j (x_j - x)^2$ в виде $\sum_j [(x_j - \hat{x}) - (x - \hat{x})]^2$, затем развернуть это выражение по формуле квадрата разности и учесть, что

$$2 \sum (x_j - \hat{x})(x - \hat{x}) = 2 (x - \hat{x}) \sum (x_j - \hat{x}) = 0$$

по свойству (1.5); тогда получится

$$\sum (x_j - \hat{x})^2 = \sum (x_j - x)^2 - N (x - \hat{x})^2,$$

откуда сразу видно, что $\sum (x_j - \hat{x})^2 < \sum (x_j - x)^2$ при $x \neq \hat{x}$. Последнее свойство придает среднему значению особую важность в статистике.

В случае совокупности из табл. 4 получаем

$$\hat{x} = \frac{1}{100} (5{,}39 + 5{,}42 + 5{,}38 + 5{,}47 + \quad + 5{,}44) = 5{,}4456.$$

Вычисления можно упростить, если не выписывать каждый раз повторяющиеся значения, а записать их по одному разу, но со множителями, указывающими, сколько раз встречаются эти значения. В табл. 4 значение 5,39 встречается 3 раза, значение 5,42 — 4 раза, значение 5,38 — 2 раза и т. д., так что можно записать

$$\hat{x} = \frac{1}{100} (3 \cdot 5{,}39 + 4 \cdot 5{,}42 + 2 \cdot 5{,}38 + \ldots) = 5{,}4456.$$

В этом случае, очевидно, число слагаемых будет уже не 100 (число отдельных значений), а 40 (число несовпадающих значений). Каждую из частот 3, 4, 2 и т. д. можно назвать *весом* соответствующего значения, так что $\hat{x}$ будет «*средневзвешенным*» значением. Если x_i суть несовпадающие значения, а n_i — частоты («веса») этих значений, то

$$\hat{x} = \frac{1}{N} \sum_{i=1}^{s} n_i x_i,$$

причем s есть число несовпадающих значений и, очевидно,

$$\sum_{i=1}^{s} n_i = N.$$

Для средневзвешенного значения можно указать простую механическую аналогию. Если на стержень насадить грузы n_i в точках с координатами x_i (предполагается, что весом стержня можно пренебречь по сравнению с весом грузов), то средневзвешенное значение укажет координату центра тяжести этой системы. Отсюда понятна существенность средневзвешенного значения в качестве эквивалента «середины» распределения: как известно, в механике рассмотрение движения центра тяжести системы позволяет достаточно полно судить о движении системы в целом.

Если число несовпадающих значений вариант велико (в разобранном примере оно равно сорока), то может быть достигнуто дальнейшее упрощение вычисления средневзвешенного — путем группировки вариант в разряды, так как при этом число слагаемых снижается до 8—12. В качестве представителей разрядов берутся середины интервалов группировки (как это имело место при построении полигона частот), так что общая формула для вычисления среднего значения имеет вид

$$\hat{x} = \frac{1}{N} \sum_{i=1}^{k} n_i x_i, \tag{1.6}$$

где x_i — значения середин интервалов, а k — число разрядов

Используя данные табл. 5 (случай *а*), находим

$$\hat{x} = \frac{1}{100}(1 \cdot 5{,}20 + 4 \cdot 5{,}25 + 7 \cdot 5{,}30 + 11 \cdot 5{,}35 + \\ + 16 \cdot 5{,}40 + 30 \cdot 5{,}45 + 14 \cdot 5{,}50 + 8 \cdot 5{,}55 + \\ + 6 \cdot 5{,}60 + 2 \cdot 5{,}65 + 1 \cdot 5{,}70) = 5{,}4405.$$

Аналогичный расчет для табл. 5 (случай *б*) дает то же значение, однако такое совпадение случайно. Обычно при изменении способа группировки получаются несколько различные значения $\hat{x}$, причем чаще всего ни одно из них не совпадает с $\hat{x}$, вычисленным по несгруппированным данным; в нашем примере 5,4405 также не совпало с найденным ранее 5,4456 для несгруппированной совокупности. Это связано с тем, что группировка вариант представляет собой по существу замену истинного распределения вариант внутри каждого из разрядов равномерным распределением, что неизбежно вносит некоторую ошибку.

В связи с этим возникает вопрос, каким числом значащих цифр надо ограничиваться при записи результатов вычислений (поскольку точного результата все равно получить нельзя). Ответ на этот вопрос будет дан в гл. 3, а пока будем придерживаться следующего правила: указывать все результаты вычислений с той точностью, с какой производили измерения. В разбираемом случае совокупости из табл. 4 будем, следовательно, считать $\hat{x} \approx 5{,}44$.

§ 4. Среднее значение функции от варьирующей величины

Измеряемые в опыте величины редко являются целью исследования; как правило, они служат для вычисления каких-то других величин, которые и будут представлять собой «выходные данные» работы. Например, непосредственно измеряют вес животных до опыта и их вес после опыта, но интересуются в действительности изменением веса. Или, скажем, взвешивается урожай на некоторой делянке определенной площади, но хотят узнать урожай в расчете на гектар. Поэтому большую важность имеет вопрос о нахождении средних значений функций от непосредственно измеряемых величин.

Рассмотрим сначала более простой случай, когда нас интересует функция от одной только статистической величины. Пусть, например, мы хотим найти среднюю урожайность $\hat{Y}$ пшеницы, обследовав 5 делянок площадью $S = 0{,}08$ *га* каждая. Если урожаи на этих делянках оказались $y_1 = 1{,}6$ *ц*, $y_2 = 1{,}1$ *ц*, $y_3 = 2{,}0$ *ц*, $y_4 = 1{,}7$ *ц*, $y_5 = 2{,}3$ *ц*, то при пересчете на 1 *га* мы получим:

$$Y_1 = \frac{1{,}6}{0{,}08} = 20{,}0; \quad Y_2 = \frac{1{,}1}{0{,}08} = 13{,}8; \quad Y_3 = \frac{2{,}0}{0{,}08} = 25{,}0;$$

$$Y_4 = \frac{1{,}7}{0{,}08} = 21{,}2; \quad Y_5 = \frac{2{,}3}{0{,}08} = 28{,}8,$$

так что

$$\hat{Y} = \frac{1}{5}(20{,}0 + 13{,}8 + 25{,}0 + 21{,}2 + 28{,}8) \approx 21{,}8.$$

Очевидно, этот же результат можно получить проще, вычислив сначала

$$\hat{y} = \frac{1}{5}(1{,}6 + 1{,}1 + 2{,}0 + 1{,}7 + 2{,}3) \approx 1{,}74,$$

а затем находя

$$\hat{Y} = \frac{\hat{y}}{S} \approx \frac{1{,}74}{0{,}08} \approx 21{,}8.$$

В общем виде можно записать

$$\hat{Y} = \frac{1}{N}(Y_1 + Y_2 + \ldots) = \frac{1}{N}\left(\frac{y_1}{S} + \frac{y_2}{S} + \ldots\right) =$$

$$= \frac{1}{S} \cdot \frac{1}{N}(y_1 + y_2 + \quad .) = \frac{\hat{y}}{S}.$$

Следовательно, если величина y связана с варьирующей величиной x соотношением $y = \alpha x$, где α — постоянный множитель, то

$\hat{y} = \alpha\hat{x}$. Иначе говоря,

$$\langle \alpha x \rangle = \alpha\hat{x} \tag{1.7}$$

(угловыми скобками будем обозначать усреднение). Аналогично можно показать, что

$$\langle x + a \rangle = \hat{x} + a, \qquad a = \text{const}. \tag{1.8}$$

Действительно,

$$\langle x + a \rangle = \frac{1}{N}[(x_1 + a) + (x_2 + a) + \quad + (x_N + a)] =$$
$$= \frac{1}{N}(x_1 + x_2 + . \quad + x_N) + \frac{1}{N}(a + a + \ldots + a) =$$
$$= \hat{x} + \frac{1}{N}(Na) = \hat{x} + a.$$

Свойства (1.7) и (1.8) позволяют упростить вычисление средних значений, которое даже после группировки вариант все еще остается достаточно громоздким. Покажем это на том же примере из табл. 4. Очевидно, каждую из вариант этой таблицы можно представить в виде

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

и тогда

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

Именно:

$$\hat{x} = 5{,}00 + \frac{1}{100}(0{,}39 + 0{,}42 + 0{,}38 + . \quad + 0{,}44) =$$
$$= 5{,}00 + \frac{1}{100}\,44{,}56 = 5{,}00 + 0{,}4456 = 5{,}4456.$$

Так же поступаем, если совокупность сгруппирована. Например, для табл. 5 (случай *а*)

$$\hat{x} = 5{,}00 + \frac{1}{100}(1 \cdot 0{,}20 + 4 \cdot 0{,}25 + 7 \cdot 0{,}30 +$$
$$+ 11 \cdot 0{,}35 + 16 \cdot 0{,}40 + 30 \cdot 0{,}45 + 14 \cdot 0{,}50 +$$
$$+ 8 \cdot 0{,}55 + 6 \cdot 0{,}60 + 2 \cdot 0{,}65 + 1 \cdot 0{,}70) =$$
$$= 5{,}00 + \frac{1}{100}\,44{,}05 = 5{,}00 + 0{,}4405 = 5{,}4405.$$

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. 5 (слу-

чай *а*) 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 $l = 0{,}05$). Теперь расчет выглядит так:

$$\bar{x} = 5{,}15 + \frac{1}{100}(1\cdot 1 + 4\cdot 2 + 7\cdot 3 + 11\cdot 4 + 16\cdot 5 + 30\cdot 6 + 14\cdot 7 + 8\cdot 8 + 6\cdot 9 + 2\cdot 10 + 1\cdot 11)\cdot 0{,}05 = 5{,}15 + \frac{0{,}05}{100}\,581 = 5{,}15 + 0{,}2905 = 5{,}4405.$$

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

$$\bar{x} = 5{,}45 + \frac{0{,}05}{100}(-1\cdot 5 - 4\cdot 4 - 7\cdot 3 - 11\cdot 2 - 16\cdot 1 + 14\cdot 1 + 8\cdot 2 + 6\cdot 3 + 2\cdot 4 + 1\cdot 5) = 5{,}45 + \frac{0{,}05}{100}(-80 + 61) = 5{,}45 - \frac{0{,}05}{100}\,19 = 5{,}4405.$$

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*.

Так как при вычислениях, связанных со статистической обработкой, возможны ошибки (более того, они почти неизбежны), то все расчеты следует выполнять по крайней мере дважды. Как показывает опыт, при повторении тех же операций с теми же числами ошибки нередко повторяются, так что проверочный расчет должен содержать либо другие операции, либо операции с другими числами. В данном случае при вычислении среднего значения удобней всего повторить расчеты при другом выборе начала отсчета, т. е. с другим кодом:

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

(начало отсчета 5,40); тогда

$$\hat{x} = 5{,}40 + \frac{0{,}05}{100}(-1\cdot 4 - 4\cdot 3 - 7\cdot 2 - 11\cdot 1 + 30\cdot 1 + \\ + 14\cdot 2 + 8\cdot 3 + 6\cdot 4 + 2\cdot 5 + 1\cdot 6) = \\ = 5{,}40 + \frac{0{,}05}{100}(-41 + 122) = 5{,}40 + \frac{0{,}05}{100}\cdot 81 = 5{,}4405.$$

Совпадение результатов в обоих случаях указывает на правильность полученной величины.

Равенства (1.7) и (1.8) справедливы только потому, что x и $y = \alpha x$, а также x и $y = x + a$ связаны между собой л и н е й н о.

Однако иначе обстоит дело, если, например, $y = ax^2$. Пусть мы хотим определить среднюю площадь эритроцитов (которые считаем дисками) в крови и для этого измеряем их диаметры (как известно, площадь круга S выражается через диаметр d формулой $S = \pi d^2/4$). Имеем

$$\hat{S} = \frac{1}{N}(S_1 + S_2 + \ldots + S_n) = \frac{1}{N}\left(\frac{\pi}{4}d_1^2 + \frac{\pi}{4}d_2^2 + \ldots + \frac{\pi}{4}d_N^2\right) = \\ = \frac{1}{N}\cdot\frac{\pi}{4}(d_1^2 + d_2^2 + \ldots + d_N^2).$$

Но совершенно очевидно, что

$$d_1^2 + d_2^2 + \ldots + d_N^2 \neq (d_1 + d_2 + \quad + d_N)^2,$$

так как в правой части, помимо суммы $d_1^2 + d_2^2 + \ldots + d_N^2$, имеется еще сумма членов

$$2d_1d_2 + 2d_1d_3 + \quad + 2d_2d_3 + \ldots + 2d_{N-1}d_N$$

Например, в простейшем случае двух слагаемых

$$(d_1 + d_2)^2 = d_1^2 + d_2^2 + 2d_1d_2 > d_1^2 + d_2^2.$$

Следовательно, вообще

$$\langle d^2\rangle \neq \hat{d}^2,$$

т. е. средний квадрат не равен квадрату среднего, точнее — средний квадрат всегда меньше квадрата среднего:

$$\langle d^2\rangle < \hat{d}^2.$$

Поэтому если $y = ax^2$, то $\hat{y} \neq a\hat{x}^2$. Это относится также ко всем $y = ax^n$, кроме случаев $n = 1$ и $n = 0$.

Теперь обратимся к тому важному случаю, когда интересующая нас величина является функцией от двух (или большего числа) варьирующих величин. Например, средний привес в стаде животных есть разность двух статистических величин — средних весов в начале и в конце периода изучения; среднее число зерен ячменя на делянке есть произведение двух статистических величин — среднего числа колосьев на делянке и среднего числа зерен в колосе.

Рассмотрим сначала первый случай. Пусть животные с номерами $i = 1, 2, \ldots, N$ имели до опыта веса $x_i = x_1, x_2, \ldots, x_N$, а после опыта — веса $y_i = y_1, y_2, \ldots, y_N$. Тогда привесы равны разностям $z_1 = y_1 - x_1$, $z_2 = y_2 - x_2$ и т. д., т. е. $z_i = y_i - x_i$, а средний привес будет

$$\hat{z} = \frac{1}{N} \sum z_i;$$

но

$$\frac{1}{N} \sum z_i = \frac{1}{N} \sum (y_i - x_i) = \frac{1}{N} \sum y_i - \frac{1}{N} \sum x_i = \hat{y} - \hat{x},$$

так что средняя разность оказывается равной разности средних. Это значит, что можно обойтись без определения индивидуальных привесов: достаточно вычислить средний вес до опыта и средний вес после опыта.

Таким образом,

$$\langle y - x \rangle = \hat{y} - \hat{x}. \tag{1.9}$$

Аналогично,

$$\langle y + x \rangle = \hat{y} + \hat{x}. \tag{1.10}$$

Однако этого нельзя сказать о случае произведения двух варьирующих величин. Равенство

$$\langle x \cdot y \rangle = \hat{x} \cdot \hat{y} \tag{1.11}$$

справедливо тогда, когда, например, величины x и y варьируют независимо одна от другой. Подробней этот вопрос обсуждается в гл. 2 и 8.

§ 5. Медиана и мода

В случае ранжированной совокупности естественно считать центром совокупности «средний ранг», принимая в качестве такового среднее арифметическое из рангов. Если рассматривать ранги

как варианты, то в этом случае

$$x_1 = 1,\ x_2 = 2,\ x_3 = 3, \qquad x_N = N$$

так что формула (1.4) дает

$$\hat{x} = \frac{1}{N}(1 + 2 + . \quad + N).$$

Но из алгебры известно, что

$$1 + 2 + \ldots N = \frac{N(N+1)}{2},$$

поэтому

$$\hat{x} = \frac{1}{N} \cdot \frac{N(N+1)}{2} = \frac{N+1}{2}.$$

Если N нечетно, то $\hat{x}$ является целым числом, если же N четно, то $\hat{x}$ будет дробным. Так, при $N = 15$ получаем

$$\hat{x} = \frac{15+1}{2} = 8,$$

а при $N = 16$

$$\hat{x} = \frac{16+1}{2} = 8{,}5.$$

Вычисленное таким образом среднее значение ранжированной совокупности имеет ту особенность, что число вариант, ранг которых меньше $\hat{x}$, равно числу вариант, ранг которых больше $\hat{x}$. Так, при $N = 15$ имеется 7 вариант с рангами меньше 8 (это варианты с рангами 1, 2, 3, 4, 5, 6, 7) и столько же вариант с рангами больше 8 (варианты с рангами 9, 10, 11, 12, 13, 14, 15). То же имеет место и при четном N: например, при $N = 16$ восемь вариант имеют ранги меньше 8,5 и восемь — ранги больше 8,5.

Собственно говоря, для ранжированной совокупности только это свойство величины $\hat{x}$ и имеет реальное значение. Ведь последовательным рангам, как мы знаем, не соответствуют равноотстоящие численные значения, а поэтому величина $\hat{x}$, хотя она вычислена по формуле среднего значения (1.4), в действительности не является «центром тяжести» стоящих за рангами численных значений.

Поэтому если заданы только ранги и неизвестны стоящие за ними численные значения, то нет никаких оснований считать, что величина $\frac{N+1}{2}$ имеет свойства среднего значения. Можно лишь утверждать, что она делит совокупность на две части та-

ким образом, что половина всех вариант имеет ранги меньшие этой величины, а другая половина — большие. Величину, обладающую такими свойствами, называют *медианой* совокупности (обозначается через Me).

Таким образом, при порядковой градации вариант медиана дает максимально возможную информацию о середине ряда. Ввиду простоты нахождения медианы ее нередко применяют в качестве характеристики середины распределения и для тех совокупностей, варианты которых характеризуются численными значениями (хотя в этом случае имеется возможность определить параметр, дающий более точную информацию о середине ряда, а именно — среднее значение). Например, в совокупности с вариантами

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

медианой будет варианта 32, ибо шесть значений (16, 19, 21, 26, 27 и 31) меньше 32 и столько же значений (34, 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:

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

Пример 2. Найдем медиану совокупности

34 46 32 58 42 21 26 35 54 25 39 54.

Прежде всего располагаем варианты в порядке возрастания:

21 25 28 32 34 35 39 42 46 54 54 58

(конечно, можно было бы расположить их и в порядке убывания). Если бы число вариант было нечетным, то медианой была бы варианта, имеющая ранг $\frac{N+1}{2}$. Но в данном случае число вариант четно ($N = 12$), а поэтому медианой будет значение, равное полусумме значений вариант с рангами $\frac{N}{2} = 6$ и $\frac{N}{2} + 1 = 7$, а именно,

$$Me = \frac{35 + 39}{2} = 37.$$

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. 5 (случай *а*) ряд накопленных частот будет иметь вид табл. 10:

Таблица 10

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. 10), $s_6 = s_5 + n_6 = 39 + 30 = 69$, $s_7 = s_6 + n_7 = 69 + 14 = 83$ 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). Чтобы более точно указать положение 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:

$$\begin{array}{l|l} 5{,}40 \sim 39\% & Me = 5{,}40 + \frac{5{,}45-5{,}40}{69-39} \cdot (50-39) = \\ 5{,}45 \sim 69\% & \\ Me \sim 50\% & = 5{,}40\,\frac{0{,}05 \cdot 11}{30} \approx 5{,}418. \end{array}$$

Такой способ нахождения промежуточных значений называется линейной интерполяцией. Он знаком каждому, кто пользовался таблицами логарифмов, тригонометрических функций и т. п. В данном случае 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. Поэтому искомый сдвиг равен

$$0{,}05 \cdot \frac{11}{30} = 0{,}018,$$

так что для медианы получается

$$5{,}40 + 0{,}018 = 5{,}418.$$

В виде формулы это можно записать так:

$$Me = x_{Me} + \Delta x \frac{\frac{N}{2} - s_{Me}}{n_{Me+1}}, \qquad (1.13)$$

где x_{Me} — начало интервала, содержащего медиану; Δx — ширина интервала; s_{Me} — накопленная частота на начало медианного интервала; n_{Me+1} — частота в медианном интервале. В данном случае

$$Me = 5{,}40 + 0{,}05 \frac{50-39}{39} = 5{,}40 + 0{,}018 = 5{,}418.$$

Применение медианы для характеристики середины распределения целесообразно в случае незамкнутой совокупности, т. е. когда не указано начало или конец ряда (или оба вместе). Например, для ряда

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

нельзя вычислить среднее значение, потому что неизвестны середины первого и последнего разрядов. Между тем медиану указать можно: $Me = 3000$ *га*.

В совокупностях, в которых вообще отсутствует возможность какой-либо градации вариант, а может лишь быть произведена их классификация по какому-нибудь качественному признаку, единственным способом указать некий «центр тяжести» совокупности является указание той группы, в которую входит больше всего вариант. Эту наиболее типичную группу называют *модой* и обозначают *Mo*. Например, в совокупности, представленной в табл. 8, модой будет группа «зерно».

Понятие моды может оказаться полезным и для совокупностей с количественной градацией. Так, на рис. 8 изображено распределение по возрасту заболевших дифтерией (на 10 тыс. населения соответствующего возраста). Очевидно, знание с р е д н е г о возраста заболевающих дифтерией (в данном случае 7,75 лет) менее интересно, чем знание возраста, в котором ч а щ е

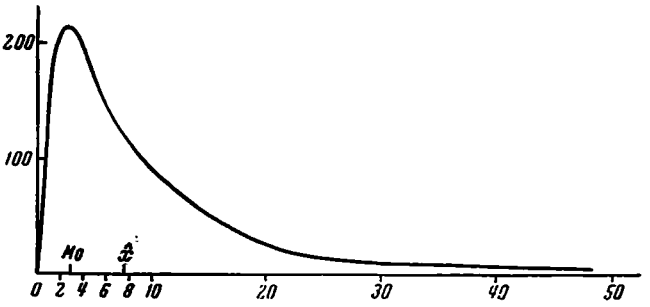

Рис. 8

в с е г о происходит заболевание (здесь — от 2 до 4 лет), в частности, при решении вопроса о том, где должны быть сосредоточены главные профилактические усилия: 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 среднее значение близки между собой. Поэтому в таких случаях мода, находимая более просто, может служить приближенной оценкой среднего значения.

§ 6. Характеристики рассеяния вариант. Дисперсия и коэффициент вариации

Сравним изображения распределения длины красных бобов для двух совокупностей одинакового объема. В первом случае (рис. 9, *а*) для посева использовались семена генетически чистой линии, а во втором (рис. 9, *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 среднего значения при помощи величины, полученной усреднением всех их отклонений от $\hat{x}$. Однако

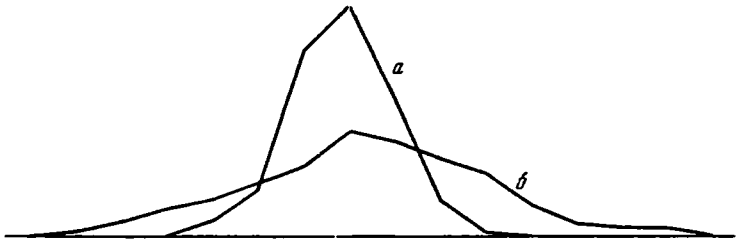

Рис. 9

эта величина не может быть средним арифметическим из отклонений, так как это среднее арифметическое всегда тождественно равно нулю. В самом деле,

$$\frac{1}{N}\sum n_i\xi_i = \frac{1}{N}\sum n_i\,(x_i - \hat{x}) = \frac{1}{N}\sum n_i x_i - \frac{1}{N}\,\hat{x}\sum n_i =$$

$$= \hat{x} - \hat{x} = 0.$$

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 положительных 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

$$\langle|\xi|\rangle = \frac{1}{N}\sum_{i=1}^{k} n_i\,|\xi_i| \qquad (1.14)$$

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

$$\langle\xi^2\rangle = \frac{1}{N}\sum_{i}^{k} n_i\xi_i^2 = \frac{1}{N}\sum_{i=1}^{k} n_i\,(x_i - \hat{x})^2, \qquad (1.15)$$

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

$$\sigma = \sqrt{\langle\xi^2\rangle} = \sqrt{\frac{1}{N}\sum n_i\,(x_i - \hat{x})^2} \qquad (1.16)$$

ее называют *средним квадратичным 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 квадратичное отклонение имеет ряд свойств, делающих очень удобным оперирование с ним; подробней об этом будет сказано ниже.

Вычисление стандартного отклонения по формуле (1. 16) представляет собой несколько трудоемкую задачу. Пусть мы хотим найти σ для распределения, приведенного в табл. 5 (случай *а*), которое запишем в виде табл. 11 (первые две строки), приняв для x_i середины разрядов. Так как $\hat{x} = 5{,}44$ (см. § 3 настоящей главы), то для отклонений получим значения, приведенные в третьей строке табл. 11.

Таблица 11

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		1
ξ_i	−0,24	−0,19	−0,14	−0,09	−0,04	0,01	0,06	0,11	0,16	0,21	0,26

Эти значения надо возводить в квадрат, после чего умножить на соответствующие n_i и суммировать. Сразу видно, что предстоят довольно громоздкие вычисления. Использование условной шкалы (начало отсчета $x_0 = 5{,}45$, масштаб $l = 0{,}05$) само по себе также не дает упрощения; в самом деле, поскольку в этой шкале $\hat{x} = -0{,}19$, мы получаем табл. 12, что приводит к еще более громоздким вычислениям.

Таблица 12

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
ξ_i	−4,81	−3,81	−2,81	−1,81	−0,81	−0,19	1,19	2,19	3,19	4,19	5,19

Однако расчет можно значительно упростить, если использовать формулу, получающуюся в результате ряда элементарных преобразований. Именно, так как

$$(x_i - \hat{x})^2 = x_i^2 - 2x_i\hat{x} + \hat{x}^2,$$

то

$$\sigma^2 = \frac{1}{N}\sum n_i (x_i - \hat{x})^2 = \frac{1}{N}\sum n_i x_i^2 - 2\frac{1}{N}\sum n_i x_i \hat{x} + \frac{1}{N}\sum n_i \hat{x}^2.$$

Но $\hat{x}$ и $\hat{x}^2$ не зависят от индекса i и поэтому их можно выносить за знак суммы; тогда

$$2\frac{1}{N}\sum n_i x_i \hat{x} = 2\hat{x}\frac{1}{N}\sum n_i x_i = 2\hat{x}\cdot\hat{x} = 2\hat{x}^2;$$

$$\frac{1}{N}\sum n_i \hat{x}^2 = \hat{x}^2\frac{1}{N}\sum n_i = \hat{x}^2\frac{1}{N}N = \hat{x}^2.$$

Поэтому окончательно

$$\sigma^2 = \frac{1}{N}\sum n_i(x_i - \hat{x})^2 = \frac{1}{N}\sum n_i x_i^2 - \hat{x}^2. \qquad (1.17)$$

Величина $\frac{1}{N}\sum n_i x_i^2$ есть средний квадрат значений x_i. Так как левая часть этого равенства не может быть отрицательной (ибо это сумма квадратов), то и правая часть всегда неотрицательна. Отсюда следует, что средний квадрат не может быть меньше квадрата среднего — вывод, который был уже сформулирован в § 4

этой главы. Формула (1.17) играет в статистике очень важную роль; в дальнейшем мы неоднократно будем на нее ссылаться.

Применим теперь эту формулу к вычислению стандартного отклонения, используя условную шкалу. Из табл. 12 имеем

$$\frac{1}{N}\sum n_i x_i^2 = \frac{1}{100}(1\cdot 25 + 4\cdot 16 + 7\cdot 9 + 11\cdot 4 + 16\cdot 1 + 14\cdot 1 + 8\cdot 4 + 6\cdot 9 + 2\cdot 16 + 1\cdot 25) = 3{,}69;$$

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

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

в единицах $l = 0{,}05$ *мм*, т. е.

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

Указанный выше способ вычисления дисперсии приводит к небольшой систематической ошибке, вызванной тем, что при группировке вариант в разряды несколько искажается истинный характер распределения. А именно: в сгруппированной совокупности распределение вариант внутри каждого из разрядов считается равномерным, тогда как на самом деле число вариант обычно больше в той части разряда, которая ближе к центру распределения. Понятно, что по этой причине показатель рассеяния получается завышенным. Как показал Шеппард, для устранения этой ошибки нужно вычисленное значение σ^2 уменьшить на величину $l^2/12$. В нашем примере поправка Шеппарда равна $(0{,}05)^2/12 \approx 0{,}0002$, т. е. она незначительна. Поправку Шеппарда вообще лучше не вносить, если объем совокупности не очень велик ($N < 500$) и группировка была не очень грубой.

Помимо описанного здесь способа нахождения дисперсии, основанного на вычислении произведений $n_i x_i$ и $n_i x_i^2$, существует также так называемый способ сумм, который в ряде случаев оказывается очень удобным. Он подробно описан в книге А. К. Митропольского (1961).

В § 4 были получены формулы для средних значений функций от варьирующих величин:

$$\langle ax \rangle = a\hat{x};$$
$$\langle x + a \rangle = \hat{x} + a;$$

$$\langle x - y\rangle = \hat{x} - \hat{y};$$
$$\langle x + y\rangle = \hat{x} + \hat{y}.$$

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 здесь не как функция от частных значений x и y, а как функция от в с е й совокупности значений аргументов.

Имеем по определению

$$\sigma^2\{\alpha x\} = \frac{1}{N}\sum n_i\,(\alpha x_i - \langle \alpha x\rangle)^2.$$

Учитывая (1.7), можно переписать это в виде

$$\sigma^2\{\alpha x\} = \frac{1}{N}\sum n_i\,(\alpha x_i - \alpha\hat{x})^2.$$

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

$$\sigma^2\{\alpha x\} = \alpha^2\frac{1}{N}\sum n_i\,(x_i - \hat{x})^2 = \alpha^2\sigma^2\{x\}. \tag{1.18}$$

Далее,

$$\sigma^2\{x + a\} = \frac{1}{N}\sum n_i\,[(x_i \quad a) - (\hat{x} + a)]^2 = \frac{1}{N}\sum n_i\,(x_i - \hat{x})^2,$$

так что

$$\sigma^2\{x + a\} = \sigma^2\{x\}. \tag{1.19}$$

Теперь найдем $\sigma^2\{x - y\}$ и $\sigma^2\{x + y\}$. Согласно (1.9) имеем

$$[(x_i - y_j) - \langle x - y\rangle]^2 = [(x_i - y_j) - (\hat{x} - \hat{y})]^2.$$

Правую часть этого равенства можно преобразовать к виду

$$[(x_i - \hat{x}) - (y_j - \hat{y})]^2,$$

что дает после развертывания

$$(x_i - \hat{x})^2 - 2\,(x_i - \hat{x})\,(y_j - \hat{y}) + (y_j - \hat{y})^2.$$

Суммируя по всем x и y и деля на объем совокупности, получаем

$$\sigma^2\{x - y\} = \frac{1}{N}\sum_{i,j}[(x_i - y_j) - \langle x - y\rangle]^2 =$$

$$= \frac{1}{N}\sum_i (x_i - \hat{x})^2 - \frac{2}{N}\sum_{i,j}(x_i - \hat{x})\,(y_j - \hat{y}) + \frac{1}{N}\sum_j (y_j - \hat{y})^2.$$

Если x и y варьируют независимо, то

$$\sum_{i,j}(x_i - \hat{x})(y_j - \hat{y}) = \sum_i (x_i - \hat{x}) \sum_j (y_j - \hat{y}) = 0$$

[ибо каждая из сумм $\sum_i (x_i - \hat{x})$ и $\sum_j (y_j - \hat{y})$ равна нулю]. Так как

$$\frac{1}{N}\sum_i (x_i - \hat{x})^2 = \sigma^2\{x\} \text{ и } \frac{1}{N}\sum_j (y_j - \hat{y})^2 = \sigma^2\{y\},$$

то окончательно

$$\sigma^2\{x - y\} = \sigma^2\{x\} + \sigma^2\{y\}. \tag{1.20}$$

Совершенно аналогично, исходя из (1.10), получим

$$\sigma^2\{x + y\} = \sigma^2\{x\} + \sigma^2\{y\}. \tag{1.21}$$

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.

Значение σ не всегда достаточно полно характеризует вариабильность 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

$$v = \frac{\sigma}{\bar{x}}; \qquad (1.22)$$

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 и в том отношении, что позволяет сравнивать вариабильности совокупностей, значения которых имеют различную размерность. Например, не имеет смысла спрашивать, в какой совокупности рассеяние больше — в распределении деревьев по толщине со среднеквадратичным 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 вариант 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° С) при среднемесячной температуре +1° С. В данном случае величина σ более адекватно отразит характер явления.

§ 7. Асимметрия распределения

В табл. 13 представлено распределение длины волокна для определенного сорта хлопка, а на рис. 10 — полигон частот этого распределения. Бросается в глаза несимметричность полигона частот.

Таблица 13

l, мм	9	12	15	18	21	24	27	30	33	36	39	Сумма
‰	8	12	21	33	56	98	183	256	214	97	22	1000

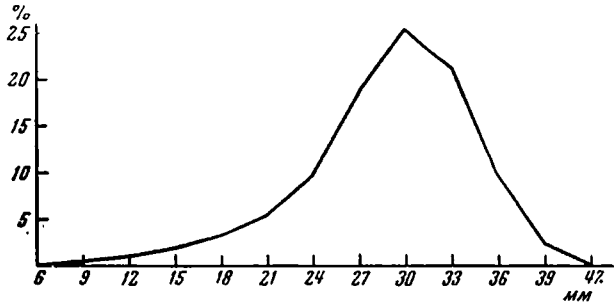

Рис. 10

Не вдаваясь пока в исследование причин такого свойства распределения (это свойство принято называть *асимметрией*), посмотрим, каким образом можно дать его количественное описание. Иными словами, мы хотим найти такой параметр, который являлся бы количественной характеристикой асимметрии распределения.

Вычислим средний куб отклонения:

$$\langle \xi^3 \rangle = \frac{1}{N} \sum n_i \xi_i^3 = \frac{1}{N} \sum n_i (x_i - \bar{x})^3. \qquad (1.23)$$

В отличие от среднего отклонения

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

которое, как это было показано выше, всегда равно нулю, средний куб отклонения 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 вычисле-

ний $\langle\xi^3\rangle$ 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 (ибо значения возводятся в нечетную степень) вклады в сумме $\sum n_i\xi_i^3$.

Очевидно, величина $\langle\xi^3\rangle$ тем больше, чем сильней 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

$$A = \frac{\langle\xi^3\rangle}{\sigma^3} \tag{1.24}$$

будет безразмерным. Этот параметр называют *коэффициентом асимметрии*.

Следует иметь в виду, что значение коэффициента A является более существенным показателем асимметрии, нежели вид графика распределения. Дело в том, что форма графика распределения в известной мере зависит от способа группировки в разряды. Правильная группировка предполагает, что среднее значение, которое потом будет вычисляться, окажется либо в середине одного из разрядов, либо на границе между разрядами. Если группировка произведена так, что среднее значение занимает положение, промежуточное между этими двумя оптимальными, то внешний вид графика окажется искаженным; в частности, может появиться кажущаяся асимметрия, в то время как распределение в действительности симметрично. Рис. 11 иллюстрирует эту особенность группировки. Пунктирные вертикали изображают границы разрядов, а сплошные вертикали — отрезки, пропорциональные площадям в этих границах, т. е. числу вариант в соответствующих разрядах. В случае *а* значение $\hat{x}$ приходится на середину одного из разрядов, а в случае *б* — на границу между двумя разрядами; в обоих случаях распределение частот симметрично. В случае же *в* величина $\hat{x}$ занимает упомянутое выше промежуточное положение, и поэтому распределение частот получи-

лось асимметричным; для большей ясности эти частоты изображены отдельно (фиг. *в'* на рис. 11).

Тем не менее коэффициент асимметрии A и в этом случае окажется равным нулю. Это объясняется тем, что меньшие частоты расположены дальше от $\hat{x}$, чем соответствующие бо́льшие

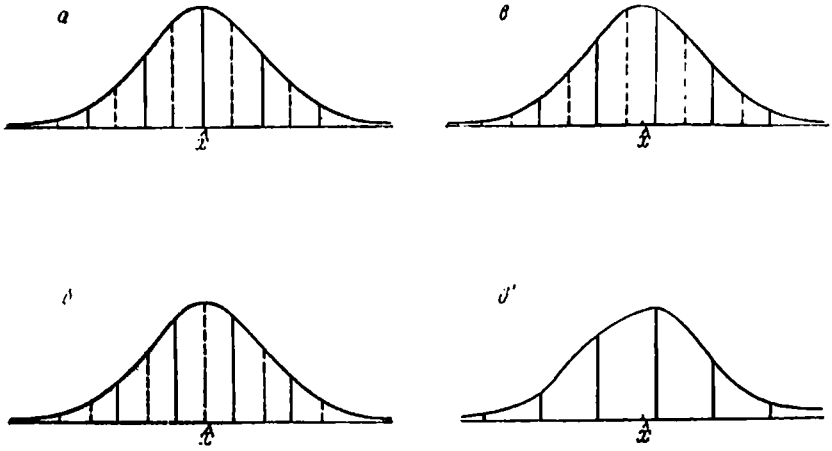

Рис. 11

частоты, так что значения $n_i (x_i - \hat{x})^3$ и здесь будут компенсироваться. Если все же желательно дать наглядное графическое представление о ряде распределения, то следует после вычисления среднего значения пересоставить ряд, выбрав границы разрядов так, чтобы среднее значение находилось возможно ближе к середине одного из разрядов.

§ 8. Статистические моменты

Выше, при вычислении среднего значения, дисперсии и коэффициента асимметрии, нам пришлось пользоваться величинами:

$$\hat{x} = \frac{1}{N} \sum n_i x_i;$$

$$\langle \xi^2 \rangle \quad \frac{1}{N} \sum n_i (x_i - \hat{x})^2;$$

$$\langle \xi^3 \rangle \quad \frac{1}{N} \sum n_i (x_i - \hat{x})^3.$$

Сопоставляя эти формулы, замечаем, что все они могут рассматриваться как частные случаи одной более общей формулы

$$M_h = \frac{1}{N} \sum n_i (x_i - x)^h. \qquad (1.25)$$

Действительно, при $h = 1$ и $x = 0$ получим $\hat{x}$, при $h = 2$ и $x = \hat{x}$ получаем $\langle \xi^2 \rangle$, при $h = 3$ и $x = \hat{x}$ получаем $\langle \xi^3 \rangle$.

Величина (1.25) называется *моментом* h-го порядка распределения (x_i, n_i) относительно значения x.

Если в качестве x выбрано начало отсчетов, т. е. положено $x = 0$, то момент называется *начальным* и обозначается m_h. Если же в качестве x выбран центр распределения $\hat{x}$, то момент называется *центральным* и обозначается μ_h. В соответствии с этой терминологией среднее значение $\hat{x}$ есть начальный момент первого порядка: $\hat{x} = m_1$; дисперсия, или средний квадрат отклонения, есть центральный момент второго порядка: $\sigma^2 = \mu_2$; средний куб отклонения есть центральный момент третьего порядка: $\langle \xi^3 \rangle = \mu_3$ Очевидно, центральный момент первого порядка всегда равен нулю: $\mu_1 = 0$, это было показано в § 6 настоящей главы. Само собой разумеется, что аналогичным образом могут быть определены начальные и центральные моменты четвертого, пятого и т. д. порядков; ниже (см. гл. 2) будет показано, что центральный момент четвертого порядка μ_4 играет в математической статистике заметную роль.

Статистические моменты имеют определенную механическую аналогию (момент силы, момент инерции и т. д.), с которой связано происхождение этого названия. Как мы сейчас увидим, использование свойств моментов существенно облегчает решение многих задач статистической обработки.

Рассмотрим прежде всего связь между центральными моментами и начальными моментами заданного распределения. Пример такой связи дает формула

$$\langle \xi^2 \rangle = \langle x^2 \rangle - \hat{x}^2,$$

вывод который был приведен в § 6 настоящей главы; на языке моментов это означает

$$\mu_2 = m_2 - m_1^2 \tag{1.26}$$

Поскольку при надлежащем выборе начала отсчета вычисление начальных моментов m_2 и m_1 много проще, чем вычисление центрального момента μ_2, использование формулы (1.26) намного упрощает расчеты. Теми же выкладками, какими была получена формула (1.26), можно получить подобные формулы и для $\langle \xi^3 \rangle$ и $\langle \xi^4 \rangle$, которые запишем в терминах моментов:

$$\mu_3 = m_3 - 3m_2m_1 + 2m_1^3; \tag{1.27}$$

$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4; \tag{1.28}$$

$$m_1 = \frac{1}{N}\sum n_i x_i; \quad m_2 = \frac{1}{N}\sum n_i x_i^2; \quad m_3 = \frac{1}{N}\sum n_i x_i^3;$$

$$m_4 = \frac{1}{N}\sum n_i x_i^4. \tag{1.29}$$

Простыми подстановками можно получить формулы

$$\mu_3 = m_3 - 3\mu_2 m_1 - m_1^3; \tag{1.30}$$

$$\mu_4 = m_4 - 4\mu_3 m_1 - 6\mu_2 m_1^2 - m_1^4 \tag{1.31}$$

для проверочных вычислений μ_3 и μ_4.

Проиллюстрируем вычисление моментов μ_2 и μ_3 на примере распределения, приведенного в табл. 13. В табл. 14 вычисляются начальные моменты, которые затем подставляются в формулы (1.26) и (1.27). В столбце (3) записаны значения x_i в условной шкале. Числа столбца (4) получаются перемножением чисел данной строки из столбцов (2) и (3). Числа столбцов (5) и (6) получаются умножением чисел из предыдущего столбца на числа столбца (3). Например, $97 \cdot 3 = 291$; $291 \cdot 3 = 873$; $873 \cdot 3 = 2619$; $21 \cdot (-4) = -84$; $(-84) \cdot (-4) = 336$; $336 \cdot (-4) = -1344$. Внизу

Таблица 14

l_i	n_i	x_i	$n_i x_i$	$n_i x_i^2$	$n_i x_i^3$	$n_i (x_i+1)^4$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
9	8	—6	—48	288	—1728	—1000
12	12	—5	—60	300	—1500	—768
15	21	—4	—84	336	—1344	—567
18	33	—3	—99	297	—891	—264
21	56	—2	—112	224	—448	—56
24	98	—1	—98	98	—98	0
27	183	0	0	0	0	183
30	256	1	256	256	256	2048
33	214	2	428	856	1712	5778
36	97	3	291	873	2619	6208
39	22	4	88	352	1408	2750
			501		6009	2655
			1063		5995	16967
Сумма	1000		562	3880	—14	14312

таблицы суммируются отдельно отрицательные и отдельно положительные числа, а еще ниже пишется общий итог. Например, сумма отрицательных чисел столбца (4) равна —501, а сумма положительных чисел равна +1063, так что получается итог +1063—501 = +562; для столбца (6) имеем +5995—6009 = —14.

В столбце (7) вычислена сумма $\Sigma n_i (x_i + 1)^3$. Она служит для проверки правильности вычислений. Действительно,

$$\Sigma n_i (x_i + 1)^3 = \Sigma n_i (x_i^3 + 3x_i^2 + 3x_i + 1) =$$
$$= \Sigma n_i x_i^3 + 3\Sigma n_i x_i^2 + 3\Sigma n_i x_i + \Sigma n_i,$$

так что должно быть

$$(7) = (6) + 3\cdot(5) + 3\cdot(4) + (2)$$

(числа в скобках обозначают суммы, стоящие в итогах соответствующих столбцов). В нашем примере

$$(6) + 3\cdot(5) + 3\cdot(4) + (2) =$$
$$= -14 + 3\cdot 3880 + 3\cdot 562 + 1000 = -14 + 11640 + 1686 + 1000 = 14312,$$

что совпадает с итогом столбца (7).

Ввиду того что при большом количестве вычислений арифметические ошибки почти неизбежны, контрольный расчет является совершенно необходимым. Выполнение его надо положить себе за правило, не допускающее никаких исключений.

Для получения начальных моментов надо разделить получившиеся итоги на объем совокупности. В данном случае

$$m_1 = \frac{+562}{1000} = +0{,}562; \quad m_2 = \frac{3880}{1000} = 3{,}880;$$
$$m_3 = \frac{-14}{1000} = -0{,}014.$$

Теперь по формулам (1.26) и (1.27) находим центральные моменты:

$$\mu_2 = 3{,}880 - (+0{,}562)^2 = 3{,}564 \approx 3{,}56;$$
$$\mu_3 = -0{,}014 - 3\cdot 3{,}880\,(+0{,}562) + 2\,(+0{,}564)^3 \approx -6{,}20.$$

Окончательно имеем:

$$\bar{x} = 27 + 0{,}562\cdot 3 = 27 + 1{,}68 = 28{,}68 \text{ мм};$$
$$\sigma = \sqrt{3{,}56}\cdot 3 = 1{,}89\cdot 3 = 5{,}67 \text{ мм};$$
$$A = \frac{-6{,}20}{3{,}56\cdot 1{,}89} = -0{,}92.$$

В терминах статистических моментов коэффициент асимметрии запишется так:

$$A = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}}. \tag{1.32}$$

Это есть частный случай величин вида

$$\rho_h = \frac{\mu_h}{\sigma^h}, \tag{1.33}$$

носящих название *основных моментов* (h-го порядка); основные моменты всегда безразмерны. Из (1.32) и (1.33) следует, что

$$A = \rho_3. \tag{1.34}$$

Основной момент четвертого порядка

$$\rho_4 = \frac{\mu^4}{\sigma^4} \tag{1.35}$$

употребляется при вычислении одной из важных характеристик распределений — так называемого эксцесса, о котором будет сказано подробней в следующей главе.

Очевидно, при любом распределении

$$\rho_1 = 0, \quad \rho_2 = 1, \tag{1.36}$$

так как $\mu_1 = 0$ и $\mu_2 = \sigma^2$.

ТЕОРЕТИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ

§ 1. Постановка задачи. Элементы теории вероятностей

Как мы уже знаем, знание статистических моментов позволяет описать немногими параметрами основные свойства найденного из опыта распределения. Однако часто бывает желательно дать аналитическое описание (т. е. в виде определенной формулы) непосредственно кривой плотности распределения. Имея уравнение кривой, можно вычислить плотность распределения для любого значения варианты, воспроизвести график распределения, выполнять разного рода вычисления и т. д.

Другое весьма важное значение аналитического описания кривой плотности распределения связано со следующим обстоятельством. Если имеется ряд точек на плоскости координат, то можно всегда подобрать такое уравнение, что изображающая его кривая пройдет через все заданные точки; существуют регулярные методы, позволяющие получить это уравнение, например, в виде многочлена. Однако такое описание данной совокупности точек было бы чисто формальным. Мы поставим другую задачу. Происхождение каждого эмпирического распределения обусловлено какими-то определенными естественными причинами. Совокупность причин, приводящих к тому или иному распределению, может быть в каждом случае иной. Задача будет состоять в том, чтобы представить себе, за счет каких причин могло получиться найденное распределение (т. е. построить подходящую математическую или физическую модель явления), а затем, исходя из сделанного предположения, вывести математически функцию распределения. После этого надо будет проверить справедливость сделанного предположения.

Очевидно, модель, приводящая к определенному распределению, всегда содержит какие-то числовые параметры, которые потом войдут в уравнение кривой плотности распределения.

Таким образом, желая получить не формальное, а теоретически обоснованное математическое описание какого-либо эмпирического распределения, мы должны решить следующие три задачи:

1) исходя из тех или иных предположений о происхождении данного распределения, подобрать надлежащую модель и тем самым выбрать определенный вид функции распределения;

2) найти численные значения параметров функции распределения, соответствующие свойствам данного эмпирического распределения;

3) проверить, правильно ли найденная функция распределения с вычисленными значениями параметров описывает эмпирическое распределение.

Сейчас мы укажем, каким образом решаются первые две задачи. Третья задача будет подробно рассмотрена в главах 4 и 6.

Основой для построения статистических моделей служит теория вероятностей. Предметом этой теории является изучение *случайных событий* и *случайных величин*.

Событие называется случайным, если оно при данных условиях может либо произойти, либо не произойти. Например, падение брошенной монеты гербом вверх может либо произойти, либо не произойти. Конечно, падение монеты той или иной стороной в действительности обусловлено вполне однозначно определенными материальными причинами. Но количество и сложность этих причин столь велики, что детерминированное описание движения монеты практически невозможно.

Было бы, однако, неправильным заключить отсюда, что в явлении падения монеты (или любом другом подобном явлении) вообще невозможно делать какие бы то ни было предсказания. Это верно лишь в отношении е д и н и ч н о г о случайного события. Если же случайное событие может многократно повторяться при одних и тех же условиях, то могут быть найдены определенные закономерности, которым подчиняются эти события. Изучением таких закономерностей и занимается теория вероятностей.

Каждое явление (бросание монеты, стрельба по цели, вынимание шара из урны и т. д.), в котором может осуществиться или не осуществиться случайное событие (выпадение герба, попадание в цель, появление шара определенного цвета и т. д.), называется *испытанием*. Предполагается, что на результат очередного испытания не влияют результаты предыдущих (и, разумеется, последующих) испытаний. Если при n испытаниях событие произошло m раз, то величину $h = m/n$ называют *относительной частотой*, или *частостью* события. Наблюдения показывают, что если производить в одинаковых условиях опыты, повторяя раз за разом серии из достаточно большого числа испытаний, то относительные частоты будут получаться довольно близкими. При этом оказывается, что чем больше число испытаний в каждой

серии, тем меньше разброс относительных частот. Это позволяет предположить, что имеется некоторая «объективная степень возможности» события, которая реализуется тем точнее, чем длиннее серия испытаний.

Пример 1. В трех опытах многократного бросания монеты были получены результаты, приведенные в табл. 15.

Таблица 15

Число испытаний (бросаний монеты)	Число событий (появлений герба)	Относительная частота
4000	2038	0,5095
12000	5981	0,4984
24000	12009	0,5004

Мы видим, что при увеличении числа испытаний относительная частота появления герба приближается к 0,5. Это число будем считать *вероятностью* появления герба при бросании монеты. Вероятность обычно обозначают буквой p или P.

Во многих случаях можно найти вероятность события чисто теоретически, не производя никаких испытаний, а лишь анализируя условия опыта. Например, если в урне имеется 10 шаров, из которых 7 белых и 3 черных, то вероятность вынуть наугад белый шар будет равна $p_б = 7 : 10 = 0{,}7$, а вероятность вынуть черный шар — $p_ч = 3\ \ 10 = 0{,}3$. Такое определение вероятности — по доле шансов, благоприятствующих данному событию, — принято называть *классическим*, в отличие от данного выше *статистического* определения вероятности (по предельному значению относительной частоты).

Если событие при данных условиях невозможно, то его вероятность равна нулю: $p = 0$. Если событие обязательно наступает при данных условиях (такое событие называют *достоверным*), то его вероятность равна единице: $p = 1$; действительно, поскольку это событие наступает при каждом испытании, то в любой серии испытаний число достоверных событий m всегда равно числу испытаний n, так что всегда $m : n = 1$. Следовательно, вероятность любого случайного события во всех случаях есть неотрицательное число, не превосходящее единицы: $0 \leqslant p \leqslant 1$.

Очевидно, если вероятность наступления какого-нибудь события равна p, то вероятность ненаступления этого события равна $1 - p$; эту величину обычно обозначают буквой q.

Два события, которые не могут осуществляться одновременно, называются *несовместными*; пример несовместных событий —

наступление и ненаступление какого-либо данного события. Можно показать, что, в силу классического определения вероятности, вероятность появления одного из двух несовместных событий A и B, безразлично какого, равна сумме вероятностей этих событий (теорема сложения вероятностей несовместных событий). Событие, состоящее в появлении либо события A, либо события B, либо обоих этих событий (если они совместны), называется суммой $A + B$ событий. Поэтому теорему сложения для несовместных событий можно записать в виде

$$p(A + B) = p(A) + p(B).$$

В общем случае k событий $A_1, A_2, \ldots, A_k$ $(k \geqslant 2)$ называют попарно несовместными, если любые два события A_i и A_j $(i \neq j)$ не могут осуществиться одновременно. Для попарно несовместных событий имеем

$$p(A_1 + A_2 + \ldots + A_k) = p(A_1) + p(A_2) + \ldots + p(A_k).$$

Два события называют *независимыми*, если вероятность одного из них не зависит от появления или непоявления другого. Точнее, события A и B независимы, если вероятность их одновременного осуществления равна произведению вероятностей событий A и B.

Событие, заключающееся в одновременном осуществлении событий A и B, называют произведением этих событий и обозначают AB. В этих обозначениях для независимых событий A и B имеем

$$p(AB) = p(A)\,p(B).$$

Отсюда следует, что несовместные события A и B, отличные от невозможных, всегда зависимы, так как в этом случае $p(AB) = 0$, но $p(A)p(B) \neq 0$.

В общем случае, когда события A и B зависимы, вероятность $p(B)$ не равна отношению $p(AB)/p(A)$. Это отношение представляет собой вероятность события B, вычисленную при условии, что событие A осуществилось.

Пример 2. В урне находится 7 белых и 3 черных шара. Вероятность вынуть белый шар при первом испытании равна $^7/_{10}$. Если после этого испытания шар не возвращается в урну, то вероятность вынуть белый шар при втором испытании зависит от результата первого испытания. А именно, если в первом испытании был вынут белый шар, то вероятность вынуть белый шар во втором испытании равна $^6/_9$; если же раньше был вынут черный шар, то эта вероятность равна $^7/_9$. Однако если после первого испытания шар возвращается в урну, то вероятность вынуть белый шар во

втором испытании не зависит от результата первого испытания; в этом случае события независимы [1].

Определение независимости можно распространить на произвольное количество событий. События A_1, A_2, . . . , A_k ($k > 2$) называются независимыми, если, во-первых, $p(A_1A_2 \ldots A_k) = p(A_1)p(A_2) \quad p(A_k)$ и, во-вторых, любые $k-1$ событий из общего числа k являются независимыми. Таким образом, например, три события A, B и C независимы тогда и только тогда, когда, во-первых, $p(ABC) = p(A)p(B)p(C)$ и, во-вторых, $p(AB) = p(A)p(B)$, $p(AC) = p(A)p(C)$ и $p(BC) = p(B)p(C)$.

Если обозначить две стороны монеты двумя числами (например 0 и 1), то число, которое появляется при бросании монеты, можно рассматривать как *случайную величину*, могущую случайным образом принимать одно из двух возможных значений. Однако существование лишь двух возможных значений — это только простейший случай. Обычно случайная величина может принимать большее (и даже бесконечное) количество значений. Так, при бросании игральной кости может случайно выпасть одно из шести значений; при толкании маховика число положений, в которых он может остановиться, бесконечно (от 0 до 360°).

Принято различать *дискретные* и *непрерывные* случайные величины. Случайную величину называют дискретной, если любые ее два значения отделены друг от друга конечным промежутком. Простейшим примером дискретной случайной величины является случайная величина, принимающая конечное число возможных значений (например, целочисленные значения 1, 2, 3, 4, 5, 6, появляющиеся при бросании игральной кости). Чтобы полностью охарактеризовать дискретную случайную величину x, нужно не только перечислить ее возможные значения x_1, x_2, , x_k, . , но и указать вероятности p_1, p_2, , p_k, ., соответствующие этим значениям (т. е. вероятности, с которыми случайная величина принимает каждое из этих значений). Перечень возможных значений и соответствующих им вероятностей

$$x : x_1, x_2, \ldots, x_k, .$$

$$p : p_1, p_2, \ldots, p_k,$$

---

[1] Очевидно, если число шаров в урне очень велико, то изъятие одного шара мало сказывается на распределении шаров по цветам. Поэтому вероятность вынуть белый шар во втором испытании почти не зависит от результата первого испытания даже в том случае, когда шар после первого испытания не возвращается в урну. Следовательно, события при испытаниях без возвращения шара можно считать в первом приближении независимыми, если число шаров в урне очень велико.

называется *законом распределения вероятностей* для дискретной случайной величины. Так как при любом испытании случайная величина x обязательно примет одно из возможных значений x_i, то сумма вероятностей событий, состоящих в том, что x примет значение x_1, или x_2, . , или x_k ,. . , равна единице:

$$p_1 + p_2 + \quad + p_k + \ldots = 1.$$

Примером непрерывной случайной величины может служить угол ω, характеризующий положение маховика в момент остановки. Вероятность попадания ω в заданный интервал углов $\Delta\omega$ тем больше, чем шире этот интервал (при $\Delta\omega = 360°$ эта вероятность, очевидно, равна единице). Если маховик хорошо центрирован, то такая вероятность пропорциональна ширине интервала $\Delta\omega$, т. е. $\Delta P = p \cdot \Delta\omega$ (коэффициент пропорциональности p в данном случае равен 1/360°, так как, согласно только что сделанному замечанию, при $\Delta\omega = 360°$ должно выполняться равенство $\Delta P = 1$).

Из формулы $\Delta P = p \cdot \Delta\omega$ следует, что событие $\omega = \omega_0$ имеет вероятность, равную нулю. Действительно, ведь условие, что величина ω примет заданное значение ω_0, эквивалентно условию, что ω попадет в интервал с «нулевой шириной» $\Delta\omega = 0$; поэтому здесь $\Delta P = 0$. Этот факт является общим для всех непрерывных случайных величин.

Отсюда ясно, что распределение вероятностей непрерывной случайной величины не имеет смысла задавать указанием вероятностей для каждого из возможных значений, как это делается для дискретных случайных величин. Объективной характеристикой распределения непрерывной случайной величины служит упомянутый выше коэффициент пропорциональности p. В предыдущем примере этот коэффициент представлял собой отношение $p = \Delta P/\Delta\omega$. По аналогии с плотностью массы однородного вещества $p = \Delta M/\Delta V$ (ΔM — масса, заключенная в объеме ΔV) величину $p = \Delta P/\Delta\omega$ называют *плотностью вероятности*.

Как уже отмечалось выше, если маховик центрирован правильно, то плотность вероятности одинакова для всех положений интервала $\Delta\omega$ и равна 1/360°. В общем же случае плотность вероятности для произвольной непрерывной случайной величины непостоянна и зависит от положения рассматриваемого интервала Δx ее возможных значений на оси x. Поэтому плотность вероятности обозначают $p(x)$. Переменную плотность вероятности (т. е. зависящую от x) определяют как предел, к которому стремится отношение $\Delta P/\Delta x$ при $\Delta x \to 0$. Иными словами, вероятность ΔP попадания случайной величины в интервал, заключенный между x и $x + \Delta x$, приближенно равна $p(x) \cdot \Delta x$, причем с уменьшением ин-

тервала Δx относительная погрешность формулы $\Delta P \approx p(x) \cdot \Delta x$ уменьшается.

Конкретный вид закона распределения зависит от характера модели явления. В последующих параграфах этой главы будут рассмотрены некоторые простейшие модели, имеющие наибольшее значение для статистической практики.

§ 2. Биномиальное распределение

Так называемое *биномиальное распределение*, или *распределение Бернулли*, получается в том случае, когда для каждого испытания существует лишь два возможных несовместных исхода. Например, в результате взаимодействия многих естественных и искусственных факторов опухоль может либо уменьшиться, либо не уменьшиться. При этом мы не принимаем во внимание, что исход «не уменьшится» может быть подразделен на исходы «останется без изменения» и «увеличится». Аналогично при бросании игральной кости мы также можем различать только две альтернативы, например, выпадение двух очков и выпадение не двух очков, не интересуясь тем, что выпадение не двух очков включает в себе исходы «выпадение единицы», «выпадение тройки» и т. д.

Пусть мы бросили одовременно ν игральных костей. Какова вероятность того, что двойка выпадет на определенных x из них? Чтобы такое событие произошло, должны одновременно осуществиться следующие ν событий: на отмеченных x костях должна выпасть двойка (вероятность каждого такого события равна $p = {}^1/_6$), а на остальных $\nu - x$ костях должна выпасть не двойка (вероятность каждого такого события равна $q = {}^5/_6$). Поэтому вероятность того, что на отмеченных x костях выпадет двойка, а на остальных $\nu - x$ костях выпадет не двойка, равна, согласно условию независимости событий, $p^x q^{\nu - x}$. Так, если брошено одновременно 12 костей, то вероятность того, что на костях № 2, 7 и 10 выпадет двойка, а на остальных 9 костях (с номерами 1, 3, 4, 5, 6, 8, 9, 11, 12) выпадет не двойка, будет

$$\left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^{12-3} = \frac{5^9}{6^{12}} \approx 0{,}00089.$$

Теперь изменим немного условия задачи. Нас будет интересовать выпадение двойки не обязательно именно на о т м е ч е н н ы х x костях, а на любых x костях. В рассмотренном примере — не обязательно на костях № 2, 7 и 10, а на любой тройке костей. Очевидно, вероятность такого события будет больше во столько раз, сколько различных троек костей можно выбрать из двенадцати костей. Как известно, число таких троек есть число сочетаний из двенадцати элементов по три, обозначаемое C_{12}^3 и равное

$$C_{12}^3 = \frac{12 \cdot 11 \cdot 10}{1 \cdot 2 \cdot 3} = 220,$$

Поэтому вероятность того, что при одновременном бросании 12 костей двойка выпадет хоть на какой-нибудь тройке костей, будет равна

$$\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^{12-3} C_{12}^3 = 0{,}00089 \cdot 220 \approx 0{,}196.$$

В общем виде формула для вероятности того, что при одновременном бросании ν костей двойка выпадет на каких-нибудь x костях, будет

$$w_x = C_\nu^x p^x q^{\nu-x}. \qquad (*)$$

Суммирование по всем возможным значениям x (очевидно, от 0 до ν) должно дать полную вероятность, т. е. 1. И действительно,

$$\sum_{x=0}^{\nu} C_\nu^x p^x q^{\nu-x} = 1,$$

ибо эта сумма есть нечто иное, как развернутая формула бинома Ньютона:

$$(p+q)^\nu = \sum_{x=0}^{\nu} C_\nu^x p^x q^{\nu-x}; \qquad (**)$$

но левая часть равенства (**) равна 1, так как $p + q = 1$.

Если произведено N бросаний по ν костей в каждом, то следует ожидать, что число бросаний, в которых двойка выпадет на x костях, будет близко к

$$n_x = Nw_x = NC_\nu^x p^x q^{\nu-x}. \qquad (2.1)$$

Например, если 1000 раз бросить горсть из 12 костей, то число бросаний, в которых двойка выпадет на трех костях из этих двенадцати, должно быть близко к

$$n_x = 1000 \cdot 0{,}196 = 196.$$

Конечно, фактически таких бросаний может оказаться несколько больше или меньше (об этом см. гл. 3), но нас интересует сейчас «теоретическое» число нужных результатов.

Распределение частот (2.1) называется *биномиальным*. Сопоставление формул (*) и (**) делает понятным происхождение этого названия. Биномиальные коэффициенты в общем виде записываются так:

$$C_\nu^x = \frac{\nu(\nu-1)(\nu-2)\ldots(\nu-[x-1])}{x!}. \qquad (2.2)$$

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 — быстрей, чем показательная функция.

Очевидно,

$$k! = (k-1)!\,k,$$

откуда

$$(k-1)! = \frac{k!}{k}. \qquad (***)$$

Так как $1! = 1$, то в соответствии с (***) естественно считать

$$0! = \frac{1!}{1} = \frac{1}{1} = 1,$$

хотя это и выглядит несколько парадоксально.

Сильно облегчает нахождение биномиальных коэффициентов при разных ν так называемый треугольник Паскаля (табл. 16). Каждый коэффициент образуется сложением двух, стоящих над ним (слева и справа), коэффициентов; например, $15 = 5 + 10$, $84 = 28 + 56$ и т. д. Пользуясь этим правилом, продолжают треугольник Паскаля до нужного значения ν.

Таблица 16

	Биномиальные коэффициенты
0	1
1	1 1
2	1 2 1
3	1 3 3 1
4	1 4 6 4 1
5	1 5 10 10 5 1
6	1 6 15 20 15 6 1
7	1 7 21 35 35 21 7 1
8	1 8 28 56 70 56 28 8 1
9	1 9 36 84 126 126 84 36 9 1

Особый интерес представляет случай, когда вероятности двух альтернативных исходов одинаковы: $p = q = 1/2$. Этот случай имеет место, например, при бросании монеты, при рождении особи

мужского или женского пола и т. д.

Если $p = q = 1/2$, то

$$p^x q^{\nu-x} = \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{\nu-x} = \left(\frac{1}{2}\right)^{\nu} = \frac{1}{2^{\nu}},$$

так что

$$n_x = \frac{N}{2^{\nu}} C_{\nu}^{x}. \qquad (2.3)$$

Пример 3. В колонках 1 и 2 табл. 17 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 6 выводках было по 3 петушка, в 16 выводках — по 7 петушков и т. д. Посмотрим, можно ли это распределение считать

Таблица 17

Число петушков	Фактическое число выводков	Биномиальные коэффициенты	Теоретическое число выводков
1	2	3	4
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

биномиальным. Для этого нужно по формуле (2. 2) или из треугольника Паскаля найти биномиальные коэффициенты C_{ν}^{x} для разных x при $\nu = 12$, после чего по формуле (2. 3) вычислить значения n_x. 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}^{x} и n_x записаны соответственно в колонках 3 и 4 табл. 17; значения n_x получены умножением C_{12}^{x} 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. Для того чтобы оценить степень совпадения более объективно, а тем более количественно, нужны какие-то определенные критерии. О них будет сказано в следующих главах (в частности, в гл. 6).

§ 3. Нормальное распределение

Пусть мы имеем некоторое биномиальное распределение с определенным значением параметра ν, скажем, $\nu = 10$. На рис. 12, *а* изображена диаграмма такого распределения при $p = {}^1/_2$; столбики этой диаграммы могут изображать, например, вероятности выпадения герба на x монетах при бросании горсти из $\nu = 10$ монет или вероятности рождения x самцов в пометах из 10 особей и т. д. За начало отсчета величин x примем значение $x_0 = \nu p = 5$, при котором вероятность максимальна, т. е. будем отсчитывать x влево и вправо от x_0.

Если бросать горсть из 100 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 $\nu = 10$.

Расчет показывает, что дисперсия σ^2 биномиального распределения примерно пропорциональна показателю степени бинома ν. Поэтому стандартное отклонение σ изменяется при изменении ν приблизительно как $\sqrt{\nu}$. Поскольку ширина распределения равна $\nu + 1$, то гистограмма, получившаяся в результате произведенной нами деформации, будет шире, чем диаграмма на рис. 12, *а*. Если мы захотим сохранить прежнюю ширину рисунка, то на нем поместится лишь часть новой диаграммы. В нашем примере ширина диаграммы увеличилась в $101 : 11 \approx$ $\approx 9{,}2$ раза, а стандартное отклонение — примерно в $\sqrt{100} : \sqrt{10} \approx$ $\approx 3{,}2$ раза. Так как мы сжали диаграмму в 3,2 раза, то рисунок вместил только $3{,}2 \quad 9{,}2 \approx 0{,}35$ всей диаграммы; ее «хвосты» не поместились на рисунке.

После того как мы произведем такие надлежащую деформацию по вертикали (чтобы обеспечить сохранение неизменной площади всей диаграммы), получится фигура, изображенная на рис. 12, *б*; повторяем, что это лишь средняя часть всей диаграммы.

На рис. 12, *в* изображена диаграмма для $\nu = 1000$, преобразованная аналогичным образом. Точнее говоря, это опять-таки лишь центральная часть диаграммы: ведь по сравнению со случаем $\nu = 10$ ширина диаграммы увеличилась в $1001 : 11 \approx 91$ раз, между тем как сужение графика было произведено во столько раз, во сколько увеличилось стандартное отклонение, т. е. лишь примерно в $\sqrt{1000} : \sqrt{10} = 10$ раз. Следовательно, непоместившиеся на рисунке «хвосты» диаграммы здесь еще больше: поместившаяся на рисунке центральная часть составляет лишь $10/91 \approx 0{,}11$ всей диаграммы (но только по ширине; по площади же эта центральная часть содержит подавляющую долю всей вероятности).

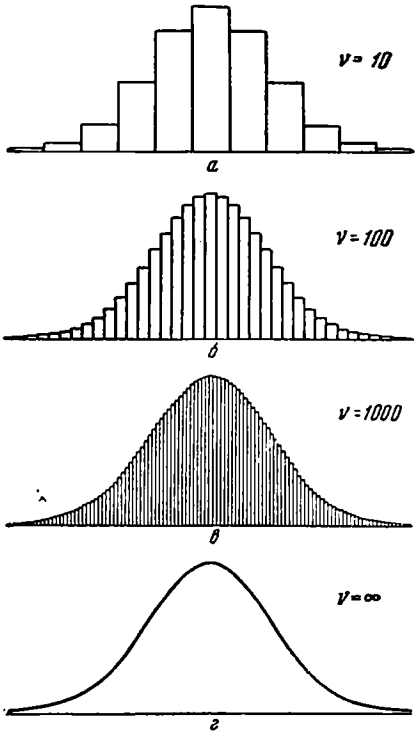

Рис. 12

Сравнение рис. 12, *а*, *б* и *в* показывает, что общий характер распределения во всех трех случаях одинаков, но это распределение тем больше детализировано, чем больше значение ν.

Если продолжить эту процедуру, беря все большие значения ν, то в пределе при $\nu \to \infty$ верхние стороны столбиков сольются в гладкую кривую (рис. 12, *г*). При этом «хвосты» графика распространятся неограниченно далеко по обе стороны от центра. Получившееся предельное распределение называется *нормальным* или *гауссовым* (*распределением Гаусса*). Это распределение обычно получается при совместном воздействии ряда малых независимых (значит, случайно сочетающихся) факторов, число которых неограниченно велико. Такое условие (одновременное воздей-

ствие 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.

Уравнение гауссовой кривой имеет вид

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

Здесь $\hat{x}$ и σ — среднее значение и стандартное отклонение. Величина $e = 2{,}718\ldots$ есть основание натуральных логарифмов. Выбор именно этой величины в качестве основания показательной функции диктуется соображениями чисто математического порядка, на которых мы не останавливаемся. Укажем лишь, что e есть предел, к которому стремится выражение $\left(1 + \frac{1}{k}\right)^k$ при $k \to \infty$; кроме того, e есть сумма бесконечного ряда

$$\sum_{k=0}^{\infty} \frac{1}{k!} = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \qquad 1 + 1 + \frac{1}{2} + \frac{1}{6} + \ldots$$

Из рис. 12 видно, что нормальное распределение является симметричным, т. е. положительное и отрицательное отклонения равной величины встречаются одинаково часто. Далее видим, 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). Далее, на форму 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, а ее максимальная высота тем ниже; обратно, при малых σ кривая уже и выше. Значит, при малых σ кривая плотности нормального распределения «стягивается» к середине, а при больших σ она «расплывается» в стороны (рис. 13).

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

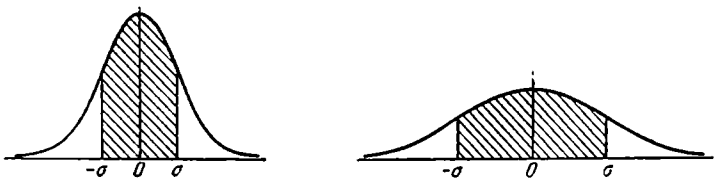

Рис. 13

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. 13. Расчет показывает, что ее величина составляет $\sim 0{,}683$ всей площади. Следовательно, в среднем 68,3% (или примерно две трети) всех вариант отклоняются от среднего значения не больше чем на величину среднего квадратичного отклонения.

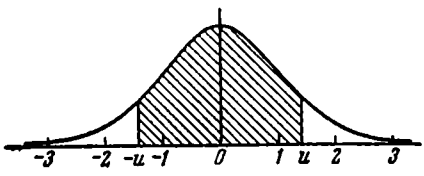

Рис. 14

Аналогично можно получить, что в пределах от -2σ до $+2\sigma$ лежит 95,5% всех вариант, в пределах от -3σ до $+3\sigma$ — 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. 14), с шагом $\Delta u = 0{,}1$ или даже 0,01; аргумент $u = \xi/\sigma$ выбран для удобства безразмерным. Эта доля вариант обозначается $\theta(u)$. Табл. I 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 опущены. Как увидим далее, табл. I приходится пользоваться довольно часто.

Во многих случаях удобно пользоваться таблицей, указывающей долю вариант, лежащих левее заданной абсциссы (заштрихованная часть на рис. 15). Такая величина обозначается $\Phi(u)$

n называется *интегралом вероятностей*; название этой функции связано с тем, что вычисление площадей, ограниченных кривыми, сводится к математической операции интегрирования (вычисления интеграла).

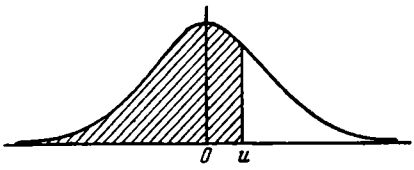

Рис. 15

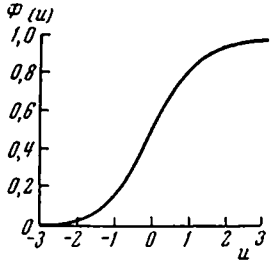

Рис. 16

Из сравнения рис. 14 и 15 следует

$$\Phi(u) = \frac{1}{2} + \frac{\theta(u)}{2} = \frac{1}{2}[1 + \theta(u)]. \qquad (2.5)$$

Это позволяет находить значения $\Phi(u)$, пользуясь табл. I Приложений. Впрочем, существуют таблицы, содержащие непосредственно значения $\Phi(u)$. График функции $\Phi(u)$ представлен на рис. 16.

Разумеется, доля вариант, лежащих правее заданной абсциссы, равна $1 - \Phi(u)$; если учесть (2.5), то эта доля может быть выражена как

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

Пример 4. Из 500 студентов 27 имеют рост, превышающий 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 на рис. 14 равна $u = 10/\sigma$. По условию задачи правый незаштрихованный «хвост» на графике приблизительно составляет $\frac{27}{500} = 0{,}054$ всей площади под кривой. Подставляя эти значения в формулу (2.6), имеем

$$0{,}054 \approx \frac{1}{2}\left[1 - \theta\left(\frac{10}{\sigma}\right)\right],$$

откуда $\theta\left(\frac{10}{\sigma}\right) \approx 1 - 2 \cdot 0{,}054 \approx 1 - 0{,}108 = 0{,}892.$

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

Многие вопросы, касающиеся нормального распределения, приобретают большую наглядность, если применить так называемый метод спрямления нормальной кривой.

Пусть мы имеем нормально распределенную совокупность, частости которой задаются формулой (2.4): $\nu_i \quad \varphi(u_i)$. Тогда для накопленных частостей будем иметь

$$z_i = \Phi(u_i) = \Phi\left(\frac{x_i - \hat{x}}{\sigma}\right), \tag{2.7}$$

где $\Phi(u)$ — интеграл вероятностей. Построим теперь график, откладывая по оси ординат (т. о. по вертикальной оси) не значения z_i, а значения

$$y_i = \Psi(z_i),$$

где Ψ есть функция, обратная к интегралу вероятностей Φ. Последнее означает, что

$$z_i = \Phi(y_i),$$

а тогда сравнение с (2.7) дает

$$y_i = \frac{x_i - \hat{x}}{\sigma}, \tag{2.8}$$

что можно переписать в виде

$$y_i = \frac{1}{\sigma} x_i - \frac{\hat{x}}{\sigma},$$

т. е. в виде уравнения прямой линии. Следовательно, если откладывать по оси абсцисс значения x_i, а по оси ординат — значения $y_i = \Psi(z_i)$, то точки расположатся вдоль прямой линии (имеющей наклон $1/\sigma$). Значения $\Psi(z_i)$ можно находить из таблицы функции $\Phi(u)$, но с целью упрощения расчетов составлена специальная таблица для функции Ψ (табл. II Приложений; аргумент этой функции обозначен там через p).

Пример 5. В табл. 18 проведен расчет для данных из табл. 9. Как видно из рис. 17, точки с координатами x_i, $\Psi(z_i)$ хорошо укладываются на прямой; самые крайние точки не надо при этом принимать во внимание, так как они отвечают очень малым частотам $n_i = 1$.

График $\Psi(z_i)$ позволяет найти приближенные значения параметров $\hat{x}$ и σ заданной совокупности. Действительно, поскольку нормальное распределение симметрично, то среднее значение является одновременно медианой; поэтому среднему значению $\hat{x}$ соответствует накопленная частость $z_i = 0{,}500$ и, следовательно, значение $\Psi(z_i) = 0$. Значит, $\hat{x}$ есть абсцисса той точки прямой,

Таблица 18

Рост, см	Частоты n_i	Частости ν_i	Накопленные частости z_i	$\Psi(z_i)$
143	1	0,001	0,001	—3,09
146	2	0,002	0,003	—2,75
149	8	0,008	0,011	—2,29
152	26	0,026	0,037	—1,79
155	65	0,065	0,102	—1,27
158	120	0,120	0,222	—0,77
161	179	0,179	0,401	—0,25
164	201	0,201	0,602	0,26
167	172	0,172	0,774	0,75
170	120	0,120	0,894	1,25
173	64	0,064	0,958	1,73
176	28	0,028	0,986	2,20
179	10	0,010	0,996	2,65
182	3	0,003	0,999	3,09
185	1	0,001	1,000	

которая имеет ординату $\Psi(z_i) = 0$ (на рис. 17 сразу находим $\hat{x} \approx 162{,}5$ *см*). Далее, точки прямой, имеющие ординаты -1 и $+1$, соответствуют значениям x_i, равным $\hat{x} - \sigma$ и $\hat{x} + \sigma$. Так,

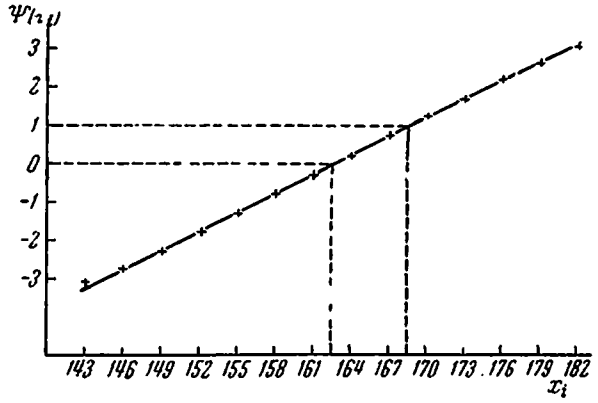

Рис. 17

на рис. 17 ординатам -1 и $+1$ отвечают абсциссы $\sim 156{,}5$ и $\sim 168{,}5$, поэтому

$$\sigma = x(+1) - x(0) \approx 168{,}5 - 162{,}5 = 6 \text{ см}$$

или

$$\sigma = x(0) - x(-1) \approx 162{,}5 - 156{,}5 = 6 \text{ см}.$$

Более точное значение σ может дать выражение

$$\sigma = \frac{1}{2}[x(+1) - x(-1)],$$

которое представляет собой по существу усреднение

$$\sigma = \frac{1}{2}\{[x(+1) - x(0)] + [x(0) - x(-1)]\}.$$

§ 4. Отклонения от нормального распределения

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. Таким образом, критерием отклонения эмпирического распределения от нормального может служить отклонение эмпирического значения ρ_3 от значения $\rho_3 = 0$, отвечающего нормальному распределению.

Однако выполнение условия $\rho_3 = 0$ еще не означает, что распределение нормально. Дело в том, что для нормального распределения характерно не только равенство нулю всех нечетных основных моментов (вследствие симметричности этого распределения), но и вполне определенные значения ч е т н ы х основных моментов; в частности, для нормального распределения имеем $\rho_4^{\text{норм}} = 3$. Поэтому, если $\rho_4^{\text{эмп}}$ заметно отличается от 3, то распределение не может считаться нормальным, даже если оно симметрично; $\rho_4^{\text{эмп}}$ вычисляется по формулам (1.28), (1.29) и (1.35).

Выясним, какой смысл имеет то обстоятельство, что ρ_4 равен не 3, а больше или меньше этой величины. Для этого мы сравним три симметричных распределения с одинаковыми объемами N, средними значениями $\hat{x}$ и стандартными отклонениями σ, но с различными ρ_4; одно из этих распределений является нормальным, так что для него $\rho_4 = 3$, для второго $\rho_4 > 3$ и для третьего $\rho_4 < 3$. Эти три распределения изображены на рис. 18. При $\rho_4 > 3$

кривая называется островершинной, а при $\rho_4 < 3$ — туповершинной. Вообще же свойство кривой, отражаемое значением ρ_4, называют *эксцессом* кривой. Обычно эксцесс кривой распределения характеризуют не значением ρ_4, а отклонением его от нормального значения $\rho_4^{\text{норм}} = 3$. Величина

$$E = \rho_4 - 3 \qquad (2.9)$$

называется *коэффициентом эксцесса* распределения. При $\rho_4 > 3$ эксцесс положителен, а при $\rho_4 < 3$ он отрицателен.

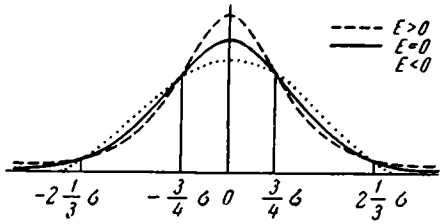

Рис. 18

Теория показывает, что отрицательные коэффициенты эксцесса имеют нижнюю границу: $E \geqslant -2$ (т. е. ρ_4 не может быть меньше единицы), а положительные эксцессы могут иметь любые значения.

Нахождение коэффициента асимметрии и эксцесса требует подчас довольно громоздких вычислений. Однако качественное заключение об отклонении распределения от нормального можно сделать сравнительно просто, используя метод спрямления нормальной кривой, изложенный в предыдущем параграфе.

Именно, если распределение асимметрично, то линия искривляется, причем знак кривизны связан со знаком асимметрии: при $A > 0$ кривая выпукла, а при $A < 0$ она вогнута (рис. 19). Действительно, пусть асимметрия положительна; тогда, если двигаться от середины распределения к краям, то для отрицательных

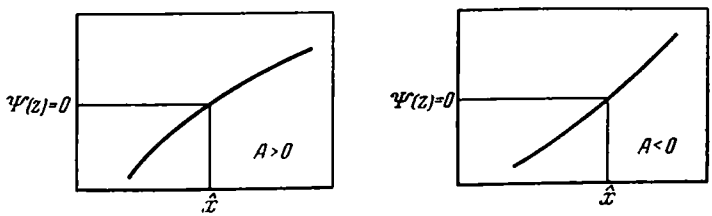

Рис. 19

отклонений накопленные частости убывают быстро, а для положительных отклонений они возрастают медленно. Если же асимметрия отрицательна, то дело обстоит наоборот: при движении в сторону отрицательных отклонений накопленные частости медленно

убывают, а при движении в сторону положительных отклонений они быстро возрастают.

При наличии эксцесса линия принимает S-образную форму, т. е. концы кривой загибаются в противоположные стороны; вид этой S-образности определяется знаком эксцесса (рис. 20): при положительном эксцессе накопленные частости на обоих концах ряда приближаются к своим предельным значениям (0 и 1) медленно, при отрицательном эксцессе — быстро.

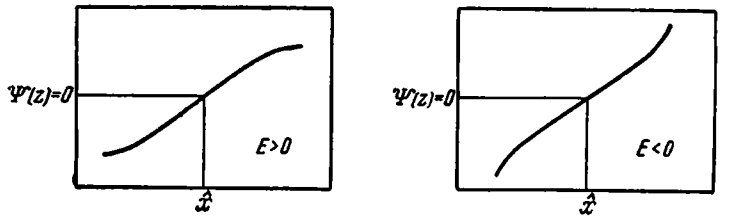

Рис. 20

Впрочем, описанный способ дает несколько преувеличенное представление об отклонении распределения от нормального. Дело в том, что накопленные частости не являются независимыми. Действительно, если какая-нибудь накопленная частость оказалась случайно завышенной (или заниженной), то следующая за ней накопленная частость, вероятней всего, тоже окажется завышенной (или, соответственно, заниженной), так как она ведь включает в себя предыдущую накопленную частость. Это и создает иногда видимость *систематического* отклонения от прямой.

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

Одной из причин наличия асимметрии и эксцесса 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 N, $\hat{x}$ и σ. Так, если мы будем строить распределение по 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 кри-

вая распределения будет двухвершинной (рис. 21, *а*); при небольшом различии средних значений кривая имеет одну, но тупую вершину (рис. 21, *б*). На рис. 22 представлены два других частных случая, приводящих к негауссову распределению (тонкие

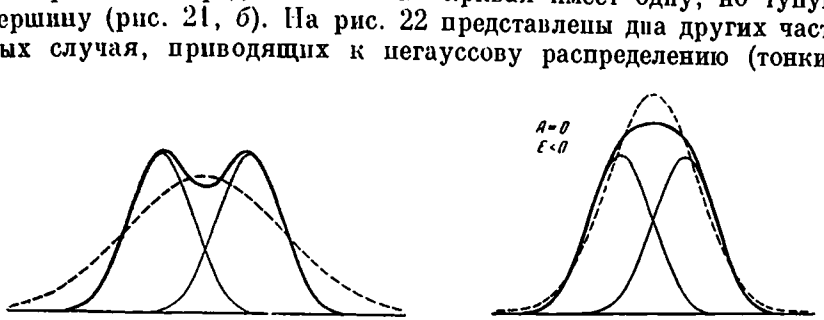

Рис. 21 а, б

линии —графики плотностей смешиваемых нормальных подсовокупностей; жирные линии—графики плотностей суммарных неоднородных совокупностей; пунктирные линии—графики нормальных совокупностей, имеющих такую же дисперсию, что и суммарные совокупности).

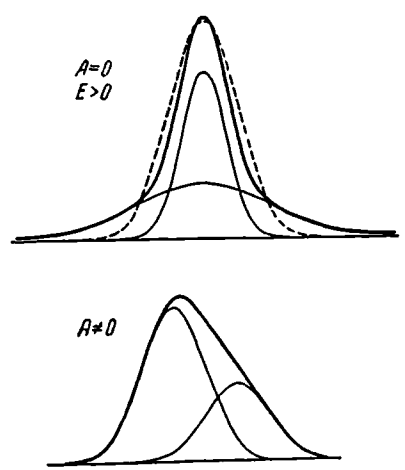

Рис. 22

Разумеется, вряд ли кто-нибудь 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 с определенным сдвигом по исследуемому признаку; ясно, что в этом случае неоднородность совокупности будет весьма желаемым результатом. Изучение характера распределения поз-

воляет выявить это расщепление популяции — выделение нового сорта, породы и т. д. — на очень ранней стадии и тем самым верно выбрать направление дальнейших поисков.

В ряде случаев (например, при отсутствии асимметрии или при некоторых упрощающих предположениях) 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). Поставленную задачу можно решить, основываясь на том, что начальный момент k-го порядка для суммарной совокупности равен сумме начальных моментов этого же порядка для составляющих подсовокупностей (причем все моменты берутся с «весами», равными объемам совокупностей); в данном случае

$$Nm_k = N'm'_k + N''m''_k. \qquad (*)$$

Дальнейшее упрощение состоит в предположении, что стандартные отклонения σ' и σ'' обеих подсовокупностей мало различаются; такой случай может иметь место, если, например, часть популяции подверглась одностороннему воздействию какого-либо фактора, приводящему к смещению среднего значения, но мало влияющему на дисперсии[1]. В этом случае задача сводится к отысканию пяти неизвестных параметров N', N'', $\hat{x}'$, $\hat{x}''$ и $\sigma' = \sigma''$, для чего требуется пять уравнений типа (*), т. е. уравнения для моментов пяти первых порядков (с нулевого по четвертый). Не приводя здесь всей цепи алгебраических выкладок, дадим сразу конечный результат — точнее, практическую процедуру вычислений[2].

Прежде всего, определяем для заданной совокупности коэффициенты асимметрии A и эксцесса E и вычисляем величину E^3/A^4. После этого из табл. 19 находим для полученного значения E^3/A^4 величину $\nu' = N'/N$ и некоторое число d, а затем вычисляем:

$$N' = \nu' N; \quad N'' = N - N';$$

$$\hat{x}' - \hat{x} = \sigma d \frac{E}{A}; \quad \hat{x}'' - \hat{x} = -\frac{N'}{N''}(\hat{x}' - \hat{x});$$

$$\sigma' = \sigma'' = \sqrt{\sigma^2 + (\hat{x}' - \hat{x})(\hat{x}'' - \hat{x})}$$

(так как $\hat{x}' - \hat{x}$ и $\hat{x}'' - \hat{x}$ имеют разные знаки, то $\sigma' = \sigma''$ м е н ь ш е, чем σ).

---

[1] Последнее будет в том случае, если мала дисперсия как самого воздействующего фактора, так и чувствительности к нему организмов.

[2] Подробней см.: В. Ю. У р б а х. Биофизика, 1961, т. VI, № 1 и 3. Там же рассмотрены некоторые другие частные случаи.

Таблица 19

E^3/A^4		d	E^3/A^4		d	E^3/A^4		d
14	044	1,15	0,01	198	10,24	—20	338	—0,618
12	050	1,20	10^{-4}	205	20,7	—25	348	—0,558
10	055	1,22	10^{-6}	210	95,5	—30	352	—0,528
8	062	1,26	-10^{-4}	215	—34,4	—40	357	—0,500
6	072	1,33	-10^{-3}	219	—17,5	—60	365	—0,441
5	079	1,38	—0,01	227	—8,12	—80	372	—0,401
4	086	1,44	—0,02	231	—6,45	—100	378	—0,370
3	096	1,53	—0,05	237	—4,71	—120	383	—0,348
2,5	102	1,60	—0,10	243	—3,71	—150	389	—0,312
2,0	110	1,69	—0,2	251	—2,97	—200	396	—0,286
1,5	118	1,80	—0,3	256	—2,56	—300	405	—0,256
1,2	125	1,92	—0,5	264	—2,13	—400	411	—0,232
1,0	131	2,03	—0,7	270	—1,84	—600	418	—0,207
0,8	136	2,14	—1,0	277	—1,61	—800	423	—0,191
0,6	144	2,36	—1,5	284	—1,41	—1000	428	—0,174
0,4	153	2,65	—2,0	289	—1,29	—2000	437	—0,148
0,3	158	2,84	—3	297	—1,14	—4000	446	—1,123
0,2	166	3,29	—5	304	—1,02	—6000	452	—0,108
0,15	171	3,68	—7	310	—0,930	-10^4	459	—0,0938
0,10	177	4,25	—10	317	—0,838	$-2\cdot10^4$	463	—0,0808
0,05	183	5,19	—12	322	—0,790	$-4\cdot10^4$	468	—0,0681
0,03	191	6,84	—16	331	—0,693	-10^5	474	—0,0544

Примечание. Для краткости в значениях v' опущены нуль и запятая.

Пример 6. В табл. 20 и на рис. 23 дано распределение $N = 5588$ красных бобов по длине. Предполагая, что отклонение этого распределения от нормального вызвано неоднородностью совокупности, найти численности, средние значения и стандартные отклонения двух подсовокупностей, из которых, как предполагается, состоит заданная совокупность.

Таблица 20

Длина, *мм* .	17,5	18,5	19,5	20,5	21,5	22,5	23,5	24,5
Частоты	12	51	184	228	530	692	855	768
Длина, *мм* .	25,5	26,5	27,5	28,5	29,5	30,5	31,5	32,5
Частоты	723	561	389	250	216	83	37	9

Для этого распределения расчеты дают $\hat{x} = 24{,}48$ *мм*, $\sigma^2 = 7{,}21$ *мм*^{2}, $\sigma = 2{,}68$ *мм*, $A = 0{,}215$, $E = -0{,}218$. Вычислив

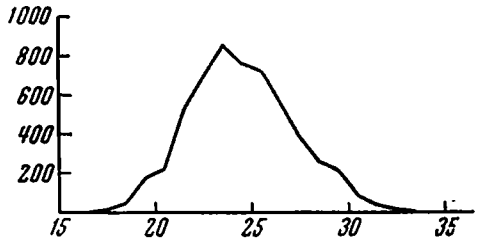

Рис. 23

$$\frac{E^3}{A^4} = \frac{(-0{,}248)^3}{0{,}215^4} = -7{,}15,$$

находим из табл. 19:

$$\nu' \approx 0{,}310; \quad d \approx -0{,}930.$$

Тогда

$$N' \approx 0{,}310 \cdot 5588 \approx 1730;$$

$$N'' \approx 5588 - 1730 \approx 3860;$$

$$\hat{x}' - \hat{x} = \frac{2{,}68\,(-0{,}930)\,(-0{,}248)}{0{,}215} \approx 2{,}88 \text{ мм};$$

$$\hat{x}'' - \hat{x} = -\frac{1730}{3860} \cdot 2{,}88 \approx -1{,}29 \text{ мм};$$

$$\hat{x}' = 24{,}48 + 2{,}88 = 27{,}36 \text{ мм};$$

$$\hat{x}'' = 24{,}48 - 1{,}29 = 23{,}19 \text{ мм};$$

$$\sigma' = \sigma'' = \sqrt{7{,}21 - 2{,}88 \cdot 1{,}29} = \sqrt{7{,}21 - 3{,}72} = \sqrt{3{,}49} = 1{,}87 \text{ мм}.$$

Необходимо подчеркнуть, что изложенный выше (или любой другой) математический анализ должен играть лишь подсобную роль — вопрос об однородности или неоднородности заданной статистической совокупности может быть решен окончательно только на основе биологического анализа исследуемого материала.

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

Пример 7. В табл. 21 приведено распределение по весу зерен пшеницы: были взяты пробы со 100 делянок и для каждой из этих проб определен средний вес зерна. Из рис. 24, изображающего полигон частот этого распределения, видно, что оно симметрично. Расчет подтверждает это: коэффициент асимметрии $A = 0{,}06$, т. е. весьма мал. Возьмем теперь в качестве признака, 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. 21 и на рис. 24 под соответствующими числами x_i. Естественно, шкала значений x_i получилась неравномерной. Если мы хотим взять за основу числа x_i, то следует сделать шкалу этих значений равномерной. Это может быть

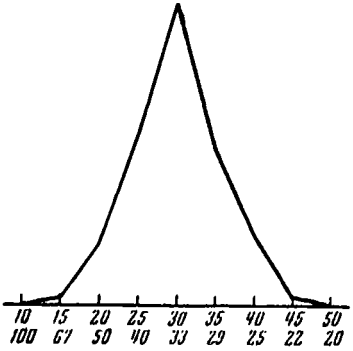

Рис. 24

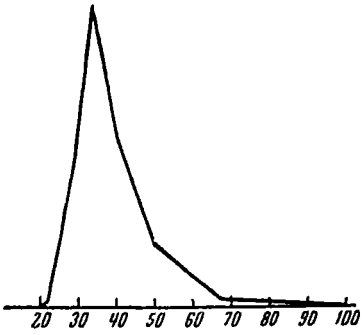

Рис. 25

достигнуто соответствующей деформацией графика, что приводит к рис. 25. Вновь получившееся распределение уже асимметрично, что подтверждается и расчетом: здесь $A = 1{,}34$.

Таблица 21

Средний вес зерна, *мг* на делянке	Число делянок	Число зерен в 1 *г*
1	2	3
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 в одном грамме, 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?

Пример 8. Пусть мы имеем некоторое распределение по диаметрам каких-либо плодов шаровидной формы (например апельсинов, арбузов и т. п.). Составим теперь распределение этих же плодов по весу. Как известно, вес шара пропорционален его объему (плотность всех плодов данного вида можно считать одинаковой), а последний пропорционален третьей степени диаметра. Отсюда ясно, что распределение упомянутых плодов по диаметрам и распределение их по весам не могут быть одновременно нормальными (и даже просто симметричными) — если одно из них симметрично, то другое неизбежно будет асимметричным.

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. При таком эмпирическом подходе выбор нужного преобразования может быть подсказан видом линии, получившейся в результате применения метода спрямления нормальной кривой; преобразование должно быть таким, чтобы эта линия превращалась в прямую.

Пример 9. В табл. 22 приведены результаты опыта над группой из 17 мух. Мухи подвергались действию яда в течение 30 *сек*, а затем определялось время реакции, т. е. время до того момента, когда проявляется паралитическое действие яда (муха падает).

Таблица 22

Время реакции x_i мин	Накопленные частости z_i	$\Psi(z_i)$	$\lg x_i$
3	0,059	—1,56	0,477
4	0,118	—1,19	0,602
5	0,176	—0,93	0,699
6	0,236	—0,72	0,778
8	0,294	—0,54	0,903
9	0,353	—0,38	0,954
10	0,411	—0,23	1,000
12	0,470	—0,08	1,079
15	0,529	0,07	1,176
16	0,588	0,22	1,204
19	0,647	0,38	1,279
26	0,706	0,54	1,415
28	0,765	0,72	1,447
36	0,824	0,93	1,556
54	0,882	1,19	1,732
71	0,941	1,56	1,851
93	1,000		

Накопленные частости равны $^1/_{17}$, $^2/_{17}$ и т. д. На рис. 26 изображен график распределения: по оси абсцисс отложены времена реакции, а по оси ординат — значения $\Psi(z_i)$, найденные из табл. 11

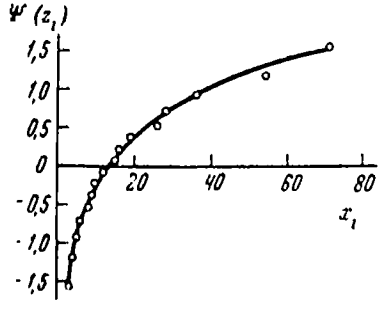

Рис. 26

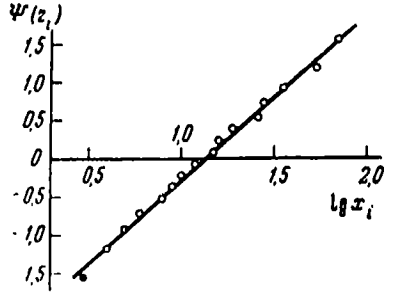

Рис. 27

Приложение. График имеет вид логарифмической кривой. Это наводит на мысль, что преобразование абсцисс по правилу $y_i = \lg x_i$ сделает график прямолинейным. Расчет подтверждает это предположение (рис. 27).

§ 5. Пробит-анализ

Соображения, изложенные в предыдущих параграфах, лежат в основе одного из методов анализа так называемых *кривых смертности* (или вообще *кривых эффекта*), играющих большую роль в радиобиологии, токсикологии, микробиологии.

Одним из существенных биологических свойств повреждающих факторов (например ионизирующих излучений) является их летальное действие. Радиочувствительность животных данного вида может характеризоваться дозой облучения, вызывающей гибель (летальной дозой LD). Само собой разумеется, что эта характеристика так же подвержена вариациям от одной особи к другой, как и любое другое свойство организма; поэтому речь может идти только об определении среднего значения $\overline{LD}$, вычисляемого по данным для некоторой совокупности животных изучаемого вида.

Однако решение этой задачи затрудняется тем, что измерение летальной дозы LD для отдельной особи практически невозможно. Это связано в основном с тем, что гибель животного, получившего дозу излучения, даже заведомо большую, чем LD, наступает не во время облучения, а лишь через несколько дней или даже недель.

Поэтому если сообщенная животному некоторая доза D недостаточна для того, чтобы вызвать его гибель, то это выяснится только через известное время. Если бы биологическое действие фракционированного (т. е. производимого с промежутками) облучения обладало свойством непосредственного накопления, то можно было бы, прибавляя каждые несколько недель небольшую порцию излучения, зафиксировать эту дозу, при которой наступает гибель. Однако полное накопление радиобиологического эффекта не имеет места из-за протекающих в организме восстановительных процессов. Если бы, с другой стороны, эти восстановительные процессы приводили организм полностью в начальное состояние, то величину LD можно было бы определить, повторяя опыт с промежутками в несколько недель и каждый раз увеличивая дозу облучения. В действительности же восстановление заведомо не является полным. Чтобы обойти это затруднение, поступают так. Облучают несколько групп животных разными дозами и для каждой группы (т. е. для каждой дозы) находят процент p погибших животных. Ту дозу, при которой погибает половина (p 50%) облученных животных, и принимают за усредненную характеристику летального действия излучения для данного вида животных; эту характеристику обычно обозначают LD_{50}.

Разумеется, при практическом проведении такого эксперимента не приходится ожидать, что одна из принятых доз окажется в точности равной LD_{50}. Но если, например, при дозе D' погибло $p' = 40\%$ животных, а при дозе D'' погибло p'' 60%, то в качестве первого приближения можно принять LD_{50} $\frac{1}{2}(D' + D'')$.

Такое приближение можно было бы считать удовлетворительным, если бы зависимость процента гибели от дозы была примерно линейной. Однако чаще всего эта зависимость отнюдь не линейна; как показали многочисленные эксперименты, она имеет обычно так называемый S-образный характер (рис. 28). Поэтому для того, чтобы сделать возможной линейную интерполяцию, нужно подобрать такое преобразование координатных осей, после которого график смертности будет изображаться прямой линией.

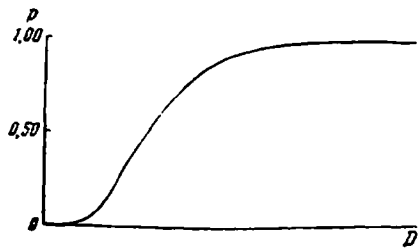

Рис. 28

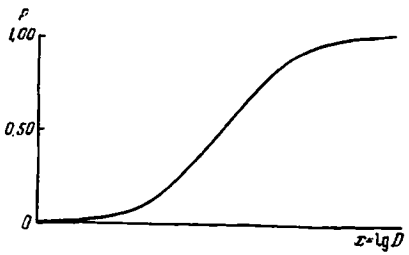

Рис. 29

Прежде всего следует, деформировав надлежащим образом ось абсцисс, сделать кривую смертности симметричной. Оказалось, что это может быть во многих случаях достигнуто при помощи преобразования

$$x = \lg D. \tag{2.10}$$

Это значит, что если по оси абсцисс откладывать не дозы D, а их логарифмы x, то получится симметричная кривая (рис. 29). Эта логарифмическая кривая смертности часто имеет такой же вид, как кривая накопленных частостей нормального распределения (см. рис. 16). Это позволяет применить к ней описанный в § 3 настоящей главы метод спрямления нормальной кривой. В данном случае уравнение (2.7) будет иметь вид

$$p = \Phi\left(\frac{x - x_{50}}{\sigma}\right), \tag{2.11}$$

где $x_{50} = \lg LD_{50}$, а σ есть стандартное отклонение того нормального распределения, которое служит моделью для нашей кривой. Далее, вместо уравнения (2.8) будем иметь

$$y = \frac{x - x_{50}}{\sigma}. \tag{2.12}$$

Таким образом, если по оси абсцисс откладывать значения $x = \lg D$, а по оси ординат — значения y $\Psi(p)$, то график смертности будет прямой линией.

Пример 10. В столбцах 1—4 табл. 23 приведены данные о гибели мышей при облучении их рентгеновскими лучами. Преобразованные величины x и y приведены в столбцах 5 и 6. Логарифмы найдены по нижней шкале логарифмической линейки, а значения $\Psi(p)$ — из табл. II Приложений.

Таблица 23

Доза в рентгенах D	Число мышей в опыте n	Число погибших n_i	Доля погибших p	$x = \lg D$	y $\Psi(p)$	$y' = y + 5$
1	2	3	4	5	6	7
350	32	1	0,031	2,544	—1,87	3,13
425	27	5	0,185	2,628	—0,90	4,10
500	39	15	0,384	2,699	—0,39	4,70
575	34	18	0,530	2,760	0,08	5,08
650	30	23	0,767	9,813	0,73	5,73

При $p < 0{,}5$ значения y отрицательны, что представляет собой известное неудобство. Чтобы избежать этого, заменяют величины y величинами $y' = y + w$, где w — 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 $w = 5$, ибо значению $\Psi(p) = +5$ соответствует очень малое значение $p \approx 0{,}0000003$. Таким образом, по оси ординат мы будем откладывать величины

$$y' = \Psi(p) + 5. \tag{2.13}$$

Эти величины называют *пробитами* (от английского probability unit — вероятностная единица), в связи с чем изложенный выше метод анализа кривых смертности, использующий в качестве модели график интеграла вероятностей, называется *пробит-методом*, или *пробит-анализом* (К. Блисс)[1].

Значения пробитов для разбираемого примера приведены в столбце 7 табл. 23. Точки x, y' отложены на рис. 30. Видно, что эти точки достаточно хорошо укладываются на прямую.

Чтобы определить x_{50} (а затем и LD_{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 x, y'. В первом приближении это можно сделать графически, пользуясь

[1] Подробней см. в книге М. Л. Беленького (1963).

прозрачной линейкой. По графику на рис. 30 теперь получаем $x_{50} = 2{,}737$, чему соответствует $LD_{50} = 545$ *рентген*.

Если работа ведется с объектом, для которого предыдущими экспериментами установлена прямолинейность графика (x, y), то нахождение LD_{50} можно упростить. В этом случае можно провести прямую только по двум точкам, т. е. использовать только две дозы облучения. Для обеспечения наилучшей точности обе точки должны лежать по разные стороны от LD_{50}.

Существуют расчетные методы, которые позволяют более точно провести прямую, наилучшим образом отвечающую экспериментальным точкам (см. § 3 гл. 9). Однако в пробит-анализе

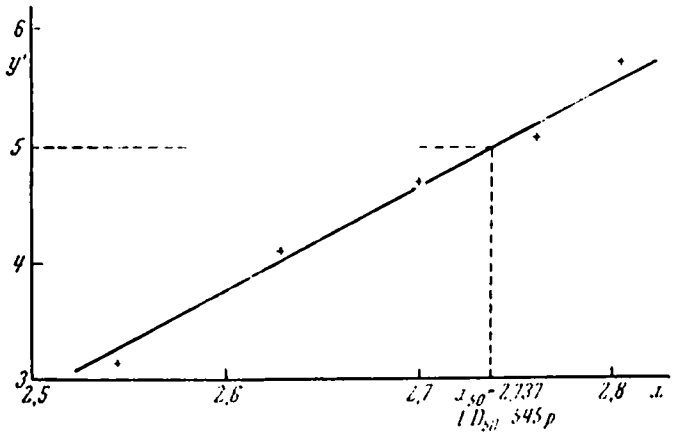

Рис. 30

главной причиной неточностей является обычно то, что логарифмическая кривая эффекта, даже для генеральной совокупности, не всегда имеет в точности нормальную форму. Поэтому использование этих более сложных вычислительных методов чаще всего не приносит большой пользы. О других осложнениях, связанных с выполнением пробит-анализа, см. в цитированной монографии М. Л. Беленького (1963).

Методы, изложенные в этом параграфе, применимы при изучении и таких результатов (эффектов) действия облучения или фармакологических препаратов, которые могут быть представлены в альтернативной форме, т. е. в форме наступления или ненаступления некоторого события — судороги, рвота, исчезновение реакций и т. д.; в этом случае говорят не о летальной дозе LD, а об эффективной дозе ED.

§ 6. Распределение Пуассона

До сих пор речь шла о распределениях, хотя и отклоняющихся от нормального, но все же более или менее близких к нему. Однако часто встречаются распределения, сильно отличающиеся по виду от нормального.

Важнейшим из них является распределение Пуассона. Подобно нормальному распределению, оно также может быть получено как предельный случай биномиального распределения.

Напомним, что в общем случае биномиальное распределение частот дается формулой

$$n_x = NC_\nu^x p^x q^{\nu-x} = NC_\nu^x p^x (1-p)^{\nu-x}, \tag{2.1}$$

где

$$C_\nu^x = \frac{\nu(\nu-1)(\nu-2)\dots(\nu-[x-1])}{x!}. \tag{2.2}$$

Формула (2.1) определяет, например, число бросаний ν костей, в которых двойка выпадает на x костях (из общего числа N бросаний).

Если осуществить много таких бросаний, а затем произвести усреднение, то окажется, что среднее число выпадений двойки близко к νp; так, при многих бросаниях 12 костей среднее (по большому числу бросаний) число костей с двойкой наверху близко к $12 \cdot {}^1/_6 = 2$. Впрочем, это непосредственно вытекает из определения вероятности: $p = \hat{x}/\nu$.

С учетом этого равенства формула (2.1) примет вид

$$n_x = N \frac{\hat{x}^x}{x!} \left\{ \frac{\nu(\nu-1)(\nu-2)\dots(\nu-[x-1])}{\nu^x} \right\} \left(1-\frac{\hat{x}}{\nu}\right)^{\nu} \left(1-\frac{\hat{x}}{\nu}\right)^{-x}$$

Выражение, стоящее в фигурных скобках, можно переписать так:

$$\frac{\nu(\nu-1)(\nu-2)\dots(\nu-[x-1])}{\nu^x} = \left(1-\frac{1}{\nu}\right)\left(1-\frac{2}{\nu}\right)\dots\left(1-\frac{x-1}{\nu}\right).$$

Тогда

$$n_x = N \frac{\hat{x}^x}{x!} \left(1-\frac{\hat{x}}{\nu}\right)^{\nu} \left\{ \left(1-\frac{1}{\nu}\right)\left(1-\frac{2}{\nu}\right)\dots \right.$$
$$\left. \dots \left(1-\frac{x-1}{\nu}\right)\left(1-\frac{\hat{x}}{\nu}\right)^{-x} \right\}. \tag{*}$$

Рассмотрим теперь случай, когда вероятность события p очень мала, но величина $\hat{x}$ все же конечна. Например, вероятность

рождения тройни в семье очень мала, но все же в большом городе ежегодно рождается в среднем несколько троен. Так как $\hat{x} = \nu p$, то $\hat{x}$ может быть величиной порядка единицы только в том случае, если при малом p велико ν; в нашем примере это означает, что общее число рождений в городе (за год) велико. Но при очень больших ν все множители в фигурных скобках формулы (*) равны примерно единице. Что касается множителя $\left(1 - \frac{\hat{x}}{\nu}\right)^{\nu}$ то при $\nu \to \infty$ он стремится к величине $e^{-\hat{x}}$ [в частности, если $\hat{x} = -1$, то получается $\left(1 + \frac{1}{\nu}\right)^{\nu} \to e$ при $\nu \to \infty$; на это уже указывалось выше, см. стр. 65]. Таким образом, окончательно получается

$$n_x = N \frac{\hat{x}^x}{x!} e^{-\hat{x}} \tag{2.14}$$

Это и есть распределение Пуассона.

Рассмотрим следующий пример. Мешок содержит 100 каких-то мерок белых бобов, причем одна мерка вмещает 100 семян; таким образом, общее число семян в мешке равно 10 000. Заменим 100 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 вовсе 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, приближенно выражается законом Пуассона. Здесь $\hat{x}$ — среднее число черных семян в одной порции — равно единице. Тогда число порций, не содержащих черных семян ($x = 0$), пропорционально

$$\frac{1^0}{0!} = 1{,}00;$$

число порций, содержащих одно черное семя ($x = 1$), пропорционально

$$\frac{1^1}{1!} = 1{,}00.$$

Далее, при $x = 2$ имеем

$$\frac{1^2}{2!} = \frac{1}{1 \cdot 2} = \frac{1}{2} = 0{,}50;$$

при $x = 3$

$$\frac{1^3}{3!} = \frac{1}{1\cdot 2\cdot 3} = \frac{1}{6} \approx 0{,}17;$$

при $x = 4$

$$\frac{1^4}{4!} = \frac{1}{1\cdot 2\cdot 3\cdot 4} = \frac{1}{24} \approx 0{,}04;$$

при $x = 5$

$$\frac{1^5}{5!} = \frac{1}{1\cdot 2\cdot 3\cdot 4\cdot 5} = \frac{1}{120} \approx 0{,}01.$$

Величины $\hat{x}^x/x!$, соответствующие значениям $x > 5$, получаются в данном случае (при $\hat{x} = 1$) настолько малыми, что ими можно пренебречь.

Сумма полученных величин равна приблизительно $1{,}00 + 1{,}00 + 0{,}50 + 0{,}17 + 0{,}04 + 0{,}01 = 2{,}72$ ($\approx e$). Для того чтобы сумма n_x равнялась $N = 100$, нужно каждую из полученных величин умножить на

$$\frac{100}{2{,}72} \approx 37,$$

так что, например,

$$n_1 = 1{,}00\cdot 37 = 37, \qquad n_2 = 0{,}17\cdot 37 \approx 6$$

и т. д. Поэтому окончательно распределение будет задаваться табл. 24.

Таблица 24

x	Число черных семян в одной порции	0	1	2		4	5
n_x	Число порций с данным числом черных семян	37	37	19	6	1	0

Полигон частот этого распределения изображен на рис. 31.

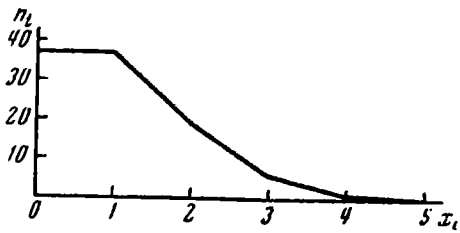

Рис. 31

Таким образом, в 37 порциях теоретически не окажется ни одного черного семени, в 37 будет по одному, в 19 — по два,

в 6 порциях — по 3, в одной порции — 4 черных семени; вероятность появления пяти и больше черных семян в одной порции очень мала, так что вернее всего из ста порций таких не окажется ни одной (но в принципе распределение Пуассона распространяется вправо неограниченно).

Форма полигона частот зависит существенно от значения параметра $\hat{x}$. В нашем примере было $\hat{x} = 1$. Теперь рассмотрим для сравнения еще два случая, когда $\hat{x} < 1$ и $\hat{x} > 1$. В первом случае, очевидно, числитель в (2.14) убывает при возрастании x, а так как знаменатель одновременно возрастает, то n_x все время уменьшается. Так, если $\hat{x} = {}^1/_2$, то мы получаем ряд:

$$\left(\frac{1}{2}\right)^0 : 0! = 1; \quad \left(\frac{1}{2}\right)^1 : 1! = \frac{1}{2}; \quad \left(\frac{1}{2}\right)^2 : 2! = \frac{1}{8}; \quad \left(\frac{1}{2}\right)^3 : 3! = \frac{1}{48};$$

$$\left(\frac{1}{2}\right)^4 : 4! = \frac{1}{384} \text{ и т. д.}$$

Если $\hat{x} > 1$, то в (2.14) возрастают и числитель и знаменатель. Но они возрастают по разному закону, так что сначала «перевешивает» числитель, а затем знаменатель. Пусть, например, $\hat{x} = 2$, тогда

$$2^0 = 1, \quad 2^1 = 2, \quad 2^2 = 4, \quad 2^3 = 8, \quad 2^4 = 16, \quad 2^5 = 32,$$
$$0! = 1, \quad 1! = 1, \quad 2! = 2, \quad 3! = 6, \quad 4! = 24, \quad 5! = 120,$$

так что получаем табл. 25.

Таблица 25

x	0	1	2	3	4	5	6	7
$\frac{\hat{x}^x}{x!}$	1	2	2	1,33	0,67	0,27	0,09	0,02

Аналогичный расчет сделаем для $x = 3$, после чего сравним все полученные распределения. Пересчитав их к объему совокупности $N = 100$, получаем табл. 26. Соответствующие полигоны частот изображены на рис. 32.

Как мы видим, при $\hat{x} < 1$ частоты монотонно убывают с возрастанием x, а при $\hat{x} > 1$ имеется максимум; значение $\hat{x} = 1$ является в этом отношении критическим. Чем больше $\hat{x}$, тем все

Таблица 26

x	0	1	2	3	4	5	6	7
1	61	30	8	1				
1	37	37	19	6	1			
$x=2$	14	27	27	18	9	4	1	
$\hat{x}=3$	5	15	22	22	17	6	2	1

дальше отодвигается максимум, и асимметричность распределения становится менее и менее заметной. При достаточно больших $\hat{x}$ распределение мало отличается от симметричного биномиального

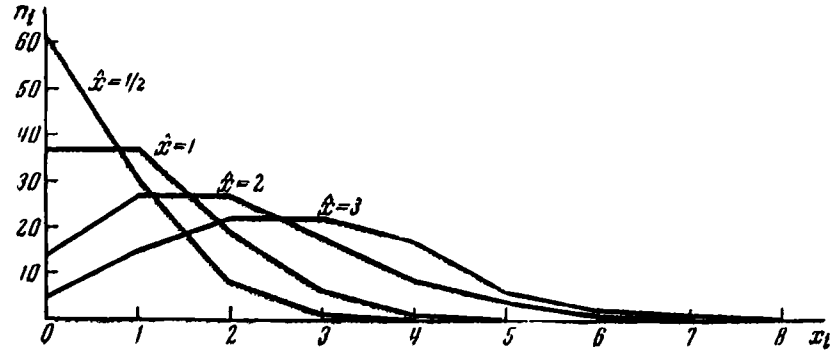

Рис. 32

распределения, близкого, как мы знаем, к нормальному распределению. Если $\hat{x} > 20$, то распределение Пуассона достаточно хорошо приближается законом Гаусса.

Формула

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!},$$

являющаяся обобщением формулы (*) из § 3 этой главы, позволяет легко вычислить моменты распределения Пуассона. Расчет пока-

зывает, что

$$m_1 = \hat{x}, \tag{2.15}$$

как и следовало ожидать. Далее оказывается, что

$$\mu_2 = \hat{x}, \quad \mu_3 = \hat{x}, \tag{2.16}$$

т. е. для распределения Пуассона μ_2 и μ_3 равны между собой и равняются значению единственного параметра $\hat{x}$. Следовательно, если для нормального распределения характерны условия

$$\mu_3 = 0, \quad \rho_4 = 3,$$

то для пуассоновского распределения характерно условие

$$m_1 = \mu_2 = \mu_3; \tag{2.17}$$

оно может быть использовано для проверки того, описывается ли данное эмпирическое распределение законом Пуассона.

Пример 11. Буква «ц» встречается в русском языке довольно редко, поэтому можно ожидать, что ее распределение будет удовлетворять закону Пуассона. В табл. 27 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 27

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

Вероятность наличия буквы «ц» в сотне слов меньше единицы, поэтому частоты монотонно убывают.

Вычисляем моменты:

$$\hat{x} = m_1 = \sum n_i x_i = \frac{1}{1000}(207 \cdot 1 + 38 \cdot 2 + 3 \cdot 3) = 0{,}292;$$

$$m_2 = \sum n_i x_i^2 = \frac{1}{1000}(207 \cdot 1 + 38 \cdot 4 + 3 \cdot 9) = 0{,}386;$$

$$m_3 = \sum n_i x_i^3 = \frac{1}{1000}(207 \cdot 1 + 38 \cdot 8 + 3 \cdot 27) = 0{,}592;$$

$$\mu_2 = m_2 - m_1^2 = 0{,}386 - 0{,}085 = 0{,}301;$$

$$\mu_3 = m_3 - 3m_2m_1 + 2m_1^3 = 0{,}592 - 0{,}338 + 0{,}050 \quad 0{,}304.$$

Мы видим, что числа $m_1 = 0{,}292$; $\mu_2 = 0{,}301$; $\mu_3 = 0{,}304$ действительно довольно близки, так что распределение можно считать пуассоновым.

Очевидно, асимметрия распределения Пуассона всегда положительна.

В заключение укажем, что помимо рассмотренных примеров, распределению Пуассона также следуют: число понизирующих частиц, попадающих в счетчик в одну минуту; число несчастных случаев в единицу времени; число клеток в квадратах цитометра и т. д.

§ 7. Равномерное распределение и композиция распределений

Всякие замеры, взвешивания, отсчеты по шкале и другие измерения всегда производятся с некоторой ограниченной точностью. Поэтому каждая полученная эмпирически величина представляет собой лишь округленное значение истинной величины. Обычно предполагается, что расстояния между истинными значениями и ближайшими к ним делениями шкалы распределены равномерно. Если расстояние между двумя соседними делениями шкалы равно h, то распределение ошибок округления имеет вид прямоугольника с основанием h и высотой N/h (так как площадь этого прямоугольника, изображающая объем совокупности, должна равняться N), т. е. это распределение может описываться равенствами:

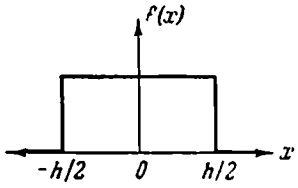

Рис. 33

$$\left.\begin{aligned} n_x &= \frac{N}{h} \quad \text{при} \quad -\frac{h}{2} < x < \frac{h}{2}; \\ n_x &= 0 \quad \text{при} \quad x < -\frac{h}{2},\ x > \frac{h}{2} \end{aligned}\right\} \tag{2.18}$$

(рис. 33). Расчет показывает, что дисперсия этого распределения равна

$$\sigma^2 = \frac{h^2}{12}. \tag{2.19}$$

Если варианты, изучаемые в некотором опыте, представляют собой результаты измерений (длина, вес, pH, сила тока и т. п.), а не результаты счета (как, например, число лепестков цветка, число самцов в помете и т. д.), то на вариации объектов по изучаемому признаку всегда накладываются вариации ошибок округ-

ления. В таком случае говорят, что распределение эмпирических результатов является *композицией* двух распределений — композицией распределений двух независимо варьирующих случайных величин, но относящихся к одному и тому же признаку объекта.

Чаще всего рассматривается случай композиции равномерного распределения ошибок округления и нормального распределения истинных величин. Если при этом $h < \sigma$ (гда σ — стандартное отклонение того нормального распределения, которое участвует в композиции), то суммарное распределение мало отличается от нормального и имеет дисперсию, равную сумме дисперсий компонируемых распределений:

$$\widetilde{\sigma}^2 = \sigma^2 + \frac{h^2}{12}. \qquad (2.20)$$

Пример 12. В табл. 9 на стр. 25 приведено распределение ростов 1000 мужчин. На сколько увеличится дисперсия этого распределения, если измерения будут производиться с точностью в 1 *см*?

Считая, что в интервале от 143 до 185 *см* укладывается примерно 6σ, получаем приблизительно

$$\sigma = \frac{185-143}{6} = 7 \text{ см}$$

и $\sigma^2 = 49$ *см*^{2}. (Конечно, выполнив надлежащий расчет, можно было бы получить более точное значение; но сейчас нас интересует только порядок величины σ^2.) Так как $h = 1$ *см*, то слагаемое $h^2/12 < 0{,}1$ *см*^{2} очень мало добавляет к первоначальной дисперсии $\sigma^2 = 49$ *см*^{2}.

§ 8. Распределения при качественной группировке

Распределение Гаусса и Пуассона представляют распределения вариант по их численным значениям. Понятно, что эти распределения совершенно неприменимы при классификации вариант по качественному признаку — хотя бы потому, что порядок расположения классов при этом произволен.

Однако и при качественной классификации можно указать распределения, которые могут быть выведены из особо простых модельных представлений. Примером может служить распределение численностей при расщеплении гибридных форм, предсказываемое теорией. Так, если у скрещиваемых форм учитывались два признака (например, окраска и длина шерсти у морских

свинок), то во втором поколении соотношение численностей форм AB, aB, Ab и ab (A и a — доминантный и рецессивный виды первого признака, B и b — то же для второго признака) будет 9 : 3 3 : 1.

В частном случае двух классов (альтернативное распределение) примерами теоретически предсказываемых распределений численностей будут: соотношение 3 : 1 для доминантных и рецессивных форм при расщеплении гибридов с одним учитываемым признаком, соотношение 1 : 1 рождения животных разного пола и т. д.

ОЦЕНКА ПАРАМЕТРОВ ПО ВЫБОРОЧНЫМ ДАННЫМ

§ 1. Составление выборок

Если эмпирическая статистическая совокупность интересует исследователя сама по себе, то описание ее при помощи методов, изложенных в гл. 1, можно считать исчерпывающим. Однако в биологических исследованиях, как уже указывалось во «Введении», полученная из наблюдения или опыта эмпирическая совокупность представляет собой обычно выборку из некоторой более обширной генеральной совокупности; эта эмпирическая совокупность представляет интерес не сама по себе, а лишь постольку, поскольку изучение ее позволяет получить информацию о свойствах представляемой ею генеральной совокупности.

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 случайным образом (*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 колосьев по высоте, и экспериментатор, составляя выборку, руководствуется только желанием, чтобы в ней были представлены в одинаковой пропорции низкорослые, средние и высокие колосья. Если ему попадутся

подряд три или четыре высоких колоса (что вполне может случиться), то он в дальнейшем невольно будет стараться «скомпенсировать» это преимущественным включением в выборку низкорослых колосьев. А это внесет в составление выборки элемент, зависящий от свойств исследователя, в то время как выборка должна отражать лишь свойства генеральной совокупности. Столь же ошибочным был бы, например, отбор для опыта тех мышей, 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 лотерен, но с возвратом вынутых номеров, так что отдельные номера могут попасться дважды или большее число раз). Полученная этим (или другим аналогичным) способом последовательность таких чисел записывается в виде таблицы (см. табл. V Приложений). Для удобства применения этой таблицы все числа записывают так, чтобы они имели одно и то же число цифр; если, например, таблица четырехзначная, то число 6 будет записано в виде 0006. Использование таблицы случайных чисел поясним на примере.

Пример 1. Из 146 животных, имеющихся в виварии, нужно отобрать для опыта 8. Было бы неправильным взять первые 8 по списку: может оказаться, что они все из одного помета или соседство их клеток могло как-то повлиять на их свойства и т. д. Поэтому мы обращаемся к таблице случайных чисел (см. табл. V Приложений). Просматривание таблицы можно начинать в любом месте и вести в произвольном направлении. Мы начнем с начала третьей колонки (с числа 3156) 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 0047 (в предпоследнем столбике); следующее — 0144 — мы найдем только в третьем столбике пятой колонки. Ясно, что такой способ составления выборки очень неэффективен — используется лишь незначительная часть чисел таблицы; более того, если выборка должна быть довольно большой, то в таблице, содержащей ограниченный набор чисел, может вообще не оказаться нужного количества чисел, меньших 146.

Поэтому мы упростим процедуру. Учитывая, что 146 есть трехзначное число, не будем обращать внимания на первые цифры

[1] Более подробно эти вопросы обсуждаются во вступительной статье проф. В. Н. Перегудова к книге Дж. У. Снедекора (1961).

табличных чисел. Тогда сравнительно быстро получим набор: 080, 105, 005, 098, 016, 112, 113, 047.

Так как большинство чисел таблицы больше 146 (речь идет теперь уже только о трехзначных числах, получившихся после отбрасывания первой цифры), то при просматривании колонок все же пришлось бо́льшую часть чисел пропускать. Если бы требовалось отобрать не 8, а, скажем, 38 животных, то эта процедура заняла бы довольно много времени. Отбор можно ускорить, если условиться считать первую цифру (из трех) нулем, когда она четная, и единицей, когда она нечетная; тогда еще больше чисел «пойдет в дело», и уже первые 14 чисел из третьей колонки дадут нам нужные восемь номеров: 077, 080, 123, 032, 105, 049, 116, 085. Животных с этими номерами мы и берем в опыт.

Другой вид выборки — *типическая*, или *зональная*. Она состоит в том, что генеральная совокупность делится на несколько классов (типических групп, или зон), исходя из изучаемого признака, а затем производится выборка, уже случайным образом, отдельно из каждого класса. Это имеет смысл делать, если в генеральной совокупности имеется какая-либо очевидная систематическая неравномерность. Пусть, например, имеется заметная уже на глаз неравномерность в густоте растений между краями поля и его серединой, между более высокой и более низкой частями 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 — в других частях. Чтобы избежать 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:

$$n_j = np_j, \tag{3.1}$$

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

Если имеются какие-нибудь предварительные сведения о вариабильности признака в каждой из зон (т. е. о величинах σ_j), то можно получить лучшую репрезентативность общей выборки, выбирая объемы зональных частей выборки так, чтобы они были пропорциональны также этим значениям вариабильности:

$$n_j = n\frac{p_j\sigma_j}{\sum\limits_j p_j\sigma_j}. \tag{3.2}$$

Очевидно, формула (3.1) является частным случаем формулы (3.2): когда о величинах σ_j ничего не известно, то делается простейшее предположение, что они все одинаковы; тогда σ_j можно вынести за знак суммы в знаменателе правой части (3.2) и сократить с σ_j в числителе, в результате чего и получится формула (3.1) — с учетом того, что $\Sigma p_j = 1$.

Пример 2. Поле, на котором желательно изучить выборочным методом 40 небольших делянок (чтобы определить их распределение по числу колосьев), разбито на 4 примерно однородные части. Площади этих частей составляют доли $p_j = 0{,}1;\ 0{,}4;\ 0{,}3;\ 0{,}2$ от всей площади поля. Приблизительная оценка (на глаз или другим упрощенным приемом) вариабильности густоты колосьев дает для этих частей поля соответственно значения $\sigma_j = 2;\ 3;\ 1;\ 4$ колоса на делянку. 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 (о количественном выражении этой «близости» будет сказано в следующем параграфе этой главы) получится, если выбирать с каждой из частей поля соответственно:

$$n_1 = 40 \cdot 0{,}1 = 4; \quad n_2 = 40 \cdot 0{,}4 = 16;$$
$$n_3 = 40 \cdot 0{,}3 = 12; \quad n_4 = 40 \cdot 0{,}2 = 8$$

делянок — по формуле (3.1). Еще лучше, если, используя данные о величинах σ_j, воспользоваться формулой (3.2). Имеем:

$$p_1\sigma_1 = 0.1 \cdot 2 = 0{,}2; \quad p_2\sigma_2 = 0{,}4 \cdot 3 = 1{,}2;$$
$$p_3\sigma_3 = 0{,}3 \cdot 1 = 0{,}3; \quad p_4\sigma_4 = 0{,}2 \cdot 4 = 0{,}8;$$
$$\Sigma p_j\sigma_j = 0{,}2 + 1{,}2 + 0{,}3 + 0{,}8 = 2{,}5;$$
$$\frac{n}{\Sigma p_j\sigma_j} = \frac{40}{2{,}5} = 16,$$

так что получаем:

$$n_1 = 16 \cdot 0{,}2 = 3{,}2; \quad n_2 = 16 \cdot 1{,}2 = 19{,}2;$$
$$n_3 = 16 \cdot 0{,}3 = 4{,}8; \quad n_4 = 16 \cdot 0{,}8 = 12{,}8;$$

поскольку число делянок должно быть целым, принимаем:

$$n_1 = 3, \quad n_2 = 19, \quad n_3 = 5, \quad n_4 = 13.$$

Иногда применяются так называемые *механические выборки*; например, отбирается каждая десятая особь или данные на каждый четвертый 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.

§ 2. Соотношение между выборочным и генеральным средними значениями

Итак, выборка составлена, причем составлена правильно. Можно ли утверждать, что вычисленные для этой выборки значения параметров (например $\hat{x}$, σ, A и т. д.) равны тем значениям, которые характеризуют генеральную совокупность?

Прямая проверка обычно невозможна, поскольку, как правило, генеральная совокупность не находится в распоряжении исследователя. Однако возможна косвенная проверка. Составим несколько независимых выборок из одной генеральной совокупности и сравним полученные для них значения параметров; если все эти значения соответственно совпадут, то это даст основание считать, что выборочные параметры дают правильные значения параметров генеральной совокупности.

Проверка показывает, что значения параметров, полученные для разных выборок из одной генеральной совокупности, обычно не совпадают. Это можно иллюстрировать следующим простым примером. Пусть мы имеем генеральную совокупность, состоящую всего 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). Среднее значение составляет

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

Заменим, однако, изучение всей генеральной совокупности изучением выборки из нее объема $n = 4$. 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 значение для выборки будем впредь обозначать $\bar{x}$, в отличие от среднего значения для генеральной совокупности $\hat{x}$. Среднее арифметическое из этих выборочных средних равно, конечно, генеральному среднему

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

Но такой способ нахождения генерального среднего 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 уж 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 17 310 309 456 440 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 лишь одной из такого громадного 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 совпа-

дает, как правило, с генеральным средним $\hat{x}$. Однако можно показать, что чем больше объем выборки, тем меньше вероятность того, что $\bar{x}$ будет значительно отличаться от $\hat{x}$. 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 $\hat{x}$, 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 обстоятельством, что чем ближе $\bar{x}$ к краю генеральной совокупности, тем больше ограничены возможности «раздвигания» вариант в паре для сохранения заданного $\bar{x}$ (рис. 34).

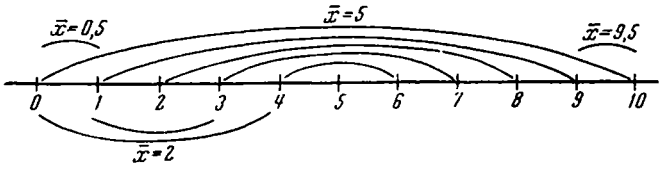

Рис. 34

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. 35 изображены распределение вариант в генеральной совокупности объемом $N = 16$ и три распределения выборочных средних, полученных из выборок с $n = 2$, $n = 4$ и $n = 8$.

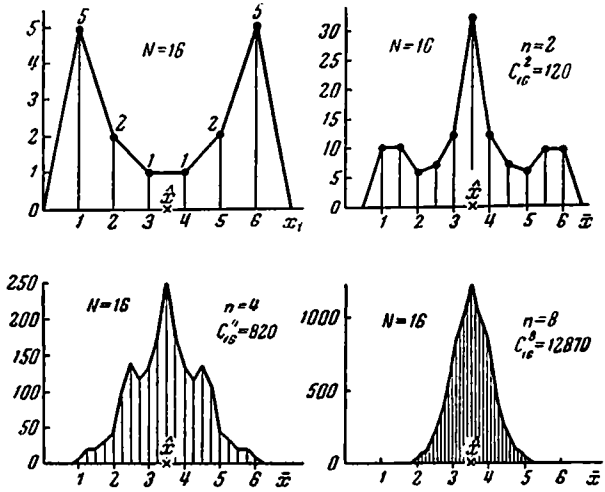

Рис. 35

Аналогичным образом можно убедиться, что распределение выборочных средних становится при больших n симметричным, если даже распределение вариант в генеральной совокупности явно асимметрично. Это показано на рис. 36.

Можно сказать, что если $N \to \infty$, то 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. гл. 2, § 3), причем эти условия реализуются тем полней, чем больше число n вариант в выборке. Из сказанного выше ясно, что центром распределения выборочных средних $\bar{x}$ является среднее значение генеральной совокупности $\hat{x}$. Но отсюда следует важный вывод: хотя выборочное среднее значение $\bar{x}$, полученное по результатам одной только выборки, и не равно генеральному среднему $\hat{x}$, оно все же указывает значение, вблизи

которого находится $\hat{x}$. Поэтому выборочное среднее $\bar{x}$ называют *оценкой* генерального среднего $\hat{x}$. Впредь мы будем рассматривать только выборки из б е с к о н е ч н о б о л ь ш о й генеральной совокупности (если не будет специальных оговорок).

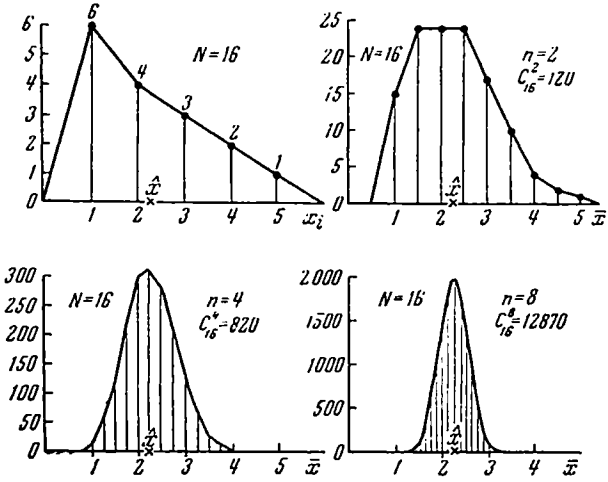

Рис. 36

Дисперсию выборочных средних относительно центра их распределения, т. е. относительно $\hat{x}$, можно найти следующим образом. Пусть мы имеем выборку

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

из n независимых вариант. Поскольку

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

то

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

так что

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

Но согласно (1.18)

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

а согласно (1.21)

$$\sigma^2\{x_1 + x_2 + \quad + x_n\} = \sigma^2\{x_1\} + \sigma^2\{x_2\} + \quad + \sigma^2\{x_n\};$$

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

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

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

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

откуда

$$\sigma^2\{\bar{x}\} = \frac{\sigma^2\{x\}}{n}. \qquad (3.3)$$

Это можно переписать в виде

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

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

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}; \qquad (3.4)$$

в биологической литературе величину $\sigma_{\bar{x}}$ часто 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 большее 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 $\hat{x}$.

Когда выборка состоит из нескольких подвыборок (в случае типической выборки), стандартная ошибка среднего значения вычисляется следующим образом:

а) при произвольных объемах n_j подвыборок — по формуле

$$\sigma_{\bar{x}} = \sqrt{\sum_j \frac{p_j^2 \sigma_j^2}{n_j}}; \qquad (3.5)$$

б) если объемы подвыборок n_j пропорциональны объемам типических групп (зон), то подстановка в (3.5) значений n_j из (3.1) дает

$$\sigma_{\bar{x}} = \sqrt{\frac{\Sigma p_j \sigma_j^2}{n}}; \qquad (3.6)$$

в) если n_j пропорциональны также стандартным отклонениям σ_j в типических группах, то в (3.5) подставляются значения n_j из (3.2); тогда получается

$$\sigma_{\bar{x}} = \frac{\Sigma p_j \sigma_j}{\sqrt{n}}. \qquad (3.7)$$

Пример 3. Вычислим по указанным формулам значения $\sigma_{\bar{x}}$ для данных из примера 2; будем считать, что фактически найденные значения σ_j для частей поля совпали с предварительными оценками ($\sigma_j = 2, 3, 1, 4$).

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

в) $\sigma_{\bar{x}} = \frac{0{,}1 \cdot 2 + 0{,}4 \cdot 3 + 0{,}3 \cdot 1 + 0{,}2 \cdot 4}{\sqrt{40}} = 0{,}395.$

Очевидно, чем меньше для данной выборки значение $\sigma_{\bar{x}}$, тем выборочное среднее ближе к генеральному. Из рассмотренного примера видно, что способ (б) составления типической выборки лучше, чем способ (а), а способ (в) еще лучше.

В заключение этого параграфа укажем, что если выборка, в отличие от принятого здесь общего правила, взята из генеральной совокупности конечного объема N, то в формулу (3.4) нужно ввести поправочный множитель

$$\sqrt{\frac{N-n}{N-1}}.$$

Очевидно, при $N \to \infty$ этот множитель стремится к единице, так как

$$\frac{N-n}{N-1} = \frac{1-n/N}{1-1/N}.$$

§ 3. Несмещенная оценка дисперсии

Итак, мы нашли, что $\bar{x}$ может служить оценкой для $\hat{x}$. Можно ли утверждать, что и выборочная дисперсия $\frac{1}{n}\sum(x_i - \bar{x})^2$ 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 § 3 гл. 1, 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 $\sum_i (x_i - x)^2$ 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, в том числе и от генерального среднего $\hat{x}$. Следовательно, при вычислении дисперсии на основании выборки по обычной формуле (1.15)

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

мы всегда получаем заниженную оценку.

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

$$\sum n_i [(x_i - \hat{x}) - (\bar{x} - \hat{x})]^2$$

(прибавив и вычтя в скобках $\hat{x}$). Проведя такие же выкладки, какие приводят к формуле (1.17), получим

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

откуда сразу видно, что

$$\sum n_i (x_i - \bar{x})^2 < \sum n_i (x_i - \hat{x})^2.$$

Учтем теперь, что

$$\langle \sum n_i (x_i - \hat{x})^2 \rangle = n\sigma^2,$$

а квадрат отклонения выборочного среднего $\bar{x}$ от генерального среднего $\hat{x}$ составляет в среднем по всем выборкам $\sigma^2_{\bar{x}}$, т. е.

$$\langle(\bar{x} - \hat{x})^2\rangle = \sigma^2_{\bar{x}} = \frac{\sigma^2}{n};$$

тогда равенство (*), усредненное по всем выборкам, будет иметь вид

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

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

$$s^2 = \frac{1}{n-1}\sum n_i (x_i - \bar{x})^2, \tag{3.8}$$

а не величину

$$\sigma^2_{\text{выб}} = \frac{1}{n}\sum n_i (x_i - \bar{x})^2. \tag{3.8'}$$

Принято говорить, что s^2 является *несмещенной* оценкой дисперсии σ^2 (в то время как $\sigma^2_{\text{выб}}$ будет *смещенной* оценкой, ибо 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 **независимыми**. 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-е отклонение должно по необходимости быть таким, чтобы выполнялось условие (**). Число независимых величин, участвующих в образовании того или иного параметра, называется *числом степеней свободы* этого параметра и обозначается через f. Оно равно общему числу величин, по которым вычисляется параметр, минус число условий, связывающих эти величины. Дисперсия 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

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

а несмещенную оценку дисперсии — по формуле

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

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, то мы о ней ничего не можем знать, если имеется только одно значение x. Но такой же ответ дает и формула (3.8):

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

Между тем по формуле (3.8′) мы бы имели

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

т. е. заключение об отсутствии дисперсии, что явно неверно.

Однако если среднее значение $\hat{x}$ генеральной совокупности известно, то полученное значение x уже позволяет судить о рассеянии вариант. Этому отвечает и написанная выше формула. Действительно, в этом случае $f = n$, так что при $n = 1$

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

Пример 4. Для проверки качества pH-метра были проведены измерения восьми образцов воды (бидистиллята). Результаты приведены в табл. 28.

Таблица 28

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 генеральное среднее $\bar{x} = 7{,}00$ 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:

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

Если дисперсия вычислялась при помощи моментов и генеральное среднее $\bar{x}$ было неизвестно, то для получения несмещенной оценки для σ^2 нужно μ_2 умножить на $\frac{n}{n-1}$

$$s^2 = \frac{n}{n-1}\mu_2. \tag{3.9}$$

Из (3.8) имеем оценку стандартного отклонения:

$$s = \sqrt{\frac{\sum n_i (x_i - \bar{x})^2}{n-1}}. \tag{3.10}$$

Эта формула позволяет найти выборочную оценку для стандартной ошибки среднего значения. Именно, если подставить в (3.4) вместо неизвестной дисперсии σ^2 ее оценку согласно (3.10), то получится

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum n_i (x_i - \bar{x})^2}{n(n-1)}}. \tag{3.11}$$

Подобно тому как выборочное значение μ_2 есть смещенная оценка генеральной дисперсии σ^2 (при неизвестном $\hat{x}$), другие центральные моменты (μ_3, μ_4 и т. д.), вычисленные по данным выборки, также представляют собой смещенные оценки соответствующих генеральных параметров. Это значит, что для получения несмещенных оценок основных моментов ρ_3, ρ_4 и т. д. нужно ввести надлежащие поправки; при этом надо учесть, что основные моменты зависят от μ_2, где тоже требуется поправочный множитель. Это приводит к следующим формулам (которые даем без вывода) для несмещенных оценок коэффициентов асимметрии и эксцесса:

$$A = \frac{\sqrt{n(n-1)}}{n-2}\rho_3, \tag{3.12}$$

$$E = \frac{(n^2-1)}{(n-2)(n-3)}\left[(\rho_4 - 3) + \frac{6}{n+1}\right]. \tag{3.13}$$

Стандартные ошибки величин A и E, при не очень малых объемах выборки, приближенно равны:

$$\sigma_A \approx \sqrt{\frac{6}{n+3}}; \quad \sigma_E \approx \sqrt{\frac{24}{n+5}}. \tag{3.14}$$

При увеличении порядка момента стандартные ошибки этих моментов быстро возрастают. Поэтому моменты порядка выше четвертого употребляются редко (когда имеются выборки очень большого объема).

В заключение приведем оценки для стандартных ошибок σ_{s^2}, σ_s и σ_v (v — выборочный коэффициент вариации):

$$s_{s^2} = s^2\sqrt{\frac{2}{n}}; \quad s_s \approx \frac{s}{\sqrt{2n}}; \tag{3.15}$$

$$s_v = \frac{1}{\sqrt{n-1}}v\sqrt{\frac{1}{2} + \left(\frac{v}{100}\right)^2} = \frac{\gamma(v)}{\sqrt{n-1}} \tag{3.16}$$

(v выражен в процентах); значения функции $\gamma(v) = v\sqrt{\frac{1}{2} + \left(\frac{v}{100}\right)^2}$ даны в табл. III Приложений.

§ 4. Доверительные интервалы

В § 2 этой главы было показано, что при больших n выборочные средние $\bar{x}$ распределены приближенно нормально вокруг $\hat{x}$ со стандартным отклонением $\sigma_{\bar{x}}$. Это значит, что относительное

отклонение выборочного среднего $\bar{x}$ от генерального среднего $\hat{x}$, т. е. величина

$$\tau = \frac{\bar{x} - \hat{x}}{\sigma_{\bar{x}}}, \tag{3.17}$$

распределена так же, как относительные отклонения вариант x_i от $\hat{x}$, т. е. величины

$$u = \frac{x_i - \hat{x}}{\sigma}$$

в нормально распределенной генеральной совокупности.

Поэтому вероятность того, что $\bar{x}$ будет находиться в пределах $\hat{x} \pm t\sigma_{\bar{x}}$, можно приближенно описывать функцией $\theta(u)$, значения которой даны в табл. I Приложений,— нужно только вместо u подставлять t.

Отсюда мы, в частности, получаем, что $\sim 68{,}3\%$ всех выборочных $\bar{x}$ находятся в пределах $\hat{x} \pm \sigma_{\bar{x}}$; иными словами, имеется вероятность 0,683, что $\bar{x}$ отличается от $\hat{x}$ 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 $\hat{x}$ (или, иными словами, ~ 68 этих интервалов будут содержать $\hat{x}$).

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

Из сказанного ясно, что указание доверительного интервала заключает в себе некоторое утверждение о среднем значении 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 $\hat{x}$, сравнительно невысока. С большей уверенностью можно утверждать, что $\hat{x}$ покрывается интервалом $(\bar{x} - 2\sigma_{\bar{x}}, x + 2\sigma_{\bar{x}})$, так как вероятность этого равна 95,5%; в том, что $\hat{x}$ 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 $\hat{x}$, тем шире должен быть интервал, могущий обеспе-

чить такую вероятность. Уровень этой вероятности принято называть *доверительной вероятностью*, а иногда *надежностью*; выбор той или иной доверительной вероятности определяет ширину доверительного интервала, который с принятой вероятностью содержит $\hat{x}$. Так, при доверительной вероятности $P = 95{,}5\%$ границы доверительного интервала для $\hat{x}$ 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.

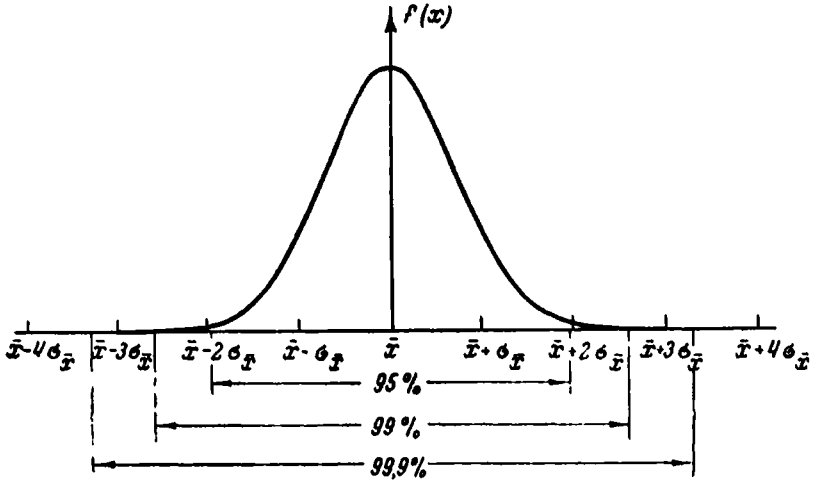

Рис. 37

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. I Приложений: это те значения $\tau = u$, при которых $\theta(u)$ имеет заданные значения 0,950 и 0,990 (рис. 37).

В биологических применениях статистики 95%-ный доверительный интервал считается достаточно надежным. Однако при исследованиях, уточняющих предыдущие результаты, а также в случаях, когда от результатов исследования зависит принятие или непринятие решения о затрате средств и труда (применение добавочных удобрений или подкормок, реорганизация производства и т. п.), целесообразно применять более высокий доверительный уровень $P = 99\%$ или даже $P = 99{,}9\%$.

Очевидно, относительное отклонение выборочного среднего от генерального среднего, т. е.

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

будет оцениваться величиной

$$t = \frac{\bar{x} - \hat{x}}{s_{\bar{x}}}, \qquad (3.18)$$

где $s_{\bar{x}}$ определяется формулой (3.11). 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 $\hat{x}$, 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 $\hat{x}$, 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.

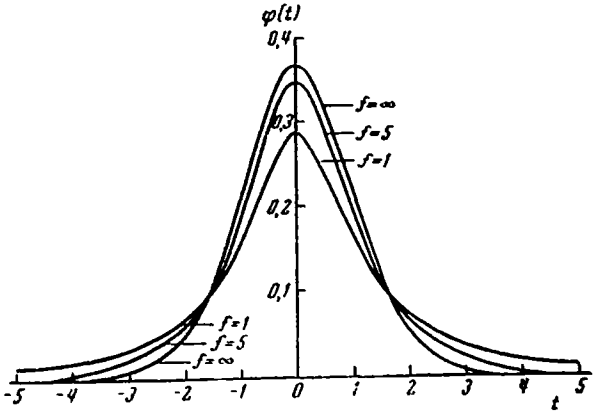

Рис. 38

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 $f = 5$ представлены на рис. 38;

там же для сравнения изображена нормальная кривая, являющаяся предельной (при $f = \infty$) для кривых Стьюдента.

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 5. При изучении урожайности нового сорта картофеля за $n = 10$ лет получены результаты: $\bar{x} = 5{,}46$ *т/га*, $s_{\bar{x}} = 0{,}66$ *т/га*. Требуется найти 95%-ный доверительный интервал для $\hat{x}$, характеризующего урожайность данного сорта.

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:

$$5{,}46 - 2{,}26 \cdot 0{,}66 = 3{,}97; \qquad 5{,}46 + 2{,}26 \cdot 0{,}66 = 6{,}95.$$

Из таблицы 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 $t_P s_{\bar{x}} = \frac{t_P}{\sqrt{n}} s$) 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 ($63{,}66/2 = 45{,}0$; $9{,}93/3 = 5{,}7$; $45{,}0/5{,}7 = 7{,}9$). При больших значениях n увеличение n на одну единицу сказывается на ширине доверительного интервала гораздо меньше.

Очевидно, во всех случаях желательно, чтобы доверительный интервал был как можно у́же, т. е. чтобы (при принятом доверительном уровне P) величина $\sigma_{\bar{x}}$ была как можно меньше. Из формулы (3.4) видно, что $\sigma_{\bar{x}}$ зависит как от σ, так и от n.

В случае биологических совокупностей численное значение величины σ определяется в основном вариабильностью самого

исследуемого материала и поэтому не поддается регулированию со стороны исследователя; поэтому там, где это возможно, желательно брать выборки как можно большего объема.

В некоторых случаях возможности увеличить объем выборки ограничены — либо из-за ограниченности экспериментального материала, находящегося в распоряжении исследователя, либо (при экспериментах, повторяемых во времени) из-за необходимости разумно ограничить общее время проведения работы. В этом случае значение $\sigma_{\bar{x}}$ не может быть сделано сколь угодно малым, т. е. значение $\hat{x}$ 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 *кг*, т. е. принять $x = 2{,}31$ *кг*. Чтобы избежать накопления ошибок в промежуточных расчетах, целесообразно проводить их 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*. Эти соображения надо всегда учитывать при планировании опыта, чтобы избавить себя от лишней и неоправданной работы (достижение большей точности измерений часто связано с существенным усложнением методики).

Если мы задаемся некоторой определенной точностью при нахождении $\hat{x}$, причем известен естественный разброс изучаемого материала (характеризуемый значением σ), то можно заранее вычислить, каков должен быть объем выборки (или повторность) для получения заданной точности. Пусть, например, мы хотим, чтобы неточность в определении $\hat{x}$ (т. е. величина $\tau_P\sigma_{\bar{x}} = u_P\sigma_{\bar{x}}$) не превышала некоторого значения Δ:

$$u_P\sigma_{\bar{x}} \leqslant \Delta. \qquad (*)$$

Так как

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

то условие (*) дает

$$u_P \frac{\sigma}{\sqrt{n}} \leqslant \Delta,$$

откуда

$$n \geqslant \frac{u_P^2 \sigma^2}{\Delta^2}. \qquad (**)$$

Например, при $\sigma = 4{,}3$ *мг* мы хотим, чтобы с вероятностью $P = 0{,}95$ неточность в определении $\hat{x}$ была не больше, чем 0,1 *мг*. Тогда по формуле (**) находим, что выборка должна содержать не менее чем

$$n = \frac{4{,}3^2}{0{,}1^2} 1{,}96^2 \approx 7100$$

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

$$\varepsilon = \frac{u_P \sigma_{\bar{x}}}{\bar{x}}, \qquad (3.19)$$

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

$$\varepsilon = \frac{u_P}{\bar{x}} \frac{\sigma}{\sqrt{n}} = \frac{u_P}{\sqrt{n}} \frac{\sigma}{\bar{x}} = \frac{u_P}{\sqrt{n}} v, \qquad (***)$$

где v — коэффициент вариации. Из (***) получаем наименьшее допустимое значение n:

$$n = \frac{u_P^2}{\varepsilon^2} v^2, \qquad (3.20)$$

причем значение u_P определяется принятой доверительной вероятностью. В табл. VIII Приложений приводятся так называемые *достаточно большие объемы выборок* n, обеспечивающие при данном значении коэффициента вариации v ошибку ε не более 5%

или 3% при u_P, соответствующем доверительной вероятности 99%. Если, например, $v = 10\%$, то для получения 99%-ной доверительной ошибки не более 5% нужно иметь выборку объемом не менее 27 вариант; если желательно, чтобы ошибка не превышала 3%, то достаточно большим числом вариант можно будет считать только $n = 75$.

Иногда приходится определять доверительные интервалы для стандартного отклонения. Проще всего это можно делать, используя формулу (3.15) 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 предпочтительней находить доверительные интервалы для σ при помощи специальной таблицы, основанной на более точном распределении выборочных s. Эта таблица (см. табл. VI Приложений) дает значения $q'_P(f)$ и $q''_P(f)$, определяющие границы P%-ного доверительного интервала (в предположении, что распределение вариант в генеральной совокупности нормально):

$$q'_P s < \sigma < q''_P s. \tag{3.21}$$

Пример 6. В примере 5 мы имели $s_{\bar{x}} = 0{,}66$, $n = 10$. Этому соответствует $s = 0{,}66\sqrt{10} = 2{,}09$. Если нас интересует 95%-ный доверительный интервал для σ, то по табл. VI Приложений $q'_{95}(9) = 0{,}688$ и $q''_{95}(9) = 1{,}826$. Тогда границы доверительного интервала будут

$$2{,}09 \cdot 0{,}688 \approx 1{,}44 \text{ и } 2{,}09 \cdot 1{,}826 \approx 3{,}82.$$

§ 5. Объединение выборок

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 генерального среднего $\hat{x}$ и несмещенные оценки $s^2_{(1)}, s^2_{(2)}, \ldots, s^2_{(w)}$ для генеральной дисперсии σ^2. Так как ширина доверительного интервала для генеральных параметров зависит от объема выборки, то, очевидно, мы получим более точные оценки для $\hat{x}$ и σ^2, если объединим все w выборок в одну выборку суммарного объема

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

и найдем $\bar{x}$ и s^2 для этой 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 шириной доверительного интервала, пропорциональной в свою очередь дисперсии выборочного среднего $\sigma^2_{\bar{x}}$. Поэтому чем меньше $\sigma^2_{\bar{x}}$, тем больше должен быть «вес» выборки; таким образом, в качестве «веса» надо брать обратные дисперсии выборочных средних величин, т. е. величины $1/\sigma^2_{\bar{x}}$.

Значит, усредненное по w выборкам среднее значение будет:

$$\bar{x} = \frac{\frac{1}{\sigma^2_{\bar{x}(1)}}\bar{x}^{(1)} + \frac{1}{\sigma^2_{\bar{x}(2)}}\bar{x}^{(2)} + \ldots + \frac{1}{\sigma^2_{\bar{x}(w)}}\bar{x}^{(w)}}{\frac{1}{\sigma^2_{\bar{x}(1)}} + \frac{1}{\sigma^2_{\bar{x}(2)}} + \ldots + \frac{1}{\sigma^2_{\bar{x}(w)}}} = \frac{\sum_{j=1}^{w}\frac{1}{\sigma^2_{\bar{x}(j)}}\bar{x}^{(j)}}{\sum_{j=1}^{w}\frac{1}{\sigma^2_{\bar{x}(j)}}} = \frac{\sum\frac{n^{(j)}}{\sigma^2_{(j)}}\bar{x}^{(j)}}{\sum\frac{n^{(j)}}{\sigma^2_{(j)}}} \qquad (3.22)$$

Но поскольку, согласно предположению, все выборки взяты из одной и той же генеральной совокупности, то все $\sigma^2_{(j)}$ одинаковы. Поэтому, вынося $\sigma^2_{(j)}$ в числителе и знаменателе за знаки суммы, получаем после сокращения на $\sigma^2_{(j)}$:

$$\bar{x} = \frac{n^{(1)}\bar{x}^{(1)} + n^{(2)}\bar{x}^{(2)} + \ldots + n^{(w)}\bar{x}^{(w)}}{n^{(1)} + n^{(2)} + \ldots + n^{(w)}} = \frac{1}{n}\Sigma n^{(j)}\bar{x}^{(j)}, \qquad (3.23)$$

т. е. весами оказываются просто объемы выборок.

Пример 7. В табл. 29 представлены данные о размерах эритроцитов крови, полученные на четырех мазках от одной мыши.

Найдем усредненное значение $\bar{x}$, пользуясь формулой (3.23):

$$\bar{x} = \frac{1}{1540}(392\cdot 8{,}33 + 364\cdot 8{,}26 + 345\cdot 8{,}30 + 448\cdot 8{,}28)$$
$$= \frac{1}{1540}(3260 + 3010 + 2860 + 3710) = \frac{12840}{1540} = 8{,}29.$$

Таблица 29

Диаметр эритроцита, мк	1-й мазок	2-й мазок	3-й мазок	4-й 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

Более сложная формула (3.22) оказывается полезной тогда, когда сводятся воедино литературные данные из разных источников, где часто не указывается объем выборки, но дается значение $s_{\bar{x}}$.

Пример 8. В литературе приводятся (по данным разных авторов) следующие значения процентного содержания воды в дрожжах одного штамма: 68 ± 4; 70 ± 4; 67 ± 3; 59 ± 8; 65 ± 5. Какова, по этим данным, наилучшая оценка генерального среднего? (Результаты записаны символически в виде $\bar{x} \pm s_{\bar{x}}$.)

Считая, что приведенные значения $s_{\bar{x}}$ являются достаточно точными оценками соответствующих $\sigma_{\bar{x}}$, имеем

$$\bar{x} = \frac{\frac{68}{16} + \frac{70}{16} + \frac{67}{9} + \frac{59}{64} + \frac{65}{25}}{\frac{1}{16} + \frac{1}{16} + \frac{1}{9} + \frac{1}{64} + \frac{1}{25}} =$$

$$= \frac{4{,}25 + 4{,}37 + 7{,}45 + 0{,}92 + 2{,}60}{0{,}0625 + 0{,}0625 + 0{,}1111 + 0{,}0156 + 0{,}0400} = \frac{19{,}59}{0{,}2917} = 67{,}2\%.$$

Это более правильная оценка, чем среднее арифметическое

$$(68 + 70 + 67 + 59 + 65) : 5 = 65{,}9\%.$$

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

$$s^2 = \frac{1}{n-w}\sum_{j=1}^{w}(n^{(j)}-1)\,s^2_{(j)} + \frac{1}{n-1}\sum_{j=1}^{w} n^{(j)}(\bar{x}^{(j)}-\bar{x})^2. \quad (3.24)$$

Первый член в (3.24) есть средневзвешенная из выборочных дисперсий; при этом в качестве «веса» для каждой из $s^2_{(j)}$ берется ее число степеней свободы $f^{(j)} = n^{(j)} - 1$ *. Необходимость второго члена видна из следующего: если бы все частные дисперсии были равны нулю (т. е. в пределах каждой отдельной совокупности все варианты были бы одинаковы), но частные средние значения были бы 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 (3.8) дало бы как раз выражение, совпадающее со вторым членом формулы (3.24). Следовательно, смысл этой формулы состоит в том, что *дисперсия суммарной совокупности равна средней частной дисперсии плюс дисперсия частных средних.* Например, для выборки из табл. 29 имеем:

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

В данном случае поправка, вносимая вторым членом формулы (3.24), 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 $s_{\bar{x}} = s/\sqrt{n}$. Для данных из табл. 29 имеем

$$s_{\bar{x}} = \sqrt{\frac{0{,}78}{1549}} = 0{,}022.$$

* 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 $\hat{x}^{(j)}$ и σ_j^2 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 — ибо мы рассматриваем $\hat{x}^{(j)}$ и σ_j^2 как характеристики именно этой совокупности, а не как оценки для всей популяции данного биологического вида.

Формула (3.24) 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 (3.5) 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. Подробней эти вопросы будут обсуждаться в главе 5 при рассмотрении основ дисперсионного анализа.

Если производится обработка литературных данных и приведены только значения $s_{\bar{x}(j)}$ (а значения $n^{(j)}$ и $s_{(j)}$ неизвестны, но известно, что $n^{(j)}$ велики), то обычно считают, что все $s^2_{(j)}$ одинаковы. Тогда формула (3.24) дает (если пренебречь вторым членом)

$$s^2 = \frac{1}{n-w} s^2_{(j)} \sum_{j=1}^{w} (n^{(j)} - 1) = s^2_{(j)},$$

так что выражение

$$s^2_{\bar{x}} = \frac{s^2}{n} = \frac{s^2}{\sum_j n^{(j)}} = \frac{1}{\sum_j \frac{n^{(j)}}{s^2}}$$

можно записать в виде

$$s^2_{\bar{x}} = \frac{1}{\sum_j \frac{n^{(j)}}{s^2_{(j)}}} = \frac{1}{\sum_j \frac{1}{s^2_{\bar{x}(j)}}}, \qquad (3.25)$$

т. е. мы получаем дробь, в числителе которой стоит единица, а знаменатель такой же, как и в формуле (3.22).

Для примера 8 находим:

$$s^2_{\bar{x}} = \frac{1}{0{,}2917} = 3{,}45; \quad s_{\bar{x}} = \sqrt{3{,}45} = 1{,}9.$$

Таким образом, окончательно имеем (67,2 ± 1,9)%.

§ 6. Стандартные ошибки сложных средних

Как уже указывалось ранее, чаще всего величина, интересующая исследователя, является некоторой функцией от величины (или ряда величин), непосредственно определяемой в опыте. В § 4

и 6 гл. 1 было показано, как вычисляются среднее значение и дисперсия для некоторых функций от статистических величин: $z = f(x)$, $z = f(x, y)$.

Поскольку обычно изучается выборка, то получаемое значение $\bar{x}$ есть лишь **оценка** генерального среднего $\hat{x}$, причем разброс выборочных $\bar{x}$ относительно $\hat{x}$ характеризуется величиной $\sigma_{\bar{x}}$. Аналогичным образом значения[1] $\tilde{z} = f(x)$ 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 (3.4), подставляя соответствующее значение σ. Так, если $z = ax$, то согласно (1.18)

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

так что

$$\sigma_{\tilde{z}} = |a|\sigma_{\bar{x}}, \quad s_{\tilde{z}} = |a|s_{\bar{x}}; \tag{3.26}$$

если $z = x + a$, то в соответствии с (1.19) получаем:

$$\sigma_{\tilde{z}} = \sigma_{\bar{x}+a} = \sigma_{\bar{x}}; \quad s_{\tilde{z}} = s_{\bar{x}}. \tag{3.27}$$

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

$$\left.\begin{aligned} &\text{если } z = \alpha e^{\beta x}, \text{ то } \sigma_{\tilde{z}} = |\beta|\bar{z}\sigma_{\bar{x}}; \\ &\text{если } z = \alpha \lg \beta x, \text{ то } \sigma_{\tilde{z}} = \frac{|\alpha|}{\bar{x}}\sigma_{\bar{x}} \end{aligned}\right\} \tag{3.28}$$

(α и β — численные множители);

$$\text{если } z = x^m, \text{ то } \sigma_{\tilde{z}} = |m|(\bar{x})^{m-1}\sigma_{\bar{x}}. \tag{3.29}$$

Перейдем к случаю, когда z зависит от двух статистических величин. Если $z = x \pm y$, то в соответствии с (1.20) и (1.21) получим (когда x и y варьируют независимо одна от другой):

$$s_{\bar{x}\pm\bar{y}} = \sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}. \tag{3.30}$$

---

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

Пример 9. Животные, подвергшиеся облучению ультрафиолетовыми лучами, за месяц прибавили в весе на 6,8 ± 0,4 *кг*; животные, не облучавшиеся (контрольная группа той же численности, что и опытная), за это же время прибавили в весе на 5,2 ± 0,3 *кг*. Какое увеличение привеса дает облучение?

В данном случае по условию эксперимента обе группы животных независимы, так что

$$\bar{z} \pm s_{\bar{z}} = (6{,}8-5{,}2) \pm \sqrt{0{,}4^2 + 0{,}3^2} = 1{,}6 \pm 0{,}5 \text{ кг}.$$

Если численности достаточно велики, чтобы можно было пользоваться нормальным приближением, то 99%-ный доверительный интервал эффекта облучения будет (1,6—2,58·0,5; 1,6 + 2,58·0,5) или (0,3 ÷ 2,9) *кг*[1].

Пример 10. В табл. 30 приведены данные об урожаях риса на опытных делянках без применения (x) и с применением (y) удобрений. Требуется найти 95%-ный доверительный интервал для влияния удобрений.

Таблица 30

	x	y
	35,50	46,38
	32,66	41,36
	30,56	52,20
	36,63	46,20
	42,28	
	34,78	
	40,20	
Сумма	252,61	186,14
Среднее	36,09	46,53
$\Sigma\xi_i^2$	99,9	59,0

Общая оценка дисперсии равна здесь $s^2 = \frac{99{,}9 + 59{,}0}{6+3} = 17{,}66$, так что $s_{\bar{x}}^2 = 17{,}66 : 7 = 2{,}52$ и $s_{\bar{y}}^2 = 17{,}66 : 4 = 4{,}41$. Тогда по формуле (3.30) имеем

$$s_{\bar{y}-\bar{x}} = \sqrt{2{,}52 + 4{,}41} = 2{,}63.$$

[1] Во всех примерах этой главы предполагается, что распределение вариант в генеральной совокупности нормально.

Общее число степеней свободы равно $f = 6 + 3 = 9$, чему соответствует $t_{95}(9) = 2{,}26$. Так как разность средних значений равна $46{,}53 - 36{,}09 = 10{,}44$ *кг*, то 95%-ный доверительный интервал для влияния удобрений будет

$$(10{,}44 - 2{,}26 \cdot 2{,}63 \div 10{,}44 + 2{,}26 \cdot 2{,}63) \text{ или } (4{,}49 \div 16{,}39) \text{ кг}.$$

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

$$s_{\tilde{z}} = |\tilde{z}| \sqrt{\left(\frac{s_{\bar{x}}}{\bar{x}}\right)^2 + \left(\frac{s_{\bar{y}}}{\bar{y}}\right)^2} \qquad (3.31)$$

Пример 11. Измерения показали, что на площади 1 $м^2$ имеется $x = 247 \pm 8$ колосьев пшеницы. Известно, что для данного сорта суммарный вес зерен в колосе составляет $y = 851 \pm 14$ *мг*. Какова урожайность этого сорта?

Имеем: $z = xy$, так что

$$\bar{z} = \bar{x}\bar{y} = 247 \cdot 851 \approx 210\,200 \text{ мг} = 210{,}2 \text{ г};$$

далее по формуле (3.31) находим

$$s_{\bar{z}} = 210{,}2 \sqrt{\left(\frac{8}{247}\right)^2 + \left(\frac{14}{851}\right)^2} = 210{,}2 \sqrt{0{,}0324^2 + 0{,}0165^2} = 7{,}63 \text{ г}.$$

При пересчете на гектар получим $21{,}02 \pm 0{,}76$ *ц/га*, т. е. 99%-ный доверительный интервал для урожайности есть $(21{,}02 - 2{,}58 \cdot 0{,}76;\ 21{,}02 + 2{,}58 \cdot 0{,}76)$, или $(19{,}06;\ 22{,}98)$ *ц/га*.

§ 7. Нахождение оценки для σ и доверительного интервала для $\hat{x}$ по размаху варьирования

Чем больше стандартное отклонение σ генеральной совокупности, тем больше, вообще говоря, будет размах значений

$$R = x_{\max} - x_{\min}$$

в выборке. Однако величина R в большей степени зависит от случайностей образования выборки, чем другие выборочные характеристики, при вычислении которых используется информация о значениях *всех* вариант выборки. Поэтому оценка рассеяния вариант в генеральной совокупности по размаху в одной только выборке является очень неточной. Но если имеется большое число (10—20) выборок, то усредненный по всем этим выборкам размах $\bar{R}$ может дать достаточно точную оценку σ.

Установлено, что наилучшая оценка при помощи размаха получается при очень малых объемах выборок (n порядка 6—8). Это позволяет применить прием искусственного разбиения выборки на «подвыборки» по 6—8 вариант в каждой.

Пример 12. В табл. 4 были приведены данные о длине 100 зерен пшеницы. Найдем средний размах $\bar{R}$ для 20 «подвыборок», каждая из которых содержит пять вариант одной из двадцати строк таблицы.

В первой строке содержатся варианты

$$5{,}39 \quad 5{,}43 \quad 5{,}49 \quad 5{,}42 \quad 5{,}45.$$

Здесь $x_{\max} = 5{,}49$; $x_{\min} = 5{,}39$, так что

$$R_1 = 5{,}49 - 5{,}39 = 0{,}10;$$

для второй строки

$$5{,}42 \quad 5{,}52 \quad 5{,}35 \quad 5{,}45 \quad 5{,}37$$

$x_{\max} = 5{,}52$; $x_{\min} = 5{,}35$ и $R_2 = 0{,}17$. Найдя таким образом все двадцать R_j, получим

$$\bar{R} \approx 0{,}24.$$

Соотношение между $\bar{R}$ и σ зависит от объема n' подвыборок, по которым находились R_j (эти подвыборки должны иметь все одинаковый объем). Оценку стандартного отклонения по размаху варьирования (обозначим ее $s^{(R)}$) вычисляют по формуле

$$s^{(R)} = \rho^{(n')}\,\bar{R}^{(n')}, \tag{3.32}$$

где $\rho^{(n')}$ — коэффициент, зависящий от n'; значения $\rho^{(n')}$ приведены в табл. IX Приложений.

В нашем случае $n' = 5$, так что $\rho^{(n')} = 0{,}430$; поэтому

$$s^{(R)} = 0{,}24 \cdot 0{,}430 = 0{,}103.$$

Более сложный расчет в § 6 гл. 1 дал близкую оценку 0,095.

Можно было бы получить оценку $s^{(R)}$, разбивая выборку на 10 подвыборок по $n' = 10$ вариант в каждой. (Рекомендуем читателю проделать этот расчет, составляя подвыборки несколькими разными способами: а) объединяя попарно 1-ю и 2-ю строки, затем 3-ю и 4-ю строки и т. д.; б) объединяя 1-ю и 20-ю строки, 2-ю и 19-ю строки и т. д.; в) объединяя 1-ю и 11-ю строки, 2-ю и 12-ю строки и т. д.).

Использование размаха варьирования для получения оценки стандартного отклонения позволяет упростить построение доверительных интервалов для генеральных средних значений, так как стандартная ошибка среднего значения $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ может быть

оценена величиной

$$s_{\bar{x}}^{(R)} = \frac{\rho^{(n')}\,\overline{R}^{(n')}}{\sqrt{n}}.$$

При этом следует учитывать, что числа n' (объемы подвыборок) и n (объем всей выборки) не совпадают.

Если подвыборки не составляются (так что $n' = n$), то полуширина доверительного интервала приближенно равна

$$u_P s_{\bar{x}}^{(R)} = \frac{u_P \rho^{(n)}}{\sqrt{n}} R.$$

Значения

$$\frac{u_P \rho^{(n)}}{\sqrt{n}} = d_P^{(n)}$$

для $P = 0{,}95$ и $P = 0{,}99$ и для разных n даны в табл. X Приложений. Полуширина доверительного интервала получается простым умножением табличного значения $d_P^{(n)}$ на размах R.

Пример 13. В табл. 31 приведены данные о времени (в *мин*) сохранения жизнеспособности в анаэробных условиях перонеальных нервов кроликов. Найдем 95%-ный доверительный интервал для среднего времени.

Таблица 31

		$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	16,2	−3,2	10,24
	22,5	3,1	9,61
	21,4	2,0	4,00
	19,6	0,2	0,04
	24,8	5,4	29,16
	21,4	2,0	4,00
	19,0	0,4	0,16
	14,7	−4,7	22,09
	13,3	−6,1	37,21
	23,0	−3,6	12,96
	16,8	2,6	6,76
	20,1	−0,7	0,49
Сумма	232,8		136,72
Среднее	19,4		

По данным табл. 31 имеем

$$s_{\bar{x}} = \sqrt{\frac{136{,}72}{12 \cdot 11}} = 1{,}02;$$

табл. IV Приложений дает $t_{95}(11) = 2{,}20$, так что $t_{95}s_{\bar{x}} = 2{,}20 \cdot 1{,}02 \approx 2{,}2$, и для доверительного интервала получается $17{,}2 \div 21{,}6$.

Воспользуемся теперь оценкой σ через размах варьирования. В данном случае $R = 24{,}8 - 13{,}3 = 11{,}5$, а для $d_P^{(n)}$ имеем из табл. X Приложений 0,174 (для $n = 12$); поэтому $d_R^{(n)} R = 0{,}174 \cdot 11{,}5 = 2{,}0$, так что доверительный интервал будет $17{,}4 \div 21{,}4$. Результаты довольно близки, а бо́льшая простота второго способа вычислений очевидна.

Следует, однако, иметь в виду, что описанный в этом параграфе метод оценки σ по размаху основан на предположении о том, что варианты в генеральной совокупности распределены нормально. Если нет уверенности в справедливости такого предположения, то этот метод применять не надо.

§ 8. Стандартная ошибка среднего значения при распределении Пуассона

Вычисление стандартной ошибки среднего значения существенно упрощается, если варианты в генеральной совокупности распределены по закону Пуассона. Как указывалось в § 6 гл. 2, по закону Пуассона распределено, в частности, число 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 все шире внедряется в биологические исследования, 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

$$s_{\bar{x}} = \sqrt{\frac{\Sigma \xi_i^2}{k(k-1)}},$$

после чего, при достаточно большом k, получим окончательный результат в виде $\bar{x} \pm u_p s_{\bar{x}}$.

То обстоятельство, что распределение скорости счета является пуассоновским, сильно упрощает вычисления. Действительно, это распределение имеет ту особенность, что для него $\mu_2 = m_1$

(см. гл. 2, § 6). Но тогда (учитывая, что при радиометрических измерениях n обычно достаточно велико) имеем приближенно:

$$\sigma = \sqrt{\bar{x}} = \sqrt{\frac{\Sigma x_i}{k}}; \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{k}} = \frac{\sqrt{\Sigma x_i}}{k},$$

так что результат имеет вид

$$\bar{x} \pm u_p \sigma_{\bar{x}} = \frac{\Sigma x_i}{k} \pm u_p \frac{\sqrt{\Sigma x_i}}{k}.$$

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

$$\bar{x} \pm u_p \sigma_{\bar{x}} = \frac{X}{\tau} \pm u_p \frac{\sqrt{X}}{\tau}. \qquad (3.33)$$

Пример 14. В табл. 32 приведены результаты десяти одноминутных измерений числа импульсов за счет фона. Мы видим,

Таблица 32

i	x_i	ξ_i	ξ_i^2	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

что действительно величины

$$\sqrt{\frac{\Sigma \xi_i^2}{k(k-1)}} = \sqrt{\frac{544}{10 \cdot 9}} = 2{,}46 \quad \text{и} \quad \frac{\sqrt{X}}{\tau} = \frac{\sqrt{550}}{10} = 2{,}34$$

достаточно близки.

Пример 15. 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:

$$24{,}8-2{,}2 = 22{,}6 \approx 23, \quad 24{,}8 + 2{,}2 = 27 \quad имп/мин.$$

Пример 16. Определим 95%-ный доверительный интервал для среднего числа букв «ц» в отрывках из 100 слов чеховских текстов (см. пример 11 из гл. 2).

В данном случае $X = 292$, $\tau = 1000$, так что

$$\frac{X}{\tau} \pm u_{95}\frac{\sqrt{X}}{\tau} = \frac{292}{1000} \pm 1{,}96\frac{\sqrt{292}}{1000} = 0{,}292 \pm 0{,}033,$$

т. е. границы доверительного интервала будут:

$$0{,}292-0{,}033 = 0{,}259; \quad 0{,}292+0{,}033 = 0{,}325.$$

Значения $\mu_2 = 0{,}301$ и $\mu_3 = 0{,}304$ находятся внутри этого интервала.

Формула (3.33), 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. XIII Приложений приводятся соответствующие значения при $P = 95\%$ и $P = 99\%$ для X от 0 до 50.

Пример 17. Найдем 99%-ные доверительные границы при $X = 49$. По формуле (3.33) имеем (при $\tau = 1$)

$$49 \pm 2{,}58\sqrt{49} = 49 \pm 18{,}06 = 30{,}94 \div 67{,}06.$$

Между тем в табл. XIII Приложений находим значения 32,85 и 70,08. Следовательно, даже при $X = 49$ формула (3.33) дает не очень точные значения доверительных границ. При меньших X различие еще заметней.

§ 9. Доверительный интервал для доли вариант при альтернативном распределении

Альтернативное распределение относится к совокупностям с качественной классификацией, поэтому здесь нельзя ввести такие параметры, как среднее значение и дисперсия. Однако можно указать численный параметр, имеющий вполне объективный смысл,— долю вариант одного из двух типов. Если, например, имеется совокупность из 46 поврежденных и 118 неповрежденных клубней картофеля, то доля поврежденных клубней будет

$$\hat{p} = \frac{46}{46+118} = \frac{48}{164} \approx 0{,}28, \text{ или } 28\%.$$

Вообще

$$\hat{p} = \frac{N_1}{N_1 + N_2} = \frac{N_1}{N},$$

где N_1 и N_2 — численности альтернатив, а $N = N_1 + N_2$ есть численность всей совокупности.

Когда рассматриваемая совокупность является выборкой из некоторой бесконечной генеральной совокупности, то величина

$$p = \frac{n_1}{n_1 + n_2} = \frac{n_1}{n} \tag{3.34}$$

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 (2.1). Соответствующие расчеты очень громоздки, поэтому были составлены таблицы, в которых можно сразу найти 95%- 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 (1961, табл. 28, стр. 181—194).

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 $\hat{p}$ вычисляются по формулам:

$$p_н = p - u_p\sigma_p; \quad p_в = p + u_p\sigma_p. \tag{3.35}$$

Что касается величины 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

$$\bar{x} = \frac{1}{n}(n_1x_1 + n_2x_2) = \frac{1}{n}(n_1 \cdot 1 + n_2 \cdot 0) = \frac{n_1}{n}.$$

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 второй начальный

момент (см. § 8 гл. 1) нашего условного распределения:

$$m_2 = \frac{1}{N}(N_1 x_1^2 + N_2 x_2^2) = \frac{1}{N}(N_1 \cdot 1^2 + N_2 \cdot 0^2) = \frac{N_1}{N} = \hat{p}.$$

Тогда второй центральный момент будет равен

$$\mu_2 = m_2 - m_1^2 = \hat{p} - \hat{p}^2 = \hat{p}(1 - \hat{p}),$$

так что

$$\sigma_p = \sqrt{\mu_2} = \sqrt{\hat{p}(1 - \hat{p})}$$

Это дает

$$\sigma_p = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}. \qquad (3.36)$$

Генеральная доля $\hat{p}$ чаще всего неизвестна, поэтому, если располагают только данными о выборке, величину $\hat{p}$ заменяют ее выборочной оценкой p, считая приближенно

$$\sigma_p \approx \sqrt{\frac{p(1-p)}{n}}; \qquad (3.37)$$

если доля выражена в процентах, то

$$\sigma_p \approx \sqrt{\frac{p(100-p)}{n}}. \qquad (3.37')$$

Пример 18. 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 (3.37′) 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 $(6{,}0 \div 11{,}2)$ %.

Пример 19. При рентгеновском облучении 10 мышей дозой в 550 *рентген* погибли 5. 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).

При планировании эксперимента желательно знать заранее, какого объема выборка понадобится, чтобы получить доверительный интервал заданной ширины. В § 4 этой главы такая задача была решена для среднего значения совокупности. Сейчас рассмотрим аналогичную задачу для доли (процента) вариант.

Если ожидаемое значение p не очень мало (или не очень близко к 100%), то неточность результата лучше всего характеризуется абсолютной погрешностью $\delta = u_P \sigma_p$. При заданном δ наименьшее допустимое значение n^* определяется равенством

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

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

$$n^* = \frac{u_P^2}{\delta^2} 2500 \quad (\delta \text{ в } \%). \tag{3.38}$$

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

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

при $P = 99\%$ и $\delta = 3\%$ получаем

$$n > \frac{2{,}58^2}{3^2} 2500 = 1843.$$

Когда ожидаемое значение p мало, лучшей характеристикой неточности результата является относительная погрешность $\varepsilon = u_p\sigma_{\hat{p}}/p$. 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

$$n^* = \frac{u_p^2}{\varepsilon^2}\cdot\frac{100}{p}; \tag{3.39}$$

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

При малых объемах выборок нормальное приближение дает слишком неточные результаты, особенно из-за замены в формуле для $\sigma_{\hat{p}}$ неизвестной доли $\hat{p}$ ее оценкой p — ведь различие между p и $\hat{p}$ равно в среднем как раз величине $\sigma_{\hat{p}}$, которая, согласно (3.36), возрастает при уменьшении n.

Чтобы исправить это положение, Р. А. Фишер предложил пользоваться вспомогательной величиной φ, связанной с p равенством

$$p = \sin^2\frac{\varphi}{2}, \tag{3.40}$$

откуда

$$\varphi = 2\arcsin\sqrt{p}. \tag{3.41}$$

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:

$$\sigma_\varphi \approx \frac{1}{\sqrt{n}} \tag{3.42}$$

(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.

Чтобы получить доверительные границы для доли вариант, нужно найти φ по формуле (3.41) и вычислить

$$\varphi + u_p\sigma_\varphi = \varphi + \frac{u_p}{\sqrt{n}},$$

а затем по формуле (3.40) 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 (3.40) и (3.41) с применением обычной таблицы тригонометрических функций очень неудобен. Поэтому была составлена специальная таблица, которая непосредственно связывает значения p и φ (табл. VII Приложений).

Пример 20. 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:

$$\varphi_н = 0{,}437 - 0{,}245 = 0{,}192; \quad \varphi_в = 0{,}437 + 0{,}245 = 0{,}682;$$

пользуясь опять таблицей VII Приложений, получаем:

$$p_н = 0{,}9\%; \quad p_в = 11{,}2\%.$$

Расчет по формулам (3.37′) и (3.35) 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 21. Уточним результат примера 18, пользуясь φ-преобразованием. Здесь

$$p = 8{,}6\%; \ \varphi = 0{,}595; \ u_P\sigma_\varphi = \frac{1{,}96}{\sqrt{430}} = 0{,}094,$$

так что

$$\varphi_н = 0{,}595 - 0{,}094 = 0{,}501; \varphi_в = 0{,}595 + 0{,}094 = 0{,}689;$$

поэтому по табл. VII Приложений

$$p_н = 6{,}2\%; \ p_в = 11{,}4\%.$$

Так как в этом случае n довольно велико, то различие между результатами, даваемыми φ-преобразованием и нормальным приближением, оказалось небольшим.

Пример 22. Пересчитаем, пользуясь φ-преобразованием, данные из примера 19:

$$p = 50\%;\ \varphi = 1{,}571;\ u_{99}\sigma_\varphi = \frac{2{,}58}{\sqrt{10}} = 0{,}816.$$

Поэтому

$$\varphi_н = 1{,}571 - 0{,}816 = 0{,}755;\ \varphi_в = 1{,}571 + 0{,}816 = 2{,}387,$$

что дает

$$p_н = 13{,}6\%;\ p_в = 86{,}4\%.$$

Конечно, этот доверительный интервал все еще слишком широк, чтобы иметь достаточное практическое значение.

При малых p можно пользоваться приближением Пуассона, считая

$$p_н = \frac{x_н}{n};\ p_в = \frac{x_в}{n} \tag{3.43}$$

и беря значения $x_н$ и $x_в$ непосредственно из табл. XIII Приложений. Так, для примера 18 находим в табл. XIII Приложений $x_н = 26{,}05$ и $x_в = 51{,}00$, так что

$$p_н = \frac{26{,}05}{430} = 0{,}061 = 6{,}1\%;\ p_в = \frac{51{,}00}{430} = 0{,}119 = 11{,}9\%.$$

Для примера 20 $x_н = 0{,}619$; $x_в = 8{,}77$ и

$$p_н = \frac{0{,}619}{64} \approx 0{,}0097 \approx 1{,}0\%;\ p_в = \frac{8{,}77}{64} = 0{,}137 = 13{,}7\%.$$

Эти результаты точнее, чем найденные другими приближенными методами, и в то же время получаются совсем просто.

Сравнение результатов разных методов расчета дано в табл. 33.

Таблица 33

Номер примера	Нормальное приближение	φ-преобразование	Приближение Пуассона	Точные значения
18	6,0÷11,2	6,2÷11,4	6,1÷11,9	6,1÷11,6
19	9,2÷90,8	13,6÷86,4	p не мало	12,8÷87,2
20	—0,5÷ 9,9	0,9÷11,2	1,0÷13,7	1,0÷13,3

ПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ РАЗЛИЧИЯ

§ 1. Смысл критериев различия

В настоящей и последующих главах будет идти речь о критериях, позволяющих ответить на следующие вопросы:

1) относится ли та или иная варианта к данной статистической совокупности?

2) соответствует ли данное эмпирическое распределение тому или иному теоретическому распределению?

3) являются ли данные эмпирические совокупности выборками из одной и той же генеральной совокупности?

Первый вопрос в общем ставится так. Имеется некоторая выборочная статистическая совокупность, распределенная определенным образом. Некоторые варианты очень далеко отстоят от среднего значения, и возникает сомнение, является ли это результатом маловероятных, но все же возможных больших отклонений от центра соответствующей генеральной совокупности, или же результатом того, что в рассматриваемую выборку почему-то оказались включенными варианты, принадлежащие в действительности к другой генеральной совокупности.

Второй вопрос возникает в связи с тем, что из-за случайности в образовании выборки распределение вариант в выборке всегда отличается от их распределения в генеральной совокупности; поэтому если в генеральной совокупности варианты распределены по определенному теоретическому закону, то распределение в выборке будет заведомо отклоняться от этого закона. (Отсюда следует, что сам факт отклонения выборочного распределения от того или иного теоретического распределения еще не дает основания утверждать, что и в генеральной совокупности распределение не подчиняется данному теоретическому закону.) Таким образом, вопрос сводится к тому, можно ли расхождение между выборочным и предположенным теоретическим распределениями отнести за счет расхождения между выборкой и генеральной совокупностью или же оно является результатом того, что сама генеральная совокупность отклоняется от данного теоретического распределения.

Сущность третьего вопроса совершенно аналогична: расхождение между эмпирическими распределениями может быть просто расхождением между разными выборками из одной и той же генеральной совокупности, но может также объясняться тем, что они выбраны из разных генеральных совокупностей.

Мы видим, что во всех трех случаях приходится решать вопрос о том, является ли наблюдаемое различие между объектами отражением какого-то реального различия или же оно есть результат случайности, сопровождающей попадание вариант в выборку. Это позволяет сформулировать общий подход к решению всех этих задач. Именно, в любом случае задача может быть сведена к проверке гипотезы об отсутствии реального различия. Эту гипотезу называют *нулевой гипотезой*; обычно для нее применяется специальное обозначение H_0.

В задаче 1 нулевая гипотеза гласит, что сомнительная варианта принадлежит к той же генеральной совокупности, что и данная эмпирическая совокупность. В задаче 2 проверке подлежит нулевая гипотеза о том, что различие между выборочным распределением и теоретическим является случайным, т. е. что нет реального различия между распределением в генеральной совокупности, из которой взята выборка, и предположенным теоретическим распределением. В задаче 3 нулевая гипотеза состоит в том, что данные эмпирические совокупности являются выборками из одной и той же генеральной совокупности, так что между генеральными совокупностями, выборками из которых являются данные эмпирические совокупности, нет реального различия.

Правильность нулевой гипотезы можно проверить следующим образом. Предположив справедливость нулевой гипотезы, т. е. отсутствие реального различия, мы вычисляем вероятность того, что из-за случайности выборки расхождение 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 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 здесь состоять в том, что генеральная совокупность симметрична ($\hat{A} = 0$), 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 з н а́ ч и м ы м (т. е. реальным), если вероятность того, что нулевая гипотеза верна, меньше уровня значимости (его обозначают буквой α).

В дальнейшем будем считать, что если вероятность нулевой гипотезы меньше $\alpha = 0{,}01 = 1\%$, то она отвергается; если вероятность лежит в пределах от $\alpha = 0{,}01 = 1\%$ до $\alpha = 0{,}05 = 5\%$, то возможность отвергнуть нулевую гипотезу сомнительна; если же $\alpha \geqslant 0{,}05 = 5\%$, то нулевая гипотеза принимается.

Очевидно, уровень значимости характеризует, в какой мере мы рискуем ошибиться, отвергая нулевую гипотезу. Условно говоря, при $\alpha \geqslant 0{,}05$ риск ошибиться (отвергнув нулевую гипотезу, в то время как в действительности она верна) становится заметным; при $\alpha < 0{,}01$ этот риск очень мал.

Однако принятый здесь выбор уровня значимости не может считаться универсальным. Этот выбор в значительной мере 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.).

Надо всегда иметь в виду, что утверждение о том, что нет достаточных оснований отвергнуть гипотезу об отсутствии различия, вовсе не равносильно утверждению, что отсутствие различия доказано. Иными словами, можно лишь утверждать, что данные наблюдений н е п р о т и в о р е ч а т предположению об отсутствии различия, но нельзя утверждать, что эти данные д о к а з ы в а ю т 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 симметрична, то мы рискуем впасть в ошибку: наблюденное значение $A = 0{,}3$ могло получиться и при генеральном значении $\hat{A}$, 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 обозначается β.

Вернемся теперь к примеру 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 входа — 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, н а с к о л ь к о она меньше. Поэтому более целесообразно составить такую таблицу, в которой для каждого значения n указывалось бы то значение A_{01}, которое удовлетворяет условию: вероятность того, что $A > A_{01}$, равна $0{,}01 = 1\%$; эта таблица будет иметь только один вход: n. Величину A_{01} назовем однопроцентным верхним *критическим значением* коэффициента асимметрии (для данного объема выборки n).

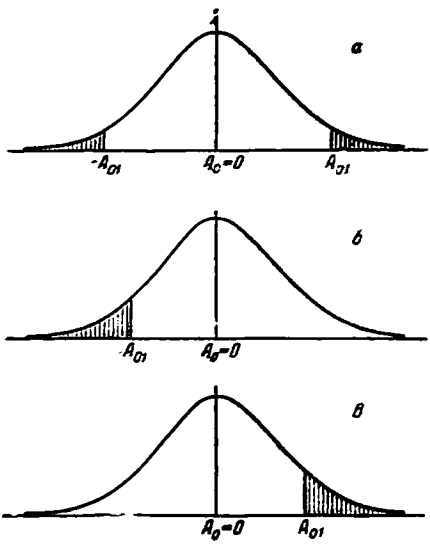

Рис. 39

Численное значение величины A_{01} (при данном n) существенно зависит от того, как формулируется нулевая гипотеза H_0. Часто эта гипотеза просто утверждает, что $\hat{A} = 0$. В этом случае H_0 будет отвергаться тогда, когда выборочный коэффициент асим-

метрии окажется либо больше A_{01}, либо меньше $-A_{01}$ (рис. 39, *а*); величина A_{01} должна здесь удовлетворять тому условию, что площади, заштрихованные на рис. 39, *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 $\hat{A} \geqslant 0$, в других случаях $\hat{A} \leqslant 0$. Тогда в первом случае H_0 будет отвергаться при $A < -A_{01}$, а во втором случае — при $A > A_{01}$ (рис. 39, *б* и *в*), причем здесь A_{01} опять удовлетворяет условию, что заштрихованная площадь на рис. 39, *б* и *в* составляет 1% от площади под кривой распределения (на рис. 39 масштаб для наглядности не соблюден); но ясно, что A_{01} из рис. 39, *б* и *в* отличается от A_{01} из рис. 39, *а*.

Критерии, проверяющие нулевую гипотезу типа $\hat{A} = 0$, называются *двусторонними*, а критерии, проверяющие нулевую гипотезу типа $\hat{A} \geqslant 0$ или $\hat{A} \leqslant 0$, называются *односторонними*; происхождение этих названий ясно из рис. 39.

В примере 1 решался вопрос о значимости различия между коэффициентами асимметрии эмпирической совокупности (A) и некоторой генеральной совокупности ($\hat{A} = 0$), т. е. вопрос о значимости различия между численными значениями некоторого параметра (в данном случае — коэффициента асимметрии) двух сравниваемых совокупностей. Критерии, проверяющие нулевую гипотезу о значениях каких-либо параметров, называются *параметрическими*; настоящая глава посвящена именно этим критериям.

§ 2. Принадлежность варианты к совокупности

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 вариабильностью случайной величины, вследствие которой 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, условия 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.

Известно, что при нормальном распределении вероятность появления вариант, отстоящих от среднего значения 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 $\hat{x}$ равна 0,01. Отсюда часто делается вывод, что при выбранном уровне значимости α можно отбрасывать варианты, для которых $|u| = \frac{|x - \hat{x}|}{\sigma} > u_\alpha$. Однако такое заключение ошибочно. Ведь ясно, что как бы ни была мала вероятность появления вариант, они вполне могут появиться при достаточно большом объеме выборки. Очевидно, при построении критерия исключения (соответствующее критическое значение обозначим τ_α) надо исходить из условия, что в выборке данного объема n из нормальной генеральной совокупности не должно содержаться, с определенной вероятностью P, н и о д н о й варианты, отклоняющейся от $\hat{x}$ больше, чем на $\tau_\alpha\sigma$. Выбранный уровень значимости $\alpha = 1 - P$ имеет здесь тот смысл, что если выбрать из нормальной генеральной совокупности большое число выборок объема n каждая, то в среднем лишь в $100\,\alpha$ процентах из них будут попадаться варианты вне пределов $\hat{x} \pm \tau_\alpha\sigma$, а $100\,(1 - \alpha) = 100P$ процентов выборок не будут содержать вариант вне этих пределов. Из сказанного ясно, что критические значения τ_α должны зависеть как от принятого уровня значимости α, так и от объема выборки n.

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

$$\tau_{\max} = \frac{x_{\max} - \bar{x}}{s}; \quad \tau_{\min} = \frac{\bar{x} - x_{\min}}{s}. \tag{4.1}$$

Пример 2. В табл. 34 дано распределение по весу зерен пшеницы определенного сорта, взятых из семенного фонда.

Таблица 34

Вес, мг	n_i	x_i	$n_i x_i$	$n_i x_i^2$
22	3	—5	15	75
24	1	—4	4	16
26	2	—3	— 6	18
28	12	—2	—24	48
30	19	—1	—19	19
32	27	0	0	0
34	22	1	22	22
36	10	2	20	40
38	3	3	9	27
40	1	4	4	16
	100		—68 +55	281
			—13	

$m_1 = -0{,}13$; $m_2 = 2{,}81$;
$\mu_2 = 2{,}81 - (-0{,}13)^2 = 2{,}79$;
$s = \sqrt{2{,}79} = 1{,}67$.

Обращает на себя внимание присутствие в этой совокупности трех зерен с весом 22 *мг*, очень сильно отклоняющимся от среднего для данного сорта веса 31,74 *мг*. Объяснение может быть двояким. С одной стороны, это может быть следствием естественной вариабильности веса зерен внутри данного сорта — ведь при нормальном распределении отклонения могут происходить в принципе сколь угодно большие. С другой стороны, появление в выборке этих трех зерен с весом 22 *мг* может являться следствием того, что в рассматриваемый семенной фонд по какой-то причине попало некоторое количество зерен другого сорта с меньшим средним весом.

Так как в данном случае отклонение сомнительных вариант от среднего значения составляет $5 - 0{,}13 = 4{,}87$ (в условных единицах), а $s = 1{,}67$ (в тех же единицах), то

$$\tau_{\min} = \frac{4{,}87}{1{,}67} = 2{,}92.$$

В табл. XII Приложений находим, что критическое значение $\tau_{05}(100)$ составляет 3,40. Поскольку $\tau_{\min} < \tau_{05}$, нулевая гипотеза о принадлежности рассматриваемых трех вариант к данной совокупности не может быть отвергнута.

Если нулевая гипотеза отвергается, то это значит, что сомнительные варианты квалифицируются как «артефакты». В таком случае их следует выбрасывать из дальнейшей обработки.

При большом объеме выборки вопрос об исключении «артефактов» не стоит особенно остро, так как относительный «вес» нескольких сомнительных вариант при вычислении усредненных параметров 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

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

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

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

то в этих обозначениях будем иметь:

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

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, как известно, ее можно оценить при помощи размаха $x_{(n)} - x_{(1)}$, используя значения $\rho^{(n)}$ из табл. IX Приложений. 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:

$$\tau' = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}; \quad \tau'' = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}. \qquad (4.2)$$

Критические значения для τ (в случае нормальной генеральной совокупности они, конечно, одинаковы для τ' и τ'') зависят как от принятого уровня значимости α, так и от объема выборки n. Значения $\tau_{01}(n)$ даются в табл. XI Приложений.

Пример 3. В приведенной выше выборке из 5 вариант

21 17 6 24 18

может вызвать сомнения варианта $x_{(1)} = 6$. По формуле (4.2) имеем

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

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

Хотя второй критерий проще первого (требует 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 (4.2) 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 (4.2) 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 (см. следующий параграф), то лучше избегать применения этих критериев.

§ 3. Критерий нормальности распределения

Последний абзац предыдущего параграфа дает пример того, что знание характера распределения вариант в генеральной совокупности может иметь существенное значение.

Как уже было упомянуто, при решении вопроса о различии между эмпирическим распределением и теоретическим нулевая гипотеза заключается в том, что генеральная совокупность, выборкой из которой является данное эмпирическое распределение, распределена по предполагаемому теоретическому закону, а отклонение эмпирического распределения от теоретического есть просто отклонение выборочного распределения от генерального (из-за случайного характера образования выборки). Нуле-

вая гипотеза отвергается только тогда, когда вероятность появления случайных отклонений такой (или еще большей) величины оказывается настолько малой, что появление такого отклонения за счет случайности образования выборки можно считать почти исключенным.

При этом прежде всего возникает вопрос, каким показателем характеризовать расхождение между двумя совокупностями. Как правило, в каждом случае можно указать несколько таких показателей, и выбор наиболее подходящего из них определяется как целями исследования, так и соображениями удобства.

Общий метод решения обсуждаемой задачи, пригодный для распределений любого вида, будет изложен в гл. 6. Здесь ограничимся случаем нормального распределения, используя некоторые его специальные свойства. А именно, как мы знаем из § 4 гл. 2, следствием нормальности распределения можно считать выполнение условий $\hat{A} = 0$ и $\hat{E} = 0$.

Однако, как мы уже видели в примере 1, было бы неправильно делать вывод, что если для выборки получено $A \neq 0$ или $E \neq 0$, то распределение в генеральной совокупности не является нормальным. Ведь известно, что выборочные параметры всегда несколько отличаются от соответствующих параметров генеральной совокупности. Значит, если даже в генеральной совокупности $A = 0$ (или $E = 0$), то выборочные A и E будут случайными величинами, распределенными примерно нормально со стандартными отклонениями $\sigma_A \approx \sqrt{\frac{6}{n+3}}$ и $\sigma_E \approx \sqrt{\frac{24}{n+5}}$. Следовательно, тот факт, что выборочная A равна σ_A (или даже больше, чем σ_A), отнюдь не противоречит тому, что генеральная $\hat{A}$ равна нулю. Только в том случае, когда

$$\frac{|A|}{|\sigma_A|} > u_\alpha,$$

можно считать маловероятным, что такое значение A получилось как случайное отклонение от генерального значения $\hat{A} = 0$; в этом случае можно считать, что асимметрия действительно имеет место (как принято говорить, является *значимой*).

Таким образом, статистически необходимым условием нормальности распределения следует считать не $A = 0$, $E = 0$, а условие

$$\frac{|A|}{\sigma_A} < u_\alpha, \quad \frac{|E|}{\sigma_E} < u_\alpha, \tag{4.3}$$

причем оба эти неравенства должны выполняться одновременно.

Согласно (3.14), величины σ_A и σ_E зависят только от n; поэтому можно вычислить для каждого n критические значения

$$A_\alpha(n) = \sigma_A(n)\, u_\alpha, \quad E_\alpha(n) = \sigma_E(n)\, u_\alpha,$$

ниже которых асимметрия и эксцесс незначимы при выбранном уровне значимости α.

Однако на практике дело обстоит сложнее. Причина заключается в том, что при не очень больших объемах выборок выборочные коэффициенты асимметрии и эксцесса не распределены нормально, а поэтому к ним нельзя применять u-критерий. Это особенно относится к коэффициенту эксцесса, для которого распределение выборочных оценок (при не очень больших n) заведомо асимметрично, поскольку значения E ограничены снизу числом -2, а сверху не ограничены (см. § 4 гл. 2).

Для коэффициента асимметрии положение исправляется тем, что критические значения A_α вычисляются исходя не из нормального, а из более точного распределения выборочных A; эти значения A_α приведены в табл. XXI Приложений. С A_α нужно сравнивать с м е щ е н н у ю оценку A, т. е. значение ρ_3, полученное подстановкой в формулу (1.24) эмпирических данных — без умножения на поправочный множитель из формулы (3.12), приводящий к несмещенной оценке.

Что касается проверки того, значим ли эксцесс распределения, то предпочитают вообще не пользоваться коэффициентом эксцесса, а применяют показатель

$$c = \frac{\overline{|\xi|}}{s}, \tag{4.4}$$

где

$$\overline{|\xi|} = \frac{1}{n}\sum n_i |\xi_i| = \frac{1}{n}\sum n_i |x_i - \bar{x}| \tag{4.5}$$

есть выборочная оценка среднего абсолютного отклонения $\langle|\xi|\rangle$ (см. § 6 гл. 1), а s — оценка стандартного отклонения. При этом используется то обстоятельство, что для нормального распределения

$$\langle|\xi|\rangle = \sigma\sqrt{\frac{2}{\pi}} \approx 0{,}798\,\sigma, \quad (\langle|\xi|\rangle/\sigma)_{\text{норм}} \approx 0{,}798,$$

а при наличии эксцесса значение $\langle|\xi|\rangle/\sigma$ отличается от 0,798. Преимуществом этого показателя является то, что его выборочные значения распределены около своих генеральных значений более симметрично, чем в случае E, и дисперсия этого распределения много меньше; это связано в основном с тем, что $\langle|\xi|\rangle/\sigma$ содержит статистические моменты более низкого порядка, чем $\hat{E}$.

Поскольку распределение выборочных c все же несколько асимметрично, нижние и верхние критические значения c_α расположены на различных расстояниях от 0,798; поэтому в табл. XXI Приложений приведены критические интервалы для c.

Нулевая гипотеза (о нормальности распределения вариант в генеральной совокупности) принимается, если одновременно $|\rho_3| \leqslant A_{05}$ и $|\overline{\xi}|/s$ находится в пределах, указанных в столбце c_{05}. Если же $|\rho_3| > A_{01}$ и л и $|\overline{\xi}|/s$ находится вне пределов, указанных в столбце c_{01}, то нулевая гипотеза отвергается. В целях сокращения в таблице везде опущен нуль целых и запятая.

Пример 4. Проверим нормальность распределения из табл. 5 (стр. 17). Расчет коэффициента асимметрии показан в табл. 35, причем

$$\mu_2 = 3{,}69 - (-0{,}19)^2 = 3{,}65, \quad s = 1{,}91;$$

$$\mu_3 = (-1{,}81) - 3 \cdot 3{,}69\,(-0{,}19) + 2\,(-0{,}19)^3 = 0{,}28;$$

$$\rho_3 = \frac{0{,}28}{3{,}65 \cdot 1{,}91} \approx 0{,}04.$$

Таблица 35

x_i	n_i	$n_i x_i$	$n_i x_i^2$	$n_i x_i^3$		n_i	$n_i x_i$	$n_i x_i^2$	$n_i x_i^3$
−5	1	− 5	25	−125	4	2	8	32	128
−4	4	−16	64	−256	5	1	5	25	125
−3	7	−21	63	−189			−80		−674
−2	11	−22	44	− 88			61		493
−1	16	−16	16	− 16					
0	30	0	0	−0	Σ	100	−19	369	−181
1	14	14	14	14					
2	8	16	32	64	m		−0,19	3,69	−1,81
3	6	18	54	162					
					μ			3,65	0,28

Теперь найдем $|\overline{\xi}|$. Так как $\bar{x} < 0$, то для отрицательных x_i (обозначим их $x_i^{(-)}$, а соответствующие им частоты $n_i^{(-)}$) модули отклонений равны

$$|\xi_i^{(-)}| = |x_i^{(-)}| - |\bar{x}|,$$

а для положительных x_i (обозначения соответственно $x_i^{(+)}$ и $n_i^{(+)}$), модули отклонений равны

$$|\xi_i^{(+)}| = |x_i^{(+)}| + |\bar{x}|;$$

для $x_i = 0$ модуль отклонения равен просто $|\bar{x}|$.

Если учесть, что

$$\Sigma n_i |\xi_i| = \Sigma n_i^{(-)} |\xi_i^{(-)}| + \Sigma n_i^{(+)} |\xi_i^{(+)}| + n^{(0)} |\xi^{(0)}|,$$

то подстановка сюда модулей отклонений даст

$$\Sigma n_i^{(-)} (|x_i^{(-)}| - |\bar{x}|) + \Sigma n_i^{(+)} (|x_i^{(+)}| + |\bar{x}|) + n^{(0)} |\bar{x}| =$$
$$= (\Sigma n_i^{(-)} |x_i^{(-)}| + \Sigma n_i^{(+)} |x_i^{(+)}|) + (-\Sigma n_i^{(-)} + \Sigma n_i^{(+)} + n^{(0)}) |\bar{x}|.$$

В данном случае (см. табл. 35):

$$\Sigma n_i^{(-)} |x_i^{(-)}| = 80; \quad \Sigma n_i^{(+)} |x_i^{(+)}| = 61;$$
$$\Sigma n_i^{(-)} = 1 + 4 + 7 + 11 + 16 = 39;$$
$$\Sigma n_i^{(+)} = 14 + 8 + 6 + 2 + 1 = 31, \; n^{(0)} = 30,$$

поэтому

$$\Sigma n_i |\xi_i| = (80 + 61) + (-39 + 31 + 30) \cdot 0{,}19 \approx 145.$$

Значит,

$$\overline{|\xi|} = \frac{145}{100} = 1{,}45, \quad \frac{\overline{|\xi|}}{s} = \frac{1{,}45}{1{,}91} = 0{,}760.$$

Обращаясь теперь к табл. XXI Приложений, видим, что $\rho_3 = 0{,}04$ много меньше 5%-ного критического значения $A_{05} = 0{,}389$, а значение $c = 0{,}760$ лишь немного выходит из 5%-ного критического интервала ($0{,}764 \div 0{,}834$) 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.

Пример 5. В табл. 13 (стр. 46) было представлено распределение длины волокна хлопка (для $n = 1000$). В § 8 гл. 1 мы вычислили коэффициент асимметрии этого распределения; получилось значение $A = -0{,}92$. Могла ли получиться такая асимметричная выборка из нормальной (т. е. симметричной) генеральной совокупности?

Так как в данном случае фактическое значение $|A| = 0{,}92$ больше, чем $A_{01}(1000) = 0{,}180$, то нулевая гипотеза отвергается — асимметрию следует считать значимой.

Если объем выборки превышает 1000, то можно пользоваться u-критерием (4.3).

Пример 6. В примере 6 гл. 2 было произведено разложение совокупности 5588 красных бобов на две нормальные подсо-

вокупности, причем в основу расчетов брались значения $A = 0{,}215$ и $E = -0{,}248$. Очевидно, эти расчеты не имели бы смысла, если бы величины A и E не были значимыми. Так как здесь $n > 1000$, то вместо табл. XXI Приложений мы воспользуемся формулами (4.3). Имеем: $\sigma_A \approx \sqrt{6/5591} \approx 0{,}033$; $\sigma_E \approx 2\sigma_A \approx 0{,}066$, так что

$$\frac{|A|}{\sigma_A} = \frac{0{,}215}{0{,}033} \approx 6{,}5; \quad \frac{|E|}{\sigma_E} = \frac{0{,}248}{0{,}066} \approx 3{,}8.$$

Оба эти числа больше, чем $u_{01} = 2{,}58$, поэтому величины $A = 0{,}215$ и $E = -0{,}248$ следует считать значимыми.

Качественное заключение о нормальности распределения можно получить графическим путем, используя метод спрямления нормальной кривой (§ 3 гл. 2), а также соображения, изложенные в § 4 гл. 2. При этом надо только учесть завышение отклонений от нормальности, возникающее из-за зависимости между соседними накопленными частотами (см. § 4 гл. 2). Этого искажения можно избежать, если разбить случайным образом выборку на две-три подвыборки и вычертить спрямленный график для каждой подвыборки отдельно: если отклонение от нормальности реально, то оно обнаружится на всех графиках. Конечно, такая операция усложняет графический анализ.

§ 4. Сравнение средних значений двух эмпирических совокупностей (критерий Стьюдента)

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

Пример 7. 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: берутся две группы животных (например мышей) одного пола и возраста — не обязательно одинаковой численности. Мышам одной группы вводится исследуемый препарат, мышам другой группы не вводится; первую группу будем называть опытной, вторую группу — контрольной. Затем мышам обеих групп вводят инфекцию и наблюдают, сколько дней переживают мыши опытной и контрольной групп. Пусть при этом получились результаты, приведенные в табл. 36. Прежде всего очевидно, что как в опыте, так и в контроле надо было исполь-

зовать именно группу животных, а не по одному животному: ведь некоторые из мышей контрольной группы переплюнули отдельных мышей из опытной группы.

Таблица 36

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

Из табл. 36 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 она заведомо дала бы другие результаты; в частности, получилось бы другое среднее значение. Так можно ли считать исключенным, что случайная выборка из незащищенной генеральной совокупности могла бы дать $\bar{x} = 5{,}22$? Следовательно, вопрос сводится к тому, не является ли расхождение между средними значениями в опыте и контроле просто расхождением между двумя выборочными средними двух выборок, взятых из одной и той же генеральной совокупности. Это 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).

В § 4 гл. 3 уже говорилось, что величина

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

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 вероят-

ностью $1 - P$ ошибиться) отвергнуть гипотезу о том, что выборка со средним значением $\bar{x}$ и стандартной ошибкой $s_{\bar{x}}$ взята из генеральной совокупности со средним значением $\hat{x}$.

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 среднем значении можно записать в виде

$$\frac{|\bar{x} - \hat{x}|}{s_{\bar{x}}} > t_{\alpha}, \tag{4.6}$$

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 8. При определении pH раствора было получено значение 7,48 + 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 $t_{05}(9) = 2{,}26$; $t_{01}(9) = 3{,}25$, то значимость щелочной реакции сомнительна: t лишь немного превышает t_{05}.

При решении вопроса о равенстве средних значений двух совокупностей задача сводится к определению значимости отношения

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

где $s_{\bar{x}-\bar{y}}$ есть оценка стандартной ошибки

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

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

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

$\sigma^2\{x\} = \sigma^2\{y\}$) выполняется автоматически в том случае, когда нулевая гипотеза гласит, что обе выборки взяты из одной генеральной совокупности. Тогда

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

где σ^2 — общая для обеих совокупностей дисперсия.

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

$$s^2\{x\} = \frac{1}{n_x - 1}\Sigma(x_i - \bar{x})^2 \text{ и } s^2\{y\} = \frac{1}{n_y - 1}\Sigma(y_j - \bar{y})^2, \quad (***)$$

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 $f_x = n_x - 1$ и $f_y = n_y - 1$. Поэтому наилучшей оценкой дисперсии σ^2 будет величина

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

так что

$$s_{\bar{x}-\bar{y}} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2 + \Sigma(y_j - \bar{y})^2}{n_x + n_y - 2}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2 + \Sigma(y_j - \bar{y})^2}{n_x + n_y - 2} \cdot \frac{n_x + n_j}{n_x n_y}}. \quad (4.7)$$

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

$$|t| = \frac{|\bar{x} - \bar{y}|}{s_{\bar{x}-\bar{y}}} > t_\alpha, \quad (4.8)$$

где $s_{\bar{x}-\bar{y}}$ дается формулой (4.7). При отыскании величины t_α в табл. IV Приложений принимается, что число степеней свободы равно

$$f = n_x + n_y - 2, \quad (4.9)$$

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. 36 получаем:

$$s_{\bar{x}-\bar{y}} = \sqrt{\frac{48{,}3+20{,}7}{32+23-2}\cdot\frac{32+23}{32\cdot 23}} = 0{,}312;$$

$$t = \frac{6{,}25-5{,}22}{0{,}312} = \frac{1{,}03}{0{,}312} = 3{,}30;$$

$$f = 32+23-2 = 53 \approx 50; \quad t_{01}(50) = 2{,}68.$$

Так как $t > t_{01}$, то нулевая гипотеза отвергается. Следовательно, расхождение между опытом и контролем можно считать значимым, т. е. препарат определенно обладает защитным действием.

Пример 9. Каждый из двух сортов ячменя высевался на пяти делянках. В табл. 37 приведены урожаи в *кг* (урожай сорта II на одной из делянок был поврежден и поэтому не включен в дальнейшую обработку).

Таблица 37

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 ($\bar{x}_I = 8{,}34$) выше, чем для сорта II ($\bar{x}_{II} = 7{,}90$), то напрашивается вывод, что вообще сорт I более урожайный. Но такой вывод может оказаться ложным: ведь каждое из чисел 8,34 и 7,90 есть не генеральное, а выборочное среднее. Другая случайная выборка, также состоящая из 5 делянок, наверняка дала бы для $\bar{x}_I$ значение, отличное от 8,34; в частности, это новое значение могло бы оказаться меньше, чем 8,34. С другой стороны, и для $\bar{x}_{II}$ выборка дала бы значение, отличное от 7,90, в частности, могло бы получиться число большее, чем 7,90. Поэтому повторение всего опыта с обоими сортами могло бы дать результат $\bar{x}_I < \bar{x}_{II}$. Очевидно, вероятность такого исхода тем меньше, чем больше отношение $\bar{x}_I - \bar{x}_{II}$ к $s_{\bar{x}_I - \bar{x}_{II}}$.

Таким образом, опять приходим к критерию Стьюдента для оценки значимости того, что $\bar{x}_I$ превышает $\bar{x}_{II}$.

Однако данная задача, в биологическом отношении, отличается от задачи в примере 7.

Действительно, там сравнивались две группы мышей, взятых из одной популяции. Нулевая гипотеза состояла в том, что препарат, который давался животным одной группы, не производит никакого действия. Если эта гипотеза верна, то мыши, получавшие

препарат, остаются в той же популяции, что и мыши, препарата не получавшие. 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: $\bar{x}_I \quad \bar{x}_{II}$ и $\sigma_I^2 = \sigma_{II}^2$.

В примере 9 с самого начала ясно, что растения принадлежат к двум различным популяциям: если даже урожайности обоих сортов одинаковы, то эти сорта, вероятней всего, различаются по большинству других признаков, например, высоте соломы, полеганию, устойчивости к болезням, продолжительности вегетации и т. д. 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 $\bar{x}_I \quad \bar{x}_{II}$. Поэтому нулевая гипотеза утверждает здесь только, что $\bar{x}_I = \bar{x}_{II}$.

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 средние значения.

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 применить и в этом случае t-критерий, если при отыскании t_α в таблице критических значений пользоваться измененным числом степеней свободы:

$$f' = (n_x + n_y - 2)\left(\frac{1}{2} + \frac{s^2\{x\}\, s^2\{y\}}{s^4\{x\} + s^4\{y\}}\right). \qquad (4.10)$$

Когда $s^2\{x\} = s^2\{y\}$, то второй множитель в (4.10) равен $^1/_2 + {}^1/_2 = 1$, так что для числа степеней свободы получается обычное значение (4.9). 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

$$s_{\bar{x}-\bar{y}} = \sqrt{\frac{s^2\{x\}}{n_x} + \frac{s^2\{y\}}{n_y}}, \qquad (4.11)$$

где $s^2\{x\}$ и $s^2\{y\}$ находятся по формулам (***). Следовательно, при $\sigma^2\{x\} \neq \sigma^2\{y\}$ условие отказа от нулевой гипотезы гласит:

$$|t'| = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s^2\{x\}}{n_x} + \frac{s^2\{y\}}{n_y}}} > t_\alpha, \qquad (4.12)$$

причем для нахождения t_α в табл. IV Приложений берется число степеней свободы из (4.10).

Применим этот критерий к данным из примера 9. Получаем:

$$s_{\text{I}}^2 = \frac{3{,}09}{4} = 0{,}772; \quad s_{\text{II}}^2 = \frac{4{,}26}{3} = 1{,}42;$$

$$s_{\bar{x}_{\text{I}} - \bar{x}_{\text{II}}} = \sqrt{\frac{0{,}772}{5} + \frac{1{,}42}{4}} = \sqrt{0{,}509} = 0{,}714;$$

$$t' = \frac{8{,}34 - 6{,}90}{0{,}714} = \frac{1{,}44}{0{,}714} = 2{,}02;$$

$$f' = (5 + 4 - 2)\left(\frac{1}{2} + \frac{0{,}772 \cdot 1{,}42}{0{,}772^2 + 1{,}42^2}\right) = 7\,(0{,}50 + 0{,}42) = 6{,}44.$$

Учитывая, что $t_{05}\,(6) = 2{,}45$ и $t_{05}\,(7) = 2{,}37$, можно принять условно $t_{05}\,(6{,}44) = 2{,}41$. Поскольку $t < t_{05}$, нулевая гипотеза не отвергается.

Когда $\frac{\bar{x} - \bar{y}}{s_{\bar{x}-\bar{y}}} < t_\alpha$, то значимость различия 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 $t > t_{01}$ (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 из-за 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 значе-

---

[1] Впрочем, если сравниваются две группы из разных популяций, то нулевая гипотеза $\hat{x} = \hat{y}$ представляет собой лишь идеализацию: в действительности разность $\hat{x} - \hat{y}$ 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.

ние t (зависящее в значительной мере от величины разности $\bar{x} - \bar{y}$) характеризует главным образом данный конкретный опыт, а не объект как таковой.

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

$$s_{\bar{x}-\bar{y}} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n_x(n_x-1)} + \frac{\Sigma(y_j - \bar{y})^2}{n_y(n_y-1)}} \qquad (4.13)$$

— ведь формула (4.7) 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 (4.7) [при неизвестном отношении дисперсий — с $s_{\bar{x}-\bar{y}}$ из (4.13)], а затем, если отличие оказалось значимым, вычисляется доверительный интервал для $\hat{x} - \hat{y}$ с соответствующим $s_{\bar{x}-\bar{y}}$ [из (4.7) или из (4.13)]. 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 (4.10).

Иногда возникает необходимость сравнить не две, а большее число эмпирических совокупностей. Пусть, например, имеются данные нескольких клиник, изучавших распределение длительностей действия анестезирующего препарата. Если бы оказалось, что данные разных клиник могут рассматриваться как выборки из одной генеральной совокупности, то все эти данные можно было бы свести в одну эмпирическую совокупность. Это привело бы к увеличению объема выборки и, тем самым, к сужению доверительных интервалов для параметров распределения.

Данная задача может быть решена при помощи t-критерия путем попарного сравнения всех совокупностей. Однако такое решение требует большого количества вычислений, так как при увеличении числа сравниваемых совокупностей число пар, которые необходимо сравнить, быстро растет (при w совокупностях число пар равно, очевидно, числу сочетаний из w по два, т. е. C_w^2; если $w = 5$, то $C_5^2 = \frac{5 \cdot 4}{1 \cdot 2} = 10$, при $w = 6$ имеем $C_6^2 = \frac{6 \cdot 5}{1 \cdot 2} = 15$ и т. д.).

Поэтому такая задача решается обычно специально разработанным для этой цели методом, который называется дисперсионным анализом. Ввиду важности этого метода для биологических приложений ему посвящена отдельная глава (пятая).

§ 5. Сравнение совокупностей с попарно связанными вариантами

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 т. д., имеющих для развития растений первостепенное значение; поэтому при сравнении двух рядов урожайностей за несколько лет всегда имеется тесная сопряженность между значениями из обоих рядов, относящихся к одинаковым годам. То же имеет место, если сравниваются 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 т. д.

Пример 10. В табл. 38 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 таблицы, 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 различным 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 в соответствии с фор-

мулой (4.6) приходим к критерию:

$$|t_{\overline{\Delta}}| = \frac{|\overline{\Delta}|}{s_{\overline{\Delta}}} > t_\alpha; \quad s_{\overline{\Delta}} = \sqrt{\frac{\Sigma(\Delta_i - \overline{\Delta})^2}{n(n-1)}}. \qquad (4.14)$$

Число степеней свободы равно здесь, конечно, $f = n - 1$.

Таблица 38

Годы	Опыт x_i	Контроль y_i	Разность $\Delta_i = x_i - y_i$	Δ_i^2
1	2	3	4	5
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	25,6	2,1	4,41
Сумма .	188,3	167,3	21,0	70,06
Среднее	26,9	23,9	3,0	

По данным табл. 38 имеем:

$$\overline{\Delta} = \frac{21,0}{7} = 3,0 \; (= 26,9 - 23,9);$$

$$\Sigma(\Delta_i - \overline{\Delta})^2 = \Sigma\Delta_i^2 - \frac{1}{n}(\Sigma\Delta_i)^2 = 70,06 - \frac{21,0^2}{7} = 7,06,$$

так что

$$s_{\overline{x-y}} = s_{\overline{\Delta}} = \sqrt{\frac{7,06}{7 \cdot 6}} = 0,41.$$

Таким образом,

$$t_{\overline{\Delta}} = \frac{3,0}{0,41} = 7,32.$$

В данном случае $f = n - 1 = 7 - 1 = 6$; так как $t_{01}(6) = 3,71$, то различие следует считать значимым.

Для предварительной быстрой оценки значимости различия между $\bar{x}$ и $\bar{y}$ можно использовать то, что входящая в $\sigma_{\overline{\Delta}} = \frac{\sigma\{\Delta\}}{\sqrt{n}}$

величина $s\{\Delta\}$ может быть оценена не только величиной

$$s\{\Delta\} = \sqrt{\frac{\Sigma(\Delta_i - \bar{\Delta})^2}{n-1}},$$

но и через размах варьирования значений Δ_i. Тогда условие значимости различия примет вид

$$\frac{\bar{\Delta}}{R_\Delta} > t_{01}^{(R\Delta)} \tag{4.15}$$

Расчеты дают для $t_{01}^{(R\Delta)}$ и $t_{05}^{(R\Delta)}$ при разных n значения, приведенные в табл. XVI Приложений.

Применим этот метод к данным из табл. 38. Имеем:

$$R_\Delta = 4{,}2 - 1{,}2 = 3{,}0; \quad t^{(R\Delta)} = \frac{\bar{\Delta}}{R_\Delta} = \frac{3{,}0}{3{,}0} = 1{,}0;$$

в табл. XVI Приложений находим для $n = 7$: $t_{05}^{(R\Delta)} = 0{,}426$; $t_{01}^{(R\Delta)} = 0{,}600$. Так как $t^{(R\Delta)} > t_{01}^{(R\Delta)}$, то различие значимо.

Если бы к данным из табл. 38 применялся обычный критерий (т. е. не учитывающий сопряженность пар), то различие не было бы обнаружено: большой разброс урожайности из-за сильной вариабильности 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 (4.14) все равно дает правильное значение t. Но в этом случае применение критерия $t_{\bar{\Delta}}$ 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 критерием $t_{\bar{\Delta}}$ целесообразно пользоваться только тогда, когда сопряженность пар несомненна; в этом случае влияние уменьшения числа степеней свободы перекроется влиянием уменьшения дисперсии разности.

§ 6. Последовательный (секвенциальный) анализ

Как мы уже говорили в § 1 этой главы, гипотеза H_1, конкурирующая с нулевой гипотезой $\hat{\pi} = \pi_0$ (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 мероприятие, направленное к повышению урожайности некоторой культуры. Поскольку проведение этого мероприятия требует известных затрат, оно окажется целесообразным только в том случае, если увеличение урожайности будет не меньше некоторой определенной величины δ. Поэтому если нулевая гипотеза имеет здесь вид $\hat{x} < x_{(0)}$ ($x_{(0)}$ — урожайность в контроле), то альтернативная гипотеза будет $\hat{x} \geqslant x_{(0)} + \delta = x_{(1)}$.

Указанная задача может быть решена обычными методами с применением t-критерия. Однако более удобным оказывается так называемый «метод *последовательного* (секвенциального) *анализа*» (А. Вальд).

При обычном анализе математическая обработка результатов производится после завершения серии наблюдений, объем которой (т. е. число наблюдений) был намечен 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 анализе.

Обозначим через $P_{0,n}$ плотность вероятности (в дискретном случае — вероятность) получить значения x_1, x_2, , x_n при условии, что выборка относится к генеральной совокупности со средним значением $\hat{x} \leqslant x_{(0)}$, и через $P_{1,n}$ — плотность вероятности получить эти значения при условии, что выборка относится к генеральной совокупности со средним значением $\hat{x} \geqslant x_{(1)}$.

Когда $P_{1,n} \leqslant \alpha$, гипотеза $\hat{x} \geqslant x_{(1)}$ может быть отвергнута; при этом риск ошибиться (если в действительности эта гипотеза верна) не превышает α. Но отвергнуть гипотезу $\hat{x} \geqslant x_{(1)}$ еще не значит, что надо обязательно принять гипотезу $\hat{x} \leqslant x_{(0)}$ можно ограничиться утверждением, что $\hat{x} < x_{(1)}$. Если же мы сделаем более определенное утверждение $\hat{x} \leqslant x_{(0)}$, 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 $\hat{x} \leqslant x_{(0)}$ и принять гипотезу $\hat{x} \geqslant x_{(1)}$:

$$\frac{P_{0,n}}{P_{1,n}} \leqslant \frac{\alpha}{1-\beta};$$

это неравенство можно заменить равносильным неравенством

$$\frac{P_{1,n}}{P_{0,n}} \geqslant \frac{1-\beta}{\alpha} \qquad (**)$$

(например, если $2 < 3$, то $1/2 > 1/3$).

Таким образом, если выполняется неравенство (*), то принимается гипотеза $\hat{x} \leqslant x_{(0)}$, а если выполняется неравенство (**), то принимается гипотеза $\hat{x} \geqslant x_{(1)}$; если же окажется, что

$$\frac{\alpha}{1-\beta} < \frac{P_{1,n}}{P_{0,n}} < \frac{1-\beta}{\alpha}, \qquad (***)$$

то это будет означать, что испытания надо продолжать.

Если распределение вариант в генеральной совокупности нормально, то в соответствии с (2.4)

$$P_{0,n} = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - x_{(0)})^2}$$

$$P_{1,n} = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - x_{(1)})^2}$$

здесь также использовано известное из § 1 гл. 2 определение независимости: если события $A, B, C, \ldots, Z$ имеют вероятности соответственно $P_A, P_B, P_C, \ldots, P_Z$ 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 неравенства (***), то можно получить равносильные неравенства

$$\frac{x_{(1)} + x_{(0)}}{2} n + \frac{s^2}{x_{(1)} - x_{(0)}} \ln \frac{\alpha}{1-\beta} < \sum_{i=1}^{n} x_i <$$

$$< \frac{x_{(1)} + x_{(0)}}{2} n + \frac{s^2}{x_{(1)} - x_{(0)}} \ln \frac{1-\beta}{\alpha}. \qquad (4.16)$$

Это значит, что испытание надо продолжать, пока сумма вариант Σx_i лежит в пределах от

$$L_0(n) = \frac{x_{(1)} + x_{(0)}}{2} n - \frac{s^2}{x_{(1)} - x_{(0)}} \ln \frac{1-\beta}{\alpha} = an - b \qquad (4.17)$$

до

$$L_1(n) = \frac{x_{(1)} + x_{(0)}}{2} n + \frac{s^2}{x_{(1)} - x_{(0)}} \ln \frac{1-\beta}{\alpha} = an + b; \qquad (4.18)$$

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 (4.17) и (4.18); очевидно, эти прямые имеют угловой коэффициент

$$a = \frac{x_{(1)} + x_{(0)}}{2} \qquad (4.19)$$

и отсекают на оси ординат соответственно

$$\pm b = \pm \frac{s^2}{x_{(1)} - x_{(0)}} \ln \frac{1-\beta}{\alpha} \qquad (4.20)$$

(рис. 40); выбирая $\alpha = 0{,}01$ и $\beta = 0{,}01$, имеем

$$\ln \frac{1-\beta}{\alpha} = \ln \frac{0{,}99}{0{,}01} = 4{,}59. \qquad (4.21)$$

Значения $\sum_{i=1}^{n} x_i$ наносятся на этот график в виде точек, абсциссами которых служат соответствующие значения n; точки соединяются отрезками прямой, которые образуют некоторую ломаную. Испытания продолжаются до тех пор, пока эта ломаная не выйдет из центральной полосы.

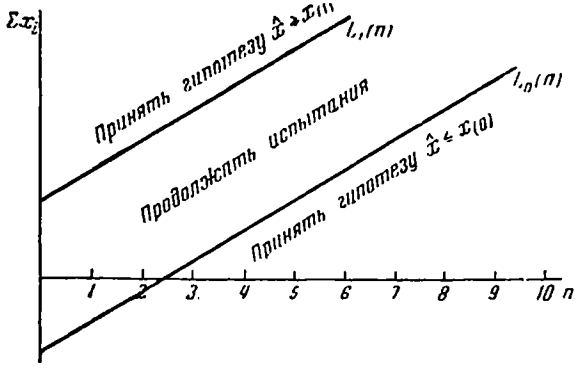

Рис. 40

Из (4.20) 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, чем тоньше различие, которое мы хотим обнаружить, и чем больше рассеяние вариант.

Пример 11. Было установлено, что затраты на предпосевную обработку семян пшеницы окупятся, если она даст средний прирост урожайности не менее $\delta = 3$ *ц с га*. Средняя урожайность необработанных семян — 20,8 *ц с 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 $x = x_{(0)} + \delta = 20{,}8 + 3{,}0 = 23{,}8$ можно считать разумной оценкой σ (во всяком случае не заниженной) величину

$$s = x \cdot v = 23{,}8 \cdot 0{,}087 = 2{,}07, \quad s^2 = 4{,}3.$$

При этих данных имеем согласно (4.20) — (4.21):

$$a = \frac{20{,}8 + 23{,}8}{2} = 22{,}3;$$

$$b = \frac{4{,}3}{3{,}0} 4{,}59 \approx 6{,}6.$$

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

$$a' = a - 20 = 2{,}3.$$

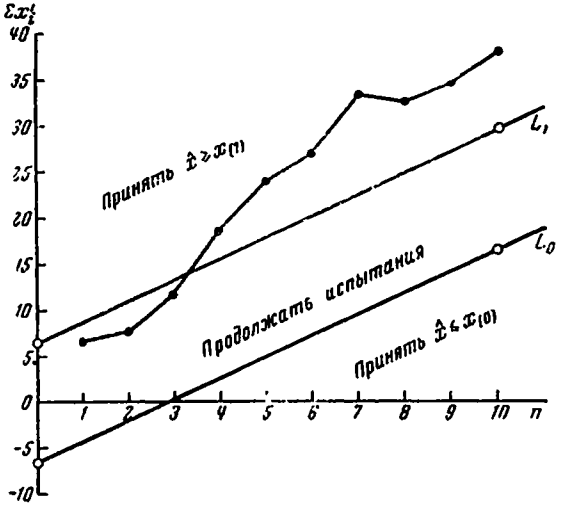

Рис. 41

Линии $L_0(n)$ и $L_1(n)$ проведем через точки:

$$L_0(0) = -6{,}6; \quad L_0(10) = -6{,}6 + 2{,}3 \cdot 10 = 16{,}4;$$
$$L_1(0) = +6{,}6; \quad L_1(10) = +6{,}6 + 2{,}3 \cdot 10 = 29{,}6$$

(рис. 41).

В нашем распоряжении имеются данные за десять лет, приведенные в табл. 39.

Таблица 39

Годы	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

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. 41). Мы видим, что уже четвертая точка $\left(n = 4, \sum_{i=1}^{4} x_i' = 18{,}9\right)$ лежит вне центральной полосы; это значит, что гипотезу $\bar{x} \geqslant x_{(1)}$ можно было бы принять на основании только четырехлетних наблюдений.

§ 7. Сравнение дисперсий (*F*-критерий)

Две выборочные совокупности, не различаясь значимо по своим средним значениям, могут различаться по стандартным отклонениям (или дисперсиям).

Пример 12. Два сорта пшеницы (табл. 40) имеют почти одинаковую среднюю урожайность ($\bar{x}_1 = 20{,}4$ *ц/га*; $\bar{x}_2 = 20{,}3$ *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 40

Годы	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

$$u_{s_1-s_2} = \frac{s_1 - s_2}{\sigma_{s_1-s_2}}$$

распределена нормально; при малых же выборках нормальное распределение должно быть заменено другим, более сложным распределением.

Вместо того чтобы искать это распределение, заметим, что $u_{s_1 s_2}$ зависит только от отношения s_1/s_2. В самом деле,

$$\sigma_{s_1-s_2} = \sqrt{\sigma^2_{s_1} + \sigma^2_{s_2}},$$

где

$$\sigma_{s_1} = \frac{s_1}{\sqrt{2(n_1-1)}}; \quad \sigma_{s_2} = \frac{s_2}{\sqrt{2(n_2-1)}}$$

Если сначала принять для простоты, что $n_1 = n_2$, то

$$u_{s_1-s_2} = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2(n-1)} + \frac{s_2^2}{2(n-1)}}} = \sqrt{2(n-1)}\,\frac{s_1 - s_2}{\sqrt{s_1^2 + s_2^2}};$$

разделив числитель и знаменатель на s_2, получим

$$u_{s_1-s_2} = \sqrt{2(n-1)}\,\frac{\frac{s_1}{s_2} - 1}{\sqrt{\left(\frac{s_1}{s_2}\right)^2 + 1}} = \sqrt{2(n-1)}\,\frac{\sqrt{\frac{s_1^2}{s_2^2}} - 1}{\sqrt{\frac{s_1^2}{s_2^2}} + 1}. \quad (*)$$

Следовательно, $u_{s_1-s_2}$ действительно определяется отношением стандартных отклонений s_1/s_2 или отношением дисперсий s_1^2/s_2^2 (и, конечно, объемом выборок n). Поэтому можно принять в качестве критерия значимости различия дисперсий отношение их оценок:

$$F = \frac{s_1^2}{s_2^2} \quad (4.22)$$

(*F-критерий Фишера*), которое нужно в каждом случае сравнивать с критическим значением $F_\alpha(n)$. В качестве критерия выбирается отношение оценок дисперсий, а не отношение оценок стандартных отклонений, так как это избавляет от извлечения квадратных корней.

Если объемы выборок различны ($n_1 \neq n_2$), то таблица для F-критерия будет иметь два входа: n_1 и n_2. Такая таблица приводится в Приложениях (табл. XIV); как и в таблице t-критерия, на входах стоят не объемы выборок n_1 и n_2, а числа степеней свободы f_1 и f_2. В таблице даны критические значения $F_\alpha(f_1, f_2)$ для двух уровней значимости: 5 и 1%. При пользовании табл. XIV Приложений надо иметь в виду, что

$$F_\alpha(f_1, f_2) \neq F_\alpha(f_2, f_1). \quad (4.23)$$

Понятно, табличные значения F_α вычислены не по формуле (*), а исходя из совсем других математических соображений, на которых мы не будем здесь останавливаться; формула (*) была приведена только для того, чтобы показать наглядно, что значимость различия между дисперсиями полностью определяется отношением оценок этих дисперсий.

В соответствии с обычными условиями применения F-критерия при составлении таблицы критических значений F_α использовался односторонний критерий (см. § 1 этой главы). Так как эта таблица содержит только значения F_α, бóльшие единицы, то при вычислении величины F надо всегда делить бóльшую дисперсию 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. XIV Приложений находим, что при числе степеней свободы числителя $f_1 = 8 - 1 = 7$ и числе степеней свободы знаменателя $f_2 = 8 - 1 = 7$

$$F_{05}\,(7;\,7) = 3{,}79.$$

Поскольку фактическое значение $F = 3{,}44$ меньше 5%-ного критического значения, то нулевая гипотеза не отвергается.

Пример 13. В примере 9 из § 4 были получены оценки дисперсий $s_{\text{I}}^2 = 0{,}772$ и $s_{\text{II}}^2 = 1{,}42$ при $f_{\text{I}} = 4$ и $f_{\text{II}} = 3$. Отношение этих оценок

$$F = \frac{s_{\text{II}}^2}{s_{\text{I}}^2} = \frac{1{,}42}{0{,}772} = 1{,}84$$

меньше критического значения $F_{05}(3;\,4) = 6{,}59$.

Если бы в формулировку нулевой гипотезы входило условие $\sigma_{\text{I}}^2 = \sigma_{\text{II}}^2$, то результат $F < F_{05}$ означал бы, что опыт не опровергает справедливость этого равенства; тогда при пользовании t-критерием нужно было бы вычислять $s_{\bar{x}-\bar{y}}$ и f по формулам (4.7) и (4.9). Но в примере 9 не было оснований выдвигать гипотезу $\sigma_{\text{I}}^2 = \sigma_{\text{II}}^2$, 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. По этой причине использование формул (4.7′) и (4.10) для $s_{\bar{x}-\bar{y}}$ и f в примере 9 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.

В некоторых случаях бывает необходимо сравнить сразу несколько дисперсий. Если все выборки имеют одинаковый объем, то эта задача решается при помощи *критерия Кохрена*, основанного на вычислении величины

$$G = \frac{s^2_{\max}}{s_1^2 + s_2^2 + \ldots + s_w^2}; \qquad (4.24)$$

полученное значение G надо сравнить с приведенными в табл. XV Приложений критическими значениями $G_\alpha(w; f)$, где α — уровень значимости; w — число сравниваемых дисперсий и f — число степеней свободы выборок (одинаковое для всех выборок).

Пример 14. Были произведены определения pH одних и тех же образцов (при повторности $n = 17$) с применением пяти разных электродов, причем дисперсии воспроизводимости для этих электродов оказались равными 0,015; 0,039; 0,027; 0,011; 0,034. Можно ли считать, что все электроды обеспечивают одинаковую воспроизводимость результатов (т. е. что различие между дисперсиями незначимо)?

Применение G-критерия дает

$$G = \frac{0{,}039}{0{,}015 + 0{,}039 + 0{,}027 + 0{,}011 + 0{,}034} = \frac{0{,}039}{0{,}126} = 0{,}31,$$

в то время как $G_{05}(5; 16) = 0{,}409$. Так как $G < G_{05}$, то различие незначимо.

Если объемы выборок различны, то приходится пользоваться более сложным критерием Бартлета [1].

§ 8. Сравнение двух выборочных долей вариант

Сравнение выборочных долей вариант лучше всего производить при помощи методов, изложенных в § 4 гл. 6. Однако, если нас интересует не только установление значимости различия двух долей, но и построение доверительного интервала для их разности, то приходится прибегать к параметрическим методам. В соответствии со сказанным в § 9 гл. 3 ясно, что в этом случае целесообразно применять φ-преобразование долей.

При сравнении двух выборок будем иметь

$$\sigma_{\varphi_1 - \varphi_2} = \sqrt{\sigma^2_{\varphi_1} + \sigma^2_{\varphi_2}} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \qquad (4.25)$$

[1] См. книги В. В. Налимова (1960) и Дж. У. Снедекора (1961).

Поскольку величины φ распределены приближенно нормально, то мы пользуемся u-критерием, т. е. вычисляем величину

$$u = \frac{|\varphi_1 - \varphi_2|}{\sigma_{\varphi_1 - \varphi_2}} = |\varphi_1 - \varphi_2| \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \qquad (4.26)$$

и сравниваем ее с критическими значениями $u_{05} = 1{,}96$ и $u_{01} = 2{,}58$ (понятно, что при вычислении u всегда из большей величины φ вычитается меньшая).

Более правильные результаты получаются при введении так называемой поправки на непрерывность. Если $p_1 > p_2$, то производится замена:

$$p_1 \to p_1' = p_1 - \frac{1}{2n_1} \text{ или } \frac{x_1}{n_1} \to \frac{x_1 - 0{,}5}{n_1};$$

$$p_2 \to p_2' = p_2 + \frac{1}{2n_2} \text{ или } \frac{x_2}{n_2} \to \frac{x_2 + 0{,}5}{n_2},$$

после чего по табл. VII Приложений находят значения φ_1 и φ_2, соответствующие p_1 и p_2.

Пример 15. Были проведены опыты иммунизации телят от туберкулеза: телятам сначала делали либо предохранительную прививку, либо прививку контрольных средств, а затем их заражали туберкулезными бациллами. Результаты получились следующие: с прививкой заболели 6 из 20, без прививки заболели 16 из 19.

Оценим значимость влияния прививки. Так как здесь проверяется гипотеза $p_2 > p_1$, 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

$$u = (2{,}255 - 1{,}213) \sqrt{\frac{20 \cdot 19}{20 + 19}} = 1{,}042 \sqrt{\frac{380}{39}} = 3{,}25.$$

Так как $u > u_{01} = 2{,}58$, то разность $p_2 - p_1$ значима.

Если требуется сравнить p с некоторым теоретическим значением p_0 (которому соответствует значение φ_0, определяемое из табл. VII Приложений), то

$$\sigma_{\varphi-\varphi_0} = \sigma_\varphi = \frac{1}{\sqrt{n}} \tag{4.27}$$

ибо $\sigma_{\varphi_0} = 0$), так что

$$u = |\varphi - \varphi_0| \sqrt{n}. \tag{4.28}$$

Пример 16. При расщеплении гибридов Primula получено 468 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:

$$p = 102 : (468 + 102) = 0{,}179 = 17{,}9\%;$$
$$p_0 = 1 : (3 + 1) = 0{,}250 = 25{,}0\%.$$

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 $u_{01} = 2{,}58$ (и даже $u_{001} = 3{,}29$), 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).

Пример 17. В 10 выводках получились 51 курочка и 37 петушков. Противоречит ли это гипотезе $\hat{p} = 0{,}5$?

Так как

$$p = 51 : (51 + 37) = 51 : 88 = 0{,}580 = 58{,}0\%,$$

то $\varphi = 1{,}731$; далее, значению $p_0 = 50\%$ соответствует $\varphi_0 = 1{,}571$. Значит,

$$u = (1{,}731 - 1{,}571)\sqrt{88} = 0{,}160 \cdot 9{,}37 = 1{,}50.$$

Это меньше, чем $u_{05} = 1{,}96$, так что отличие p от 0,5 незначимо.

Для любого значения p_0 можно составить таблицу, в которой будет указано, какова должна быть максимальная численность меньшей группы, чтобы при имеющемся объеме выборки n откло-

нение от заданного p_0 могло считаться значимым. Например, отклонение от $p_0 = 0{,}25$ будет значимым только тогда, когда при $n = 470$ меньшая группа содержит менее 117 вариант (в примере 16 она содержала 102 варианты), при $n = 32$ она содержит менее трех вариант и т. д. Табл. XXIII Приложений представляет пример такой таблицы — для $p_0 = 0{,}5$. Выбор именно этого значения p_0 определялся главным образом особой важностью этого случая для биологической практики; кроме того, эта таблица нужна также для непараметрического критерия знаков (см. § 6 гл. 7).

Применим табл. XXIII к примеру 17. При $n = 88$ имеем критические значения $Z_{05} = 35$, $Z_{01} = 32$. Так как фактическое значение $Z = 37$ больше, чем Z_{05}, то нулевая гипотеза не отвергается. Здесь этот результат получен без всяких вычислений.

Табл. XXIII содержит значения n до 100. При бо́льших объемах выборки приходится пользоваться расчетным методом (с применением табл. VII). 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:

$$Z_\alpha = \frac{1}{2}(n - u_\alpha\sqrt{n}), \tag{4.29}$$

причем нулевая гипотеза принимается при $Z \geqslant Z_{05}$ и отвергается при $Z < Z_{01}$. Как и ранее, Z есть численность меньшей группы.

Пример 18. 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 $\hat{p} = 0{,}5$ о вероятности рождения девочки?

Имеем:

$$p = 1\,285\,047 : (1\,359\,814 + 1\,285\,047) =$$

$$= 1\,285\,047 : 2\,644\,861 \approx 0{,}486, \text{ или } 48{,}6\%,$$

что дает $\varphi = 1{,}543$. Тогда

$$u = (1{,}571 - 1{,}543)\sqrt{2\,644\,861} \approx 0{,}028 \cdot 1{,}624 \cdot 10^3 \approx 45{,}5.$$

Это во много раз превышает даже $u_{001} = 3{,}29$, поэтому гипотеза $\hat{p} = 0{,}5$ определенно отвергается.

Применение u-критерия к этим же данным приводит к тому же результату:

$$p = \frac{1\,285\,047}{2\,644\,861} = 0{,}486;$$

$$\sigma_p = \sqrt{\frac{0{,}486 \cdot 0{,}514}{2\,644\,860}} \approx 0{,}000307; \quad u = \frac{0{,}500 - 0{,}486}{0{,}000307} \approx 45{,}5.$$

Наконец, по формуле (4.29) получаем

$$Z_{01} = \frac{1}{2}\left(2\,644\,861 - 2{,}58\sqrt{2\,644\,861}\right) = 1\,320\,335,$$

в то время как фактическое значение $Z = 1\,285\,047$ меньше.

Если разность долей вариант оказалась значимой, то вычисляется доверительный интервал для этой разности. Из (4.25) очевидно, что ширина доверительного интервала определяется величиной

$$u_P \sigma_{\varphi_1 - \varphi_2} = u_P \sqrt{\frac{n_1 + n_2}{n_1 n_2}}, \tag{4.30}$$

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 15 имеем при $P = 99\%$:

$$u_P \sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 2{,}58 \sqrt{\frac{39}{380}} = 0{,}826;$$

поэтому

$$\Delta\varphi_н = 1{,}165 - 0{,}826 = 0{,}339; \; \Delta\varphi_в = 1{,}165 + 0{,}826 = 1{,}991.$$

Этому соответствуют значения (см. табл. VII Приложений):

$$\Delta p_н = 2{,}8\%; \quad \Delta p_в = 70{,}4\%.$$

Если одна из долей задана, то в соответствии с формулой (4.28) доверительные границы определяются выражением

$$\varphi = \varphi_0 \pm \frac{u_P}{\sqrt{n}}. \tag{4.31}$$

Так, для примера 18 получаем при $P = 99\%$:

$$\frac{u_{99}}{\sqrt{n}} = \frac{2{,}58}{\sqrt{2\,644\,860}} = \frac{2{,}58}{1\,626} \approx 0{,}002,$$

так что

$$\varphi_н = 1{,}599 - 0{,}002 = 1{,}597; \quad p_н = 51{,}3;$$
$$\varphi_в = 1{,}599 + 0{,}002 = 1{,}601; \quad p_в = 51{,}5.$$

Если желательно получить более точные значения доверительных границ, то нужно пользоваться формулой (3.35); в данном случае это возможно, поскольку n велико. Расчет дает:

$$p = \frac{1\,359\,814}{2\,644\,861} \approx 51{,}413\,\%; \quad \sigma_p = \sqrt{\frac{51{,}413 \cdot 48{,}587}{2\,644\,860}} \approx 0{,}0307\,\%;$$
$$u_{99}\sigma_p = 2{,}58 \cdot 0{,}0307 \approx 0{,}079,$$

так что

$$p_н = 51{,}413 - 0{,}079 = 51{,}334; \qquad p_в = 51{,}413 + 0{,}079 = 51{,}492.$$

ДИСПЕРСИОННЫЙ АНАЛИЗ

§ 1. Задачи дисперсионного анализа

Во многих случаях биологический эксперимент может быть повторен на одной особи многократно, например при изучении физиологических функций; определенная повторность может быть получена при изучении одной особи и в тех случаях, когда непосредственным объектом исследования являются какие-либо повторяющиеся элементы организма — листья, семена, клетки какой-то ткани и т. д.

Подвергая имеющийся материал статистической обработке, можно получить те или иные обобщенные характеристики (например, средние значения). Однако, ввиду вариабильности организмов в популяции среднее значение по особи еще не есть среднее значение по виду. Поэтому приходится, помимо многократного повторения опыта на одной особи или рассмотрения многих элементов одного организма, ставить повторные опыты с другими особями данного вида.

При этом всегда возникает вопрос, в каком отношении должны находиться повторность «внутри» особи и повторность с разными особями. Качественно можно сказать, что если вариабильность между особями велика по сравнению с вариабильностью внутри особи, то повторять опыт несколько раз на каждой особи нет смысла; наоборот, если очень велика вариабильность внутри особей, то повторение опытов со многими особями нецелесообразно.

Так как вариабильность мы характеризуем количественно дисперсией, то решение поставленного выше вопроса сводится к сравнению дисперсий соответствующих распределений. В этом и состоит задача *дисперсионного анализа*, разработанного Р. Фишером.

Пример 1. Изучалось образование условного рефлекса у собаки под действием некоторого ранее индифферентного раздражителя. Количественным признаком служило время (в *сек*) между моментом включения условного раздражителя и моментом начала слюноотделения. Исследование было поставлено на пяти собаках, причем на каждой проделано шесть опытов. Результаты представлены в табл. 41.

Таблица 41

Номера животных (a)	Номера опытов (i)							
	1	2	3	4		6		$s^2_{\bar{x}_a}$
1	6	4	9	8	15	12	9,0	2,67
2	9	7	3	4	11	14	8,0	2,93
3	6	8	10	14	13	15	11,0	2,13
4	2	3	7	4	9	11	6,0	2,13
5	6	5	4	10	14	9	8,0	2,33
					Сумма .		42,0	12,9
					Среднее .		8,4	2,44

Мы видим, что средние значения $\bar{x}_a$ у разных собак неодинаковы. Однако было бы неверным относить расхождения только за счет различия между особями и делать на основании этого вывод, что для повышения точности результата нужно увеличить число подопытных животных. Дело в том, что разброс значений $\bar{x}_a$ имел бы место и при условии, что все пять собак совершенно одинаковы. Действительно, в этом случае можно было бы считать, что было поставлено пять серий опытов на одной собаке, так что строки были бы просто отдельными выборками из одной генеральной совокупности. Но выборочные средние, как мы знаем, обычно не совпадают между собой.

Таким образом, дисперсия средних значений $s^2\{\bar{x}_a\} = 3{,}30$. вычисленная в табл. 42, может оказаться в данном случае просто

Таблица 42

$\bar{x}_a$	ξ_a	ξ_a^2	
9,0	0,6	0,36	
8,0	−0,4	0,16	
11,0	2,6	6,76	$\xi_a = \bar{x}_a - \bar{x}, \quad \bar{x} = \frac{42{,}0}{5} = 8{,}4$
6,0	−2,4	5,76	
8,0	−0,4	0,16	$s^2\{\bar{x}_a\} = \frac{\Sigma\xi_a^2}{n_A - 1} = \frac{13{,}20}{4} = 3{,}30$
42,0		13,20	

оценкой дисперсии выборочного среднего $\sigma^2_{\bar{x}_a}$ для выборки из 6 вариант. Если это действительно так, то значения $s^2\{\bar{x}_a\}$ и $s^2_{\bar{x}_a}$ не должны значимо различаться; но если животные и в самом деле различны (в отношении исследуемого признака), т. е. если соответствующие им генеральные средние значения $\hat{x}_a$ не совпадают, то значение величины $s^2\{\bar{x}_a\}$ должно быть значимо больше, чем $s^2_{\bar{x}_a}$. Значения $s^2_{\bar{x}_a}$ вычисляются обычным образом — по формуле

$$s^2_{\bar{x}_a} = \frac{\sum_{i=1}^{n_a} (x_{ai} - \bar{x}_a)^2}{n_a (n_a - 1)}, \qquad (*)$$

где n_a — число вариант в строке; x_{ai} — значения вариант, входящих в строку с номером a. Они приведены в последнем столбце табл. 41; в табл. 43 дан в качестве примера расчет для одной из строк (первой).

Таблица 43

Номера опытов	x_{1i}	ξ_{1i}	ξ^2_{1i}	
1	6	−3	9	
2	4	−5	25	$\bar{x}_1 = \frac{\Sigma x_{1i}}{n_1} = \frac{54}{6} = 9{,}0$
3	9	0	0	
4	8	−1	1	
5	15	6	36	
6	12	3	9	$s^2_{\bar{x}_1} = \frac{\Sigma \xi^2_{1i}}{n_1 (n_1 - 1)} = \frac{80}{6 \cdot 5} = 2{,}67$
Сумма	54		80	

Поскольку значения $s^2_{\bar{x}_a}$ у разных собак не совсем одинаковы, то вычислим усредненное значение

$$\langle s^2_{\bar{x}} \rangle_{\text{ср}} = \sum_{a=1}^{n_A} (n_a - 1)\, s^2_{\bar{x}_a} \Big/ \sum_{a=1}^{n_A} (n_a - 1), \qquad (**)$$

где n_A — число строк (т. е. в данном случае число подопытных животных); в нашем примере все n_a одинаковы, так что $\Sigma (n_a - 1) = n_A(n_a - 1)$ и

$$\langle s^2_{\bar{x}} \rangle_{\text{ср}} = \frac{2{,}67 + 2{,}93 + 2{,}13 + 2{,}13 + 2{,}33}{5} = 2{,}44.$$

Теперь можно оценить дисперсию, связанную с индивидуальными различиями подопытных животных; она оценивается разностью

$$s^2\{\bar{x}_a\} - \langle s^2_{\bar{x}}\rangle_{\text{ср}}.$$

Однако вычисление такой разности будет иметь смысл лишь в том случае, если различие между $s^2\{\bar{x}_a\}$ и $\langle s^2_{\bar{x}}\rangle_{\text{ср}}$ является значимым, а не случайным. Последнее можно оценить, как мы знаем, с помощью F-критерия. Поскольку разность $\sigma^2\{\bar{x}_a\} - \sigma^2_{\bar{x}}$ может быть только положительной, то альтернативой к нулевой гипотезе будет $\sigma^2\{\bar{x}_a\} > \sigma^2_{\bar{x}}$; поэтому здесь должен применяться о д н о с т о р о н н и й критерий. Как было сказано в § 7 гл. 4, табл. XIV Приложений составлена как раз для такого критерия.

В нашем примере $s^2\{\bar{x}_a\} = 3{,}30$, так что $F = 3{,}30 \quad 2{,}44 = 1{,}35$. Это значение слишком мало, чтобы различие можно было считать значимым (подробней см. ниже). Отсюда следует, что нулевая гипотеза о равенстве генеральных средних $\hat{x}_a$ для разных собак не отвергается. Это значит, что нет оснований считать, что для повышения точности общего результата нужно увеличить число подопытных животных, а не число опытов на каждом из них. Такой вывод может иметь существенное значение, если учесть, что выполнение добавочных опытов на имеющихся собаках требует меньших затрат, чем приобретение и содержание дополнительного числа животных [1].

§ 2. Схема однофакторного дисперсионного анализа

Проведенный выше дисперсионный анализ мы назовем *однофакторным*, поскольку рассматривалось влияние на дисперсию одного лишь известного фактора (помимо случайных причин) — индивидуальных свойств животных.

Изложенная методика анализа очень проста в принципе, но она требует большого количества вычислений. Однако эти вычисления могут быть значительно сокращены при помощи ряда чисто технических приемов.

Выпишем в развернутом виде величины $s^2\{\bar{x}_a\}$ и $\langle s^2_{\bar{x}}\rangle_{\text{ср}}$, которые нам нужно сравнить между собой. Используя формулу (1.17),

[1] Рассматриваемый пример условный: численные значения выбраны так, чтобы лучше иллюстрировать излагаемый метод; в действительности различие между животными по показателям условных рефлексов чаще всего бывает значимым.

имеем

$$s^2\{\bar{x}_a\} = \frac{1}{n_A - 1}\sum_{a=1}^{n_A}(\bar{x}_a - \bar{x})^2 = \frac{1}{n_A - 1}\left(\sum_{a=1}^{a_A}\bar{x}_a^2 - n_A\bar{x}^2\right). \qquad (*)$$

Введем обозначения: X_a — сумма значений вариант строки a; X — сумма всех значений вариант (очевидно, $X = \sum_{a=1}^{n_A} X_a$). Тогда

$$\bar{x}_a = \frac{1}{n_a}\sum_{a=1}^{n_a} x_{ai} = \frac{1}{n_a}X_a; \quad \bar{x} = \frac{1}{n_A n_a}\sum_{a=1}^{n_A}\sum_{i=1}^{n_a} x_{ai} = \frac{1}{n_A n_a}X, \quad (5.1)$$

так что

$$s^2\{\bar{x}_a\} = \frac{1}{n_a(n_A - 1)}\left(\frac{1}{n_a}\sum_{a=1}^{n_A} X_a^2 - \frac{1}{n_A n_a}X^2\right).$$

Далее, после подстановки в равенство (**) предыдущего параграфа значений $s^2_{\bar{x}_a}$ из равенства (*) того же параграфа, величина $\langle s^2_{\bar{x}}\rangle_{\text{ср}}$ примет вид

$$\langle s^2_{\bar{x}}\rangle_{\text{ср}} = \frac{1}{n_A n_a(n_a - 1)}\sum_{a=1}^{n_A}\sum_{i=1}^{n_a}(x_{ai} - \bar{x}_a)^2$$

(с учетом того, что все n_a здесь одинаковы). Это выражение при помощи формулы (1.17) приводится к виду

$$\langle s^2_{\bar{x}}\rangle_{\text{ср}} = \frac{1}{n_A n_a(n_a - 1)}\sum_{a=1}^{n_A}\left(\sum_{i=1}^{n_a} x_{ai}^2 - n_a\bar{x}_a^2\right);$$

с учетом только что введенных обозначений это дает

$$\langle s^2_{\bar{x}}\rangle_{\text{ср}} = \frac{1}{n_A n_a(n_a - 1)}\left(\sum_{a=1}^{n_A}\sum_{i=1}^{n_a} x_{ai}^2 - \frac{1}{n_a}\sum_{a=1}^{n_A} X_a^2\right).$$

Величины $s^2\{\bar{x}_a\}$ и $\langle s^2_{\bar{x}}\rangle_{\text{ср}}$ содержат общий множитель $1/n_a$. Так как нас интересует только их отношение F, то этот множитель можно опустить. Если ввести обозначения

$$S_{ai} = \sum_{a=1}^{n_A}\sum_{i=1}^{n_a} x_{ai}^2, \quad S_a = \frac{1}{n_a}\sum_{a=1}^{n_A} X_a^2, \quad S_x = \frac{1}{n_A n_a}X^2, \qquad (5.2)$$

то задача сведется к вычислению величин

$$n_a s^2\{\bar{x}_a\} = s_A^2 = \frac{S_a - S_x}{n_A - 1}; \tag{5.3}$$

$$n_a \langle s_{\bar{x}}^2 \rangle_{\text{ср}} = s_Z^2 = \frac{S_{ai} - S_a}{n_A (n_a - 1)}, \tag{5.4}$$

которые и надо сравнивать между собой по F-критерию; расчет показывает, что величины (5.3) и (5.4) статистически независимы.

Поскольку величина s_A^2 отражает вариации от строки к строке, то ее обычно называют дисперсией по фактору A, т. е. по тому фактору, влияние которого на изучаемый признак подозревается (если таких факторов несколько, то их обозначают A, B, C и т. д.); в нашем примере фактором A являются индивидуальные свойства животных. Однако такое название величины s_A^2 не совсем точно: ведь влияние фактора A приводит лишь к превышению $s^2\{\bar{x}_a\}$ над $\langle s_{\bar{x}}^2 \rangle_{\text{ср}}$, так что это влияние, в чистом виде, характеризуется разностью

$$s^2\{A\} = s^2\{\bar{x}_a\} - \langle s_{\bar{x}}^2 \rangle_{\text{ср}} = \frac{1}{n_a}(s_A^2 - s_Z^2);$$

именно эту величину, строго говоря, и следовало бы называть оценкой дисперсии по фактору A.

Величины $S_a - S_x = S_A$ и $S_{ai} - S_a = S_Z$, пропорциональные тем суммам квадратов отклонений, которые входят в (*) и (**), мы будем называть *вариациями*.

Итак, схема однофакторного дисперсионного анализа выглядит теперь следующим образом. Прежде всего составляем таблицу числовых данных, придав ей вид табл. 44. Значения X_a суть суммы

Таблица 44

Номера собак	Номера опытов						X_a	X_a^2
	1	2	3	4	5	6		
1	6_{36}	4_{16}	9_{81}	8_{64}	15_{225}	12_{144}	54	2916
2	9_{81}	7_{49}	3_{9}	4_{16}	11_{121}	14_{196}	48	2304
3	6_{36}	8_{64}	10_{100}	14_{196}	13_{169}	15_{225}	66	4356
4	2_{4}	3_{9}	7_{49}	4_{16}	9_{81}	11_{121}	36	1296
5	6_{36}	5_{25}	4_{16}	10_{100}	14_{196}	9_{81}	48	2304
					Сумма .	.	252	13176

вариант по строкам (например, $54 = 6 + 4 + 9 + 8 + 15 + 12$), а X_a^2 — квадраты этих сумм ($54^2 = 2916$, $48^2 = 2304$ и т. д.). Внизу записаны суммы $\Sigma X_a = X$ и ΣX_a^2. В углу каждой клетки вписываются квадраты соответствующих вариант[1]; сумма этих квадратов в нашем примере равна $\sum\limits_{a,\,i} x_{ai}^2 = 2562$.

Используя данные из этой таблицы, находим:

$$S_{ai} = 2562; \qquad S_a = \frac{13176}{6} = 2196; \qquad S_x = \frac{252^2}{5 \cdot 6} = 2116{,}8,$$

так что

$$s_A^2 = \frac{2196 - 2116{,}8}{4} = \frac{79{,}2}{4} = 19{,}8; \quad s_Z^2 = \frac{2562 - 2196}{5 \cdot 5} = \frac{366}{25} = 14{,}6$$

и

$$F = \frac{s_A^2}{s_Z^2} = \frac{19{,}8}{14{,}6} = 1{,}35,$$

как и ранее.

Чтобы оценить значимость найденного значения F, надо воспользоваться табл. XIV Приложений. При этом для числа степеней свободы величин s_A^2 и s_Z^2 следует принять значения $n_A - 1$ и $n_A(n_a - 1)$, входящие в эти величины согласно формулам (5.3) и (5.4). Действительно, из (*) видно, что $s_A^2 = n_a s^2\{\bar{x}_a\}$ вычисляется по $n_A - 1$ независимым отклонениям, так как n_A отклонений $\bar{x}_a - \bar{x}$ связаны одним соотношением $\sum\limits_{a=1}^{n_A} (\bar{x}_a - \bar{x}) = 0$; далее, из (**) следует, что $s_Z^2 = n_a \langle s_x^2 \rangle_{\text{ср}}$ вычисляется по $n_A(n_a - 1) = n_A n_a - n_A$ независимым отклонениям, так как $n_A n_a$ отклонений $x_{ai} - \bar{x}_a$ (где $i = 1, 2, \ldots, n_a$; $a = 1, 2, \ldots, n_A$) связаны n_A соотношениями $\sum\limits_{i=1}^{n_A} (x_{ai} - \bar{x}_a) = 0$ (для $a = 1, 2, \ldots, n_A$).

Из табл. XIV Приложений находим, что для $n_A - 1 = 4$ степеней свободы числителя и $n_A(n_a - 1) = 5 \cdot 5 = 25$ степеней свободы знаменателя $F_{05} = 2{,}76$. Поскольку в нашем случае $F = 1{,}35$, то нулевая гипотеза не отвергается.

Результаты дисперсионного анализа обычно записывают в виде таблицы, имеющей вид табл. 45.

---

[1] В табл. 44 эти квадраты напечатаны мелким шрифтом правей и ниже значений вариант.

Таблица 45

Источник разброса	Вариации	Число степеней свободы	Дисперсии	Отношение дисперсий	Критические значения	
					$\alpha = 5\%$	$\alpha = \%$
Фактор A	$S_A = S_a - S_x$	$f_A = n_A - 1$	$s_A^2 = \frac{S_A}{f_A}$	$F = \frac{s_A^2}{s_Z^2}$		
Случайный Z	$S_Z = S_{ai} - S_a$	$f_Z = n_A(n_a - 1)$	$s_Z^2 = \frac{S_Z}{f_Z}$			

Для разбираемого примера таблица будет иметь следующий вид (табл. 46).

Таблица 46

Источник разброса	Вариации	Число степеней свободы	Дисперсии	Отношение дисперсий	Критические значения	
					$\alpha = 5\%$	$\alpha = 1\%$
Фактор A	79,2	4	19,8	1,35	2,76	4,18
Случайный Z	366	25	14,6			

Поскольку оказалось, что между s_A^2 и s_Z^2 нет значимого различия, то следует считать, что фактор A не оказывает существенного влияния на разброс значений вариант, так что $s_A^2 = 19{,}8$ может служить оценкой разброса вариант в той же мере, что и $s_Z^2 = 14{,}6$. Поэтому мы должны объединить обе оценки в одну усредненную, причем усреднение следует производить с учетом «веса» каждой из них, каковым является число степеней свободы:

$$s^2 = \frac{f_A s_A^2 + f_Z s_Z^2}{f_A + f_Z}. \tag{5.5}$$

Но $f_A s_A^2 = S_A$, $f_Z s_Z^2 = S_Z$, а

$$S_A + S_Z = (S_a - S_x) + (S_{ai} - S_a) = S_{ai} - S_x = S; \tag{5.6}$$

$$f_A + f_Z = (n_A - 1) + n_A(n - 1) = n_A n - 1 = f. \tag{5.7}$$

Таким образом, при $F < F_{05}$ принимаем в качестве оценки дисперсии величину

$$s_{Z^*}^2 = \frac{S}{f} = \frac{S_{ai} - S_x}{n_A n_a - 1}. \qquad (5.8)$$

В данном примере

$$S = 79{,}2 + 366 = 445{,}2; \quad f = 4 + 25 = 29$$

или по формулам (5.6) и (5.7)

$$S = 2562 - 2116{,}8 = 445{,}2; \quad f = 5 \cdot 6 - 1 = 29;$$

поэтому

$$s_{Z^*}^2 = \frac{445{,}2}{29} = 15{,}35.$$

Пример 2. В табл. 47 приведены данные об урожайности четырех сортов ячменя, каждый из которых высевался на пяти делянках (числа указывают урожай в килограммах с делянки). Можно ли считать, что различие средних урожаев сортов есть различие между выборочными средними, или же сорта действительно имеют разную урожайность?

Таблица 47

Номера сортов	Номера делянок					x_a	x_a^2	Доверительный интервал
	1	2	3	4	5			
1	8_{64}	6_{36}	7_{49}	6_{36}	8_{64}	35	1225	6,04÷7,96
2	9_{81}	10_{100}	7_{49}	9_{81}	8_{64}	43	1849	7,64÷9,56
3	5_{25}	5_{25}	4_{16}	3_{9}	6_{36}	23	529	3,64÷5,56
4	6_{36}	4_{16}	5_{25}	5_{25}	6_{36}	25	625	4,04÷5,96
				Сумма		126	4228	

$S_{ai} = 864;$

$S_a = \frac{4228}{5} = 845{,}6;$

$S_A = 845{,}6 - 793{,}8 = 51{,}8;$

$S_Z = 864 - 845{,}6 = 18{,}4;$

$s_A^2 = \frac{51{,}8}{3} = 17{,}27;$

$F = \frac{17{,}27}{1{,}15} = 15{,}0.$

$S_x = \frac{126^2}{5 \cdot 4} = \frac{15876}{20} = 793{,}8;$

$f_A = 4 - 1 = 3;$

$f_Z = 4(5 - 1) = 16;$

$s_Z^2 = \frac{18{,}4}{16} = 1{,}15.$

Выполнив расчеты, как это показано в табл. 47, получаем $F = 15{,}0$, что превышает граничное значение 5,29 для $\alpha = 1\%$. Следовательно, с вероятностью более 99% можно отвергнуть гипотезу о том, что все сорта имеют одинаковую урожайность.

В таком случае приобретает смысл нахождение доверительных интервалов для построчных средних (в данном случае — для средних урожайностей сортов). Очевидно, эти доверительные интервалы будут, в соответствии с (5.4):

$$\bar{x}_a \pm t_P \langle s_{\bar{x}} \rangle_{\text{ср}} = \frac{X_a}{n_a} \pm t_P \frac{s_Z}{\sqrt{n_a}}.$$

В нашем примере $n_a = 5$. Если принять $P = 95\%$, то при $f_Z = 16$ получим из табл. IV Приложений $t_{95}(16) = 2{,}02$, так что

$$t_P \frac{s_Z}{\sqrt{n_a}} = 2{,}02 \cdot \frac{\sqrt{1{,}15}}{\sqrt{5}} = 0{,}96.$$

Учитывая, что $\bar{x}_1 = X_1/n_1 = 35/5 = 7{,}00$; $\bar{x}_2 = 43/5 = 8{,}60$ и т. д., получаем доверительные интервалы, записанные в последнем столбце табл. 47.

Схема дисперсионного анализа представляет собой проверку нулевой гипотезы о том, что распределения вариант во всех строках одинаковы (даже если они относятся к разным популяциям). Поэтому в формулировку H_0 входит не только условие равенства всех средних значений, но и условие равенства строчных дисперсий.

Если варианты, записанные в клетках таблицы, представляют собой результаты измерения скорости счета радиоактивных препаратов, то предположение об одинаковости дисперсий заведомо неприемлемо. Причина в том, что скорость счета подчиняется распределению Пуассона, для которого дисперсия равна среднему значению (см. § 6 гл. 2). Анализ показывает, что если от значений x_i перейти к преобразованным значениям $y_i = \sqrt{x_i}$, то для величин y_i дисперсия почти не зависит от среднего значения. Условие постоянства дисперсий выполняется еще лучше, если пользоваться преобразованными величинами ψ_i из табл. XXII Приложений. Дальнейшие расчеты производятся с величинами y_i или ψ_i обычным образом.

Дисперсия зависит от среднего значения и в том случае, когда вариантами в дисперсионном анализе являются доли какого-нибудь альтернативного распределения, причем либо очень малые, либо близкие к 1; примером может служить процент всхожести

семян (который обычно больше 90%) или процент зараженности их вредителем (малые значения). Тогда целесообразно перейти от долей p_i к величинам φ_i (см. § 9 гл. 3), пользуясь табл. VII Приложений.

Если желательно, для упрощения расчетов, произвести кодирование вариант, то его нужно выполнять п о с л е преобразования (т. е. перехода к величинам φ или, в случае распределения Пуассона, к величинам $y = \sqrt{x}$ или ψ).

В заключение этого параграфа выпишем в одном месте все расчетные формулы однофакторного дисперсионного анализа:

$$\left.\begin{gathered} X_a = \sum_{i=1}^{n_a} x_{ai}, \quad X = \sum_{a=1}^{n_A} X_a; \\ S_x = \frac{1}{n_a n_A} X^2, \quad S_a = \frac{1}{n_a} \sum_{a=1}^{n_A} X_a^2, \quad S_{ai} = \sum_{a=1}^{n_A} \sum_{i=1}^{n_a} x_{ai}^2; \\ S_A = S_a - S_x, \quad S_Z = S_{ai} - S_a; \\ f_A = n_A - 1, \quad f_Z = n_A (n_a - 1); \\ s_A^2 = \frac{S_A}{f_A}, \quad s_Z^2 = \frac{S_Z}{f_Z}; \\ F = \frac{s_A^2}{s^2}. \end{gathered}\right\} \qquad (5.9)$$

§ 3. Однофакторный дисперсионный анализ при неодинаковых или больших объемах выборок

В примере 1 на каждой собаке было поставлено одно и то же число опытов (шесть). Это, конечно, не обязательно.

Если число повторностей в разных строках различно, то говорят, что анализируемый комплекс является *неравномерным*. В этом случае расчетные формулы несколько видоизменяются.

Пусть опять n_a есть число опытов в строке a (в примере 1 все n_a были одинаковы и равнялись шести). Тогда прежнее выражение

$$S_a = \frac{1}{n_a} \sum_{a=1}^{n_A} X_a^2$$

принимает вид

$$S_a = \sum_{a=1}^{n_A} \frac{X_a^2}{n_a}; \qquad (5.10)$$

полное число вариант, входящее в S_x, будет не $n_A n_a$, а $\sum_{a=1}^{n_A} n_a$, так что теперь

$$S_x = \frac{X^2}{\sum_{a=1}^{n_A} n_a}; \tag{5.11}$$

величина S_{ai} остается без изменения:

$$S_{ai} = \sum_{a=1}^{n_A} \sum_{i=1}^{n_a} x_{ai}^2.$$

Соответственно

$$f_Z = n_A (n_a - 1) = n_A n_a - n_A$$

переходит в

$$f_Z = \sum_{a=1}^{n_A} n_a - n_A, \tag{5.12}$$

а $f_A = n_A - 1$ остается без изменения. Когда все n_a равны между собой, мы получаем прежние формулы.

Пример 3. Каждый из четырех сортов пшеницы размещался на нескольких делянках примерно одинакового почвенного типа. Полученные урожаи (в пересчете на центнеры с гектара) представлены в табл. 48. Можно ли считать, что эти сорта имеют разную урожайность?

Таблица 48

Номера сортов	Номера делянок					
	1	2	3	4	5	6
1	17,0	17,2	16,1	17,0	16,8	
2	15,8	17,0	16,4			
3	17,4	16,6	16,2	15,6	15,5	17,2
4	15,7	16,8	15,1	15,2		

Для упрощения расчетов примем 16 за начало отсчета и изменим масштаб в 10 раз, чтобы избавиться от дробей. Тогда, например, вместо 17,2 мы запишем $(17{,}2 - 16)\cdot 10 = 12$, а вместо 15,6 запишем $(15{,}6 - 16)\cdot 10 = -4$. На результат дисперсионного

анализа такое кодирование, очевидно, не повлияет: с одной стороны, как это следует из формулы (1.19), дисперсии не зависят от одновременного прибавления или вычитания во всех вариантах одной и той же величины; с другой стороны, при изменении масштаба все дисперсии изменятся в соответствии с формулой (1.18) в одно и то же число раз, так что отношение дисперсий, которое нас интересует, останется без изменения.

Трансформированные указанным выше способом числа записаны в табл. 49, где приведены также все промежуточные данные; к таблице прибавлены столбцы со значениями n_a и X_a^2/n_a.

Таблица 49

Номера сортов	Номера делянок						X_a	X_a^2	n_a	$\frac{X_a^2}{n_a}$
	1	2	3	4	5	6				
1	10_{100}	12_{144}	1_1	10_{100}	8_{64}		41	1681	5	336,2
2	-2_4	10_{100}	4_{16}				12	144	3	48,0
3	14_{196}	6_{36}	2_4	-4_{16}	-5_{25}	12_{144}	25	625	6	104,2
4	-3_9	8_{64}	-9_{81}	-8_{64}			−12	144	4	36,0
						Сумма	66		18	524,4

$$S_a = \frac{1681}{5} + \frac{144}{3} + \frac{625}{6} + \frac{144}{2} = 336{,}2 + 48 + 104{,}2 + 36 = 524{,}4;$$

$$S_x = \frac{66^2}{5+3+6+4} = \frac{4356}{28} = 242; \quad S_{ai} = 1168.$$

Теперь имеем:

$$S_A = 524{,}4 - 242 = 282{,}4;$$
$$S_Z = 1168 - 524{,}4 = 643{,}6.$$

Число степеней свободы равно:

$$f_A = 4 - 1 = 3; \quad f_Z = (5 + 3 + 6 + 4) - 4 = 14,$$

так что

$$s_A^2 = \frac{282{,}4}{3} = 94{,}1, \quad s_Z^2 = \frac{643{,}6}{14} = 46{,}0, \quad F_{A/Z} = \frac{94{,}1}{46{,}0} = 2{,}05.$$

Поскольку это число меньше, чем $F_{05} = 3{,}34$ для $f_1 = 3$, $f_2 = 14$, нет оснований считать сорта различными (вероятность этого меньше 95%). Результаты сведем в табл. 50.

Таблица 50

Разброс	S	f		F	F_a	
					5%	1%
По сортам A .	282,4	3	94,1	2,05	3,34	5,56
Случайный Z .	643,6	14	46,0			

Если число вариант в каждой серии велико, то построенная обычным образом таблица дисперсионного анализа становится очень громоздкой. В этом случае имеет смысл произвести группировку вариант в разряды, так что в заголовках столбцов (или строк) записываются не номера опытов, а середины (или границы) разрядов, а в клетках — разрядные частоты вместо значений отдельных вариант.

Пример 4. С каждого из четырех участков почвы был произведен посев бактерий на 20 пластинках. В результате на пластинках были получены количества колоний, представленные в табл. 51.

Таблица 51

Участки почвы	Количество колоний на пластинках																			
1	7	4	8	10	10	7	16	11	7	12	14	3	6	3	5	17	8	8	6	12
2	5	10	9	4	7	5	1	11	12	15	7	17	15	10	5	3	6	7	16	6
3	6	7	9	10	15	14	12	12	4	7	12	8	4	11	4	10	7	12	8	10
4	7	7	11	19	8	8	12	7	12	8	4	12	10	5	11	4	10	10	10	7

Являются ли колебания средних в четырех выборках чисто случайными?

Группируя варианты в разряды (1—4, 5—8, 9—12, 13—16, 17—20), получаем табл. 52.

Удобней, конечно, весь расчет вести в условных единицах; они проставлены рядом с обозначениями разрядов.

Значения X_a теперь получаются так же, как при обработке обычных статистических рядов:

$$X_1 = 3\,(-2) + 9\,(-1) + 5 \cdot 0 + 2 \cdot 1 + 1 \cdot 2 = \\ = -15 + 4 = -11$$

и т. д.

Таблица 52

Значения вариант		Участки почвы				Сумма	$\sum x_{ai}^2$
количество колоний	x_i	1	2	3	4		
1—4	—2	3	3	3	2	11	44
5—8	—1	9	8	6	8	31	31
9—12	0	5	5	9	9	28	0
13—16	1	2	3	2	0	7	7
17—20	2	1	1	0	1	3	12
n		20	20	20	20	80	94
X_a		—11	—9	—10	—10	—40	
X_a^2		121	81	100	100	402	

Величина $S_{ai} = \Sigma\Sigma x_{ai}^2$ получается суммированием произведений суммарных разрядных частот на квадраты соответствующих разрядных значений:

$$S_{ai} = 11\cdot(-2)^2 + 31\,(-1)^2 + 28\cdot 0^2 + 7\cdot 1^2 + 3\cdot 2^2 = \\ = 44 + 31 + 0 + 7 + 12 = 94.$$

После этого находим обычным образом

$$S_a = \frac{1}{n_a}\sum_{a=1}^{n_A} X_a^2 = \frac{402}{20} = 20{,}1;\quad S_x = \frac{1}{n_A n_a}(\Sigma X_a)^2 = \frac{1600}{4\cdot 20} = 20,$$

так что

$$s_A^2 = \frac{S_a - S_x}{n_A - 1} = \frac{20{,}1 - 20}{3} = 0{,}33;$$

$$s_Z^2 = \frac{S_{ai} - S_a}{n_A(n-1)} = \frac{94 - 20{,}1}{4\cdot 19} = 0{,}97.$$

Поскольку $s_A^2 < s_Z^2$, то различие незначимо.

§ 4. Факторная доля вариабильности

При решении некоторых задач дисперсионного анализа может представить интерес определение той доли общей вариабильности, которая обусловлена действием изучаемого фактора A.

Напрашивается мысль определять эту долю как отношение S_A/S, где

$$S = S_A + S_Z. \tag{5.13}$$

Смысл величины S раскрывается следующим образом.

Учитывая, что $S_A = S_a - S_x$ и $S_Z = S_{ai} - S_a$, так что

$$S = (S_a - S_x) + (S_{ai} - S_a) = S_{ai} - S_x, \tag{5.14}$$

получаем после подстановки S_{ai} и S_x из (5.2):

$$S = \sum_{a=1}^{n_A} \sum_{i-1}^{n_a} x_{ai}^2 - \frac{1}{n_A n_a} X^2. \tag{5.15}$$

Применим опять формулу (1.17), но в обратном направлении. Это дает с учетом (5.1)

$$S = \sum_{a=1}^{n_A} \sum_{i=1}^{n_a} (x_{ai} - \bar{x})^2;$$

следовательно, S есть сумма квадратов отклонений всех значений вариант от общего среднего значения. Эту сумму назовем *полной вариацией*.

Таким образом, формула (5.13) имеет тот смысл, что вычисление S_A и S_Z представляет собой разбиение полной суммы квадратов отклонений S на две части, из которых одна — S_A — отражает влияние некоторого известного фактора A, а другая — S_Z — определяется только случайными вариациями.

Последнее обстоятельство и наводит на мысль о том, что отношение S_A/S могло бы характеризовать факторную долю вариабильности (т. е. долю, обусловленную фактором A). Но в действительности это не так. В самом деле, хотя S_A и отражает разброс построчных средних от общего среднего значения, нужно иметь в виду, что сами эти построчные средние определяются не только действием фактора A: так как это в ы б о р о ч н ы е средние, они всегда содержат в себе некоторую «неопределенность», связанную со случайностями вхождения вариант в выборку. В частности, даже когда фактор A заведомо отсутствует, эмпирические построчные средние, будучи выборочными средними, все равно не одинаковы, так что $S_A \neq 0$; между тем правильный показатель факторной доли вариабильности должен в этом случае давать нуль.

Чтобы пояснить это более выпукло, запишем равенство $S = S_Z + S_A$ в развернутом виде

$$\sum_{a,i} (x_{ai} - \bar{x})^2 = \sum_{a,i} (x_{ai} - \bar{x}_a)^2 + \sum_{a,i} (\bar{x}_a - \bar{x})^2. \tag{*}$$

Второе слагаемое правой части, т. е. величину $\Sigma(\bar{x}_a - \bar{\bar{x}})^2$, равную S_A, можно в свою очередь представить в виде суммы двух членов:

$$\sum(\bar{x}_a - \bar{\bar{x}})^2 = \sum(\bar{x}_a - \hat{x}_a)^2 + \sum(\hat{x}_a - \bar{\bar{x}})^2; \qquad (**)$$

это значит, что вариация эмпирических (выборочных) построчных средних $\bar{x}_a$ относительно общего среднего $\bar{\bar{x}}$ разлагается на две части: вариацию эмпирических средних $\bar{x}_a$ относительно «истинных» (не искаженных случайностями выборки) построчных средних $\hat{x}_a$ и вариацию «истинных» построчных средних $\hat{x}_a$ относительно общего среднего $\bar{\bar{x}}$. Очевидно, только вторая часть должна входить в показатель факторной доли, так как именно она связана с действием фактора A; первая же часть, как и величина $\Sigma(x_{ai} - \bar{x}_a)^2$ в (*), связана со случайностями образования выборки. Таким образом, в качестве показателя факторной доли вариабильности следует принять величину [1]

$$e_A^2 = \frac{\sum(\hat{x}_a - \bar{\bar{x}})^2}{\sum(x_{ai} - \bar{\bar{x}})^2}.$$

Подставляя равенство (**) в (*), имеем

$$\sum(x_{ai} - \bar{\bar{x}})^2 = \sum(x_{ai} - \bar{x}_a)^2 + \sum(\bar{x}_a - \hat{x}_a)^2 + \sum(\hat{x}_a - \bar{\bar{x}})$$

откуда

$$\sum(\hat{x}_a - \bar{\bar{x}})^2 = \sum(x_{ai} - \bar{\bar{x}})^2 - \sum(x_{ai} - \bar{x}_a)^2 - \sum(\bar{x}_a - \hat{x}_a)^2$$

Поэтому

$$e_A^2 = 1 - \frac{\sum(x_{ai} - \bar{x}_a)^2 + \sum(\bar{x}_a - \hat{x}_a)^2}{\sum(x_{ai} - \bar{\bar{x}})^2}. \qquad (***)$$

Если бы все данные относились не к выборке, а к генеральной совокупности, то мы бы имели $\bar{x}_a = \hat{x}_a$ и

$$\eta_A^2 = 1 - \frac{\sum(x_{ai} - \hat{x}_a)^2}{\sum(x_{ai} - \hat{\bar{x}})^2},$$

где η_A^2 — генеральное значение e_A^2.

---

[1] Точнее, величину $\Sigma(\hat{x}_a - \hat{\bar{x}})^2/\Sigma(x_{ai} - \hat{\bar{x}})^2$, где $\hat{\bar{x}}$ — «истинное» значение общего среднего. Однако в дальнейшем, при получении несмещенной оценки e_A^2, эта неточность будет устранена.

После деления обеих входящих сюда сумм на объем совокупности N мы получили бы соответствующие дисперсии $\sigma^2\{x_{ai} - \hat{x}_a\}$ и $\sigma^2\{x_{ai} - \hat{x}\}$, которые можно было бы обозначить через σ_Z^2 и σ^2; тогда показатель факторной доли вариабильности был бы равен

$$\eta_A^2 = 1 - \frac{\sigma_Z^2}{\sigma^2}.$$

Если же мы имеем дело с выборкой, как это всегда бывает в дисперсионном анализе, то деление сумм $\Sigma(x_{ai} - \bar{x}_a)^2 = S_Z$ и $\Sigma(x_{ai} - \bar{x})^2 = S$ на объем выборки n дает смещенные оценки соответствующих дисперсий. Для получения несмещенных оценок s_Z^2 и s^2 нужно суммы S_Z и S разделить на соответствующие числа степеней свободы f_Z и f. Но применение несмещенных оценок дисперсии означает учет вариабильности тех выборочных средних, относительно которых имеется разброс [см. вывод формулы (3.8) в § 3 гл. 3]. В случае s^2 речь идет о вариабильности общего среднего значения $\bar{x}$, а в случае s_Z^2 — о вариабильности эмпирических построчных средних $\bar{x}_a$ относительно «истинных» построчных средних $\hat{x}_a$. Следовательно, когда мы делим $\Sigma(x_{ai} - \bar{x}_a)^2$ не на полное число вариант n, а на число степеней свободы $f_Z = n - n_A$, мы тем самым автоматически учитываем вклад слагаемого $\Sigma(\bar{x}_a - \hat{x}_a)^2$ в (***). Поэтому «почти несмещенной» оценкой[1] величины η_A^2 будет

$$e_A^2 = 1 - \frac{s_Z^2}{s^2} = 1 - \frac{f}{f_Z}\ \frac{S_Z}{S} \tag{5.16}$$

или

$$e_A^2 = 1 - \frac{n-1}{n-n_A}\ \frac{S_Z}{S}. \tag{5.17}$$

«Почти несмещенной» оценкой доли случайной вариабильности будет, очевидно,

$$e_Z^2 = \frac{s_Z^2}{s^2} = \frac{f}{f_Z}\ \frac{S_Z}{S} = \frac{n-1}{n-n_A} \cdot \frac{S_Z}{S}. \tag{5.18}$$

Чем больше, при данном общем объеме выборки n, число групп n_A, тем малочисленней каждая из этих групп и тем больше вариабильность построчных средних. Этому отвечает увеличение поправочного множителя $\frac{n-1}{n-n_A}$ при увеличении числа групп n_A.

[1] Более строгий расчет показывает, что оценка e_A^2, даваемая формулой (5.16), все еще смещенная, но ее смещение весьма мало (при $n > 10$ это смещение не более $0{,}1/n$).

Заметим, что $^{*}e_A^2 = S_A/S$ можно было бы записать в виде

$$^{*}e_A^2 = 1 - \frac{S_Z}{S}; \qquad (5.16')$$

это была бы систематически завышенная оценка.

Для примера 2 имеем

$$e_A^2 = 1 - \frac{19}{16} \cdot \frac{18{,}4}{51{,}8 + 18{,}4} = 0{,}688,$$

т. е. несколько более двух третей общей вариабильности приходится на действие фактора сорта.

Величина e_A^2 будет рассмотрена в другом аспекте в § 3 гл. 8. Тогда станет ясно, почему мы обозначили оценку факторной доли вариабильности через e_A^2, а не e_A.

§ 5. Двухфакторный дисперсионный анализ без повторности

В предыдущем параграфе было показано, что полная вариация S представляет собой сумму

$$S = S_A + S_Z.$$

Так как S_Z есть остаток от вычитания S_A из S, то S_Z называют *остаточной вариацией*; соответствующую ей дисперсию s_Z^2 — *остаточной дисперсией*).

Влияние случайных причин сказывается непосредственно в том, что имеет место вариация значений в каждой строке таблицы. Когда мы квалифицируем эти причины как случайные, то это просто означает, что они нам неизвестны. Однако можно попытаться, основываясь на знании биологии изучаемого объекта, условий опыта и т. д., сделать те или иные предположения о причине вариаций внутри строк; эту причину мы можем назвать фактором B.

Пример 5. Зная, что описанные в примере 1 опыты производились на разных собаках, в разные дни, но в одинаковые часы, — начиная с 8 часов утра с двухчасовым интервалом между опытами, — можно предположить, что на результат опыта может влиять время суток. Это значит, что генеральные средние значения (для всех животных данного вида) для утренних и вечерних опытов различны, а так как результаты опытов сведены вместе, то появляется дополнительный разброс, помимо вызываемого фактором A (индивидуальными свойствами животных).

При сделанном предположении варианты, стоящие в одной строке, уже не являются повторностями, которые можно произ-

вольно переставлять. Каждая из них теперь соответствует как определенному номеру строки, так и определенному номеру столбца. Это приводит к тому, что столбцы и строки становятся совершенно равноправными. Поэтому как таблица, так и все вычисления приобретают полную симметрию относительно столбцов и строк, так что, обозначая общий номер столбцов вместо i через b ($b = 1, 2, \ldots, n_B$), мы можем записать следующие формулы:

$$\left.\begin{aligned} &S_A = S_a - S_x, \quad S_B = S_b - S_x; \\ &S_a = \frac{1}{n_B}\sum_{a=1}^{n_A} X_a^2, \quad S_b = \frac{1}{n_A}\sum_{b=1}^{n_B} X_b^2; \\ &X_a = \sum_{b=1}^{n_B} x_{ab}, \quad X_b = \sum_{a=1}^{n_A} x_{ab}; \\ &f_A = n_A - 1, \quad f_B = n_B - 1; \\ &s_A^2 = \frac{S_A}{f_A}, \quad s_B^2 = \frac{S_B}{f_B}; \end{aligned}\right\} \tag{5.19}$$

кроме того,

$$\left.\begin{aligned} &S = S_{ab} - S_x; \quad S_Z = S - (S_A + S_B); \\ &S_{ab} = \sum_{a=1}^{n_A}\sum_{b=1}^{n_B} x_{ab}^2; \quad S_x = \frac{1}{n_A n_B} X^2; \\ &X = \sum_{a=1}^{n_A} X_a = \sum_{b=1}^{n_B} X_b. \end{aligned}\right\} \tag{5.20}$$

Мы видим, что в S, S_A и S_B входит в качестве вычитаемого одна и та же величина S_x; поэтому ее часто называют *корректирующим фактором*.

Применим эти формулы к нашему примеру. Величины S_a и S_x были вычислены ранее ($S_a = 2196$, $S_x = 2116{,}8$); S_A, а также s_A^2 остаются без изменения ($S_A = 79{,}2$, $s_A^2 = 19{,}8$). Пользуясь данными из табл. 53, дополнительно находим:

$$S_b = \frac{11824}{5} = 2364{,}8;$$

$$S_B = 2364{,}8 - 2116{,}8 = 248;$$

$$f_B = 6 - 1 = 5;$$

$$s_B^2 = 49{,}6.$$

Таблица 63

Номера собак (a)	Номера опытов (b)						X_a	X_a^2
	1	2	3	4	5	6		
1	6_{36}	4_{16}	9_{81}	8_{64}	15_{225}	12_{144}	54	2916
2	9_{81}	7_{49}	3_{9}	4_{16}	11_{121}	14_{196}	48	2304
3	6_{36}	8_{64}	10_{100}	14_{196}	13_{169}	15_{225}	66	4356
4	2_{4}	3_{9}	7_{49}	4_{16}	9_{81}	11_{121}	36	1296
5	6_{36}	5_{25}	4_{16}	10_{100}	14_{196}	9_{81}	48	2304
X_b	29	27	33	40	62	61	252	13176
X_b^2	841	729	1089	1600	3844	3721	11824	2562

$$S_a = \frac{13176}{6} = 2196; \qquad S_x = \frac{252^2}{5 \cdot 6} = \frac{63504}{30} = 2116{,}8;$$

$$S_b = \frac{11824}{5} = 2364{,}8; \quad S_{ab} = 2562;$$

$$S_A = 2196 - 2116{,}8 = 79{,}2; \; S_B = 2364{,}8 - 2116{,}8 = 248;$$

$$f_A = 5 - 1 = 4; \; s_A^2 = \frac{79{,}2}{4} = 19{,}8; \; f_B = 6 - 1 = 5; \; s_B^2 = \frac{248}{5} = 49{,}6.$$

Мы видим, что s_B^2 значительно превышает s_A^2, т. е. время суток оказывает большее влияние на значения вариант, чем индивидуальные свойства животных, и это нужно учесть при окончательной интерпретации результатов.

Установление наличия определенного фактора, вызывающего вариацию значений внутри строк, заставляет пересмотреть вывод об одинаковости генеральных средних у использованных в опытах животных — вывод, который основывался на том, что s_A^2 не превышает значимо величину s_Z^2. Поскольку выяснилось, что вариации внутри строк не являются чисто случайными, нельзя считать, что делением величины

$$s^2\{x_{ab}\} = \frac{\sum_{b=1}^{n_B} (x_{ab} - \bar{x}_a)^2}{n_B - 1}$$

— дисперсии вариант внутри строки — на число вариант в строке n_B мы получим дисперсию строчного среднего значения, т. е.

величину $s^2_{\bar{x}_a}$. Очевидно, истинная величина $s^2_{\bar{x}_a}$ (а вместе с тем и усредненная величина $\langle s^2_{\bar{x}_a} \rangle$) будет гораздо меньше; соответственно меньше будет и величина s^2_Z, с которой нам нужно сравнить s^2_A.

Чтобы найти исправленное значение s^2_Z, нужно получить исправленное значение S_Z. Последняя величина получится, если мы из прежнего значения $S_Z = S_{ab} - S_a$ вычтем величину S_B, отражающую влияние фактора B. Действительно, как указывалось выше, дисперсионный анализ представляет собой разбиение полной суммы квадратов отклонений

$$S = \sum_{a=1}^{n_A} \sum_{b=1}^{n_B} (x_{ab} - \bar{x})^2$$

на части, связанные с упорядоченными и случайными факторами. В случае однофакторного анализа это разбиение имело вид

$$S = S_A + S_Z. \tag{5.21}$$

При двухфакторном анализе это выражение должно, естественно, замениться выражением

$$S = S_A + S_B + S_Z. \tag{5.22}$$

Отсюда и получается значение S_Z, записанное в (5.20). В нашем примере

$$S = S_{ab} - S_x = 2562 - 2116{,}8 = 445{,}2;$$
$$S_Z = S - (S_A + S_B) = 445{,}2 - (79{,}2 + 248) = 118.$$

Чтобы вычислить s^2_Z, нужно S_Z разделить на f_Z — число степеней свободы величины S_Z. При вычислении S_Z используется $n_A n_B$ отклонений $x_{ab} - \bar{x}$, связанных $n_A + n_B - 1$ независимыми соотношениями:

$$n_A \text{ соотношениями } \sum_{b=1}^{n_B} (x_{ab} - \bar{x}_a) = 0 \quad (a = 1, 2, \ldots, n_A),$$

$$n_B \text{ соотношениями } \sum_{a=1}^{n_A} (x_{ab} - \bar{x}_b) = 0 \quad (b = 1, 2, \ldots, n_B),$$

из которых одно есть следствие остальных в силу условия $\sum_{a=1}^{n_A} \sum_{b=1}^{n_B} (x_{ab} - \bar{x}) = 0$. Поэтому

$$f_Z = n_A n_B - (n_A + n_B - 1) = (n_A - 1)(n_B - 1).$$

Заметим, что по аналогии с (5.22) имеет место соотношение

$$f = f_A + f_B + f_Z. \tag{5.23}$$

Действительно,

$$f = (n_A - 1) + (n_B - 1) + (n_A - 1)(n_B - 1) = n_A n_B - 1$$

есть число степеней свободы величины S, поскольку S вычисляется по $n_A n_B$ отклонениям $x_{ab} - \bar{x}$, связанным одним соотношением

$$\sum_{a=1}^{n_A} \sum_{b=1}^{n_B} (x_{ab} - \bar{x}) = 0.$$

Таким образом,

$$s_Z^2 = \frac{S_Z}{(n_A - 1)(n_B - 1)},$$

что дает для рассматриваемого примера

$$s_Z^2 = \frac{118}{4 \cdot 5} = 5{,}90.$$

Поэтому значения F равны:

$$F_{A/Z} = \frac{19{,}8}{5{,}90} = 3{,}36; \quad F_{B/Z} = \frac{49{,}6}{5{,}90} = 8{,}41.$$

Теперь мы можем записать результат анализа в виде табл. 54.

Таблица 54

Разброс	Сумма квадратов отклонений S	Число степеней свободы f	Дисперсии s^2	Отношение дисперсий F	Критические значения F	
					$\alpha = 5\%$	$\alpha = 1\%$
По фактору A	79,2	4	19,8	3,36	2,87	4,43
По фактору B	248	5	49,6	8,41	2,71	4,10
Остаточный Z	118	20	5,90			

Из сравнения фактических значений F с критическими значениями видим, что различие генеральных средних по столбцам (т. е. по времени проведения опыта) значимо ($P > 0{,}99$); таким образом, предположение о том, что разброс вариант внутри строк не случаен, подтвердилось.

Пример 6. Каждый из четырех сортов картофеля был размещен на пяти делянках одинакового размера и почвенного типа; для каждого сорта были применены пять различных удобрений. Полученные урожаи (в тоннах) представлены в табл. 55.

Таблица 55

Сорта	Удобрения				
	1	2	3	4	
1	1,9	2,2	2,6	1,8	2,1
2	2,5	1,9	2,3	2,6	2,2
3	1,7	1,9	2,2	2,0	2,1
4	2,1	1,8	2,5	2,3	2,4

Требуется установить, есть ли основания считать, что урожаи разных сортов в среднем различны (т. е. независимо от вида удобрений) и что эффективность разных удобрений различна независимо от сорта.

Таблица 56

Сорта	Удобрения					X_a	X_a^2
	1	2	3	4	5		
1	-1_1	2_4	6_{36}	-2_4	1_1	6	36
2	5_{25}	-1_1	3_9	6_{36}	2_4	15	225
3	-3_9	-1_1	2_4	0_0	1_1	-1	1
4	1_1	-2_4	5_{25}	3_9	4_{16}	11	121
X_b	2	-2	16	7	8	31	383
X_b^2	4	4	256	49	64	377	191

$$S_a = \frac{383}{5} = 76{,}6;\quad S_b = \frac{377}{4} = 94{,}25;\quad S_x = \frac{31^2}{4\cdot 5} = \frac{961}{20} = 48{,}05;$$

$$S_A = 76{,}6 - 48{,}05 = 28{,}55;\quad S_B = 94{,}25 - 48{,}05 = 46{,}20;$$

$$S = 191 - 48{,}05 = 142{,}95;\quad S_Z = 142{,}95 - (28{,}55 + 46{,}20) = 68{,}20;$$

$$s_A^2 = \frac{28{,}55}{3} = 9{,}52;\quad s_B^2 = \frac{46{,}20}{4} = 11{,}55;\quad s_Z^2 = \frac{68{,}20}{12} = 5{,}68;$$

$$F_{A/Z} = \frac{9{,}52}{5{,}68} = 1{,}68;\quad F_{B/Z} = \frac{11{,}55}{5{,}68} = 2{,}03.$$

Чтобы упростить вычисления, будем измерять урожаи в центнерах, причем за начало отсчета примем 20 ц. Тогда получим значения, записанные в табл. 56. Там же приведены все остальные расчеты. Результат анализа приведен в табл. 57.

Таблица 57

Разброс	S	f		F	F_α	
					5%	1%
По сортам .	28,55	3	9,52	1,68	3,49	5,95
По удобрениям	46,20	4	11,55	2,03	3,26	5,41
Остаточный .	68,20	12	5,68			
Полный .	142,95	19				

Поскольку для обоих факторов $F < F_{05}$, следует считать, что нулевая гипотеза не отвергается, т. е. что вариации могли иметь чисто случайное происхождение.

§ 6. Метод случайных (рэндомизированных) блоков

Сравним табл. 46 и 54. Предположение о влиянии на результаты опыта времени суток (фактор B) позволило выделить из вариации $S_Z = 366$ (табл. 46) величину $S_B = 248$ (табл. 54), так что на долю случайного разброса осталось лишь $S = 366 - 248 = 118$ (табл. 54). Иными словами, оказалось, что в первом расчете было получено завышенное значение случайной вариации S_Z — в нее был включен разброс, вызываемый неодинаковым влиянием разного времени суток.

После того как влияние этого фактора было исключено из S_Z, случайная дисперсия резко снизилась — с 14,6 в табл. 46 до 5,90 в табл. 54. И если раньше влияние фактора A выглядело как явно незначимое ($F_{A/Z} = 1{,}35$ против $F_{05} = 2{,}76$), то теперь это влияние стало значимым (по крайней мере на 5%-ном уровне: $F_{A/Z} = 3{,}36 > F_{05} = 2{,}87$).

Это обстоятельство можно использовать в тех случаях, когда имеется какой-нибудь фактор, который сам по себе не представляет интереса, но который, если влияние его не выделить из случайной вариации, может замаскировать влияние интересующего нас фактора.

Пусть, например, мы хотим сравнить эффективность различных видов удобрений и для этого ставим опыт на ряде делянок, повторяя его несколько лет. Если результаты, полученные в разные годы, считать просто повторностями, то случайная дисперсия окажется очень большой прежде всего вследствие колебаний погодных условий от года к году. Ясно, что влияние разных удобрений выявится гораздо лучше, если фактор годичных колебаний условий будет исключен. В то же время очевидно, что сам по себе этот фактор совершенно неинтересен — он был заведомо известен.

Другой вариант этой задачи возникает, когда опыт по действию разных удобрений ставится в пределах одного года, а повторность обеспечивается использованием для каждого вида удобрения нескольких делянок. Если, например, изучается 4 вида удобрения и желательна повторность не менее 6, то в опыте должно участвовать не меньше 24 делянок. Понятно, что при таком числе делянок разброс в их почвенных свойствах, условиях освещения и увлажнения будет довольно велик. Но если разбить эти 24 делянки на 6 блоков по 4 делянки в каждом, то можно подобрать эти блоки так, что в пределах каждого из них делянки будут более или менее близки по своим свойствам, различие же между блоками будет рассматриваться как некоторый дополнительный фактор, который может быть оценен при дисперсионном анализе. Ясно, что этот фактор будет интересовать нас не сам по себе, а лишь постольку, поскольку оценка его позволит очистить от него случайную вариацию. Этот же метод можно использовать и в опытах с животными — здесь роль более или менее однородных блоков могут играть пометы.

Такой способ постановки опыта является обобщением того, что было описано в § 5 гл. 4 (сравнение совокупностей с попарно связанными вариантами), — вместо двух сравниваются сразу несколько выборок. Этот способ называют *методом случайных (рэндомизированных) блоков*, в отличие от *метода полной рэндомизации* (от английского random — случайный), когда в пределах каждой градации фактора A условия подбираются совершенно случайно.

Пример 7. Для сравнения урожайности четырех сортов ржи каждый из них был высеян на 5 делянках, причем так, что отобранные для опыта 20 делянок были разбиты на 5 примерно однородных блоков по 4 делянки в каждом. Результаты (в *ц/га*) представлены в табл. 58.

Возьмем за начало отсчета 15 *ц/га* и уменьшим масштаб в 10 раз (чтобы избавиться от дробей). Тогда получатся числа, записанные в табл. 59. Там же приведен и весь дальнейший расчет. Результат дисперсионного анализа (табл. 60) показывает, что

Таблица 58

Сорта	Блоки				
	1	2	3	4	5
1	13,4	14,1	15,4	16,7	17,7
2	15,2	15,0	17,6	17,3	17,7
3	12,8	11,5	14,4	13,6	17,2
4	13,2	15,2	15,7	17,1	18,9

различие урожайности по сортам значимо (9,95>5,95). Значимо также различие по блокам, но этого можно было ожидать заранее (и не для этого ставился опыт).

Таблица 59

С \ Б	1	2	3	4		X_a	X_a^2
1	-16_{256}	-9_{81}	4_{16}	17_{289}	27_{729}	23	529
2	2_4	0_0	26_{676}	23_{529}	27_{729}	78	6084
3	-22_{484}	-35_{1225}	-6_{36}	-14_{196}	21_{441}	−56	3136
4	-18_{324}	2_4	7_{49}	21_{441}	39_{1521}	51	2601
X_b	−54	−42	31	47	114	96	12 350
X_b^2	2916	1764	961	2209	12 996	20 846	8030

$$S_a = \frac{12350}{5} = 2470;\ S_b = \frac{20846}{4} = 5211{,}5;\ S_x = \frac{96^2}{20} = \frac{9216}{20} = 460{,}8;$$

$$S_A = 2470-460{,}8=2009{,}2;\ S_B = 5211{,}5-460{,}8=4750{,}7;$$

$$S = 8030-460{,}8=7569{,}2;\ S_Z = 7569{,}2-(2009{,}2+4750{,}7) = 809{,}3;$$

$$s_A^2 = \frac{2009{,}2}{3} = 669{,}7;\ s_B^2 = \frac{4750{,}7}{4} = 1187{,}7;\ s_Z^2 = \frac{809{,}3}{12} = 67{,}44;$$

$$F_{A/Z} = \frac{669{,}7}{67{,}44} \approx 9{,}95;\ F_{B/Z} = \frac{1187{,}7}{67{,}44} \approx 17{,}6.$$

Если бы разбиение на блоки не производилось, т. е. опыт был бы полностью рэндомизирован, то распределение сортов по делянкам было бы другим. Это несколько изменило бы результат за счет того, что один сорт мог бы получить, скажем, две «хороших» делянки и ни одной «плохой», а другой сорт — наоборот.

Таблица 60

Разброс	S	f		F	F_α	
					5%	1%
По сортам .	2009,2	3	669,7	9,95	3,49	5,95
По блокам .	4750,7	4	1187,7	17,6	3,26	5,41
Остаточный .	809,3	12	67,44			
Полный .	7569,2	19				

Но еще важней, что в случайный (т. е. неконтролируемый) разброс оказался бы включенным в е с ь разброс, связанный с неодинаковостью делянок, а не та небольшая его часть, которая остается при разбиении делянок на блоки. Это сильно увеличило бы s_Z^2, что могло бы сделать незначимым отношение s_A^2/s_Z^2. В этом легко убедиться, если переставить местами (случайным образом, например с помощью игральной кости) клетки в пределах строк табл. 59:

$$\begin{pmatrix} 1&2&3&4&5\\ 1&2&3&4&5\\ 1&2&3&4&5\\ 1&2&3&4&5 \end{pmatrix} \rightarrow \begin{pmatrix} 2&1&5&4&3\\ 1&3&4&5&2\\ 3&1&5&2&4\\ 2&4&3&1&5 \end{pmatrix}$$

Таблица 61

С \ Б	1	2	3	4		X_a	X_a^2
1	-9_{81}	-16_{256}	27_{729}	17_{289}	4_{16}	23	529
2	2_4	26_{676}	23_{529}	27_{729}	0_0	78	6084
3	-6_{36}	-22_{484}	21_{441}	-35_{1225}	-14_{196}	−56	3136
4	2_4	21_{441}	7_{49}	-18_{324}	-39_{1521}	51	2601
X_b	−11	9	78	−9	29	96	12 350
X_b^2	121	81	6084	81	841	7208	8030

Тогда получится табл. 61. Значения S_a, S_{ab} и S_x остаются, конечно, прежними; поэтому остаются прежними S_A и S и, как и ранее, $s_A^2 = 669{,}7$. Но $S_B = 7208 : 4 = 1802$, что дает $S_B = 1802 -$

— 460,8 = 1341,2 и $s_B^2 = 1341,2 \quad 4 = 335,3$, т. е. гораздо меньшее значение, чем прежде. Зато сильно возрастает s_Z^2: так как величина $S - S_A$ осталась без изменения, а S_B уменьшилась, то разность $(S - S_A) - S_B = S_Z$ возросла до величины 4218,8; поэтому теперь $s_Z^2 = 4218,8 \quad 12 = 351,6$. Как и следовало ожидать, отношение $s_B^2/s_Z^2 = 0,95$ стало незначимым. Но незначимым стало и отношение

$$F_{A/Z} = \frac{669,7}{351,6} = 1,91 < F_{05}(3; 12) = 3,49.$$

Поскольку дисперсия s_B^2 оказалась незначимой, мы можем считать ее еще одной оценкой случайной дисперсии. Тогда в соответствии с формулой (5.5) из § 2 наилучшей оценкой случайной дисперсии будет

$$s_{Z^*}^2 = \frac{S_B + S_Z}{f_B + f_Z} = \frac{1341,2 + 4218,8}{4 + 12} = \frac{5560}{16} = 347,5.$$

Тот же результат получился бы и из табл. 59, если бы мы игнорировали фактор B и объединили бы s_B^2 и s_Z^2:

$$s_{Z^*}^2 = \frac{S_B + S_Z}{f_B + f_Z} = \frac{4750,7 + 809,3}{4 + 12} = \frac{5560}{16} = 347,5.$$

При сравнении с этой улучшенной оценкой $s_{Z^*}^2$ фактор A все равно оказывается незначимым:

$$F_{A/Z^*} = \frac{669,7}{347,5} = 1,93 < F_{05}(3; 16) = 3,24.$$

Применение метода случайных блоков требует, чтобы в каждом блоке было одинаковое число элементов. Однако иногда случается, что хотя при закладке опыта это условие было соблюдено, результаты получились не для всех клеток таблицы — на некоторых делянках погиб урожай, заболели отдельные животные или произошли еще какие-нибудь изменения, заведомо исказившие результат.

В таком случае приходится заменять недостающую варианту x_{ab} какой-либо ее разумной оценкой. Расчеты показывают, что лучшей такой оценкой является

$$x'_{ab} = \frac{n_A X'_a + n_B X'_b - X'}{(n_A - 1)(n_B - 1)} \qquad (5.24)$$

где X'_a, X'_b и X' вычислены по обычным формулам (5.19) и (5.20), но по действительно наблюденным значениям (т. е. с одной выпавшей вариантой). Например, если бы в табл. 59 выпала варианта $x_{24} = 23$ (т. е. из второй строки и четвертого столбца), то мы бы имели:

$$X'_{a=2} = 2 + 0 + 26 + 27 = 55 \ (= 78 - 23);$$
$$X'_{b=4} = 17 - 14 + 21 = 24 \ (= 47 - 23);$$
$$X' = 23 + 55 - 56 + 51 = 73 \ (= 96 - 23),$$

так что оценка варианты x_{24} была бы

$$x'_{24} = \frac{4 \cdot 55 + 5 \cdot 24 - 73}{3 \cdot 4} = \frac{267}{12} = 22{,}25 \approx 22.$$

Для варианты $x_{12} = -9$ аналогично получим:

$$X'_{a=1} = -16 + 4 + 17 + 27 = 32;$$
$$X'_{b=2} = 0 - 35 + 2 = -33;$$
$$X' = 96 - (-9) = 105;$$
$$x'_{12} = \frac{4 \cdot 32 + 5(-33) - 105}{3 \cdot 4} = \frac{-142}{12} = -11{,}83 \approx -12$$

и т. д. Указанный способ можно применять только в том случае, если не хватает одной варианты. Если же отсутствует большее число данных, то приходится пользоваться несколько более сложными формулами [1].

Дальнейший дисперсионный анализ проводят обычным образом, считая восстановленные варианты как бы полученными экспериментально. Однако поскольку в действительности для их получения были использованы соотношения типа (5.24), связывающие x'_{ab} с другими вариантами, величины x'_{ab} нельзя считать независимыми вариантами. Поэтому число степеней свободы остаточной (и полной) вариации уменьшают на число восстановленных вариант. Числа же степеней свободы для s^2_A и s^2_B остаются без изменений, так как они определяются соответственно числом строк и столбцов.

Само собой разумеется, что указанный способ восстановления недостающих вариант можно применять и при обычном двухфакторном анализе, когда оценка значимости фактора B является самостоятельной целью исследования.

[1] См. книгу Дж. У. Снедекора (1961), стр. 292.

§ 7. Двухфакторный дисперсионный анализ с повторными данными

Величину S_Z, даваемую выражением (5.20), мы считали рассеянием, связанным лишь со случайными вариациями. В действительности это не всегда так. Дело в том, что если имеются два каких-нибудь упорядоченных фактора A и B, каждый из которых вызывает некоторое увеличение полного рассеяния S, то может иметь место известный «перекрестный эффект», связанный, как принято говорить, со «взаимодействием» факторов A и B. В случае, описанном в примере 1, этот эффект может состоять в том, что не у всех собак скорость реакции зависит одинаково от времени суток; в примере 6 взаимодействие факторов может проявиться в том, что разные виды удобрения наиболее эффективны для различных сортов картофеля (так что нельзя сказать, что одно удобрение «вообще» хуже другого).

Другой яркий пример — влияние влажности на рост растений: известно, что повышенная влажность стимулирует рост растений при высоких температурах, но замедляет его при низких; поэтому если двумя изучаемыми факторами будут температура и влажность почвы, то «эффект взаимодействия» окажется очень большим.

Наличие такого эффекта приводит к тому, что разложение полной вариации S должно иметь вид не (5.22), а

$$S = S_A + S_B + S_{AB} + S_Z, \tag{5.25}$$

где S_{AB} — вариация, включающая эффект взаимодействия.

Сравнение выражений (5.22) и (5.25) показывает, что величина S_Z, которую мы ранее находили по формуле (5.20), есть не только рассеяние за счет случайных вариаций, но содержит еще величину[1] S_{AB}. Чтобы вычленить последнюю из S_Z, недостаточно данных, имеющихся в табл. 53 или 55. Для вычисления «чистого» эффекта случайных вариаций нужно, чтобы в каждой клетке таблицы была не одна, а несколько вариант; тогда средний разброс вариант в клетках определит значение s_Z^2, поскольку в данном приближении можно считать, что он не зависит от значений факторов A и B и комбинации[2] AB.

[1] Это, между прочим, означает, что величина S_Z, находимая в бесповторном двухфакторном анализе, всегда несколько завышена, так что значения $F_{A/Z}$ и $F_{B/Z}$ оказываются заниженными. Поэтому при не очень малых выборках можно удовлетворяться менее жестким уровнем значимости.

[2] Впрочем, некоторые заключения об S_{AB} можно сделать и в случае бесповторного комплекса. Именно, было показано, что при определенных, достаточно естественных добавочных предположениях удается построить критерий для проверки гипотезы $S_{AB} = 0$ при отсутствии повторностей (см. Г. Шеффе. Дисперсионный анализ. М., Физматгиз, 1963).

Приведем без вывода расчетные формулы, являющиеся естественным обобщением формул для бесповторного комплекса:

$$\left.\begin{aligned} S &= S_{abi} - S_x; & f &= n_A n_B n - 1;\\ S_A &= S_a - S_x; & f_A &= n_A - 1;\\ S_B &= S_b - S_x; & f_B &= n_B - 1;\\ S_Z &= S_{abi} - S_{ab/i}; & f_Z &= n_A n_B (n-1);\\ S_{AB} &= S - (S_A + S_B + S_Z); & f_{AB} &= (n_A - 1)(n_B - 1) = f_A f_B, \end{aligned}\right\} \quad (5.26)$$

где

$$\left.\begin{aligned} S_x &= \frac{1}{n_A n_B n} X^2;\\ S_a &= \frac{1}{n_B n} \sum_{a=1}^{n_A} X_a^2;\\ S_b &= \frac{1}{n_A n} \sum_{b=1}^{n_B} X_b^2;\\ S_{ab/i} &= \frac{1}{n} \sum_{a=1}^{n_A} \sum_{b=1}^{n_B} \left(\sum_{i=1}^{n} x_{abi} \right)^2\\ S_{abi} &= \sum_{a=1}^{n_A} \sum_{b=1}^{n_B} \sum_{i=1}^{n} x_{abi}^2, \end{aligned}\right\} \quad (5.27)$$

причем X_a — сумма всех значений вариант, относящихся к строке a; X_b — сумма всех значений вариант, относящихся к столбцу b; X — сумма всех значений вариант таблицы; n — число вариант в каждой клетке.

Эти величины удовлетворяют как соотношению (5.25), так и аналогичному соотношению

$$f = f_A + f_B + f_{AB} + f_Z. \quad (5.28)$$

Если $n = 1$, т. е. в каждой клетке находится одна варианта, то, как видно из (5.27),

$$S_{ab/i} = S_{abi} = S_{ab};$$

тогда получается $S_Z = 0$, а S_{AB} принимает вид (5.20), т. е. получается выражение, которое мы ранее обозначали как S_Z.

Пример 8. Были поставлены опыты, аналогичные описанным в примере 6, но для каждого сорта и каждого вида удобрения

использовались четыре делянки ($n = 4$). Вопросы, подлежащие выяснению при помощи дисперсионного анализа, — те же, что и в предыдущем примере.

При небольших значениях n исходные и промежуточные данные удобно разместить в виде табл. 62. Каждая клетка, стоящая на пересечении строки и столбца, разбивается на 4 части (в табл. 63

Таблица 62

A \ B	1		2		3		4		X_a	X_a^2
		x^2	x	x^2	x	x^2		x^2		
1	2	4	3	9	2	4	0	0		
	2	4	2	4	3	9	4	16		
	2	4	2	4	−2	4	2	4		
	2	4	4	16	2	4	3	9		
	8_{64}	16	11_{121}	33	5_{25}	21	9_{81}	29	33	1089
2	0	0	1	1	1	1	−1	1		
	−1	1	0	0	0	0	0	0		
	1	1	−1	1	1	1	0	0		
	2	4	0	0	0	0	−1	1		
	2_4	6	0_0	2	2_4	2	-2_4	2	2	4
3	2	4	3	9	2	4	2	1		
	1	1	3	9	−1	1	1	1		
	1	1	3	9	−1	1	1	1		
	1	1	4	16	−1	1	0	0		
	5_{25}	7	13_{169}	43	-1_1	7	3_9	3	20	400
X_b	15		24		6		10		55	1493
X_b^2	225		576		36		100		937	$\frac{507}{171}$

отдельно показана клетка с $a = 1$, $b = 2$). В левой части записываются значения вариант: вверху — сами значения (в данном случае числа 3, 2, 2, 4), внизу — их сумма (11). В правой верхней части проставляются квадраты значений — в данном случае $9 = 3^2$, $4 = 2^2$, $16 = 4^2$; в правой нижней части — сумма этих квадратов ($33 = 9 + 4 + 4 + 16$). Число 11 из левой нижней части клетки вой-

дет—по горизонтали—в X_a для своей строки и—по вертикали—в X_b для своего столбца. Число 33 из правой нижней части войдет в S_{abi} — сумму всех квадратов значений. В левой нижней части записано также число $121 = 11^2$; оно войдет в $S_{ab/i}$ — среднюю сумму квадратов поклеточных сумм.

Таблица 63

3	9
2	4
2	4
4	16
11	33

Затем обычным образом вычисляются X_a и X_b (например, $X_a = 8 + 11 + 5 + 9 = 33$ для $a = 1$, $X_b = 11 + 0 + 13 = 24$ для $b = 2$) и их квадраты, после чего используются формулы (5.26) и (5.27). В данном случае имеем:

$$X = \sum_{a=1}^{n_A} X_a = \sum_{b=1}^{n_B} X_b = 33 + 2 + 20 = 15 + 24 + 6 + 10 = 55;$$

$$\sum_{a=1}^{n_A} X_a^2 = 33^2 + 2^2 + 20^2 = 1089 + 4 + 400 = 1493;$$

$$\sum_{b=1}^{n_B} X_b^2 = 15^2 + 24^2 + 6^2 + 10^2 = 225 + 576 + 36 + 100 = 937;$$

$$\sum_{a=1}^{n_A}\sum_{b=1}^{n_B}\left(\sum_{i=1}^{n} x_{abi}\right)^2 = 64 + 121 + 25 + \quad + 1 + 9 = 507,$$

$$\sum_{a=1}^{n_A}\sum_{b=1}^{n_B}\sum_{i=1}^{n} x_{abi}^2 = 16 + 33 + 21 + \quad + 7 + 3 = 171$$

— все эти числа записаны в табл. 62.

Теперь вычисляем:

$$S_x = \frac{55^2}{3\cdot 4\cdot 4} = 63; \quad S_a = \frac{1493}{4\cdot 4} = 93{,}33; \quad S_b = \frac{937}{3\cdot 4} = 78{,}08;$$

$$S_{ab/i} = \frac{507}{4} = 126{,}75,$$

после чего находим:

$$S = 171 - 63 = 108;$$

$$S_A = 93{,}33 - 63 = 30{,}33; \quad S_B = 78{,}08 - 63 = 15{,}08;$$

$$S_Z = 171 - 126{,}75 = 44{,}25;$$

$$S_{AB} = 108 - (30{,}33 + 15{,}08 + 44{,}25) = 18{,}34.$$

Для числа степеней свободы имеем:

$$f_A = 3 - 1 = 2;\ f_B = 4 - 1 = 3;\ f_{AB} = 2 \cdot 3 = 6;$$
$$f_Z = 3 \cdot 4\ (4 - 1) = 36;$$

поэтому

$$s_A^2 = \frac{30{,}33}{2} = 15{,}16; \qquad s_B^2 = \frac{15{,}08}{3} = 5{,}03;$$
$$s_{AB}^2 = \frac{18{,}34}{6} = 3{,}06; \qquad s_Z^2 = \frac{44{,}25}{36} = 1{,}23.$$

Методика оценки значимости этих дисперсий зависит от свойств факторов A и B. Часто выбор градаций факторов определяется экспериментатором; так, в примере 6 это определенные, назначенные к исследованию сорта картофеля, определенные виды удобрений. В других экспериментах выбор градаций факторов носит случайный характер. Так, в примере 5 градациями фактора индивидуальных свойств животных являются 5 собак, представляющие собой выборку из популяции, произведенную случайным образом. Так же случайны 5 блоков делянок в примере 7.

Модель, в которой градации факторов заранее фиксированы, называются *моделью* I. Модель со случайными градациями факторов называется *моделью* II. Если же один фактор имеет фиксированные, а другой — случайные градации, то модель называют *смешанной*. Очевидно, смешанной моделью описываются эксперименты из примеров 5 (5 собак — случайные градации, 6 времен суток — фиксированные градации) и 7 (4 сорта ржи — фиксированные градации, 5 блоков делянок — случайные градации). Опыт из примера 6 описывается моделью I (4 сорта картофеля и 5 видов удобрения).

Если изучаемый комплекс описывается моделью I (фиксированные градации), то дисперсии s_A^2, s_B^2 и s_{AB}^2 взаимно независимы. Тогда значимость факторов A, B и AB проверяется по отношениям:

$$F_{A/Z} = \frac{s_A^2}{s_Z^2}; \quad F_{B/Z} = \frac{s_B^2}{s_Z^2}; \quad F_{AB/Z} = \frac{s_{AB}^2}{s_Z^2}. \qquad (5.29)$$

Если же правильна модель II (случайные градации факторов), то s_A^2 и s_B^2 содержат в себе часть, зависящую от AB. «Чистое» влияние факторов A и B определяется в этом случае величинами $s_A^2 - s_{AB}^2$ и $s_B^2 - s_{AB}^2$. Очевидно, вычисление таких разностей имеет смысл лишь тогда, когда s_A^2 и s_B^2 значимо отличаются от s_{AB}^2. Поэтому ясно, что условием значимости влияния факторов A и B

является значимость отношений

$$F_{A/AB} = \frac{s_A^2}{s_{AB}^2} \quad \text{и} \quad F_{B/AB} = \frac{s_B^2}{s_{AB}^2}. \tag{5.30}$$

При этом, конечно, само s_{AB}^2 должно значимо отличаться от s_Z^2. Если это не имеет места, то s_{AB}^2 может считаться оценкой случайной дисперсии σ_Z^2 наравне с s_Z^2. Тогда в качестве наилучшей оценки величины σ_Z^2 принимают среднее из s_{AB}^2 и s_Z^2, т. е. величину

$$s_{Z^*}^2 = \frac{f_{AB}\, s_{AB}^2 + f_Z\, s_Z^2}{f_{AB} + f_Z} = \frac{S_{AB} + S_Z}{f_{AB} + f_Z}, \tag{5.31}$$

с которой затем сравнивают дисперсии s_A^2 и s_B^2.

В случае смешанной модели значимость фактора с фиксированными градациями проверяется сравнением соответствующей дисперсии с s_Z^2, а значимость фактора со случайными градациями — сравнением с s_{AB}^2.

В нашем примере справедлива модель I. Поэтому вычисляем отношения (5.29). Результаты приведены в табл. 64. Мы видим,

Таблица 64

Разброс	S	f		F	F_α	
					5%	1%
По сортам (A) . .	30,33	2	15,16	12,68	3,26	5,25
По удобрениям (B) . .	15,08	3	5,03	4,09	2,86	4,38
По взаимодействию (AB)	18,34	6	3,06	2,49	2,36	3,35
Случайный (Z)	44,25	36	1,23			
Полный	108,00					

что $F_{AB/Z}$ лишь немного превышает F_{05}, так что взаимодействие AB можно считать незначимым. Разброс по сортам почти достоверен, ибо $F_{A/Z}$ намного больше, чем F_{01}. Что касается разброса по удобрениям, то он весьма вероятен, так как $F_{B/Z}$ гораздо ближе к F_{01}, чем к F_{05}.

§ 8. Многофакторный дисперсионный анализ

Выше были рассмотрены три схемы дисперсионного анализа для последовательно усложняющихся комплексов — однофакторного, двухфакторного без повторности и двухфакторного с повторностью.

В практике обработки биологических данных могут встретиться еще более сложные случаи. Усложнение двухфакторного анализа может быть двояким: а) неравномерный комплекс и б) большое число вариант; может иметь место и то и другое. Во всех этих случаях дисперсионный анализ становится весьма громоздким; мы не будем здесь разбирать эти схемы, а отошлем читателя к специальной литературе [1].

Наконец, может возникнуть необходимость провести трехфакторный, четырехфакторный и вообще многофакторный дисперсионный анализ. Не входя в детали, укажем лишь, что во всех этих случаях дело сводится к последовательному применению двухфакторного анализа. Например, при анализе по трем факторам A, B и C численности расположатся в виде трехмерной таблицы — в кубических ячейках параллелепипеда со сторонами n_A, n_B, n_C. Процедура состоит в том, что сначала это трехмерное распределение проектируется на одну из граней, например AB; при этом в каждую клетку вписывается сумма численностей, имеющих одинаковые «координаты» a и b:

$$x_{ab} = \sum_{c=1}^{n_C} x_{abc}.$$

Теперь, произведя двухфакторный анализ, находят S_A, S_B и S_{AB}. Затем распределение проектируется на другую грань, например AC, и после нового двухфакторного анализа получают, кроме S_A, еще S_C и S_{AC}.

Таким же образом, проектируя на грань BC, находят, кроме найденных ранее S_B и S_C, также S_{BC} [2].

При наличии трех факторов могут, в принципе, иметь место эффекты взаимодействия второго порядка — зависимость эффекта AB от уровня C, эффекта AC от уровня B и эффекта BC от уровня A. Однако такие эффекты второго порядка обычно очень малы и ими можно пренебречь (т. е. не выделять их из S_Z).

Величина S_Z вычисляется из равенства

$$S = S_A + S_B + S_C + S_{AB} + S_{BC} + S_{AC} + S_Z \qquad (5.32)$$

(эффекты второго порядка включены в S_Z), являющегося непосредственным обобщением (5.25). Число степеней свободы для

---

[1] См. монографию Н. А. Плохинского (1960).

[2] Впрочем, все это — лишь изложение принципа расчета. Для практического выполнения пользуются более простой вычислительной схемой, которую можно найти в указанной выше книге Н. А. Плохинского (1960).

каждой суммы находится обычным образом:

$$\left.\begin{gathered} f = n_A n_B n_C - 1;\ f_A = n_A - 1;\ f_B = n_B - 1;\ f_C = n_C - 1; \\ f_{AB} = f_A f_B;\ f_{BC} = f_B f_C;\ f_{AC} = f_A f_C, \end{gathered}\right\} \quad (5.33)$$

причем f_Z находится из

$$f = f_A + f_B + f_C + f_{AB} + f_{BC} + f_{AC} + f_Z. \quad (5.34)$$

Многофакторный анализ значительно упрощается в том частном случае, когда каждый фактор имеет только два уровня (например, наличие и отсутствие влияния). Сокращение вычислений происходит в этом случае не просто из-за того, что имеется меньше строк и столбцов и вообще меньше вариант. Основная причина состоит в том, что для указанного специального случая можно построить совершенно другую, гораздо более простую схему расчетов.

Чтобы пояснить эту схему, рассмотрим сначала бесповторный двухфакторный анализ, для которого в § 5 был дан общий метод. Пусть мы имеем варианты, записанные в табл. 65 (это могут быть урожаи на делянках, в почву которых вносились или не вносились калийные и фосфатные удобрения; время образования рефлекса у животных, подвергавшихся или не подвергавшихся вибрации и облучению и т. д.). В табл. 66 эти варианты записаны в общем виде. Удобно обозначать два уровня фактора A через 1 и a, а два уровня фактора B — через 1 и b, как это сделано в табл. 67.

Таблица 65

A \ B	1	2
1	10	25
2	20	30

Таблица 66

A \ B	1	2
1	x_{11}	x_{12}
2	x_{21}	x_{22}

Таблица 67

A \ B	1	b
1	(1)	(b)
a	(a)	(ab)

Тогда для каждой из вариант можно будет принять обозначение, представляющее собой как бы произведение названий строки и столбца, на пересечении которых стоит эта варианта. Так, в нашем случае варианта $x_{12} = 25$ получит обозначение b (как бы $1 \cdot b$). Символы в клетках табл. 67 взяты в скобки, чтобы показать, что там стоят не сами числа, а лишь обозначения этих чисел.

Таблица 68

	(0)	(1)	(2)
1	x_{11}	$x_{11}+x_{21}$	$x_{11}+x_{21}+x_{12}+x_{22}$
a	x_{21}	$x_{12}+x_{22}$	$x_{21}-x_{11}+x_{22}-x_{12}$
b	x_{12}	$x_{21}-x_{11}$	$x_{12}+x_{22}-x_{11}-x_{21}$
ab	x_{22}	$x_{22}-x_{12}$	$x_{22}-x_{12}-x_{21}+x_{11}$

Таблица 69

	(0)	(1)	(2)	t
1	10	30	85	
a	20	55	15	3,0
b	25	10	25	5,0
ab	30	5	−5	

Упрощенная схема дисперсионного анализа представлена в табл. 68 и 69, где расчеты приведены параллельно в общем виде и в числах. В левом столбце перечислены способы воздействия: отсутствие всякого воздействия (строка 1), действие одного фактора A (строка a), действие одного фактора B (строка b) и совместное действие факторов A и B (строка ab). В столбце (0) записаны соответствующие варианты, взятые из табл. 67 и 65. Столбец (1) заполняется так: в первой строке пишется сумма двух первых чисел из столбца (0), во второй строке — сумма второй пары чисел из столбца (0), в третьей и четвертой строках столбца (1) пишутся вместо сумм соответствующие разности, причем каждый раз верхнее число вычитается из нижнего. Числа столбца (2) получаются точно таким же образом из чисел столбца (1). Теперь оказывается, что квадраты чисел строк a, b и ab из столбца (2) равны соответственно $4S_a$, $4S_b$ и $4S_{ab}$, где S_a, S_b и S_{ab} вычислены по обычной схеме из § 5[1].

[1] Как всегда в бесповторном комплексе, считается, что s^2_{AB} играет роль случайной дисперсии s^2_Z, с которой надо сравнивать факториальные дисперсии s^2_A и s^2_B.

Например, формулы (5.19) и (5.20) дают (с учетом того, что в данном случае $n_A = n_B = 2$):

$$4S_A = 4(S_a - S_x) =$$

$$= 4\left\{\frac{1}{2}[(x_{11}+x_{12})^2 + (x_{21}+x_{22})^2] - \frac{1}{4}(x_{11}+x_{12}+x_{21}+x_{22})^2\right\}.$$

Но, как легко проверить, развертывание этого выражения дает тот же результат, что и развертывание выражения $(x_{21} - x_{11} + x_{22} - x_{12})^2$ из строки a табл. 68. Так как в рассматриваемом случае $f_A = f_B = f_Z = 1$, то квадраты последних трех чисел из столбца (2) представляют собой одновременно $4s_A^2$, $4s_B^2$ и $4s_Z^2$, а сами эти числа, следовательно, равны $2s_A$, $2s_B$ и $2s_Z$. Значимость факторов A и B определяется по F-критерию, т. е. сравнением величин $F_{A/Z} = s_A^2/s_Z^2$ и $F_{B/Z} = s_B^2/s_Z^2$ с F_α. Но из табл. XIV и IV Приложений видно, что если число степеней свободы числителя равно единице, то $F_\alpha(f_1; f_2) = t_\alpha^2(f_2)$; например, F_{05} (1;18) = 4,41; t_{05} (18) = 2,10, причем 4,41 = (2,10)^2.

Так как в нашем случае $f_A = 1$ и $f_B = 1$, то критерий $s_A^2/s_Z^2 > F_\alpha$ можно заменить критерием $s_A/s_Z > t_\alpha$ или $2s_A/2s_Z > t_\alpha$, т. е. нужно просто разделить числа из строк a и b столбца (2) на число из строки ab этого же столбца, пренебрегая знаками. В данном примере получаем $t_{A/Z} = 15 \quad 5 = 3{,}0$; $t_{B/Z} = 25 \quad 5 = 5{,}0$ (числа, записанные в последнем столбце табл. 69). Теперь остается только сравнить эти числа с $t_\alpha(1)$, и на этом дисперсионный анализ заканчивается.

Переход к трехфакторному и вообще многофакторному анализу не вносит никаких принципиальных усложнений в эту схему. Просто в таблицу типа табл. 69 добавляется соответствующее число столбцов, заполнение которых производится точно так же, как заполнение столбцов (1) и (2) в табл. 69.

Пример 9. В табл. 70 приведены данные, относящиеся к предпосевному облучению семян моркови двух сортов, высеянных на почве двух типов; числа означают урожаи в килограммах

Таблица 70

	Кислая почва		Известкованная почва	
	необлученная	облученная	необлученная	облученная
Сорт Д . .	8	7	20	23
Сорт П .	14	16	13	17

(округленно) на делянке. Здесь мы имеем три фактора: сорт (фактор A, градации 1 и a), тип почвы (фактор B, градации 1 и b) и облучение (фактор C, градации 1 и c).

Непосредственное рассмотрение таблицы позволяет сделать следующие заключения: 1) фактор сорта, по-видимому, отсутствует, так как суммарные урожаи для обоих сортов примерно одинаковы ($8+7+20+23=58$, $14+16+13+17=60$); 2) тип почвы влияет на урожай ($8+7+14+16=45$, $20+23+13+17=73$); 3) поскольку сорт $Д$ дает явно меньший урожай на кислой почве, а сорт $П$ безразличен к типу почвы, то фактор взаимодействия AB довольно велик; 4) облучение семян дает почти всегда некоторую прибавку урожая. Однако вопрос о значимости этих факторов может решить только количественный анализ.

Таблица 71

	(0)	(1)	(2)	(3)	t
1	8	22	55	118	
a	14	33	63	2	1,0
b	20	23	—1	28	14,0
ab	13	40	3	—28	14,0
c	7	6	11	8	4,0
ac	16	—7	17	4	2,0
bc	23	9	—13	6	3,0
abc	17	—6	—15	—2	

В данном случае табл. 71 факторного анализа будет содержать 8 строк с наименованиями $1\cdot1\cdot1=1$, $a\cdot1\cdot1=a$, $1\cdot b\cdot1=b$, $a\cdot b\cdot1=ab$, $1\cdot1\cdot c=c$, $a\cdot1\cdot c=ac$, $1\cdot b\cdot c=bc$, $a\cdot b\cdot c=abc$. (Читатель должен обратить внимание на то, что c стоит п о с л е ab: сначала пишется все, что относится к a и b, и лишь потом добавляется новый фактор c и все его комбинации с записанными ранее факторами.)

В столбце (0) записаны числа, соответствующие указанным комбинациям факторов. Верхняя половина столбца (1), т. е. первые его четыре строки, содержит попарные суммы $1+a$, $b+ab$, $c+ac$, $bc+abc$, а нижняя половина (последние четыре строки) — соответствующие разности $a-1$, $ab-b$, $ac-c$, $abc-bc$ (напоминаем, что всегда верхнее число пары вычитается из нижнего). Аналогично заполняются остальные столбцы (2) и (3).

Последний столбец содержит значения t. Они получаются делением (без учета знака) на число из столбца (3), стоящее в строке abc (в данном случае это —2), всех остальных чисел из столбца (3). Исключением является число, стоящее в строке 1 (в нашем примере 118), которое есть просто сумма вариант комплекса, т. е. сумма всех чисел из столбца (0); кстати, несовпадение суммы чисел столбца (0) с числом на пересечении строки 1 и столбца (3) указывает, что в расчете допущена ошибка.

Если каждый фактор имеет m уровней, то переход от двухфакторного комплекса к трехфакторному означает переход от таблицы вида $m \times m$ к таблице вида $m \times m \times m$. Таблицу первого вида можно условно обозначить m^2, а таблицу второго вида — m^3; при k факторах таблица будет обозначаться m^k. В частности, в примере 9 мы имели таблицу 2^3.

В бесповторном комплексе типа 2^k «случайная» дисперсия всегда имеет одну степень свободы: для ab при $k = 2$, для abc при $k = 3$ и т. д. Это значит, что получившиеся значения t надо сравнивать с $t_\alpha(1)$. Но из табл. IV Приложений видно, что значения $t_\alpha(1)$ очень велики: $t_{05}(1) = 12{,}71$ и $t_{01}(1) = 63{,}66$; поэтому условия для получения значимого результата $t > t_\alpha$ очень неблагоприятны. Так, в примере 9 будут значимы, да и то только на 5%-ном уровне, лишь факторы B (тип почвы) и AB (взаимодействие сорта и типа почвы).

Но положение радикально меняется, если опыт был хотя бы двукратным. Это хорошо видно из табл. 72. При наличии повтор-

Таблица 72

Число факторов k	Число повторностей r	Общая численность комплекса $r \cdot 2^k$	Полное число степеней свободы $r \cdot 2^k - 1$	Учитываемые факторы	Число учитываемых факторов	Число степеней свободы различия повторностей $r-1$	Число степеней свободы остаточной дисперсии
2	1	$2^2=4$	4—1=3	a, b	2	0	3—2=1
	2	$2 \cdot 2^2=8$	8—1=7	a, b, ab	3	1	7—(3+1)=3
3	1	$2^3=8$	8—1=7	a, b, c, ab, ac, bc	6	0	7—6=1
	2	$2 \cdot 2^3=16$	16—1=15	a, b, c ab, ac, bc, abc	7	1	15—(7+1)=7

ности могут быть выделены комбинации ab в случае 2^2 и abc в случае 2^3 (о том, как при этом вычисляются остаточная дисперсия и вообще все дисперсии, будет сказано ниже). И тем не менее, число степеней свободы остаточной дисперсии резко возрастает. Так, для комплекса 2^3 уже при повторности $r = 2$ получается $f_Z = 7$, поэтому имеем $t_{05}(7) = 2{,}37$ и $t_{01}(7) = 3{,}50$, т. е. значения, во много раз меньшие, чем прежде.

Пример 10. Присоединим к данным из табл. 70 результаты, получившиеся при повторении всего опыта (табл. 73). Теперь весь комплекс состоит из двух блоков (или *реплик*, как их

Таблица 73

	Кислая почва		Известкованная почва	
	необлученная	облученная	необлученная	облученная
Сорт *Д*	6	9	19	20
Сорт *П*	10	17	11	15

часто называют) по 8 делянок в каждом. Общее число вариант равно $2 \cdot 2^3 = 16$. Данные сводим в табл. 74. Столбцы (1), (2) и (3) заполняются так же, как в табл. 71, но на основе факторных значений, просуммированных по обоим блокам (т. е. на основании

Таблица 74

Факторы	Блок 1		Блок 2		Сумма по блокам		(1)	(2)	(3)	t
	x_1^2	x_1	x_2^2	x_2	x^2	x				
1	64	8	36	6	196	14	38	101	225	
a	196	14	100	10	576	24	63	124	1	<1
b	400	20	361	19	1521	39	49	—5	51	7,78
ab	169	13	121	11	576	24	75	6	—53	8,08
c	49	7	81	9	256	16	10	25	23	3,51
ac	256	16	289	17	1089	33	—15	26	11	1,68
bc	529	23	400	20	1849	43	17	—25	1	<1
abc	289	17	225	15	1024	32	—11	—28	—3	<1
Сумма	1952	118	1613	107	7087	225				

чисел 8+6=14, 14+10=24 и т. д.). Остальные столбцы нужны для вычисления случайной вариации. Она получается как остаток после вычитания из полной вариации S двух вариаций — между блоками $S_{бл}$ и между факторами $S_ф$. Поскольку все отклонения отсчитываются от общего среднего для комплекса, то при вычислении всех вариаций надо будет из сумм квадратов вычитать корректирующий фактор

$$S_x = \frac{1}{16} X^2 = \frac{1}{16} 225^2 = 3164{,}06,$$

где 16 — число всех вариант в комплексе, а $X = 225$ — их сумма. Последняя записана внизу столбца x; она же получается суммированием чисел 118 и 107 из итоговой строки столбцов x_1 и x_2.

Теперь находим:

1) полную вариацию, т. е. полную сумму квадратов отклонений от среднего значения,

$$S = (8^2 + 14^2 + \ldots + 17^2) + (6^2 + 10^2 + \quad + 15^2) - 3164{,}06,$$

или просто

$$S = 1952 + 1613 - 3164{,}06 = 400{,}94,$$

так как суммы квадратов вариант для каждого из блоков указаны в итоговой строке столбцов x_1^2 и x_2^2;

2) вариацию для блоков

$$S_{бл} = \frac{1}{8} (118^2 + 107^2) - 3164{,}06 = 7{,}56,$$

где 118 и 107 — суммы вариант в каждом из блоков (итоговая строка столбцов x_1 и x_2), а 8 есть число делянок в блоке, т. е. число комбинаций факторов;

3) вариацию для факторов

$$S_ф = \frac{1}{2} (14^2 + 24^2 + \quad + 32^2) - 3164{,}06$$

или просто

$$S_ф = \frac{1}{3} 7087 - 3164{,}06 = 379{,}44,$$

так как сумма квадратов факторных сумм записана в итоговой строке столбца x^2.

Результаты этой стадии анализа записываем в виде табл. 75. Как и следовало ожидать, различие по факторам оказалось

Таблица 75

Разброс	S	f		F	F_{05}	F_{01}
По блокам .	7,56	1	7,56	3,80	5,59	
По факторам .	379,44	7	54,71	27,50		7,00
Остаточный .	13,94	7	1,99			
Полный . .	400,94	15				

значимым ($27{,}50 > 7{,}00$). Что касается различия между блоками, то оно здесь незначимо: $3{,}80 < F_{05}$. Это позволяет считать $s^2_{бл} = 7{,}56$ оценкой случайной дисперсии наряду с $s^2_{ост} = 1{,}99$, так что в качестве наилучшей оценки σ^2_Z примем величину

$$s^2_Z = \frac{7{,}56 + 13{,}94}{1 + 7} = 2{,}69$$

с 8 степенями свободы. Эту оценку используем теперь для вычисления значений t по факторам и их комбинациям. При этом надо учесть, что числа, стоящие в столбце (3) табл. 74 (за исключением первого числа 225, которое есть просто сумма всех вариант), представляют собой значения $\sqrt{16\ s^2}$, где 16 есть общее число вариант в комплексе. Поэтому относить их надо к числу

$$\sqrt{16 s^2_Z} = 4\sqrt{2{,}69} = 6{,}56.$$

Полученные отношения записаны в столбце t.

Поскольку $t_{05}\,(8) = 2{,}31$ и $t_{01}\,(8) = 3{,}36$, то факторы B (тип почвы) и C (облучение), а также комбинация AB (взаимодействие сорта и типа почвы) значимы на 1%-ном уровне. Это в общем согласуется с теми качественными выводами, которые были сделаны выше. Остальные факторы оказались незначимыми.

При увеличении числа факторов размеры блоков (реплик) быстро возрастают. Однако имеются приемы, позволяющие обходиться меньшими репликами за счет отказа от выявления взаимодействий высших порядков [1].

[1] См. книги Н. Бейли (1962) и Дж. У. Снедекора (1961).

В заключение сделаем замечание общего характера. Изложенные в настоящей главе методы показывают, что не обязательно ставить опыты так, чтобы каждому уровню одного какого-нибудь фактора соответствовала отдельная серия наблюдений при постоянном уровне других факторов. Статистический анализ позволяет извлечь необходимую информацию и при такой постановке опытов, когда одновременно варьируют значения нескольких факторов. Как правило, при этом требуется выполнить меньшее число опытов; кроме того, удается оценить эффект взаимодействия факторов.

Эти преимущества приобретают особенно большое значение при испытании новых сортов сельскохозяйственных культур, когда необходимо получить информацию не только о сравнительных достоинствах того или иного сорта, но и о методах обработки почвы, нормах посева, количестве и видах удобрений и т. д. Постановка отдельных опытов для решения каждого из этих вопросов была бы чрезвычайно трудоемкой. Кроме того, не было бы учтено такое важное обстоятельство, что один сорт, лучший при одних условиях, может при других условиях оказаться худшим; данное удобрение на одной почве дает положительный эффект, а на другой — отрицательный и т. д.

КРИТЕРИЙ РАЗЛИЧИЯ «ХИ-КВАДРАТ»

§ 1. Сравнение эмпирического распределения с теоретическим

В главе 4 были указаны критерии, позволяющие установить совпадение или различие между параметрами двух распределений — их средними значениями, дисперсиями, коэффициентами асимметрии и т. д. По этой причине, как уже говорилось, эти критерии называются *параметрическими*.

При пользовании параметрическими критериями 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 относится к определенному теоретическому виду, либо гипотеза о том, что два эмпирических распределения относятся к одному и тому же виду.

В частном случае, когда нужно проверить нормальность заданного эмпирического распределения, можно обойтись при помощи параметрических критериев, как это было показано в § 2 гл. 4. Однако в общем случае такой способ проверки, основанный на сопоставлении параметров, оказывается слишком громоздким. Более просто эту задачу решают критерии различия, в основе которых лежит сравнение частот распределения. Одним из таких критериев является «*критерий хи-квадрат*» (К. Пирсон).

Прежде чем переходить к описанию этого критерия, 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 выборочными и генеральными распределениями (поскольку при данном объеме выборки вероятность такого или даже еще большего расхождения между последними не так уж мала).

Пример 1. Рассмотрим распределение длины окружности груди у студентов, представленное в табл. 76.

Для этого распределения $\bar{x} = 87{,}89$ *см* и $s = 4{,}56$ *см*. По общему виду данное статистическое распределение похоже на нормальное, но возникает вопрос о том, можно ли имеющееся

Таблица 76

x_i	<79	79	82	85	88	91	94	97	100	103	>103	Сумма
n_i	3	19	63	104	138	101	43	22	4	2	1	500

расхождение между данным распределением и нормальным считать результатом случайности образования выборки.

Указанный ранее способ решения этого вопроса сводится к тому, что надо было вычислить коэффициенты асимметрии A и эксцесса E данного распределения и сравнить полученные значения с σ_A и σ_E по критериям (4.3). Но можно поступить иначе — вычислить частоты нормального распределения с параметрами $\hat{x} = 87{,}89$ и $\sigma = 4{,}56$ для значений 79, 82, 85, . и сравнить эти вычисленные частоты с эмпирическими.

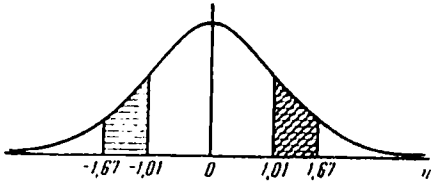

Рис. 42

Вычисление теоретических частот произведем следующим образом. Эмпирические частоты, соответствующие значениям 79, 82, 85 и т. д., суть сгруппированные частоты для разрядов с границами 80,5; 83,5; 86,5 и т. д., центрами которых являются значения 79, 82, 85 и т. д. Поэтому, чтобы вычислить теоретическую частоту, соответствующую, например, значению 94, мы должны найти число вариант, содержащихся в границах от 92,5 до 95,5. Но 92,5 отклоняется от $\hat{x} = 87{,}89$ (в единицах σ) на

$$u_1 = \frac{x_1 - \hat{x}}{\sigma} = \frac{92{,}5 - 87{,}89}{4{,}56} = \frac{4{,}61}{4{,}56} = 1{,}01,$$

а 95,5 — на

$$u_2 = \frac{x_2 - \hat{x}}{\sigma} = \frac{95{,}5 - 87{,}89}{4{,}56} = \frac{7{,}61}{4{,}56} = 1{,}67;$$

поэтому, как видно из рис. 42, значению 94 будет соответствовать теоретическая частота

$$\hat{n}_{x=94} = \frac{1}{2}\,[\Phi(1{,}67) - \Phi(1{,}01)]\,n.$$

Из табл. I Приложений находим, что $\Phi(1{,}67) = 0{,}905$, $\Phi(1{,}01) =$

$= 0{,}688$, так что в интервале (92,5 — 95,5) содержится

$$\hat{n}_{x=94} = (0{,}905 - 0{,}688)\,250 = 54{,}3,$$

или округленно 54 варианты.

Аналогично вычисляем теоретические частоты для других значений x_i. В табл. 77 этот расчет показан полностью.

Крайние числа в столбце 6 представляют собой 1,000 минус соответствующие крайние числа столбца 5. Среднее число столбца 6, т. е. 0,510, есть сумма (0,236 + 0,274), так как числа 0,236 и 0,274 относятся к отклонениям в разные стороны от $\bar{x}$.

Таблица 77

Середины разрядов	Границы разрядов	Отклонения границ разрядов от $\bar{x}$	Модули нормированных отклонений границ разрядов u	$\theta(u)$	$\Delta\theta(u)$	$\hat{n}_i = \frac{n}{2}\Delta\theta(u)$	n_i
1	2	3	4	5	6	7	8
<79					0,023	5,8	3
	77,5	—10,39	2,28	0,977			
79					0,082	20,5	19
	80,5	— 7,39	1,62	0,895			
82					0,233	58,3	63
	83,5	— 4,39	0,96	0,663			
85					0,427	106,6	104
	86,5	— 1,39	0,30	0,236			
88					0,510	127,4	138
	89,5	1,61	0,35	0,274			
91					0,414	103,4	101
	92,5	4,61	1,01	0,688			
94					0,217	54,3	43
	95,5	7,61	1,67	0,905			
97					0,075	18,7	22
	98,5	10,61	2,33	0,980			
100					0,017	4,2	4
	101,5	13,61	2,99	0,997			
>100					0,003	0,8	3
Сумма					2,001	500,0	500

Полученные таким образом теоретические частоты $\hat{n}_i$ записаны в столбце 7. Мы видим, что хотя общий характер распределения в столбцах 7 и 8 одинаков, полного совпадения в значениях частот нет.

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 т. д., которые описываются статистически распределением Пуассона (см. § 6 гл. 2). Но для этого распределения $\mu_2 = \hat{x}$ и $\sigma = \sqrt{\mu_2} = \sqrt{\hat{x}}$. 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) в каждом разряде будет равно

$$\frac{n_i - \hat{n}_i}{\sqrt{\hat{n}_i}}.$$

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

$$\chi^2 = \sum_i \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \tag{6.1}$$

(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, мы будем условно считать, что нулевая гипотеза отвергается, если $\chi^2 > \chi^2_{01}$, и принимается, если $\chi^2 \leqslant \chi^2_{05}$.

Как и для других критериев, критические значения χ^2_α зависят от числа степеней свободы f. Таблица критических значений $\chi^2_\alpha(f)$ для уровней значимости 5 и 1% приведена в Приложениях (табл. XIX).

Было бы, однако, ошибкой думать, что чем меньшее 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 невольн-

ной). Например, при $f = 26$ величина χ^2 лишь в одном проценте случаев может превысить значение 54,1 (см. табл. XIX Приложений); но, как показывает расчет, точно также лишь в одном проценте случаев χ^2 может из-за случайных выборочных вариаций оказаться меньше, чем 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 сборнике таблиц Я. Янко (стр. 115).

Критерий χ^2 может привести к неправильному результату, если имеются очень малые теоретические частоты: слагаемые в (6.1), 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, в табл. 77 получилось, что при $x_i > 100$ теоретическая частота равна 0,8. Поэтому мы объединяем разряды «100» и «>100» в один разряд «>97» с общей теоретической частотой $\hat{n}_i = 5{,}0$, считая соответственно также $n_i = 4 + 2 + 1 = 7$.

Так как точность нахождения величины χ^2 сильно 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.

Вычисление величины χ^2 показано в табл. 78. Например,

$$\frac{(-2{,}8)^2}{5{,}8} = \frac{7{,}84}{5{,}8} = 1{,}35;$$

$$\frac{(-1{,}5)^2}{20{,}5} = \frac{2{,}25}{20{,}5} = 0{,}11$$

и т. д. Сумма этих слагаемых равна $\chi^2 = 6{,}57$.

Формулу (6. 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 $\Sigma n_i = \Sigma \hat{n}_i = n$, так что

$$\chi^2 = \sum_i \frac{n_i^2}{\hat{n}_i} - n. \qquad (6.2)$$

Это делает ненужным вычисление разностей $n_i - \hat{n}_i$.

Таблица 78

x_i	n_i		$n_i - \hat{n}_i$	$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
1	2	3	4	5
<79	3	5,8	— 2,8	1,35
79	19	20,5	— 1,5	0,11
82	63	58,3	4,7	0,38
85	104	106,6	— 2,6	0,06
88	138	127,4	10,6	0,88
91	101	103,4	— 2,4	0,06
94	43	54,3	—11,3	2,35
97	22	18,7	3,3	0,58
>97	7	5,0	2,0	0,80
	500 $= n$	500,0 $= \hat{n}$	—20,6 +20,6	6,57 $= \chi^2$

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); так, в нашем примере $n = 500$, поэтому мы должны были бы иметь

$$\chi^2 = 506{,}57 - 500 = 6{,}57.$$

Это предъявляет повышенные требования к точности промежуточных вычислений и, в частности, делает невозможным использование логарифмической линейки. Но при наличии арифмометра этот способ вычисления χ^2 предпочтительней.

Теперь найдем число степеней свободы для нашего случая. Как мы знаем (см. § 3 гл. 3), число степеней свободы какой-либо статистической оценки равно числу независимых величин, использованных при вычислении этой оценки, т. е. общему числу таких величин минус число условий, связывающих эти величины. При вычислении χ^2 используются величины разрядных частот n_i, число которых равно k. Но эти частоты связаны здесь тремя условиями:

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

определяющими те значения характеристик (n, $\bar{x}$ и σ), по которым строилось теоретическое нормальное распределение (и вычисля-

лись теоретические частоты $\hat{n}_i$). Поэтому если в качестве теоретического берется нормальное распределение с неизвестными параметрами $\bar{x}$ и σ, то $f = k - 3$.

Замена числа разрядов k числом степеней свободы f имеет 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. В данном случае число степеней свободы равно $f = 9 - 3 = 6$. Из табл. XIX Приложений находим $\chi^2_{05}(6) = 12{,}6$. Так как $\chi^2 < \chi^2_{05}$, то различие между двумя распределениями следует считать незначимым.

Пример 2. Проверим при помощи χ^2-критерия, что распределение из табл. 26 незначимо отличается от распределения Пуассона.

Таблица 79

x	$\bar{x}^x$		$\frac{\bar{x}^x}{x!}$	$\hat{n}_x$
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 $\hat{n}_x$ показано в табл. 79, причем учтено, что $\bar{x} = 0{,}292$ и что

$$\hat{n}_x = n \frac{\bar{x}^x}{x!} \Big/ \sum_{x=0}^{4} \frac{\bar{x}^x}{x!},$$

так как здесь

$$\sum_{x=0}^{4} \frac{\bar{x}^x}{x!} \approx e^{\bar{x}}.$$

Расчет величины χ^2 показан в табл. 80; эмпирические частоты взяты из табл. 26.

Таблица 80

			$n_x - \hat{n}_x$	(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 четыре частоты связаны двумя условиями

$$\Sigma n_x = n; \quad \frac{1}{n}\Sigma n_x x = \bar{x},$$

так что $f = 4 - 2 = 2$. Из табл. XIX Приложений находим $\chi^2_{05} = 5{,}99$. Так как $\chi^2 < \chi^2_{05}$, то различие между эмпирическим распределением и теоретическим незначимо — распределение из табл. 26 можно считать пуассоновским.

Пример 3. Применим χ^2-критерий к данным из табл. 17 (§ 2 гл. 2). Расчет дан в табл. 81; малочисленные классы объединены.

Таблица 81

	n_x	$\hat{n}_x$	$n_x - \hat{n}_x$	(χ^2)
0—3	7	5	2	0,80
4	11	10	1	0,10
5	13	16	—3	0,56
6	19	18	1	0,56
7	16	16	0	0,00
8	7	10	—3	0,90
9—12	7	5	2	0,80
Сумма	80	80	0	3,72

Так как параметр $\hat{p}$ предполагаемого бипомиального распределения неизвестен и оценивается по данным выборки и, кроме того, задан ее объем, то здесь, как и в случае распределения Пуассона, $f = k - 2$. В нашем примере $k = 7$ и $f = 5$, так что

$\chi^2_{05} = 11{,}1$. Фактическое значение $\chi^2 = 3{,}72$ много меньше, поэтому гипотеза о биномиальном распределении не отвергается.

В рассмотренных случаях сравнение эмпирического и теоретического распределений по χ^2-критерию требовало большой вычислительной работы, связанной в основном с расчетом теоретических частот (особенно при нормальном распределении). Но если теоретические частоты известны или нахождение их не представляет большого труда, то применение χ^2-критерия упрощается.

Пример 4. В табл. 82 содержатся данные за десять 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 $\chi^2 = 15{,}23$. В данном случае частоты связаны единственным условием

$$\Sigma n_x = n,$$

так что $f = k - 1 = 10 - 1 = 9$. Поскольку $\chi^2_{05}(9) = 16{,}92$, то $\chi^2 < \chi^2_{05}$, поэтому гипотеза о постоянстве доли рождений троен не отвергается.

Таблица 82

Годы	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
$\chi^2_{05}(9) = 16{,}92$				

Пример 5. 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. 83. Расчет дает $\chi^2 = 18{,}29$ при $f = 4 - 1 = 3$ (4 численности связаны одним условием, что сумма равна 560).

Таблица 83

Признаки	n_i	$\hat{n}_i$	$n_i - \hat{n}_i$	(χ^2)
Плоские листья:				
нормальный глазок .	338	315	23	1,68
розовый » .	122	105	17	2,75
Сморщенные листья:				
нормальный глазок .	67	105	—38	13,75
розовый » .	33	35	— 2	0,11
Сумма .	560	560	0	18,29

По табл. XIX Приложений находим $\chi^2_{01}(3) = 11{,}3$; так как $\chi^2 > \chi^2_{01}$, то нулевая гипотеза опровергается.

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 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. 84. Здесь $f = 2 - 1 = 1$ и $\chi^2_{05}(1) = 3{,}84$. Мы видим, что $\chi^2 < \chi^2_{05}$, так что нулевая гипотеза (о том, что окраска глазка не влияет на жизнеспособность растения) не отвергается.

Таблица 84

Признаки		$\hat{n}_i$	$n_i - \hat{n}_i$	(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: что жизнеспособность 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. 85.

Таблица 85

Признаки	n_i	$\hat{n}_i$	$n_i - \hat{n}_i$	(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 $\chi^2_{01}(1) = 6{,}63$, так что нулевая гипотеза отвергается[1]. 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.

§ 2. Сравнение двух эмпирических распределений

Задача о сравнении двух эмпирических распределений (не просто их средних значений или других параметров, но всего хода кривых распределения) возникает обычно тогда, когда хотят проверить однородность эмпирического материала: если окажется, что две эмпирические совокупности распределены одинаково,

[1] Этот результат был получен в § 8 гл. 4 другим способом.

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 поразному 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 6. В табл. 86 (столбцы n_i' и n_i'') приведены два эмпирических распределения, в отношении которых предполагается, что они взяты из одной и той же генеральной совокупности.

Таблица 86

x_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 полусуммы эмпирических частот, относящихся к одинаковым значениям вариант (т. е. средние значения этих частот); это даст числа, записанные в четвертом столбце таблицы.

Теперь мы можем составить табл. 87. 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 87

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 распределениями, о чем упоминалось выше) сказывается в следующем. С одной стороны, в χ^2 входят не разности двух эмпирических частот, а п о л у р а з н о с т и (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 четыре раза меньше, чем при сравнении эмпирического и теоретического распределений. С другой стороны, 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 сравнива-

Таблица 88

x_i	n_i'	n_i''	Δn_i	$(\Delta n_i)^2/n_i'$
2	5	9	—4	3,20
4	6	4	+2	0,67
6	8	10	—2	0,50
8	10	6	+4	1,60
10	6	6	0	0,00
				5,97

ются не эмпирическое и теоретическое, а два эмпирических распределения. Если бы, например, в табл. 86 n_i' было теоретическим распределением и n_i'' — эмпирическим, то мы бы имели $\chi^2 = 5{,}97$ (табл. 88).

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 после несложных преобразований получим

$$\frac{(n_i' - n_i'')^2}{2(n_i' + n_i'')} + \frac{(n_i'' - n_i')^2}{2(n_i' + n_i'')} = \frac{(n_i' - n_i'')^2}{n_i' + n_i''}.$$

Поэтому

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i' - n_i'')^2}{n_i' + n_i''}. \qquad (6.3)$$

Расчет χ^2 по этой формуле для примера 6 показан в табл. 89.

Таблица 89

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) столбца 4 в табл. 86 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. XIX Приложений находим $\chi^2_{05}(4) = 9{,}49$. Так как $\chi^2 < \chi^2_{05}$, то нулевая гипотеза не опровергается — обе эмпирические совокупности можно считать выборками из одной генеральной совокупности.

§ 3. Сравнение выборок разного объема

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 7. В табл. 90 представлены 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 90

	n_i'	n_i''	$n_i' + n_i''$		$\hat{n}_i''$
(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. 91), выписывая рядом эмпирические и «теоретические» частоты, а затем обычным образом вычисляем χ^2.

Таблица 91

n_i	$\hat{n}_i$	$n_i - \hat{n}_i$	(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

Как и в случае равных объемов, здесь тоже можно получить упрощенную формулу для вычисления χ^2. Подставив в выражение (*) предыдущего параграфа значения $\hat{n}'_i$ и $\hat{n}''_i$, которые имеют вид

$$\hat{n}'_i = n' \frac{n'_i + n''_i}{n' + n''}, \quad \hat{n}''_i = n'' \frac{n'_i + n''_i}{n' + n''},$$

получим после простых преобразований формулу

$$\chi^2 = \frac{1}{n'n''} \sum_{i=1}^{k} \frac{(n'_i n'' - n''_i n')^2}{n'_i + n''_i}; \qquad (6.4)$$

очевидно, при $n' = n''$ эта формула переходит в (6.3).

В табл. 92 показано вычисление χ^2 по формуле (6.4) для примера 7.

Таблица 92

n'_i	n''_i	$n'_i + n''_i$	$n'_i n''$	$n''_i n'$	$n'_i n'' - n''_i n'$	$\frac{(n'_i n'' - n''_i n')^2}{n'_i + n''_i}$
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. 90.

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. 29, рассматривавшейся в § 5 гл. 3 (эти данные повторены в табл. 93).

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

Таблица 93

Диаметр эритроцита, мк	1-й мазок	2-й мазок	3-й мазок	4-й 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
Сумма	392	364	345	448	1549

из того, что поскольку она больше по объему, то она репрезентативней. 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 к ее общей численности 1549:

$$\frac{\hat{n}_9^{(1)}}{392} = \frac{333}{1549}, \text{ откуда } \hat{n}_9^{(1)} = 392 \cdot \frac{333}{1549} = 84.$$

Аналогично находим все остальные $n_i^{(j)}$; в общем виде

$$\hat{n}_i^{(j)} = n^{(j)} \frac{n_i}{\Sigma n_i}. \tag{6.5}$$

Теперь составляем табл. 94. В каждой клетке записано три числа. Первое есть эмпирическая численность $n_i^{(j)}$; второе — «теоретическая» численность $\hat{n}_i^{(j)}$, найденная по формуле (6.5); третье число — разность частот $n_i^{(j)} - \hat{n}_i^{(j)}$.

Величину χ^2 вычисляем обычным образом по формуле (6.1):

$0^2 : 12 = 0{,}00$	$(-3)^2 : 12 = 0{,}75$	$0^2 : 11 = 0{,}00$	$3^2 : 14 = 0{,}64$
$4^2 : 50 = 0{,}32$	$2^2 : 46 = 0{,}09$	$(-12)^2 : 44 = 3{,}28$	$6^2 : 56 = 0{,}64$
$(-19)^2 : 202 = 1{,}79$	$17^2 : 187 = 1{,}54$	$14^2 : 177 = 1{,}11$	$(-12)^2 : 231 = 0{,}62$
$12^2 : 84 = 1{,}72$	$(-12)^2 : 78 = 1{,}85$	$0^2 : 74 = 0{,}00$	$0^2 : 97 = 0{,}00$
$1^2 : 30 = 0{,}03$	$(-4)^2 : 28 = 0{,}57$	$2^2 : 27 = 0{,}15$	$1^2 : 35 = 0{,}03$
$2^2 : 14 = 0{,}29$	$0^2 : 13 = 0{,}00$	$(-4)^2 : 12 = 1{,}33$	$2^2 : 15 = 0{,}27$
$4{,}15$	$4{,}80$	$5{,}87$	$2{,}20$

$$\chi^2 = 4{,}15 + 4{,}80 + 5{,}87 + 2{,}20 = 17{,}02.$$

Таблица 94

x_i	$n_i^{(1)}$	$n_i^{(2)}$	$n_i^{(3)}$	$n_i^{(4)}$	n_i
6	12	9	11	17	49
	12	12	11	14	
	0	—3	0	3	0
7	54	48	32	62	196
	50	46	44	56	
	4	2	—12	6	0
8	183	204	191	219	797
	202	187	177	231	
	—19	17	14	—12	0
9	96	66	74	97	333
	84	78	74	97	
	12	—12	0	0	0
10	31	24	29	36	120
	30	28	27	35	
	1	—4	2	1	0
11	16	13	8	17	54
	14	13	12	15	
	2	0	—4	2	0
Сумма	392	364	345	448	1549

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 однозначно из итоговой частоты в этой строке, а оставшаяся в каждом столбце — из итоговой частоты по столбцу. Поэтому число степеней свободы равно

$$f = (k-1)(m-1). \tag{6.6}$$

То же самое 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 столб-

цам (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 $\chi^2_{05}(15) = 25{,}0$, а фактически получилось $\chi^2 = 17{,}0$, то нулевая гипотеза не опровергается.

Когда количество классов группировки и число сравниваемых рядов велико, число степеней свободы (6.6) может оказаться больше предусмотренного в табл. XIX Приложений. В таких случаях применяют приближения, описанные в § 4 гл. 10.

§ 4. Сравнение двух альтернативных распределений

При сравнении двух альтернативных распределений расчеты сильно упрощаются. В этом случае таблица содержит 2 строки и 2 столбца (поэтому ее обычно называют таблицей 2×2), причем

$$n_1 - \hat{n}_1 = -(n_2 - \hat{n}_2) = -(n_3 - \hat{n}_3) = n_4 - \hat{n}_4,$$

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 (6.1) 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. 95. При пользовании этой формулой отпадает необходимость в вычислении «теоретических» частот.

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

$$\chi^2 = \frac{(|ad - bc| - n/2)^2}{(a + b)(c + d)(a + c)(b + d)} n. \qquad (6.7)$$

Таблица 95

	B_1	B_2	Сумма
A_1	a	b	$a \quad b$
A_2	c	d	$c + d$
Сумма	$a + c$	$b + d$	n

Число степеней свободы равно

$$(k_A - 1)(k_B - 1) = (2 - 1)(2 - 1) = 1,$$

поэтому в соответствии с табл. XIX Приложений

$$\chi^2_{05} = 3{,}84; \quad \chi^2_{01} = 6{,}63.$$

Пример 8. В примере 16 из гл. 4 разбирался опыт иммунизации телят от туберкулеза; результаты приведены снова в табл. 96. Для выяснения значимости действия иммунизации применим к этим данным χ^2-критерий.

Таблица 96

	Заболевшие	Незаболевшие	Сумма
С прививкой .	6	14	20
Без прививки .	16	3	19
Сумма . .	22	17	39

Расчет по формуле (6.7) 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 $\chi^2_{01} = 6{,}63$, поэтому результат надо считать значимым.

Как было указано в § 1 настоящей главы, 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:

$$P = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}. \qquad (6.8)$$

Если окажется, что $p < 0{,}01$, то нулевая гипотеза отвергается; при $p \geqslant 0{,}05$ нулевая гипотеза принимается.

Пример 9. Из десяти больных, которых лечили способом A, у трех состояние улучшилось, у семи осталось без изменения. Из семи больных, которых лечили способом $Б$, соответствующие числа составляют 5 и 2. Можно ли считать доказанным преимущество способа $Б$?

В данном случае таблица имеет вид табл. 97

Таблица 97

Наличие улучшения	Способ лечения		Сумма
	A	$Б$	
+	3	5	8
—	7	2	9
Сумма	10	7	17

Подставляем данные в формулу (6.8):

$$P = \frac{10!\,7!\,8!\,9!}{17!\,3!\,5!\,7!\,2!}.$$

При недостатке опыта в вычислениях полезно выписать все множители (произведя очевидное сокращение на 7!):

$$P = \frac{(1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7\cdot 8\cdot 9\cdot 10)\,(1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7\cdot 8)\,(1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7\cdot 8\cdot 9)}{(1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7\cdot 8\cdot 9\cdot 10\cdot 11\cdot 12\cdot 13\cdot 14\cdot 15\cdot 16\cdot 17)\,(1\cdot 2\cdot 3)\,(1\cdot 2\cdot 3\cdot 4\cdot 5)\,(1\cdot 2)}.$$

Теперь сразу видно, что большинство чисел сокращается, после чего остается

$$P = \frac{4 \cdot 7 \cdot 9}{11 \cdot 13 \cdot 17} = 0{,}104.$$

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 (6.8) в числителе, минус сумма логарифмов чисел, стоящих в знаменателе. Так, для примера 8 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	ln $n!$	n	ln $n!$	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	3,702	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 8 имеем (сократив числитель и знаменатель на 7!):

Числитель	Знаменатель	Разность
lg 10! = 6,560	lg17! = 14,551	— 17,709
lg 8! = 4,606	lg 3! = 0,778	+ 16,726
lg 9! = 5,560	lg 5! = 2,079	— 0,983
16,726	lg 2! = 0,301	
	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 $|\lg P| > 2{,}0$ и принимается при

$|\lg P| < 1{,}3$. В примере 8 получилось $|\lg P| = 0{,}983 < 1{,}3$, поэтому нулевая гипотеза не отвергается.

Сделаем еще расчет для примера 7:

Числитель	Знаменатель	Разность
$\lg 20! = 18{,}386$	$\lg 39! = 46{,}310$	$-74{,}206$
$\lg 19! = 17{,}085$	$\lg 6! = 2{,}857$	$+71{,}073$
$\lg 22! = 21{,}051$	$\lg 14! = 10{,}940$	$-3{,}133$
$\lg 17! = 14{,}551$	$\lg 16! = 13{,}321$	
$71{,}073$	$\lg 3! = 0{,}778$	
	$74{,}206$	

Так как $|\lg P| > 2{,}0$, то нулевая гипотеза отвергается.

Строго говоря, нулевая гипотеза должна отвергаться только тогда, когда окажется меньше, чем 0,01, вероятность наличного и еще больших отклонений от равенства распределений. Так, в примере 9 еще больше, чем наличное распределение

$$\begin{pmatrix} 3 & 5 \\ 7 & 2 \end{pmatrix}, \qquad (*)$$

отклоняются от равенства такие два распределения:

$$\begin{pmatrix} 2 & 6 \\ 8 & 1 \end{pmatrix} \text{ и } \begin{pmatrix} 1 & 7 \\ 9 & 0 \end{pmatrix}.$$

Они получены последовательным уменьшением самой малой частоты (т. е. 2), причем остальные три частоты меняются каждый раз так, чтобы суммы по строкам и столбцам оставались неизменными; иначе говоря, меньшие частоты [в распределении (*) это 3 и 2] каждый раз уменьшаются на единицу, а большие частоты (в данном случае 5 и 7) увеличиваются на единицу — до тех пор, пока одна из частот не станет равной нулю.

Применяя теперь формулу (6.8), получаем для вероятности наличного и еще больших отклонений

$$P = \frac{10!\,7!\,8!\,9!}{17!\,3!\,5!\,7!\,2!} + \frac{10!\,7!\,8!\,9!}{17!\,2!\,6!\,8!\,1!} + \frac{10!\,7!\,8!\,9!}{17!\,1!\,7!\,9!\,0!} =$$
$$= \frac{10!\,7!\,8!\,9!}{17!}\left(\frac{1}{3!\,5!\,7!\,2!} + \frac{1}{2!\,6!\,8!\,1!} + \frac{1}{1!\,7!\,9!\,0!}\right). \qquad (**)$$

Это обстоятельство усложняет расчет, но на практике к такому усложнению редко приходится прибегать. Дело в том, что каждое последующее слагаемое в выражении типа (**), как

правило, на порядок меньше предыдущего слагаемого; так, в данном случае

$$\frac{1}{2!\,6!\,8!\,1!} : \frac{1}{3!\,5!\,7!\,2!} = \frac{3!\,5!\,7!\,2!}{2!\,6!\,8!\,1!} = \frac{3\cdot 2}{6\cdot 8} = 0{,}125;$$

$$\frac{1}{1!\,7!\,9!\,0!} : \frac{1}{2!\,6!\,8!\,1!} = \frac{2!\,6!\,8!\,1!}{1!\,7!\,9!\,0!} = \frac{2}{7\cdot 9} \approx 0{,}032;$$

поэтому поправки могут играть заметную роль только тогда, когда P по (6.8) окажется лишь немного меньше уровня значимости α [например, если бы в примере 9 оказалось $P = 0{,}00920$, то введение поправки дало бы $P \approx 0{,}00920\ (1 + 0{,}125) = 0{,}01035$, что уже превышало бы $\alpha = 0{,}01$]. Кроме того, поправки заведомо ненужны, когда (6.8) дает $P \geqslant 0{,}05$: ведь они могут только увеличить значение вероятности отклонений, в то время как уже из (6.8) следует, что различие незначимо.

Имеются специальные таблицы [1], позволяющие проверять значимость различия для четырехклеточных комплексов без всяких вычислений (при сравнительно малых численностях).

Анализ четырехклеточных таблиц может быть использован для сравнения двух порядковых совокупностей. Идея состоит в том, что отыскивается медиана суммарной совокупности (т. е. получившейся сведением обоих заданных рядов в один), после чего составляется табл. 98.

Таблица 98

	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

Далее применяется критерий χ^2 или критерий Фишера (в зависимости от объема n). Ввиду того что в основу классификации здесь кладется 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 значений крайних вариант.

[1] См. «Таблицы» В. С. Генеса (1964) и «Statistical Tables...», Fisher and Yates (1957).

Пример 10. Две группы мышей из двух генетических линий получили одинаковую смертельную дозу облучения. Табл. 99 показывает число дней между облучением и гибелью. Можно ли считать, что мыши одной из линий более радиочувствительны?

Таблица 99

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 2×2-таблица будет иметь вид табл. 100:

Таблица 100

	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. 100 оказывается на единицу меньше, чем сумма объемов выборок. Например, если в табл. 99 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. 101.

Таблица 101

	Ниже медианы	Выше медианы	Сумма
Первый ряд	4	1	5
Второй ряд	2	5	7
Сумма	6	6	12

К этой таблице и нужно применять критерий.

В заключение этой главы о критерии χ^2 еще раз подчеркнем, что в формулы для χ^2 должны подставляться только частоты, а не величины, получаемые измерением, взвешиванием, отсчетом 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 объемов выборок разности $n_i'-n_i''$ (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 предпоследнем разделе § 4 гл. 4.

НЕПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ РАЗЛИЧИЯ

§ 1. Назначение непараметрических критериев

В § 4 гл. 4 мы уже говорили, что одним из важнейших в биологических приложениях математической статистики является вопрос о значимости различия между двумя эмпирическими совокупностями; на языке статистики это вопрос о том, являются ли две заданные эмпирические совокупности выборками из одной генеральной совокупности или же из двух разных генеральных совокупностей.

Выше были разобраны два способа решения этой задачи — при помощи параметрического t-критерия Стьюдента (и связанного с ним F-критерия Фишера) и при помощи χ^2-критерия Пирсона. Однако далеко не все задачи такого рода, возникающие в биологических исследованиях, могут быть решены при помощи этих критериев. Как известно, t-критерий применим только тогда, когда распределение вариант в генеральной совокупности не очень отклоняется от нормального; кроме того, этот критерий не применим к совокупностям, варианты которых характеризуются условными рангами, а не точными численными значениями. Что касается χ^2-критерия, то применимость его ограничена совокупностями достаточно большого объема (не менее 20—30 вариант), причем отдельные разряды должны содержать не меньше 3—4 вариант.

Разбираемые в настоящей главе критерии приложимы как к численно определенным, так и к порядковым совокупностям; поэтому их иногда называют *порядковыми критериями*. Использование их не нуждается в каких-либо предположениях о характере распределения вариант в генеральной совокупности. Так как применение этих критериев не требует вычисления параметров распределений (средних значений, дисперсий и т. д.), то их еще называют *непараметрическими критериями*.

В задаче сравнения двух эмпирических совокупностей можно различать следующие частные случаи. Во-первых, можно поставить вопрос о том, различаются ли сравниваемые совокупности по своей *центральной тенденции* (в качестве характеристики которой может выступать среднее значение, медиана и т. д.).

Во-вторых, вопрос может состоять в том, различаются ли обе совокупности вообще в каком-нибудь отношении — по центральной тенденции, по рассеянию вариант и т. д. Для решения каждого из этих вопросов имеются отдельные критерии, которые будут рассмотрены ниже.

При выборе подходящего к своему случаю критерия исследователь должен учесть следующее обстоятельство. Критерии первой группы обнаружат различие между двумя совокупностями только тогда, когда это различие касается центральной тенденции; различие в рассеянии или каких-либо других характеристиках формы распределения игнорируется этими критериями, и, например, две совокупности с существенно различной дисперсией, но с одинаковым средним значением будут квалифицироваться критериями первой группы как неразличимые. С другой стороны, критерии второй группы позволяют обнаружить различие между двумя совокупностями, в чем бы ни заключалось это различие, но зато мы не будем знать, в чем именно различаются совокупности.

Наконец, отдельно должен быть рассмотрен случай, когда варианты из двух совокупностей попарно сопряжены.

Таким образом, мы должны иметь критерии трех видов для трех разных задач. Но прежде чем заняться конкретными критериями, необходимо остановиться еще на одном важном свойстве критериев различия: критерии, предназначенные для решения одной и той же задачи, могут иметь разную «чувствительность». Это свойство, называемое в математической статистике *мощностью* критерия, в некотором смысле аналогично разрешающей способности измерительных приборов. Как и в последнем случае, повышение разрешающей способности достигается ценой увеличения сложности метода. В соответствии с практическими целями применения критериев мы приводим здесь по два критерия для каждой из трех упомянутых выше задач. Один критерий — меньшей мощности и более простой; с него и следует начинать анализ. Если этот критерий опровергает нулевую гипотезу, то на этом анализ заканчивается. Если же нулевая гипотеза этим критерием не опровергается, то следует попытаться проверить ее при помощи более мощного, но зато более громоздкого критерия [1].

Как будет видно из дальнейшего, непараметрические критерии (даже более мощные) требуют, как правило, меньше вычислений чем параметрические. Поэтому в случае малых выборок целесообразно начинать анализ всегда с проверки по соответствующему непараметрическому критерию.

[1] Однако если значение характеристики, вычисленной для менее мощного критерия, оказалось очень далеким от критического значения, то мало надежды, что более мощный критерий опровергнет нулевую гипотезу.

§ 2. Критерий Вилкоксона

Перейдем к разбору отдельных критериев в соответствии с намеченной программой. Начинаем с более простого критерия для сравнения выборок по центральной тенденции — критерия Вилкоксона, который представляет собой модификацию критерия инверсий [1].

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; примером может служить ряд

$$x_1\ x_2\ x_3\ y_1\ y_2\ x_4\ y_3\ x_5\ x_6\ y_4\ y_5.$$

Если числа x_i и y_j представляют собой некоторые численные значения вариант, то различие обоих рядов по центральной тенденции будет определяться разностью средних значений:

$$\frac{x_1 + x_2 + \ldots + x_l}{n_x} \text{ и } \frac{y_1 + y_2 + \ldots + y_m}{n_y}. \qquad (*)$$

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

$$R_{x_1} + R_{x_2} + \ldots + R_{x_l} = T_x;$$
$$R_{y_1} + R_{y_2} + \ldots + R_{y_m} = T_y.$$

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

$$1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2},$$

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

[1] О последнем см. в книгах Б. Л. ван дер Вардена (1960) и Дж. У. Снедекора (1961). Этот критерий называют еще критерием Манна-Уитни, а также критерием Уайта.

определяет $\frac{T}{n_x}$, $T_y = \frac{n(n+1)}{2} - T_x$ и $\frac{T_y}{n_y}$, а тем самым и разность $\frac{T}{n_x} - \frac{T_y}{n_y}$. Поэтому для каждой пары чисел n_x, n_y можно указать определенное критическое значение T_α, отвечающее выбранному уровню значимости α.

Значения T_{05} и T_{01} приведены в табл. XVII Приложений, причем с T_α должна сравниваться меньшая из сумм T_x и T_y; обычно это сумма, отвечающая меньшему из чисел n_x и n_y. Если $T < T_{01}$, то нулевая гипотеза отвергается; если же $T \geqslant T_{05}$, то H_0 принимается. При этом нужно тщательно следить за тем, чтобы сравниваемое число T было действительно меньшим из чисел T_x и T_y; чтобы убедиться в этом, нужно после нахождения T вычислить

$$T' = \frac{n(n+1)}{2} - T.$$

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 1. 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. 102.

Таблица 102

Опыт . .	340	343	322	349	332	320	313	304	329
Контроль . .	318	321	318	301	312				

Проверим значимость различия при помощи критерия Стьюдента, предполагая, что распределение вариант нормально. Для упрощения вычислений применяем кодирование, уменьшая все числа на 300. Тогда расчет выглядит так, как в табл. 103. Поскольку $t_{05}(12) = 2{,}18$, то различие незначимо ($t < t_{05}$).

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
R_y	1		3		5	6		8						

Так как мы имеем здесь дело с числами, а не с буквами x и y, то во избежание путаницы записываем их не в одну строку, а в две. 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

$$T_x + T_y = \frac{n(n+1)}{2}, \text{ т. е. } 82 + 23 = \frac{14 \cdot 15}{2} = 105.$$

Из табл. XVII Приложений 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) $T_{05} = 22$; $T_{01} = 18$. Поскольку $T > T_{05}$, нулевая гипотеза не отвергается.

В данном случае можно было не вычислять T_x, так как сразу видно, что сумма пяти рангов будет меньшей.

Таблица 103

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
43	15	225	21	7	49
22	−6	36	18	4	16
49	21	441	1	−13	169
32	4	16	12	−2	4
20	−8	64			
13	−15	225	70	0	254
4	−24	576			
29	1	1			
252	0	1728			

$$\bar{x} = \frac{252}{9} = 28;$$

$$\bar{y} = \frac{70}{5} = 14;$$

$$\bar{x} - \bar{y} = 28 - 14 = 14;$$

$$s_{\bar{x}-\bar{y}} = \sqrt{\frac{1728+254}{9+5-2} \cdot \frac{9+5}{9 \cdot 5}} =$$

$$= \sqrt{\frac{1982 \cdot 14}{12 \cdot 9 \cdot 5}} = 7{,}18;$$

$$t = \frac{14}{7{,}18} = 1{,}95,$$

$$f = 12.$$

Когда значения n_x и n_y выходят за пределы табл. XVII Приложений, можно воспользоваться тем обстоятельством, что при достаточно больших объемах выборок распределение величин T_x

и T_y приближается к нормальному со средними значениями

$$\hat{T}_x = \frac{n_x(n+1)}{2}; \quad \hat{T}_y = \frac{n_y(n+1)}{2}$$

и дисперсиями

$$\sigma^2\{T_x\} = \sigma^2\{T_y\} = \frac{n_x n_y (n+1)}{12}$$

Поэтому значимость отклонения T_x от $\hat{T}_x$ можно оценить по u-критерию, вычислив

$$u = \frac{\hat{T}_x - T_x}{\sigma\{T_x\}} = \frac{\frac{n_x(n+1)}{2} - T_x}{\sqrt{\frac{n_x n_y (n+1)}{12}}}.$$

Это можно переписать в виде

$$u = \sqrt{3}\,\frac{n_x(n+1) - 2T_x}{\sqrt{n_x n_y (n+1)}}.$$

Удобней пользоваться величиной

$$w = \frac{u}{\sqrt{3}} = \frac{n_x(n+1) - 2T_x}{\sqrt{n_x n_y (n+1)}}, \tag{7.1}$$

сравнивая ее с критическими значениями

$$w_{05} = \frac{u_{05}}{\sqrt{3}} = \frac{1{,}96}{3} = 1{,}13; \quad w_{01} = \frac{u_{01}}{\sqrt{3}} = \frac{2{,}58}{3} = 1{,}49.$$

Для примера 1 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 $w < w_{05}$, то нулевая гипотеза не отвергается — результат, полученный ранее по табл. XVII Приложений. Однако в данном случае применение формулы (7.1) не совсем правомерно, так как числа n_x и n_y недостаточно велики.

§ 3. Критерий X (ван дер Вардена)

Мощность («разрешающая способность») критерия Вилкоксона сравнительно невелика, что можно усмотреть из следующего примера.

Пример 2. Рассмотрим две задачи о сравнении рядов [1]. В первом случае сравним ряды

$$x_i \; 17 \; 14 \; 28 \; 9 \; 6 \; 21 \; 11$$
$$y_j \; 26 \; 31 \; 20 \; 26 \; 34 \; 24 \; 13 \; 27$$

Составляя общий упорядоченный ряд, имеем:

x_i	6	9	11			14	17			21					28		
y_j				13				20			24	26	26	27		31	34
R_x	1	2	3			5	6			8					13		
R_y				4				7			9	10	11	12		14	15

Теперь получаем

$$T_x = 1 + 2 + 3 + 5 + 6 + 8 + 13 =$$
$$= 38 \left(< T_y = \frac{15 \cdot 16}{2} - 38 = 120 - 38 = 82 \right).$$

По табл. XVII Приложений имеем для $n_x = 7$ $n_y = 8$ критические значения: $T_{05} = 44$, $T_{01} = 38$. Фактическое число $T = 38$ не удовлетворяет ни одному из условий $T > T_{05}$ или $T < T_{01}$, поэтому мы не можем сделать надежных заключений о согласии или различии заданных рядов.

Сравним теперь ряды (которые сразу запишем в упорядоченном виде):

x_i		9	11	13	14	17	20				27				
y_i	6							21	24	26		28	29	31	34
R_x		2	3	4	5	6	7				11				
R_y	1							8	9	10		12	13	14	15

Меньшая сумма рангов по-прежнему равна 38: $T = 2 + 3 + 4 + 5 + 6 + 7 + 11 = 38$; здесь также $n_x = 7$, $n_y = 8$. Поэтому в отношении этой пары рядов мы тоже не можем сделать определенных заключений об их согласии или различии.

Между тем применение более мощных критериев показывает, что во второй задаче ситуация более определенна.

Один из таких критериев — так называемый критерий X (ван дер Вардена). При сравнении средних значений в нормальных выборках он оказывается мощнее критерия Вилкоксона.

---

[1] Речь будет идти о содержании витамина С в образцах сока томатов двух сортов (в *мг* на 100 *г* сока).

Не вдаваясь в теорию этого метода, укажем здесь лишь процедуру его применения. Прежде всего, записываем числа обоих сравниваемых рядов в один возрастающий ряд, как и в методе Вилкоксона; так, ряды

x_i	3,8	3,6	4,6	3,9	
y_j	5,4	6,1	6,3	4,2	5,7

запишем в виде:

x_i	3,6	3,8	3,9		4,6			
y_j				4,2		5,4	5,7	6,1	6,3

В этом общем ряде значения x_i имеют номера $\nu_k = 1, 2, 3, 5$. Теперь нужно вычислить сумму

$$X = \sum_k \Psi\left(\frac{\nu_k}{n+1}\right), \tag{7.2}$$

где $n = n_x + n_y$ — общее число вариант (в нашем случае $n = 4 + 5 = 9$), а Ψ — функция, обратная интегралу вероятностей $\Phi(u)$; об этой функции говорилось в § 5 гл. 2, а значения ее даны в табл. II Приложений. В нашем примере отношения $\frac{\nu_k}{n+1}$ равны

$$\frac{1}{10} = 0{,}10; \quad \frac{2}{10} = 0{,}20; \quad \frac{3}{10} = 0{,}30; \quad \frac{5}{10} = 0{,}50;$$

этим значениям $\frac{\nu_k}{n+1}$ соответствуют значения $\Psi\left(\frac{\nu_k}{n+1}\right)$: $-1{,}28$; $-0{,}84$; $-0{,}52$; $0{,}00$ (см. табл. II Приложений). Поэтому

$$X = (-1{,}28) + (-0{,}84) + (-0{,}52) + (0{,}00) = -2{,}64.$$

Применение критерия состоит теперь в том, что нулевая гипотеза отвергается (т. е. выборки считаются различными), если полученное значение X окажется больше, чем соответствующее критическое число для данного числа вариант и принятого уровня значимости; при этом знак величины X не принимается во внимание. Таблица критических значений критерия X помещена в Приложениях (табл. XVIII); как видно из таблицы, нужно учитывать не только сумму $n_x + n_y = n$, но и разность этих чисел (опять-таки, не обращая внимания на знак).

В нашем случае $n = n_x + n_y = 4 + 5 = 9$, $n_y - n_x = 5 - 4 = 1$, так что критические значения равны: 2,38 при $\alpha = 5\%$ и ∞ при $\alpha = 1\%$. Так как полученное значение $X = 2{,}64$ больше, чем 2,38, и меньше, чем ∞, то $1\% < P < 5\%$; это значит, что возможность отвергнуть нулевую гипотезу сомнительна.

Разумеется, результат не должен зависеть от того, какой ряд обозначен через x и какой через y. Это требование выполняется: если мы изменим обозначения (т. е. переставим местами x и y), то x_i будут иметь номера 4, 6, 7, 8, 9, так что получится

$$X = \Psi\left(\frac{4}{10}\right) + \Psi\left(\frac{6}{10}\right) + \Psi\left(\frac{7}{10}\right) + \Psi\left(\frac{8}{10}\right) + \Psi\left(\frac{9}{10}\right) =$$
$$= (-0{,}25) + (0{,}25) + (0{,}84) + (1{,}28) = 2{,}64,$$

т. е. то же самое число (но с обратным знаком, который мы все равно не принимаем во внимание). Понятно, для сокращения вычислений следует принимать в качестве x_i тот ряд, который содержит меньше вариант.

Чтобы продемонстрировать бóльшую мощность критерия X по сравнению с критерием Вилкоксона, применим критерий X к задачам из примера 2.

В первой задаче значения x_i имеют номера 1, 2, 3, 5, 6, 8, 13, а $n + 1 = 7 + 8 + 1 = 16$, так что

$$X = \Psi\left(\frac{1}{16}\right) + \Psi\left(\frac{2}{16}\right) + \Psi\left(\frac{3}{16}\right) + \Psi\left(\frac{5}{16}\right) + \Psi\left(\frac{6}{16}\right) + \Psi\left(\frac{8}{16}\right) +$$
$$+ \Psi\left(\frac{13}{16}\right) = \Psi(0{,}062) + \Psi(0{,}125) + \Psi(0{,}187) + \Psi(0{,}312) +$$
$$+ \Psi(0{,}375) + \Psi(0{,}500) + \Psi(0{,}812) = (-1{,}54) + (-1{,}15) +$$
$$+ (-0{,}89) + (-0{,}49) + (-0{,}32) + (0{,}00) + (+0{,}89) =$$
$$= -3{,}50.$$

По табл. XVIII Приложений находим для $n = 15$ и $n_y - n_x = 1$ критические значения $X_{05} = 3{,}24$ и $X_{01} = 4{,}07$. Так как фактическое значение $|X| = 3{,}50$ находится внутри интервала (X_{05}, X_{01}), то результат анализа остается неопределенным. Значит, в этом случае критерий X не дает ничего нового по сравнению с критерием Вилкоксона.

Обратимся, однако, ко второй задаче из примера 2. Здесь x_i имеют номера 2, 3, 4, 5, 6, 7, 11, поэтому

$$X = \Psi\left(\frac{2}{16}\right) + \Psi\left(\frac{3}{16}\right) + \Psi\left(\frac{4}{16}\right) + \Psi\left(\frac{5}{16}\right) + \Psi\left(\frac{6}{16}\right) + \Psi\left(\frac{7}{16}\right) +$$
$$+ \Psi\left(\frac{11}{16}\right) = \Psi(0{,}125) + \Psi(0{,}187) + \Psi(0{,}250) + \Psi(0{,}312) +$$
$$+ \Psi(0{,}375) + \Psi(0{,}437) + \Psi(0{,}688) = (-1{,}15) + (-0{,}89) +$$
$$+ (-0{,}67) + (-0{,}49) + (-0{,}32) + (-0{,}16) + (+0{,}49) =$$
$$= -3{,}19.$$

В данном случае получилось $|X| < X_{05}$, так что нулевая гипотеза определенно не отвергается.

Таким образом, критерий X позволяет заметить разницу между случаями, которые с точки зрения критерия Вилкоксона неразличимы (ибо в обоих случаях получилось одно и то же значение $T = 38$).

Конечно, критерий X требует больше вычислений, чем критерий Вилкоксона (где нужно лишь подсчитать сумму рангов). Поэтому следует начинать с применения критерия Вилкоксона, и только если окажется, что $T_{01} \leqslant T < T_{05}$, применить дополнительно критерий X. Однако подчеркнем еще раз, что преимущества критерия X (по сравнению с критерием Вилкоксона) имеют место лишь при сравнении выборок из совокупностей, близких к нормальным.

§ 4. Серийный критерий

Теперь рассмотрим критерии, позволяющие обнаружить различие между двумя совокупностями не только по центральной тенденции, но и по другим свойствам.

Как и прежде, нулевая гипотеза H_0 состоит здесь в том, что ряды x и y являются двумя выборками из одной генеральной совокупности. Если это так, то отдельные ранги (или значения) из обоих рядов должны, вообще говоря, чередоваться, когда эти два ряда объединены в один общий ряд. Пусть, например, ранги (или значения) рядов x и y суть соответственно x_1, x_2, x_3, x_4, x_5 и y_1, y_2, y_3, y_4, y_5, y_6; тогда расположение

$$x_1x_2x_3x_4x_5y_1y_2y_3y_4y_5y_6$$

явно противоречит нулевой гипотезе. Гораздо больше соответствует H_0 расположение вида

$$x_1x_2y_1x_3x_4y_2y_3y_4y_5x_5y_6.$$

Количественным показателем, по которому можно отличить оба эти расположения друг от друга, может служить число серий S, каждая из которых есть непрерывная последовательность вариант, принадлежащих к одному из двух рядов. Так, первое расположение состоит из двух серий:

$$\underbrace{x_1x_2x_3x_4x_5}_{1}\ \underbrace{y_1y_2y_3y_4y_5y_6}_{2},$$

а второе расположение — из шести серий:

$$\underbrace{x_1x_2}_{1}\ \underbrace{y_1}_{2}\ \underbrace{x_3x_4}_{3}\ \underbrace{y_2y_3y_4y_5}_{4}\ \underbrace{x_5}_{5}\ \underbrace{y_6}_{6}$$

Серийный критерий различия между двумя совокупностями основан на том, что нулевая гипотеза должна отвергаться, если число серий слишком мало (отсюда ясно, что в данном случае должен применяться односторонний критерий). Методы теории вероятностей позволяют вычислить вероятность того или иного числа серий при заданных численностях вариант в каждом из рядов (при справедливости нулевой гипотезы). Отсюда, обратно, можно указать число серий, отвечающих тому или иному уровню значимости (например, 0,05 или 0,01). В табл. XXIV Приложений приводятся граничные значения числа серий при $\alpha = 0{,}05$; приближенно можно считать, что при n_x, $n_y \leqslant 20$

$$S_{01} = S_{05} - 2.$$

Нулевая гипотеза принимается при $S \geqslant S_{05}$ и отвергается при $S < S_{05} - 2$.

Пример 3. Требуется сравнить два ряда [1]:

x_i	11,5	26,0	29,1	19,7	2,3	22,6	30,9	10,8	23,2	38,8	21,5
y_j	18,4	15,5	25,2	16,9	24,0	13,3	17,9	13,2			

Располагаем все значения в один возрастающий ряд, помещая для ясности варианты x_i и y_j в разных строках:

x_i	2,3	10,8	11,5							
y_j				13,2	13,3	15,5	16,9	17,9	18,4	

(продолжение)

x_i	19,7	21,5	22,6	23,2			26,0	29,1	30,9	38,8
y_j					24,0	25,2				

Мы имеем здесь пять серий. Из табл. XXIV Приложений находим $S_{05}(11; 8) = 6$. Поскольку S меньше, чем S_{05}, но не меньше, чем $S_{05} - 2$, вопрос о справедливости нулевой гипотезы остается открытым.

Заметим, что значимого различия между центральными тенденциями этих рядов заведомо нет. Действительно, подсчет меньшей суммы рангов дает $T = 68$, в то время как для $n_x = 8$, $n_y = 11$ имеем из табл. XVII Приложений $T_{05} = 55$.

Если число вариант в одном из рядов больше 20, то можно воспользоваться тем, что в условиях нулевой гипотезы значения S

[1] Числа выражают привес свиней в пересчете на 100 *кг* кормов при двух разных рационах.

распределены почти нормально со средним значением $\hat{S}$ и дисперсией σ_S^2, равными

$$\hat{S} = \frac{a}{b} + 1; \quad \sigma_S^2 = \frac{a(a-b)}{b^2(b-1)},$$

где

$$a = 2n_x n_y; \quad b = n_x + n_y.$$

Составив выражение

$$u_S = \frac{\hat{S} - S - 0{,}5}{\sigma_S}, \tag{7.3}$$

принимаем нулевую гипотезу при $u_S \leqslant 1{,}96$ и отвергаем ее при $u_S > 2{,}58$; величина 0,5 есть поправка на дискретность.

Пример 4. При сравнении двух рядов численностью $n_x = 14$, $n_y = 26$ в объединенном ряде оказалось $S = 11$ серий. Имеем:

$$a = 2 \cdot 14 \cdot 26 = 728; \quad b = 14 + 26 = 40;$$

$$\hat{S} = \frac{728}{40} + 1 = 18{,}2 + 1 = 19{,}2; \quad \sigma_S^2 = \frac{728(728-40)}{40^2(40-1)} = 8{,}03,$$

так что

$$u_S = \frac{19{,}2 - 11 - 0{,}5}{\sqrt{8{,}03}} = \frac{7{,}7}{2{,}84} = 2{,}72.$$

Поскольку $u_S > 2{,}58$, нулевая гипотеза отвергается.

§ 5. Критерий Колмогорова — Смирнова

При малых объемах совокупностей серийный критерий оказывается недостаточно чувствительным. В таких случаях, если S равно S_α или отличается от S_α не более чем на единицу, следует проверить результат при помощи более чувствительного (мощного) критерия. В качестве такового может быть использован критерий Колмогорова — Смирнова.

Этот критерий основан на сравнении рядов накопленных частостей обеих совокупностей. Именно, пусть $z_i\{x\}$ и $z_i\{y\}$ — накопленные частости рядов, расположенных в порядке возрастания. Составив разности $\eta_i = z_i\{x\} - z_i\{y\}$, находим наибольшую по абсолютной величине разность:

$$D = \max |z_i\{x\} - z_i\{y\}|.$$

Согласно рассматриваемому критерию, нулевая гипотеза отвергается, если эта максимальная разность оказывается слишком

большой. Какое именно значение D должно считаться «слишком большим», зависит, конечно, в первую очередь от принятого уровня значимости. Кроме того, оно зависит и от объемов совокупностей n_x и n_y. Как показывает расчет, в качестве критического значения D можно принять

$$D_\alpha = \sqrt{\frac{1}{2}\ln\frac{2}{\alpha}\cdot\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}.$$

Введя обозначение

$$\sqrt{\frac{1}{2}\ln\frac{2}{\alpha}} = \lambda_\alpha,$$

имеем

$$D_\alpha = \lambda_\alpha\sqrt{\frac{1}{n_x}+\frac{1}{n_y}} = \lambda_\alpha\sqrt{\frac{n_x+n_y}{n_x n_y}}. \qquad (*)$$

Вместо того чтобы сравнивать эмпирическое значение D с D_α из (*), можно сравнивать

$$\lambda^2 = D^2\frac{n_x n_y}{n_x+n_y} \qquad (7.4)$$

с λ_α^2, причем

$$\lambda_{0,5}^2 = \frac{1}{2}\ln\frac{2}{0,05} = \frac{1}{2}\ln 40 = 1,84;$$

$$\lambda_{01}^2 = \frac{1}{2}\ln\frac{2}{0,01} = \frac{1}{2}\ln 200 = 2,65.$$

Если $\lambda^2 > \lambda_{01}^2$, то нулевая гипотеза отвергается, а если $\lambda^2 \leqslant \lambda_{0,5}^2$, то она принимается.

Пример 5. Применим критерий Колмогорова — Смирнова к сравнению рядов из примера 3.

Составим табл. 104. В столбце 1 выписаны в порядке возрастания все встречающиеся здесь значения. В столбцах 2 и 3 проставлены частоты обоих рядов. Значения ряда x будут иметь в столбце 2 частоты 1 (они встречаются там по одному разу), а в столбце 3 — частоты 0 (в ряде y этих значений нет); обратно, значения ряда y имеют в столбце 2 частоты 0 и в столбце 3 частоты 1. Затем составляем ряды накопленных частот — столбцы 4 и 5. Деля эти накопленные частоты в каждом ряде на объем соответствующего ряда ($n_x = 11$, $n_y = 8$), получаем ряды накопленных частостей; они записаны в столбцах 6 и 7. После этого находим разности

$$|\eta_i| = |z_i(x) - z_i(y)|$$

Таблица 104

x_i, y_i	$n_i\{x\}$	$n_i\{y\}$	$s_i\{x\}$	$s_i\{y\}$	$z_i\{x\}$	$z_i\{y\}$	$\lvert\eta_i\rvert$
1	2	3	4	5	6	7	8
2,3	1	0	1	0	0,091	0,000	0,091
10,8	1	0	2	0	0,182	0,000	0,182
11,5	1	0	3	0	0,273	0,000	0,273
13,2	0	1	3	1	0,273	0,125	0,148
13,3	0	1	3	2	0,273	0,250	0,023
15,5	0	1	3	3	0,273	0,375	0,102
16,9	0	1	3	4	0,273	0,500	0,227
17,9	0	1	3	5	0,273	0,625	0,352
18,4	0	1	3	6	0,273	0,750	0,477
19,7	1	0	4	6	0,364	0,750	0,386
21,5	1	0	5	6	0,455	0,750	0,295
22,6	1	0	6	6	0,546	0,750	0,204
23,2	1	0	7	6	0,636	0,750	0,114
24,0	0	1	7	7	0,636	0,875	0,239
25,2	0	1	7	8	0,636	1,000	0,364
26,1	1	0	8	8	0,727	1,000	0,273
29,1	1	0	9	8	0,818	1,000	0,182
30,9	1	0	10	8	0,910	1,000	0,090
38,8	1	0	11	8	1,000	1,000	0,000
Сумма	11	8					

(пренебрегая знаком этой разности). При этом некоторые, заведомо не максимальные, разности можно не вычислять. Так, η_1 и η_2 наверняка меньше $\eta_3 = 0{,}273$; аналогично, η_6, η_7 и η_8 наверняка меньше $\eta_9 = 0{,}477$ и т. д. Вообще не трудно убедиться, что максимальная разность $|\eta_i|$ может получиться только на стыке между сериями. Поэтому достаточно вычислить и записать частости только для этих значений (в табл. 104 они даны жирным шрифтом). Просмотрев полученные значения $|\eta_i|$, находим наибольшее из них; в данном случае это будет

$$D = 0{,}477.$$

Теперь остается вычислить

$$\lambda^2 = D^2 \frac{n_x n_y}{n_x + n_y} = 0{,}477^2 \cdot \frac{11 \cdot 8}{11 + 8} = 1{,}06.$$

Поскольку это значение меньше, чем $\lambda_{05}^2 = 1{,}84$, то нулевая гипотеза не отвергается. Это является уточнением результата, полученного ранее при помощи менее мощного серийного критерия.

Если объемы обеих выборок одинаковы: $n_x = n_y = n$, расчет сильно упрощается. Прежде всего, формула (7.4) принимает вид

$$\lambda^2 = D^2 \frac{n}{2}. \qquad (**)$$

Далее, незачем вычислять накопленные частости. Достаточно найти максимальную разность накопленных частот $\max |s_i\{x\} - s_j\{y\}|$ (которую обозначим Δ), после чего сразу получаем $D = \Delta/n$. Подставляя это в (**), имеем окончательно

$$\lambda^2 = \frac{\Delta^2}{2n}. \qquad (7.5)$$

Пример 6. Если бы в предыдущем примере отсутствовали последние три значения x_i, то ряд накопленных частостей имел бы вид:

x_i	2,3	10,8	11,5	13,2	13,3	15,5	16,9	17,9
$z_i\{x\}$	0,125	0,250	0,375	0,375	0,375	0,375	0,375	0,375

(продолжение)

18,4	19,7	21,5	22,6	23,2	24,0	25,2	26,1
0,375	0,500	0,625	0,750	0,875	0,875	0,875	1,000

Сравнение с данными из столбца 7 табл. 104 показывает, что максимальная разность равна 0,375 (для x_i, $y_j = 11{,}5$ и 18,4). Тогда по формуле (7.4) получается

$$\lambda^2 = 0{,}375^2 \cdot \frac{8 \cdot 8}{8 + 8} = 0{,}562.$$

Между тем этот результат можно было бы получить, не вычисляя частостей. Именно, из сравнения столбцов 4 и 5 табл. 104 находим $\Delta = 3$ (конечно, для тех же x_i, $y_j = 11{,}5$ и 18,4), после чего по формуле (7.5) получаем

$$\lambda^2 = \frac{3^2}{2 \cdot 8} = \frac{9}{16} = 0{,}562.$$

Большое преимущество критерия Колмогорова—Смирнова перед критерием χ^2 состоит в том, что сравниваются функции распределения $F(x)$ двух совокупностей, а не их плотности

распределения $f(x)$. Поэтому отпадает необходимость в группировке вариант, вносящей всегда искажения в распределение (особенно при малых объемах совокупностей).

В заключение отметим, что критерий Колмогорова—Смирнова можно применять только тогда, когда распределение вариант в генеральной совокупности непрерывно, а выборочные значения не подвергнуты группировке.

§ 6. Критерий знаков

Теперь нам остается рассмотреть критерии для сравнения совокупностей с попарно сопряженными вариантами. Этот случай может иметь большое значение в биологических исследованиях. Дело в том, что когда сравниваются две группы независимых вариант, то требуется, чтобы каждая из этих групп была как можно более однородной. Например, если ставится некоторый эксперимент над животными, то нужно, чтобы и в опытной, и в контрольной группах все животные были одного пола, одного возраста и вообще как можно более «одинаковыми». При работе с крупными животными это требование часто оказывается весьма стеснительным. Здесь-то и приходит на помощь методика «спаренных» вариант. Именно, можно включить в опытную группу животных разного пола, разного возраста, разного экстерьера — надо только, чтобы каждое из этих животных имело свою «пару» (т. е. животное того же пола, возраста, экстерьера) в контрольной группе. Часто контролем может служить сам подопытный объект (скажем, состояние до опыта и после опыта); сопряженность вариант опытной и контрольной совокупностей в этом случае очевидна.

Тогда, в случае употребления параметрических методов, сравнение опыта с контролем заключается не в определении разности средних (для групп) значений, а в определении среднего значения разностей для пар (см. гл. 4, § 5). Аналогичным образом обстоит дело и в непараметрической статистике.

В § 5, гл.4 было показано, как решается задача при помощи t-критерия. Сейчас рассмотрим непараметрический метод решения этой задачи.

Пусть мы хотим убедиться в том, что вновь предлагаемый лечебный препарат не меняет состава крови (в частности, числа лейкоцитов). Опыт был поставлен на десяти особях, и анализ дал следующие результаты:

0,97 1,05 1,09 0,88 1,01, 1,14 1,03 1,07 0,94 1,02

(числа выражают отношение числа лейкоцитов в опыте и в норме).

Мы видим, что в семи случаях число лейкоцитов увеличилось (этим случаям мы припишем знак плюс), а в трех случаях —

уменьшилось (эти случаи будем обозначать знаком минус). Это как будто дает основания считать, что применение изучаемого препарата ведет в большинстве случаев к увеличению числа лейкоцитов в крови. Однако такое заключение было бы неверным. В самом деле, поставим такой вопрос: какого результата следовало бы ожидать, если бы препарат заведомо не влиял на содержание лейкоцитов (или если бы вообще этот препарат не вводился)?

Содержание лейкоцитов в крови подвержено случайным вариациям во времени, причем отклонения в ту и другую стороны одинаково вероятны. Поэтому следует ожидать, что в среднем будут одинаково часто происходить отклонения как в сторону увеличения, так и в сторону уменьшения (т. е. получится одинаковое число плюсов и минусов).

Но слова «в среднем» имеют тот смысл, что такая тенденция будет проявляться лишь при достаточно большом числе проб. Если же проб сделано мало, то могут иметь место сколь угодно большие отклонения от этой тенденции. Например, мы хорошо знаем, что при бросании правильной монеты одинаково вероятно выпадение каждой из сторон. Однако это вовсе не значит, что при практическом выполнении этой процедуры стороны монеты будут строго чередоваться. Гербы могут повторяться и два, и три, и большее число раз подряд; имея достаточное терпение, можно дождаться серии одних гербов весьма большой длины. Отсюда ясно, что если бы в нашем опыте увеличение содержания лейкоцитов наблюдалось бы даже все десять раз, то это еще не давало бы гарантии, что имеется связь между введением препарата и содержанием лейкоцитов. Такой результат мог, вообще говоря, получиться и при условии, что содержание лейкоцитов меняется случайным образом — так же, как может случайно появиться серия из десяти гербов при бросании монеты. Правда, такой результат имеет малую вероятность (последняя равна, очевидно, $1/2^{10} = 1/1024 \approx 0{,}001$), но он не исключен. Тем более мог случайно получиться результат, что в семи случаях содержание лейкоцитов увеличивается, а в трех — уменьшается. Но это означает, что в другой серии таких же опытов могло бы получиться наоборот: в семи случаях уменьшение и в трех случаях увеличение.

Как и ранее, будем считать тот или иной результат случайным, если вероятность его равна или больше 0,05, и значимо неслучайным, если его вероятность меньше 0,01. Пользуясь законом биномиального распределения (с учетом того, что нулевая гипотеза соответствует условию равной вероятности плюсов и минусов), можно вычислить, насколько соотношение численностей альтернатив может отклоняться от ожидаемого соотношения 1 : 1, чтобы вероятность его была бы не меньше 0,05 или 0,01.

Соответствующие критические значения даны в табл. XXIII

Приложений для разных объемов n сравниваемых совокупностей (очевидно, при наличии сопряженных пар всегда $n_x = n_y = n$). В этой таблице помещены критические значения для числа знаков Z, встречающихся менее часто. Если при сравнении двух рядов значения в каких-либо парах совпадают (так что нет ни плюса, ни минуса), то эти пары исключаются из рассмотрения; соответственно уменьшается n.

В приведенном выше примере $n = 10$ и $Z = 3$. Из табл. XXIII Приложений находим $Z_{05}(10) = 2$; так как $Z > Z_{05}$, то нулевая гипотеза не отвергается — результат (7 плюсов и 3 минуса) мог получиться случайно.

Заметим, что если бы опыт был поставлен на ста особях и получилось бы то же соотношение 7 : 3 числа плюсов и минусов (т. е. 70 плюсов и 30 минусов), то нулевую гипотезу следовало бы отвергнуть, ибо $Z_{01}(1000) = 37$.

Критерий знаков является частным случаем более общего биномиального критерия. Последний служит для проверки значимости отклонения от ожидаемого соотношения двух численностей. Например, ожидается, что из n родившихся детей $\frac{n}{2}$ будет мальчиков и $\frac{n}{2}$ девочек. Если из 100 детей число мальчиков и девочек составляет соответственно 54 и 46, то налицо отклонение от ожидаемого соотношения 1 1. Из табл. XXIII Приложений видим, что это отклонение незначимо, т. е. может быть объяснено случайностями выборки.

§ 7. Критерий Вилкоксона для сопряженных пар

Недостаточную чувствительность критерия знаков продемонстрируем на следующем примере.

Пример 7. Один из видов предпосевной обработки семян дал следующие изменения урожайности (табл. 105).

Восемь знаков из десяти положительны. Из табл. XXIII Приложений видим, что $Z_{05}(10) = 2$, причем при $Z \geqslant Z_{05}$ нулевая гипотеза не отвергается. Таким образом, в данном случае нет оснований отвергнуть нулевую гипотезу. Применение t-критерия подтверждает этот вывод: при средней разности $\bar{\Delta} = 0{,}4$ ее стандартная ошибка равна $\sqrt{\frac{6{,}66}{90}} = 0{,}272$, так что

$$t = \frac{\bar{\Delta}}{\sigma_{\bar{\Delta}}} = \frac{0{,}4}{0{,}272} = 1{,}47,$$

в то время как $t_{05}(9) = 2{,}26$.

Таблица 105

Годы	$x_{контр}$	$x_{опыт}$	$\Delta = \Delta x$	$\Delta_i - \bar{\Delta}$	$(\Delta_i - \bar{\Delta})^2$	R^{Δ}	(T^{Δ})	Знак Δ
1950	20,0	22,1	2,1	1,7	2,89	10		+
1951	17,9	18,5	0,6	0,2	0,04	6,5		+
1952	20,6	19,4	—1,2	—1,6	2,56	9	9	—
1953	22,0	22,1	0,1	—0,3	0,09	1		+
1954	21,4	21,7	0,3	—0,1	0,01	3,5		+
1955	23,8	24,9	1,1	0,7	0,49	8		+
1956	21,4	21,6	0,2	—0,2	0,04	2		+
1957	19,8	20,3	0,5	0,1	0,01	5		+
1958	18,4	18,7	—0,3	—0,7	0,49	3,5	3,5	—
1959	22,5	23,1	0,6	0,2	0,04	6,5		+
Сумма			4,0	0	6,66		12,5	

Проделанный одновременно другой вид предпосевной обработки дал результат, который мы сравниваем с тем же контролем, что и в предыдущем случае (табл. 106).

Таблица 106

Годы	$x_{контр}$	$x_{опыт}$	$\Delta = \Delta x$	$\Delta_i - \bar{\Delta}$	$(\Delta_i - \bar{\Delta})^2$	R^{Δ}	(T^{Δ})	Знак Δ
1950	20,0	26,7	6,7	3,7	13,69	10		+
1951	17,9	21,0	3,1	0,1	0,01	4		+
1952	20,6	24,1	3,5	0,5	0,25	6		+
1953	22,0	27,1	5,1	2,1	4,41	9		+
1954	21,4	25,1	3,7	0,7	0,49	7		+
1955	23,8	23,0	—0,8	—3,8	14,44	2	2	—
1956	21,4	26,2	4,8	1,8	3,24	8		+
1957	19,8	19,4	—0,4	—3,4	11,56	1	1	—
1958	18,4	21,8	3,4	0,4	0,16	5		+
1959	22,5	23,4	0,9	—2,1	4,41	3		+
Сумма			30,0	0	52,66		3	

Критерий знаков опять не дает оснований отвергнуть нулевую гипотезу. Однако, если в предыдущем случае мы могли примириться с таким ответом, то здесь с этим трудно согласиться, так как сдвиги очень велики. И действительно, применяя критерий

Стьюдента, получаем следующий результат:

$$\bar{\Delta} = 3{,}0;\ s_{\bar{\Delta}} = 0{,}765;\ t = 3{,}0 : 0{,}765 = 3{,}92,$$

что превышает $t_{01}(9) = 3{,}25$.

Причина расхождения между ответами, даваемыми двумя критериями, состоит в том, что при применении критерия знаков мы использовали не всю информацию, содержавшуюся в экспериментальных данных; именно, мы учли только знаки разностей, но не их величины, которые в последнем примере были очень велики. Поэтому-то критерий Стьюдента позволяет заметить различие между выборками, которые для критерия знаков неразличимы.

Имеется также непараметрический критерий для сравнения выборок с сопряженными вариантами, обладающий бо́льшей мощностью, чем критерий знаков, — это критерий Вилкоксона. Критерий Вилкоксона учитывает не только знак разности между сопряженными членами рядов, но и величину этой разности. 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. Из-за 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

$$1 + 2 + \ldots + n = \frac{n(n+1)}{2}.$$

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. XX Приложений.

Если какие-либо разности равны нулю, то они просто исключаются из рассмотрения, причем соответственно уменьшается n.

Пример 8. Сравнивалось действие двух экстрактов вируса табачной мозаики. Для этого каждая из половин листа натира-

лась соответствующим препаратом. Число мест поражений записано в табл. 107. Поскольку два числа в каждой строке относятся к двум половинам одного и того же листа, то варианты можно считать попарно сопряженными.

Таблица 107

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 $T^{\Delta}_{05}(8) = 5$, указанное в табл. XX Приложений, то нулевая гипотеза не отвергается.

Применим этот критерий к примеру 7. Для табл. 105 получаем $T^{\Delta} = 12{,}5$, а для табл. 106 $T^{\Delta} = 3$. Обращаясь теперь к табл. XX Приложений, заключаем, что в первом случае нулевая гипотеза может быть принята ($T^{\Delta} > T^{\Delta}_{05} = 9$), а во втором случае она отвергается ($T^{\Delta} < T^{\Delta}_{01} = 4$). В последнем случае критерий Вилкоксона учел то обстоятельство, что обе отрицательные разности принадлежат к наименьшим разностям.

Если $n > 25$ (так что табл. XX Приложений пользоваться нельзя), то можно применить u-критерий, ибо 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

$$u_{T^{\Delta}} = \frac{\hat{T}^{\Delta} - T^{\Delta}}{\sigma_{T^{\Delta}}} \tag{7.6}$$

меньше или равно $u_{05} = 1{,}96$, и отвергается, если $u_{T\Delta} > u_{01} = 2{,}58$. 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 $u_{T\Delta} > u_{01}$, то нулевая гипотеза отвергается.

В настоящей главе были рассмотрены лишь некоторые непараметрические критерии, причем только такие, которые отвечают на вопрос о значимости различия между двумя эмпирическими совокупностями. Наряду с другими непараметрическими критериями, служащими для той же цели, существует еще большое число других непараметрических критериев, позволяющих решать такие задачи, как сравнение эмпирического распределения с теоретическим или одновременное сравнение нескольких распределений [см. S. Siegel (1956)].

КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

§ 1. Связь между признаками

Отличительной чертой биологических объектов является многообразие признаков, характеризующих каждый из них. Так, животное можно характеризовать возрастом, размерами, 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 размеры животного, тем обычно больше его вес; известно также, что в однородном стаде те коровы, молоко которых имеет 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. Такого рода 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 1. Измерение диаметров и высот 250 экземпляров сосны дало результаты:

Диаметр, *см* (x)	Высота, *м* (y)
22	19,3
37	24,1
14	20,6
26 и т. д.	19,0 и т. д.

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 (точнее, соответствующие числовые значения) в одну общую таблицу. Это будет, естественно, двумерная таблица. Поскольку число особей в данном случае велико,

должна быть проведена группировка. Диаметры принимают значения от ~15 до ~50 *см*, так что интервал 50 — 15 = 35 можно разбить на 7 частей (разрядов) по 5 *см* в каждом; впрочем, можно было бы взять 9 разрядов по 4 *см*. Высоты оказались в пределах от ~18 до ~28 *м*, поэтому произведем группировку по 10 разрядам.

Результат такой группировки представлен в табл. 108. Первое дерево (диаметр 22 *см*, высота 19,3 *м*) попало в клетку на пересечении второй строки (диаметры от 17,5 до 22,5 *см*, т. е. середина разряда 20 *см*) и второго столбца (высота от 18,5 до 19,5 *м*, середина разряда — 19 *м*), т. е. в число 3; второе — в клетку на пересечении пятой строки и седьмого столбца (в число 6) и т. д.

Таблица 108

Диаметр, *см* (x)	Высота, *м* (y)										
	18	19	20	21	22	23	24	25	26	27	n_x
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		15
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 высоки, 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 обуслов-

ливаются 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.

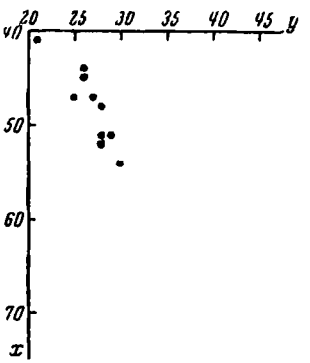

Рис. 43

В случае несгруппированной совокупности может быть получено наглядное представление о наличии или отсутствии корреляции путем построения так называемого *корреляционного поля*.

Пример 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

Нанося точки на графике в выбранном масштабе, получаем картину, изображенную на рис. 43 (точечную диаграмму). Вытянутость корреляционного поля в диагональном направлении свидетельствует о несомненном наличии корреляции между обоими признаками.

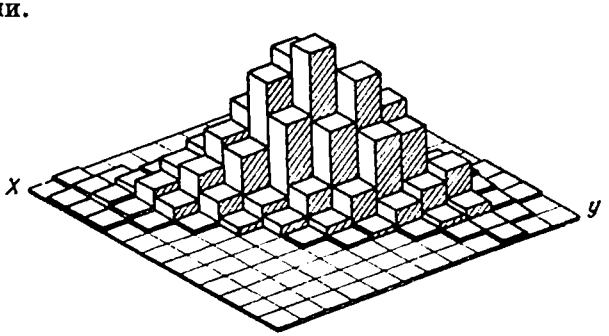

Рис. 44

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 является менее объективным показателем, ибо она в большей мере зависит от принятых масштабов по осям координат.

При большом числе вариант, когда произведена группировка, можно также получить графический образ двумерного распределения — в виде трехмерной фигуры, являющейся обобщением гистограммы или полигона частот (см., например, рис. 44). Однако выполнение такого графика сложно, и сама фигура недостаточно наглядна; она не дает особых преимуществ по сравнению с корреляционной решеткой. При некотором опыте последняя позволяет достаточно хорошо судить о характере корреляции.

§ 2. Регрессия

При двумерной группировке объектов варианты, приходящиеся на один разряд одного из признаков, обычно распределены между несколькими разрядами другого признака. Так, среди деревьев с диаметрами 15 *см* (в действительности с диаметрами в интервале 12,5 — 17,5 *см*) имеются экземпляры с высотами 19 *м* (1 шт.), 20 *м* (6 шт.), 21 *м* (4 шт.) и 22 *м* (3 шт.), как это видно из

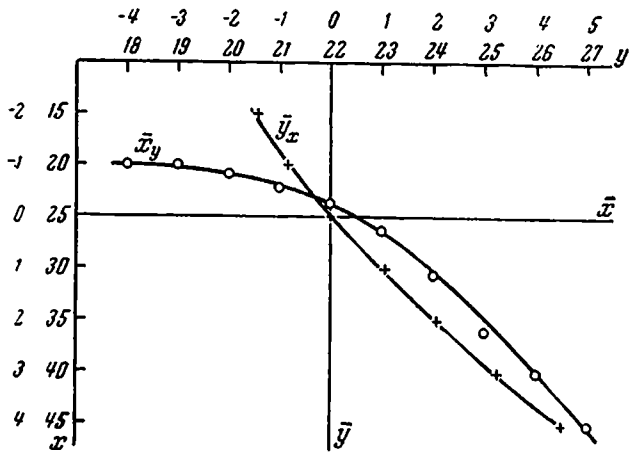

Рис. 45

табл. 27. Точно так же среди деревьев с высотой 23 *м* имеются экземпляры с диаметрами 20 *см* (8 шт.), 25 *см* (20 шт.), 30 *см* (12 шт.) и 35 *см* (3 шт.). Мы видим, что в данном случае каждая строка представляет собой некоторое распределение по высотам для деревьев с одинаковыми диаметрами, так же, как каждый столбец

есть распределение по диаметрам для деревьев с одинаковыми высотами (напомним, что варианты, попавшие в один разряд группировки, считаются одинаковыми).

В каждом из этих частных распределений может быть найдено свое среднее значение, например, средняя высота деревьев, имеющих диаметр 20 *см*. Эти средние значения $\bar{y}_x$ называют иногда *условными средними*: $\bar{y}_x$ есть среднее значение y при условии, что x имеет заданную величину.

Условные средние $\bar{y}_x$, как правило, не совпадают между собой. Однако если исследуемая совокупность является выборкой, то о корреляции между двумя признаками можно говорить только тогда, когда изменения выборочных условных средних $\bar{y}_x$ при переходе от одного значения x к другому носят более или менее систематический характер.

Изобразим на графике значения $\bar{y}_x$ и $\bar{x}_y$, отвечающие табл. 108 (рис. 45); для лучшего различения мы будем обозначать точки $(x, \bar{y}_x)$ крестиками, а точки $(y, \bar{x}_y)$ — кружками. По этим точкам можно провести две сравнительно плавные линии, изображающие зависимость $\bar{y}_x$ от x и $\bar{x}_y$ от y. Эти линии принято называть *линиями регрессии*. Как правило, линии регрессии y по x и x по y не совпадают между собой, что связано с упомянутой «размазанностью» корреляции.

В общем случае линии регрессии являются кривыми, вид которых отражает ту или иную биологическую закономерность, связывающую признаки между собой (или с каким-либо третьим признаком). Ограничимся пока рассмотрением простейшего частного (но весьма распространенного) случая, когда линии регрессии представляют собой прямые, т. е. когда средние значения одного признака зависят от значений другого признака линейно; этот случай называют *линейной регрессией*.

Из аналитической геометрии известно, что если прямая проходит через некоторую точку с координатами x_0, y_0, то уравнение этой прямой может быть записано в виде

$$y - y_0 = k\,(x - x_0),$$

где k — так называемый угловой коэффициент (так как он равен тангенсу угла между прямой и осью абсцисс). Поэтому уравнения прямых регрессии (для генеральной совокупности) будут:

$$\hat{y}_x - \hat{y} = \beta_{y/x}\,(x - \hat{x}); \quad \hat{x}_y - \hat{x} = \beta_{x/y}\,(y - \hat{y}). \qquad (8.1)$$

Для угловых коэффициентов здесь приняты обозначения $\beta_{y/x}$ и $\beta_{x/y}$; их называют *коэффициентами регрессии* (точнее, коэффициентами линейной регрессии). Уравнения (8.1) называются *уравнениями линейной регрессии*.

Для практических целей могут представлять интерес оба уравнения регрессии. Пусть, например, изучается связь между привесом свиней y и количеством затраченного корма x. Тогда, с одной стороны, интересно знать, насколько в среднем возрастает привес при увеличении количества корма на одну единицу (величина $\beta_{y/x}$), а с другой стороны, насколько в среднем нужно увеличить количество корма, чтобы привес возрос на одну единицу (величина $\beta_{x/y}$). Как уже говорилось, при «размазанности» корреляции величины $\beta_{y/x}$ и $\beta_{x/y}$ не являются просто обратными (т. е. $\beta_{x/y} \neq 1/\beta_{y/x}$).

Разумеется, в случае эмпирической совокупности точки, изображающие зависимость $\bar{y}_x$ от x (или $\bar{x}_y$ от y), никогда не ложатся на одну прямую. Поэтому речь может идти только о нахождении такой прямой, которая проходила бы наиболее близко ко всем точкам. При этом, конечно, смысл этой «близости» можно понимать по-разному. Можно, например, считать наилучшей ту прямую, 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 прямой. Можно, далее, считать наилучшей ту прямую, при которой площадь между прямой и ломаной, соединяющей эмпирические точки $(x,\ \bar{y}_x)$, 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 прямой (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 $\bar{y}_x$, которые представляют меньшее число вариант, оказывали соответственно меньшее влияние на расположение прямой 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. Отметим еще, что при отыскании $\beta_{y/x}$ и $\beta_{x/y}$ отклонения измеряются по направлениям соответствующих осей координат, т. е. y и x.

Мы не будем излагать здесь схемы расчетов. Эта чисто математическая задача сводится к системе линейных уравнений, решение которой имеет вид:

$$\beta_{y/x} = \frac{\operatorname{cov}\{x,y\}}{\sigma_x^2}; \quad \beta_{x/y} = \frac{\operatorname{cov}\{x,y\}}{\sigma_y^2}, \tag{8.2}$$

где через σ_x^2 и σ_y^2 обозначены $\sigma^2\{x\} = \frac{1}{N}\sum_x n_x (x - \hat{x})^2$ и $\sigma^2\{y\} = \frac{1}{N}\sum_y n_y (y - \hat{y})^2$, а

$$\operatorname{cov}\{x, y\} = \frac{1}{N}\sum_{x,y} n_{xy} (x - \hat{x})(y - \hat{y}) \tag{8.3}$$

есть величина, называемая *ковариацией* признаков x и y. В развернутом виде

$$\beta_{y/x} = \frac{\sum n_{xy} (x - \hat{x})(y - \hat{y})}{\sum n_x (x - \hat{x})^2}; \quad \beta_{x/y} = \frac{\sum n_{xy} (x - \hat{x})(y - \hat{y})}{\sum n_y (y - \hat{y})^2}.$$

Если рассматриваемая эмпирическая совокупность является выборкой из какой-то генеральной совокупности, то центрами рассеяния считаются выборочные средние $\bar{x}$ и $\bar{y}$. Тогда величины

$$b_{y/x} = \frac{\sum n_{xy} (x - \bar{x})(y - \bar{y})}{\sum n_x (x - \bar{x})^2}; \quad b_{x/y} = \frac{\sum n_{xy} (x - x)(\bar{y} - \bar{y})}{\sum n_y (y - \bar{y})^2} \tag{8.4}$$

будут выборочными оценками коэффициентов регрессии $\beta_{y/x}$ и $\beta_{x/y}$.

Если ввести обозначения:

$$\left.\begin{aligned} \sum_{xx} = \sum_x n_x (x - \bar{x})^2; \quad \sum_{yy} = \sum_y n_y (y - \bar{y})^2; \\ \sum_{xy} = \sum_{x,y} n_{xy} (x - \bar{x})(y - \bar{y}), \end{aligned}\right\} \tag{8.5}$$

то можно будет сокращенно записать:

$$b_{y/x} = \frac{\Sigma_{xy}}{\Sigma_{xx}}; \quad b_{x/y} = \frac{\Sigma_{xy}}{\Sigma_{yy}}. \tag{8.6}$$

Покажем теперь, как вычисляются величины (8.5). Применяя формулу (1.17), получаем:

$$\left.\begin{aligned} \sum_{xx} = \sum_x n_x (x - \bar{x})^2 = \sum_x n_x x^2 - n\bar{x}^2; \\ \sum_{yy} = \sum_y n_y (y - \bar{y})^2 = \sum_y n_y y^2 - n\bar{y}^2. \end{aligned}\right\} \tag{*}$$

Рассуждая так же, как при выводе формулы (1.17), находим аналогичное соотношение

$$\sum_{xy} = \sum_x \sum_y n_{xy} (x - \bar{x})(y - \bar{y}) = \sum_x \sum_y n_{xy} xy - n\overline{xy}. \tag{**}$$

Если теперь ввести обозначения:

$$\left.\begin{array}{c} \sum\limits_x n_x x = X_{(1)}; \quad \sum\limits_x n_x x^2 = X_{(2)}; \\ \sum\limits_y n_y y = Y_{(1)}; \quad \sum\limits_y n_y y^2 = Y_{(2)}; \\ \sum\limits_x \sum\limits_y n_{xy} xy = (XY), \end{array}\right\} \tag{8.7}$$

то получим после подстановки этих выражений в (*) и (**):

$$\Sigma_{xx} = X_{(2)} - \frac{X_{(1)}^2}{n}; \quad \Sigma_{yy} = Y_{(2)} - \frac{Y_{(1)}^2}{n}, \quad \Sigma_{xy} = (XY) - \frac{X_{(1)} Y_{(1)}}{n}, \tag{8.8}$$

поскольку в этих обозначениях

$$\bar{x} = \frac{X_{(1)}}{n}, \quad \bar{y} = \frac{Y_{(1)}}{n}. \tag{8.9}$$

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.

Пример 3. В табл. 109 представлена корреляция между высотой трехлетней сосны и длиной ее вершинного побега. Опишем подробно заполнение столбцов с 1 по 4; соответствующие строки заполняются совершенно аналогично.

Столбец (1) содержит значения n_x; они получаются суммированием частот в данной строке. Например, $8 = 5 + 3$, $25 = 7 + 15 + 1 + 2$, $42 = 2 + 21 + 17 + 2$ и т. д. Сумма значений n_x дает объем выборки n. Эта же величина должна получиться при сложении значений n_y из строки (1), что служит проверкой правильности расчета на данной стадии. После этого записывается условная шкала — столбец (2). Нуль обычно выбирают против максимальной частоты n_x, n_y. Но в данном случае, ввиду явной асимметрии каждого из распределений, нуль условной шкалы поставлен не против максимальных частот (75 и 97), а смещен к среднему разряду.

Числа столбца (3) получаются перемножением чисел из столбцов (1) и (2) в соответствующей строке. Так, $-32 = 8(-4)$, $60 = 20 \cdot 3$ и т. д. Алгебраическая сумма чисел столбца (3), в соответствии с определением, равна $X_{(1)}$. Эта сумма (в данном случае это —43) записывается под столбцом чисел, а еще ниже ставится ее обозначение $X_{(1)}$.

Числа столбца (4) опять получаются перемножением чисел двух предыдущих столбцов; например, $(-32) \cdot (-4) = 128$, $60 \cdot 3 = 180$ и т. д. Сумма этих чисел есть $X_{(2)}$, она также записывается внизу вместе с ее обозначением. Аналогично находим $Y_{(1)}$ и $Y_{(2)}$.

Таблица 109

Высота сосны, см (x)	Длина вершинного побега, см (y)							n_x	x	$n_x x$	$n_x x^2$	Y_x	$\bar{y}_x$	$Y_x \bar{y}_x$
	3	7	11	15	19	23	27	(1)	(2)	(3)	(4)	(5)	(6)	(7)
6	5	3						8	—4	—32	128	—21	—2,62	55,0
16	7	15	1	2				25	—3	—75	225	—52	—2,08	108,0
26	2	21	17	2				42	—2	—84	168	—65	—1,55	100,8
36		15	37	20	3			75	—1	—75	75	—64	—0,85	54,4
46		1	31	26	4			62	0	0	0	—29	—0,47	13,6
56		2	8	27	12	1		50	1	50	50	2	0,04	0,1
66			2	9	24	4	2	41	2	82	164	36	0,88	31,7
76			1	4	7	6	2	20	3	60	180	24	1,20	28,8
86				1		3		4	4	16	64	6	1,50	9,0
96						1	2	3	5	15	75	8	2,66	21,3
n_y (1)	14	57	97	91	50	15	6	330		—43	1129			422,7
y (2)	—3	—2	—1	0	1	2	3			$X_{(1)}$	$X_{(2)}$			
$n_y y$ (3)	—42	—114	—97	0	50	30	18	—155	$Y_{(1)}$					
$n_y y^2$ (4)	126	228	97	0	50	60	54	615	$Y_{(2)}$					
X_y (5)	—45	—112	—59	31	78	44	20							
$\bar{x}_y$ (6)	—3,21	—1,96	—0,61	0,34	1,56	2,93	3,33							
$X_y \bar{x}_y$ (7)	144,4	219,5	36,0	10,5	121,6	129,0	66,6	727,6						

Несколько сложней вычисляются числа из столбца (5) — значения $Y_x = \sum\limits_y n_{xy}y$. Так, для строки $x = -4$ получается

$$Y_{x=-4} = 5\,(-3) + 3\,(-2) = -15 - 6 = -21,$$

для строки $x = 1$

$$Y_{x=1} = 2\,(-2) + 8(-1) + 27\cdot 0 + 12\cdot 1 + 1\cdot 2 = \\ = -4 - 8 + 0 + 12 + 2 = 2$$

и т. д. Аналогично получаются числа из строки (5), т. е. значения $X_y = \sum\limits_x n_{xy}x$; например, для столбца $y = 0$ имеем

$$X_{y=0} = 2\,(-3) + 2\,(-2) + 20\,(-1) + 26\cdot 0 + 27\cdot 1 + 9\cdot 2 + \\ + 4\cdot 3 + 1\cdot 4 = \\ -6 - 4 - 20 + 0 + 27 + 18 + 12 + 4 = 31.$$

По данным из столбца и строки (5) находим (XY). Именно, $(XY) = \sum\limits_x Y_x x = \sum\limits_y X_y y$; действительно,

$$\sum_x Y_x x = \sum_x \left(\sum_y n_{xy}y\right) x = \sum_{x,y} n_{xy}xy;$$

$$\sum_y X_y y = \sum_y \left(\sum_x n_{xy}x\right) y = \sum_{x,y} n_{xy}xy.$$

В нашем примере получаем:

$$\sum_x Y_x x = (-21)\,(-4) + (-52)(-3) + (-65)(-2) + \\ + (-64)(-1) + (-29)\cdot 0 + 2\cdot 1 + 36\cdot 2 + 24\cdot 3 + 6\cdot 4 + \\ + 8\cdot 5 = 84 + 156 + 130 + 64 + 0 + 2 + 72 + 72 + 24 + \\ + 40 = 644;$$

$$\sum_y X_y y = (-45)(-3) + (-112)(-2) + (-59)(-1) + \\ + 31\cdot 0 + 78\cdot 1 + 44\cdot 2 + 20\cdot 3 = 135 + 224 + 59 + 0 + 78 + \\ + 88 + 60 = 644.$$

Совпадение обеих этих сумм показывает, что величина (XY) найдена правильно.

Числа столбца (6) получаются делением чисел столбца (5) на числа столбца (1). Это дает значения $\bar{y}_x$ — условные средние, которые могут понадобиться для построения графика. Так,

$-21:8=-2{,}62$; $-52:25=-2{,}08$ и т. д. Аналогично находятся условные средние $\bar{x}_y$ $-45:14=-3{,}21$; $-112:57=-1{,}96$ и. т. д.

После того как таблица заполнена [о столбце (7) будет сказано ниже], можно приступить к окончательному расчету. В данном случае получается:

$$\Sigma_{xy}=(XY)-\frac{X_{(1)}Y_{(1)}}{n}=644-\frac{(-43)(-155)}{331}=624;$$

$$\Sigma_{xx}=X_{(2)}-\frac{X_{(1)}^2}{n}=1129-\frac{(-43)^2}{330}=1123;$$

$$\Sigma_{yy}=Y_{(2)}-\frac{Y_{(1)}^2}{n}=615-\frac{(-155)^2}{330}=542.$$

Но эти величины выражены в единицах условной шкалы. Так как $l_x=10$ *см*, а $l_y=4$ *см*, то $\sum_{xx}$ надо умножить на $l_x^2=100$ *см*^{2}, $\sum_{yy}$ — на $l_y^2=16$ *см*^{2}, а $\sum_{xy}$ — на $l_xl_y=40$ *см*^{2}. Поэтому

$$b_{y/x}=\frac{\Sigma_{xy}\,l_x\,l_y}{\Sigma_{xx}\,l_x^2}=\frac{624\cdot4}{1123\cdot10}=0{,}222;$$

$$b_{x/y}=\frac{\Sigma_{xy}\,l_x\,l_y}{\Sigma_{yy}\,l_y^2}=\frac{624\cdot10}{542\cdot4}=2{,}88.$$

Таким образом, увеличению высоты дерева на 1 *см* соответствует увеличение средней длины вершинного побега на 0,222 *см*; обратно, увеличению длины вершинного побега на 1 *см* соответствует увеличение средней высоты всего дерева на 2,88 *см*.

Поскольку для эмпирических совокупностей условные средние $\bar{y}_x$ и $\bar{x}_y$ не лежат на прямых регрессии, то для них не выполняются уравнения (8.1). Поэтому подстановка в выражения

$$\bar{y}+b_{y/x}(x-\bar{x}).\quad \bar{x}+b_{x/y}(y-\bar{y}) \tag{8.10}$$

опытных значений x и y будет давать не $\bar{y}_x$ и $\bar{x}_y$, а какие-то другие величины, которые обозначим $\tilde{y}_x$ и $\tilde{x}_y$. Эти величины являются ординатами точек, лежащих на прямых регрессии; мы будем называть их *выравненными условными средними*. В нашем примере:

$$\tilde{x}=46+\frac{X_{(1)}\,l_x}{n}=46-\frac{43\cdot10}{330}=46-1{,}30=44{,}70,$$

$$\tilde{y}=15+\frac{Y_{(1)}\,l_y}{n}=15-\frac{155\cdot4}{330}=15-1{,}88=13{,}12,$$

так что выравненные условные средние $\tilde{y}_x$ и $\tilde{x}_y$ определяются уравнениями:

$$\tilde{y}_x = 13{,}12 + 0{,}222\,(x - 44{,}70) = 0{,}222x + 3{,}2;$$
$$\tilde{x}_y = 44{,}70 + 2{,}88\,(y - 13{,}12) = 2{,}88y + 6{,}9.$$

По этим уравнениям можно получить значения $\tilde{y}_x$ и $\tilde{x}_y$ для любых значений x и y, включая те, которые не содержатся в исходной таблице. Однако следует избегать использования уравнений регрессии для экстраполяции за пределы таблицы.

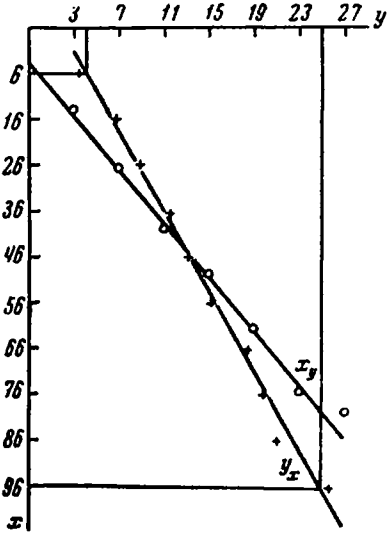

Рис. 46

На рис. 46 показаны прямые регрессии, отвечающие этим уравнениям. Построены они следующим образом. Крайние значения x равны здесь 6 и 96. Подставив их в уравнение для $\tilde{y}_x$, находим два значения:

$$\tilde{y}_{x=6} = 0{,}222 \cdot 6 + 3{,}2 \approx 4{,}5 \text{ и}$$
$$\tilde{y}_{x=96} = 0{,}222 \cdot 96 + 3{,}2 \approx 24{,}5;$$

следовательно, линия регрессии $\tilde{y}_x$ проходит через точки (6; 4,5) и (96; 24,5). Построив на графике эти точки, проводим через них прямую. Аналогично, крайние значения y (3 и 27) подставляются во второе уравнение регрессии, что дает:

$$\tilde{x}_{y=3} = 2{,}88 \cdot 3 + 6{,}9 \approx 15{,}5 \text{ и } \tilde{x}_{y=27} = 2{,}88 \cdot 27 + 6{,}9 \approx 84{,}7;$$

значит, линия регрессии $\tilde{x}_y$ проходит через точки (15,5; 3) и (84,7; 27).

На рис. 46 нанесены также точки $(x;\ \bar{y}_x)$ и $(y;\ \bar{x}_y)$, отвечающие центрам клеток таблицы; они обозначены соответственно крестиками и кружка́ми.

§ 3. Корреляционные отношения

Важной задачей теории корреляции является построение численного параметра, который давал бы количественное выражение степени или силы корреляции (связи) между признаками.

Мы будем считать корреляцию тем более сильной, чем теснее точки корреляционного поля группируются около линии регрессии.

Очевидно, полная дисперсия значений y относительно общего среднего значения $\hat{y}$, т. е. величина $\sigma^2\{y - \hat{y}\}$, определяется двумя факторами. Это, во-первых, различия в значениях y, вызываемые регрессионной зависимостью y от x; количественно эти различия могут быть описаны величиной $\sigma^2\{\hat{y}_x - \hat{y}\}$. Во-вторых, это средний разброс значений y в каждом из столбцов корреляционной решетки около своего условного среднего $\hat{y}_x$; этот разброс вызывается совокупностью неучитываемых влияний и количественно описывается дисперсией $\langle\sigma^2\{y - \hat{y}_x\}\rangle$, где угловые скобки означают усреднение по всем столбцам. Таким образом,

$$\sigma^2\{y - \hat{y}\} = \sigma^2\{\hat{y}_x - \hat{y}\} + \langle\sigma^2\{y - \hat{y}_x\}\rangle. \qquad (*)$$

В соответствии с принятым определением корреляция должна считаться тем сильней, чем меньше разброс значений y около линии регрессии, т. е. чем меньше доля величины $\langle\sigma^2\{y - \hat{y}_x\}\rangle$ в полной дисперсии $\sigma^2\{y - \hat{y}_x\}$. Это можно выразить иначе: корреляция тем сильней, чем б о л ь ш е доля дисперсии $\sigma^2\{\hat{y}_x - \hat{y}\}$. Поэтому показателем силы корреляции может служить величина

$$\eta^2_{y/x} = \frac{\sigma^2\{\hat{y}_x - \hat{y}\}}{\sigma^2\{y - \hat{y}\}}. \qquad (**)$$

Если корреляция отсутствует (значения y в общем не зависят от значений x), то полная дисперсия определяется только неучитываемыми влияниями, так что $\sigma^2\{y - \hat{y}\} = \langle\sigma^2\{y - \hat{y}_x\}\rangle$; тогда $\sigma^2\{\hat{y}_x - \hat{y}\} = 0$ и $\eta^2_{y/x} = 0$. Если же, в другом крайнем случае, корреляция является полной, т. е. неучитываемых влияний нет, то $\langle\sigma^2\{y - \hat{y}_x\}\rangle = 0$, и полная дисперсия определяется целиком регрессионной зависимостью y от x: $\sigma^2\{y - \hat{y}\} = \sigma^2\{\hat{y}_x - \hat{y}\}$; тогда получается $\eta^2_{y/x} = 1$. Очевидно, в этом случае имеет место простая функциональная зависимость.

В дальнейшем будем характеризовать степень коррелированности не отношением дисперсий, а отношением средних квадратических отклонений, т. е. величиной

$$\eta_{y/x} = \frac{\sigma\{\hat{y}_x - \hat{y}\}}{\sigma\{y - \hat{y}\}}. \qquad (8.11)$$

Эта величина называется *корреляционным отношением* y к x. Очевидно, $\eta_{y/x}$ также имеет предельные значения 0 (при отсутствии корреляции) и 1 (при полной корреляции). Аналогично опре-

деляется корреляционное отношение x к y:

$$\eta_{x/y} = \frac{\sigma\{\hat{x}_y - \hat{x}\}}{\sigma\{x - \hat{x}\}}. \tag{8.12}$$

Нетрудно видеть, что величина $\eta^2_{y/x}$, как она определена в (**), вполне аналогична факторной доле вариабильности, введенной в дисперсионном анализе (см. § 3 гл. 5)[1].

Указанная аналогия приводит к тому, что корреляционное отношение может быть вычислено по формуле

$$\eta^2_{y/x} = 1 - \frac{S_Z}{S}. \tag{8.13}$$

Если изучаемая совокупность является выборкой из генеральной совокупности, то мало смещенной оценкой величины $\eta^2_{y/x}$ будет, как показано в § 3 гл. 5,

$$e^2_{y/x} = 1 - \frac{n-1}{n-k} \cdot \frac{S_Z}{S}, \tag{8.14}$$

где k — число разрядов группировки; поправочный множитель $(n-1)/(n-k)$ появляется из-за разброса k выборочных условных средних $\bar{y}_x$ около линии регрессии (т. е. относительно истинных условных средних $\hat{y}_x$) и из-за разброса общей выборочной средней $\bar{y}$ относительно генеральной средней $\hat{y}$.

Так как приходится вычислять два корреляционных отношения — $\eta_{y/x}$ и $\eta_{x/y}$, — то следует различать $S_Z\{y\}$ и $S_Z\{x\}$, а также $S\{y\}$ и $S\{x\}$. В главе 5 мы имели

$$S_Z = S_{ai} - S_a,$$

где

$$S_{ai} = \sum_{a=1}^{n_A}\sum_{i=1}^{n_a} x^2_{ai}; \quad S_a = \sum_{a=1}^{n_A} \frac{X^2_a}{n_a}; \quad X_a = \sum_{i=1}^{n_a} x_{ai}.$$

Здесь мы должны вместо S_{ai} вычислить два разных выражения:

$$\sum_x\sum_y n_{xy}x^2 = \sum_x x^2 \sum_y n_{xy} = \sum_x n_x x^2 = X_{(2)};$$

$$\sum_x\sum_y n_{xy}y^2 = \sum_y y^2 \sum_x n_{xy} = \sum_y n_y y^2 = Y_{(2)}.$$

---

[1] Разница только в том, что здесь, в регрессионном и корреляционном анализе, учитываемый фактор (изменение значений x) является количественным, а в дисперсионном анализе — качественным (разные сорта, различные удобрения и т. д.).

Далее, вместо X_a имеем величины $X_y = \sum_x n_{xy}x$ и $Y_x = \sum_y n_{xy}y$, так что вместо S_a получаем пару величин: $\sum_y X_y^2/n_y$ и $\sum_x Y_x^2/n_x$; они равны соответственно $\sum_y X_y\bar{x}_y$ и $\sum_x Y_x\bar{y}_x$, поскольку $\bar{x}_y = X_y/n_y$ и $\bar{y}_x = Y_x/n_x$. Это приводит к формулам:

$$S_Z\{x\} = X_{(2)} - \sum_y X_y\bar{x}_y; \quad S_Z\{y\} = Y_{(2)} - \sum_x Y_x\bar{y}_x. \qquad (8.15)$$

Что касается величин $S\{x\}$ и $S\{y\}$, то они представляют собой не что иное, как Σ_{xx} и Σ_{yy}, поэтому их можно вычислять по формулам (8.8). Окончательно имеем:

$$e^2_{x/y} = 1 - \frac{n-1}{n-k_x}\cdot\frac{S_Z\{x\}}{\Sigma_{xx}}; \quad e^2_{y/x} = 1 - \frac{n-1}{n-k_y}\cdot\frac{S_Z\{y\}}{\Sigma_{yy}}, \qquad (8.16)$$

где $S_Z\{x\}$ и $S_Z\{y\}$ даются формулами (8.15), а Σ_{xx} и Σ_{yy} — формулами (8.8).

Пример 4. Вычислим корреляционные отношения для совокупности из табл. 109, считая ее выборкой из бесконечной генеральной совокупности.

Значения $X_y\bar{x}_y$ и $Y_x\bar{y}_x$, входящие в формулы (8.15), записаны в столбце и строке (7) таблицы: (—45) (—3,21) = 144,4; (—112) (—1,96) = 219,5 и т. д., (—21) (—2,62) = 55,0; (—52)·(—2,08) = 108,0 и т. д. Там же записаны суммы $\sum_y X_y\bar{x}_y = 727{,}6$ и $\sum_x Y_x\bar{y}_x = 422{,}7$.

Величины $X_{(2)}$ и $Y_{(2)}$ были вычислены ранее (см. табл. 109); они равны: $X_{(2)} = 1129$, $Y_{(2)} = 615$. Тогда формулы (8.15) дают:

$$S_Z\{x\} = 1129 - 727{,}6 = 401{,}4; \quad S_Z\{y\} = 615 - 422{,}7 = 192{,}3.$$

Величины Σ_{xx} и Σ_{yy} также были вычислены: $\Sigma_{xx} = 1123$, $\Sigma_{yy} = 542$. Поэтому окончательно:

$$e_{x/y} = \sqrt{1 - \frac{329}{320}\cdot\frac{401{,}4}{1123}} = \sqrt{0{,}633} = 0{,}795;$$

$$e_{y/x} = \sqrt{1 - \frac{329}{323}\cdot\frac{192.3}{542}} = \sqrt{0{,}639} = 0{,}799.$$

Пример 5. Найдем корреляционные отношения для связи между диаметром и высотой сосны (см. табл. 108).

Все предварительные результаты записаны в табл. 110. По этим данным имеем:

$$S_Z\{x\} = 313 - 169{,}6 = 143{,}4;$$
$$S_Z\{y\} = 569 - 285{,}4 = 283{,}6;$$
$$\Sigma_{xx} = 313 - \frac{(-13)^2}{250} = 312{,}3;$$
$$\Sigma_{yy} = 569 - \frac{11^2}{250} = 568{,}5.$$

Следовательно, в соответствии с формулами (8.16),

$$e_{x/y} = \sqrt{1 - \frac{249}{243}\,\frac{143{,}4}{312{,}3}} = \sqrt{0{,}529} = 0{,}727;$$
$$e_{y/x} = \sqrt{1 - \frac{249}{240}\,\frac{283{,}6}{568{,}5}} = \sqrt{0{,}482} = 0{,}695.$$

§ 4. Коэффициент корреляции

Если корреляция линейна, то вычисление корреляционных отношений

$$\eta_{y/x} = \frac{\sigma\{\hat{y}_x\}}{\sigma_y}, \quad \eta_{x/y} = \frac{\sigma\{\hat{x}_y\}}{\sigma_x}$$

упрощается. Именно, тогда можно в

$$\sigma^2\{\hat{y}_x\} = \frac{1}{N}\sum_x n_x (\hat{y}_x - \hat{y})^2$$

подставить

$$\hat{y}_x - \hat{y} = \beta_{y/x}(x - \hat{x})$$

из уравнения регрессии (8.1), так что получится

$$\sigma^2\{\hat{y}_x\} = \frac{1}{N}\sum_x n_x \beta_{y/x}^2 (x - \hat{x})^2 = \beta_{y/x}^2 \cdot \frac{1}{N}\sum n_x (x - \hat{x})^2 = \beta_{y/x}^2 \sigma_x^2,$$

а поэтому

$$\eta_{y/x} = \frac{\beta_{y/x}\sigma_x}{\sigma_y}.$$

Аналогично получим

$$\sigma^2\{\hat{x}_y\} = \beta_{x/y}^2 \sigma_y^2 \text{ и } \eta_{x/y} = \frac{\beta_{x/y}\sigma_y}{\sigma_x}.$$

Таблица 110

Диаметр, см (x)	Высота, см (y)										n_x	x	$n_x x$	$n_x x^2$	Y_x	$\bar{y}_x$	$Y_x \bar{y}_x$
	18	19	20	21	22	23	24	25	26	27	(1)	(2)	(3)	(4)	(5)	(6)	(7)
15		1	6	4	3						14	—2	—28	56	—19	—1,46	27,8
20	1	3	15	29	20	8					76	—1	—76	76	—64	—0,84	53,7
25		1	8	18	49	20	6	1			103	0	0	0	—2	—0,02	0,0
30			1	4	5	12	8	5			35	1	35	35	37	1,06	39,2
35					1	3	6	4	1		15	2	30	60	31	2,06	63,9
40							1	3	2		6	3	18	54	19	3,17	60,3
45									1	1	2	4	8	32	9	4,50	40,5
n_y (1)	1	5	30	55	78	43	21	13	4	1	250		—13	313			285,4
y (2)	—4	—3	—2	—1	0	1	2	3	4	5			$X_{(1)}$	$X_{(2)}$			
$n_y y$ (3)	—4	—15	—60	—55	0	43	42	39	16	5	11	$Y_{(1)}$					
$n_y y^2$ (4)	16	45	120	55	0	43	84	117	64	25	569	$Y_{(2)}$					
X_y (5)	—1	—5	—26	—33	—19	10	23	22	12	4							
$\bar{x}_y$ (6)	—1,00	—1,00	—0,87	—0,60	—0,24	0,23	1,09	1,69	3,00	4,00							
$X_y \bar{x}_y$ (7)	1,0	5,0	22,6	19,8	4,6	2,3	25,1	37,2	36,0	16,0	169,6						

Подставляя в $\eta_{x/y}$ и $\eta_{y/x}$ значения $\beta_{x/y}$ и $\beta_{y/x}$ из (8.2), находим:

$$\eta_{x/y} = \frac{\operatorname{cov}\{x, y\}}{\sigma_y^2} \cdot \frac{\sigma_y}{\sigma_x} = \frac{\operatorname{cov}\{x, y\}}{\sigma_x \sigma_y}; \quad \eta_{y/x} = \frac{\operatorname{cov}\{x, y\}}{\sigma_x^2} \cdot \frac{\sigma_x}{\sigma_y} = \frac{\operatorname{cov}\{x, y\}}{\sigma_x \sigma_y},$$

т. е. в случае линейной регрессии значения $\eta_{x/y}$ и $\eta_{y/x}$ совпадают. Эту общую меру корреляции называют *коэффициентом корреляции* и обозначают буквой ρ (греческое «ро»):

$$\rho = \frac{\operatorname{cov}\{x, y\}}{\sigma_x \sigma_y}. \tag{8.17}$$

Из этого определения коэффициента корреляции ρ вытекает, что он имеет смысл только для линейной связи, когда $\eta_{x/y} = \eta_{y/x}$. Так, сравнение рис. 45 и 46 показывает, что коэффициент корреляции можно применять для описания связи в табл. 109, но не в табл. 108.

К величине (8.17) можно прийти другим путем, чем это было сделано выше. Именно, представляется естественным, что подходящей мерой степени коррелированности двух признаков может служить ковариация

$$\operatorname{cov}\{x, y\} = \frac{1}{N} \sum_{x, y} n_{xy} (x - \hat{x})(y - \hat{y}). \tag{8.3}$$

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 (8.3) 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 (8.3) 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.

Искомая 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 корре-

ляции мы можем принять величину

$$\rho = \frac{1}{N}\sum n_{xy}\,\frac{x-\hat{x}}{\sigma_x}\cdot\frac{y-\hat{y}}{\sigma_y} = \frac{\frac{1}{N}\sum n_{xy}(x-\hat{x})(y-\hat{y})}{\sigma_x\sigma_y},$$

т. е.

$$\rho = \frac{\operatorname{cov}\{x, y\}}{\sigma_x\sigma_y}.$$

Этот второй способ получения величины ρ указывает, что ρ есть как бы «совместная дисперсия x и y» (т. е. $\operatorname{cov}\{x, y\}$), выраженная в единицах «усредненной дисперсии x и y», причем усреднение произведено геометрически:

$$\sqrt{\sigma_x^2\sigma_y^2} = \sigma_x\sigma_y.$$

Сравнивая (5.13) с (5.2), видим, что

$$\rho^2 = \beta_{x/y}\cdot\beta_{y/x}; \qquad \rho = \sqrt{\beta_{x/y}\cdot\beta_{y/x}}, \tag{8.18}$$

т. е. коэффициент корреляции представляет собой среднее геометрическое из коэффициентов регрессии. С другой стороны, сравнение (8.17) с (8.2) дает также

$$\beta_{y/x} = \rho\frac{\sigma_y}{\sigma_x}; \qquad \beta_{x/y} = \rho\frac{\sigma_x}{\sigma_y}. \tag{8.19}$$

В отличие от нелинейной корреляции, где регрессия может быть немонотонной (т. е. на одних участках возрастающей, а на других убывающей), при линейной корреляции можно говорить о положительной и отрицательной корреляциях. Именно, если при увеличении значений одного признака средние значения другого признака возрастают, то корреляцию считают положительной. Если же увеличение значений первого признака сопровождается убыванием средних значений второго признака, то корреляция считается отрицательной. Очевидно, это соответствует в обоих случаях знакам коэффициентов регрессии: $\beta_{x/y} > 0$, $\beta_{y/x} > 0$ при положительной корреляции и $\beta_{x/y} < 0$, $\beta_{y/x} < 0$ при отрицательной корреляции; заметим, что в любом случае $\beta_{x/y}$ и $\beta_{y/x}$ должны иметь одинаковый знак: либо оба положительны, либо оба отрицательны, так как подкоренное выражение в (8.18) не должно быть отрицательным.

При полной корреляции в корреляционной решетке заполнены только диагональные клетки. При этом, очевидно, линии ре-

грессии y на x и x на y сливаются в одну прямую; тогда угловые коэффициенты $\beta_{y/x}$ и $\beta_{x/y}$ в уравнениях (8.1) являются обратными величинами:

$$\beta_{y/x} = 1 : \beta_{x/y}.$$

Но отсюда следует

$$\beta_{y/x}\beta_{x/y} = 1, \text{ т. е. } \rho^2 = 1.$$

Следовательно, 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 $\rho = -1$ — полная отрицательная корреляция. Если же $\rho = 0$, то корреляция

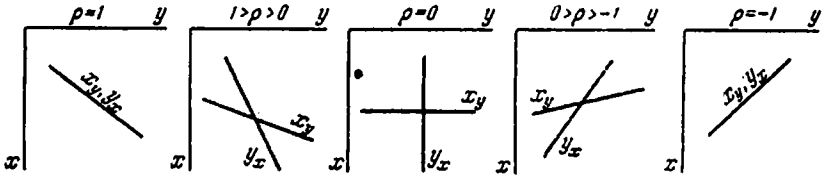

Рис. 47

отсутствует; при этом линии регрессии представляют собой прямые, параллельные осям координат, — $\hat{y}_x$ не зависят от x, а $\hat{x}_y$ не зависят от y. При промежуточных значениях ρ линии регрессии образуют некоторый угол $0 < \gamma < 90°$ (рис. 47). Этот угол тем меньше, чем больше ρ, т. е. чем сильней корреляция.

При вычислении знак ρ получается автоматически — он совпадает со знаком $\operatorname{cov}\{x, y\}$. Но о знаке и примерной величине корреляции можно судить и без вычислений — просто по виду корреляционной таблицы. Так, схематические таблички, объединенные под общим номером табл. 111, представляют случаи, изображенные на рис. 47.

Таблица 111

4	1 2 1	1 2 1	1 2 1	4
12	2 4 4 2	1 3 4 3 1	2 4 4 2	12
18	1 4 8 4 1	2 4 6 4 2	1 4 8 4 1	18
12	2 4 4 2	1 3 4 3 1	2 4 4 2	12
4	1 2 1	1 2 1	1 2 1	4

Примером отрицательной корреляции может служить корреляция между весом семян и процентным содержанием жира у ячменя (табл. 112).

Таблица 112

Вес зерна, мг \ % жира	4,5—5,0	5,0—5,5	5,5—6,0	6,0—6,5	6,5—7,0	7,0—7,5	7,5—8,0	8,0—8,5	n_x
30—35					8	2	1		11
35—40		1	6	22	33	10	2	1	75
40—45	1	2	10	48	37	8	1		107
45—50		1	12	11	2				26
50—55		2	1	1					4
55—60			1						1
n_y	1	6	30	82	80	20	4	1	

Перейдем к вопросу о вычислении коэффициента корреляции. Если использовать введенные ранее обозначения, то можно записать

$$\rho = \frac{\Sigma_{xy}}{\sqrt{\Sigma_{xx}\Sigma_{yy}}}, \tag{8.20}$$

что позволяет применить для вычисления коэффициента корреляции изложенную выше схему расчетов. При этом очевидно, что, пользуясь формулой (8.20), можно подставлять значения Σ_{xy}, Σ_{xx} и Σ_{yy} прямо в условных единицах, не переходя к фактическим единицам; действительно,

$$\rho = \frac{\Sigma_{xy} l_x l_y}{\sqrt{\Sigma_{xx} l_x^2 \Sigma_{yy} l_y^2}} = \frac{\Sigma_{xy}}{\sqrt{\Sigma_{xx}\Sigma_{yy}}}.$$

Подставляя в (8.20) значения Σ_{xy}, Σ_{xx} и Σ_{yy}, вычисленные по выборочным данным, получаем выборочную оценку r коэффициента корреляции. Эта оценка является несмещенной, в чем можно убедиться, если подставить в (8.17) вместо σ^2_{xy}, σ_x и σ_y их несмещенные выборочные оценки

$$s^2_{xy} = \frac{\Sigma_{xy}}{n-1}; \qquad s_x = \sqrt{\frac{\Sigma_{xx}}{n-1}}; \qquad s_y = \sqrt{\frac{\Sigma_{yy}}{n-1}}.$$

Очевидно, делители $n - 1$ в числителе и знаменателе выражения (8.17) сократятся.

Для совокупности из табл. 109 получим, используя значения $\Sigma_{xy} = 624$, $\Sigma_{xx} = 1123$, $\Sigma_{yy} = 542$ со стр. 277:

$$r = \frac{624}{\sqrt{1123 \cdot 542}} = 0{,}798.$$

В примере 4 мы для этой совокупности получили оценки корреляционных отношений $e_{x/y} = 0{,}795$ и $e_{y/x} = 0{,}799$.

Если число вариант невелико ($n \leqslant 25$), то группировка вариант не производится. В этом случае, очевидно, не нужна корреляционная решетка — значения обоих признаков можно записать просто рядом, так что исходная таблица состоит из трех столбцов:

Номера вариант	Значения признака x	Значения признака y

Вычисление оценок коэффициента корреляции и коэффициентов регрессии производится по обычным формулам:

$$r = \frac{\Sigma_{xy}}{\sqrt{\Sigma_{xx}\Sigma_{yy}}}, \quad b_{y/x} = \frac{\Sigma_{xy}}{\Sigma_{xx}}, \quad b_{x/y} = \frac{\Sigma_{xy}}{\Sigma_{yy}},$$

$$\sum_{xy} = (XY) - \frac{X_{(1)}Y_{(1)}}{n}, \quad \sum_{xx} = X_{(2)} - \frac{X_{(1)}^2}{n}, \quad \sum_{yy} = Y_{(2)} - \frac{Y_{(1)}^2}{n},$$

но здесь

$$X_{(1)} = \sum_{i=1}^{n} x_i, \quad X_{(2)} = \sum_{i=1}^{n} x_i^2, \quad Y_{(1)} = \sum_{i=1}^{n} y_i, \quad Y_{(2)} = \sum_{i=1}^{n} y_i^2,$$

$$(XY) = \sum_{i=1}^{n} x_i y_i, \tag{8.21}$$

так как $n_x = n_y = n_{xy} = 1$.

Поскольку в этом случае значения вариант являются, как правило, не равноотстоящими, условную шкалу применить нельзя. Но целесообразно перенести начало отсчета (как для x, так и для y) поближе к середине интервала вариации (конечно, это оценивается на глаз); такой прием избавляет от оперирования большими числами. Иногда достаточно просто уменьшить первую цифру.

Пример 6. Данные табл. 113 показывают урожайность пшеницы и картофеля за разные годы на расположенных рядом полях. Требуется определить коэффициент корреляции между урожайностями этих двух культур.

Таблица 113

Годы	Пшеница, u	Картофель, m	x	y	x^2	y^2	xy	$x+y$	$(x+y)^2$
1	2	3	4	5	6	7	8	9	10
1926	20,1	7,2	—4,9	—0,8	24,0	0,64	3,92	—5,7	32,5
1927	23,6	7,1	—1,4	—0,9	2,0	0,81	1,26	—2,3	5,3
1928	26,3	7,4	+1,3	—0,6	1,7	0,36	—0,78	0,7	0,5
1929	19,9	6,1	—5,1	—1,9	26,0	3,61	9,69	—7,0	49,0
1930	16,7	6,0	—8,3	—2,0	68,9	4,00	16,60	—10,3	106,1
1931	23,2	7,3	—1,8	—0,7	3,2	0,49	1,26	—2,5	6,2
1932	31,4	9,4	6,4	1,4	41,0	1,96	8,96	7,8	60,8
1933	33,5	9,2	8,5	1,2	72,2	1,44	10,20	9,7	94,1
1934	28,2	8,8	3,2	0,8	10,2	0,64	2,56	4,0	16,0
1935	35,3	10,4	10,3	2,4	106,1	5,76	24,72	12,7	161,3
1936	29,3	8,0	4,3	0	18,5	0	0	4,3	18,5
1937	30,5	9,7	5,5	1,7	30,2	2,89	9,35	7,2	51,8
—			—21,5	—6,9			—0,78	—27,8	
+			39,5	7,5			88,52	46,4	
Сумма	318,0	96,6	18,0	0,6	404,0	22,69	87,74	18,6	602,1
$N=12$			$X_{(1)}$	$Y_{(1)}$	$X_{(2)}$	$Y_{(2)}$	(XY)		

Расчеты становятся менее громоздкими, если каждое из чисел столбца 2 уменьшить на 25, а числа столбца 3 — на 8; полученные новые числа записаны в столбцах 4 и 5. Для проверки правильности вычислений на этой стадии можно воспользоваться тем, что если начало смещается на величину a, т. е. значения x^0 заменяются значениями $x = x^0 - a$, то после суммирования должно иметь место равенство

$$X_{(1)} = \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^0 - na = X_{(1)}^0 - na. \qquad (8.22)$$

В нашем случае $a = 25$, $n = 12$, так что должно быть

$$X_{(1)} = X_{(1)}^0 - 25 \cdot 12;$$

действительно,

$$18{,}0 = 318{,}0 - 300.$$

Аналогично

$$Y_{(1)} = Y^0_{(1)} - nb; \tag{8.23}$$

в нашем случае

$$0{,}6 = 96{,}6 - 12 \cdot 8 = 96{,}6 - 96.$$

По полученным в таблице данным имеем:

$$\sum_{xx} = 404{,}0 - \frac{18{,}0^2}{12} = 404{,}0 - 27{,}0 = 377{,}0;$$

$$\sum_{yy} = 22{,}60 - \frac{0{,}6^2}{12} = 22{,}60 - 0{,}03 = 22{,}57;$$

$$\sum_{xy} = 87{,}74 - \frac{18{,}0 \cdot 0{,}6}{12} = 87{,}74 - 0{,}90 = 86{,}84,$$

так что

$$r = \frac{86{,}84}{\sqrt{377{,}0 \cdot 22{,}57}} = 0{,}941.$$

Все расчеты можно производить на линейке; это обеспечивает вполне достаточную точность.

Столбцы 9 и 10 служат для проверки правильности вычислений: должно выполняться равенство

$$\sum (x + y)^2 = \sum x^2 + 2 \sum xy + \sum y^2 = X_{(2)} + 2\,(XY) + Y_{(2)};$$

в данном случае $404{,}0 + 2 \cdot 87{,}74 + 22{,}6 = 602{,}1$, что совпадает с $\sum (x + y)^2$. Кроме того, $\sum (x + y) = X_{(1)} + Y_{(1)}$. Если обнаружится расхождение, надо произвести проверку по формуле $(x + y)^2 = x^2 + 2xy + y^2$ для каждой строки; например, для пятой строки $106{,}1 = 68{,}9 + 2 \cdot 16{,}6 + 4{,}0$.

Пример 7. В примере 10 гл. 2 мы нашли LD_{50} для мышей при помощи графической обработки пробитов. Считая, что такой способ не обеспечивает достаточной точности, найдем LD_{50} из уравнения регрессии

$$\bar{x}_y - \bar{x} = b_{x/y}\,(y - \bar{y}).$$

Составим табл. 114; перенесем в нее исходные данные из табл. 23 и заменим для удобства величины x величинами $x^0 = x - 2{,}500$. По данным этой таблицы вычисляем:

$$\sum_{xy} = -0{,}029 - \frac{0{,}944\,(-2{,}26)}{5} = -0{,}029 + 0{,}427 = 0.398;$$

$$\sum_{yy} = 4{,}93 - \frac{(-2{,}26)^2}{5} = 4{,}93 - 1{,}02 = 3{,}91,$$

так что

$$b_{x/y} = \frac{\Sigma_{xy}}{\Sigma_{yy}} = \frac{0{,}398}{3{,}91} = 0{,}1018.$$

В нашем случае ищется $\bar{x}_y$ при $y = 0$, поэтому уравнение регрессии принимает вид

$$\bar{x}_{y=0} = \bar{x} - b_{x/y}\bar{y}.$$

Так как

$$\bar{x} = 2{,}500 + \frac{X_{(1)}}{n} = 2{,}500 + \frac{0{,}944}{n} = 2{,}689;$$

$$\bar{y} = \frac{Y_{(1)}}{n} = \frac{-2{,}26}{5} = -0{,}452,$$

то

$$\bar{x}_{y=0} = 2{,}689 - 0{,}1018\,(-0{,}452) = 2{,}689 + 0{,}046 = 2{,}735,$$

чему соответствует $LD_{50} = 543$ *рентген*.

Таблица 114

$x' = x - 2{,}500$	y	y^2	$x'y$
0,044	—1,87	3,50	—0,082
0,128	—0,90	0,81	—0,115
0,199	—0,30	0,09	—0,060
0,260	0,08	0,00	0,000
0,313	0,73	0,53	0,228
0,994	—2,26	4,93	—0,029
$X_{(1)}$	$Y_{(1)}$	$Y_{(2)}$	(XY)

Мы видим, что отличие этого результата от полученного ранее графически ($x_{50} = 2{,}737$, $LD_{50} = 545$) незначительно. Поэтому в данном случае вычислительная работа, связанная с нахождением уравнения регрессии, не оправдывается достигнутым уточнением результата.

§ 5. Доверительный интервал для коэффициента корреляции. Сравнение коэффициентов корреляции

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 $(\bar{x} - \hat{x})/s_x$ 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 коэффициент ρ к единице (рис. 48); в то же время именно последний случай представляет наибольший практический интерес.

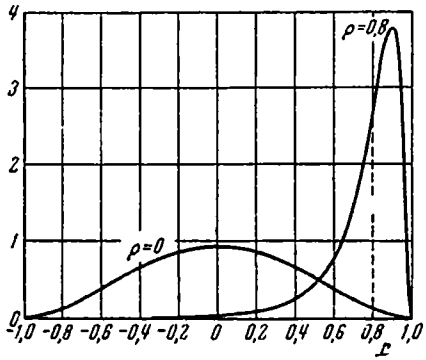

Рис. 48

Чтобы обойти это затруднение, было предложено (Р. Фишером) ввести вспомогательную величину

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}, \qquad (8.24)$$

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. 49 изображен график плотности распределения z при $\rho = 0$ и $\rho = 0{,}8$). Доказывается также, что стандартная ошибка величины z равна

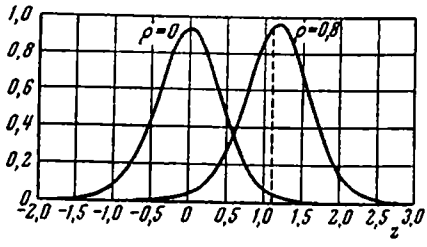

Рис. 49

$$\sigma_z = \frac{1}{\sqrt{n-3}}, \qquad (8.25)$$

где n — объем выборки.

Для перехода от r к z и обратно составлены таблицы, облегчающие вычисления (табл. XXV и XXVI Приложений).

Пример 8. При изучении корреляции между привесом свиней (в стаде из 500 особей) и количеством использованного корма получено значение $r = 0{,}814$. Каковы доверительные границы для ρ?

По формуле (8.25) получаем

$$\sigma_z = \frac{1}{\sqrt{500-3}} = \frac{1}{\sqrt{497}} = 0{,}045.$$

Так как значению $r = 0{,}814$ соответствует $z = 1{,}139$ (по табл. XXV Приложений — с применением интерполяции), то

$$z - u_{99}\sigma_z = 1{,}139 - 2{,}58 \cdot 0{,}045 = 1{,}023;$$
$$z + u_{99}\sigma_z = 1{,}139 + 2{,}58 \cdot 0{,}045 = 1{,}255.$$

Этим значениям z соответствуют (по табл. XXVI Приложений) значения $r_н = 0{,}771$; $r_в = 0{,}850$; это и будут 99%-ные доверительные границы генерального коэффициента корреляции.

Пример 9. Для двух групп свиней одной породы ($n' = 80$, $n'' = 50$) найдены коэффициенты корреляции между привесом и количеством использованного корма: $r' = 0{,}830$; $r'' = 0{,}794$. Найдем среднее по обеим группам значение коэффициента корреляции и доверительные границы для ρ.

Пользуясь табл. XXV Приложений, находим

$$z' = 1{,}188; \quad z'' = 1{,}081.$$

Усреднение производим, взвешивая по обратным дисперсиям $\left(\frac{1}{\sigma^2_{z'}} = n' - 3 = 77; \quad \frac{1}{\sigma^2_{z''}} = n'' - 3 = 47\right)$:

$$z = \frac{77 \cdot 1{,}188 + 47 \cdot 1{,}081}{77 + 47} = \frac{142{,}28}{124} = 1{,}147.$$

Далее вычисляем

$$\sigma_z = \sqrt{\sigma^2_{z'} + \sigma^2_{z''}} = \sqrt{\frac{1}{n'-3} + \frac{1}{n''-3}} =$$
$$= \sqrt{0{,}0130 + 0{,}0213} = 0{,}185.$$

Значит,

$$z_н = 1{,}147 - 2{,}58 \cdot 0{,}185 = 1{,}147 - 0{,}478 = 0{,}669;$$
$$z_в = 1{,}147 + 2{,}58 \cdot 0{,}185 = 1{,}147 + 0{,}478 = 1{,}625.$$

Теперь по табл. XXVI Приложений получаем окончательно

$$r_н = 0{,}585; \quad r_в = 0{,}925.$$

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

$$u_{z'-z''} = \frac{|z' - z''|}{\sigma_{z'-z''}} > u_\alpha, \tag{8.26}$$

где согласно (8.25)

$$\sigma_{z'-z''} = \sqrt{\frac{1}{n'-3} + \frac{1}{n''-3}}. \tag{8.27}$$

Пример 10. В примере 9 расчет среднего для двух выборок ($n' = 80$; $n'' = 50$) коэффициента корреляции исходил из допущения, что различие между обоими выборочными коэффициентами корреляции ($r' = 0{,}830$; $r'' = 0{,}794$) незначимо. Верно ли это допущение?

Как было найдено в примере 9, $z' = 1{,}188$; $z'' = 1{,}081$ и $\sigma_{z'-z''} = 0{,}185$. Поэтому

$$u_{z'-z''} = \frac{1{,}188-1{,}081}{0{,}185} = \frac{0{,}107}{0{,}185} = 0{,}58,$$

в то время как $u_{0б} = 1{,}96$; значит, различие незначимо.

Критерий (8.26) 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 $\sigma_{z(выб)-z(теор)}$ дается просто формулой (8.25). Таким образом, коэффициент корреляции может считаться значимо отличным от нуля, если

$$z\sqrt{n-3} > u_\alpha, \tag{8.28}$$

или, иными словами, если

$$z > \frac{u_\alpha}{\sqrt{n-3}} \; (\equiv z_\alpha(n)), \tag{8.29}$$

причем здесь применяется двусторонний критерий. Так как r и z связаны взаимно однозначно равенством (8.24), то можно вычислить значения r_α, соответствующие каждому из значений z_α. Критические значения r_α даны в табл. XXVII Приложений.

§ 6. Критерий линейности корреляции

Как было указано в предыдущих параграфах, уравнение регрессии в виде (8.1), а также коэффициент корреляции ρ могут употребляться только в том случае, если связь между признаками линейна. Так как при линейной связи корреляционные отношения $\eta_{y/x}$ и $\eta_{x/y}$ совпадают между собой (и совпадают с вычисленным по формуле (8.17) коэффициентом корреляции), то несовпадение между собой значений $\eta_{y/x}$ и $\eta_{x/y}$ (или несовпадение хотя бы одного из них с ρ) может служить указанием на нелинейность связи. Если (как это обычно бывает) мы имеем дело не с какой-либо генеральной совокупностью, а с выборкой, то встает вопрос о значимости обнаруженного несовпадения.

Самым простым было бы применение t-критерия в виде

$$t = \frac{e_{y/x} - e_{x/y}}{s\{e_{y/x} - e_{x/y}\}}, \quad \text{или } t = \frac{e - r}{s\{e - r\}},$$

считая

$$s\{e_{y/x} - e_{x/y}\} = \sqrt{s^2\{e_{y/x}\} + s^2\{e_{x/y}\}};$$
$$s\{e - r\} = \sqrt{s^2\{e\} + s^2\{r\}}.$$

Однако в предыдущем параграфе мы уже говорили, что распределение выборочных r сильно отличается от нормального (особенно при ρ, близких к ± 1); в равной мере это относится и к выборочным $e_{y/x}$ и $e_{x/y}$, поскольку они также ограничены значениями (0, 1). Поэтому в данном случае t-критерий оказывается неприменимым.

На помощь приходит дисперсионный анализ. Идея применения для этой цели дисперсионного анализа такова (см. схему на стр. 296). Полная вариация по переменной y, т. е. величина

$$S = \sum_{x,\,y} n_{xy}\,(y - \hat{y})^2 = N\sigma_y^2,$$

может быть разложена на две части [см. формулу (*) на стр. 279]: на вариацию между строками

$$S_{\text{м. с}} = \sum_{x} n_x\,(\hat{y}_x - \hat{y})^2 = N\sigma^2\{\hat{y}_x - \hat{y}\} \qquad (*)$$

и на вариацию внутри строк

$$S_{\text{в. с}} = N\langle\sigma^2\{y - \hat{y}_x\}\rangle.$$

Вариация между строками $S_{\text{м. с}}$ отражает наличие регрессии, т. е. систематическое изменение $\hat{y}_x$ при переходе от одного

Полная вариация

$$S = \sum_{x,\,y} n_{xy}(y - \hat{y})^2 = N\sigma_y^2$$

$$f = N - 1$$

Вариация между строками

$$S_{\text{м.с}} = \sum_x n_x(\hat{y_x} - \hat{y})^2 = N\sigma^2\{\hat{y}_x - \hat{y}\} = N\sigma_y^2\eta_{y/x}^2$$

$$f_{\text{м.с}} = k_x - 1$$

Вариация внутри строк

$$S_{\text{в.с}} = S - S_{\text{м.с}} = N\sigma_y^2 - N\sigma^2\{\hat{y}_x - \hat{y}\} = N\sigma_y^2(1 - \eta_{y/x}^2)$$

$$f_{\text{в.с}} = f - f_{\text{м.с}} = N - k_x$$

Вариация между строками за счет линейной регрессии

$$(S_{\text{м.с}})_{\text{лин}} = N\sigma_y^2\rho^2$$

Вариации между строками за счет нелинейности

$$(S_{\text{м.с}})_{\text{нелин}} = N\sigma_y^2(\eta_{y/x}^2 - \rho^2)$$

$$(f_{\text{м.с}})_{\text{нелин}} = f_{\text{м.с}} - 1 = k_x - 2$$

Дисперсия по фактору нелинейности

$$s^2_{\text{нелин}} = \frac{(S_{\text{м.с}})_{\text{нелин}}}{(f_{\text{м.с}})_{\text{нелин}}} = \frac{N\sigma_y^2(\eta_{y/x}^2 - \rho^2)}{k_x - 2}$$

Дисперсия внутри строк (остаточная)

$$s^2_{\text{ост}} = S_{\text{в.с}} : f_{\text{в.с}} = \frac{N\sigma_y^2(1 - \eta_{y/x}^2)}{N - k_x}$$

значения x к другому. В общем случае

$$S_{\text{м. с}} = N\sigma_y^2 \frac{\sigma^2\{\hat{y}_x - \hat{y}\}}{\sigma_y^2} = N\sigma_y^2\eta_{y,x}^2,$$

согласно определению (**) на стр. 279; если же регрессия линейна, то в выражение (*) можно подставить

$$\hat{y}_x - \hat{y} = \beta_{y/x}(x - \hat{x})$$

из уравнения линейной регрессии, и тогда, пользуясь формулой (5.15), получим

$$(S_{\text{м. с}})_{\text{лин}} = N\sigma_y^2\rho^2.$$

Поэтому можно считать, что часть вариации между строками, обязанная отклонению регрессии от линейной, равна

$$(S_{\text{м. с}})_{\text{нелин}} = S_{\text{м. с}} - (S_{\text{м. с}})_{\text{лин}} = N\sigma_y^2(\eta_{y/x}^2 - \rho^2).$$

Эта «нелинейная» часть вариации определяет оценку дисперсии «по фактору нелинейности»:

$$s^2_{\text{нелин}} = (S_{\text{м. с}})_{\text{нелин}} \quad (f_{\text{м. с}})_{\text{нелин}}.$$

Как видно из (*), величина $S_{\text{м. с}}$ имеет $k_x - 1$ степеней свободы, где k_x — число строк (т. е. число градаций признака x). Поскольку при переходе от $S_{\text{м. с}}$ к $(S_{\text{м. с}})_{\text{нелин}}$ было использовано еще одно соотношение — уравнение линейной регрессии, то $(S_{\text{м. с}})_{\text{нелин}}$ имеет $k_x - 2$ степеней свободы. Поэтому получаем

$$s^2_{\text{нелин}} = \frac{N\sigma^2(\eta_{y/x}^2 - \rho^2)}{k_x - 2}. \qquad (**)$$

Эту величину надо сравнивать с остаточной дисперсией (по терминологии гл. 5). Роль остаточного, т. е. полностью случайного, фактора здесь играет вариация внутри строк $S_{\text{в. с}}$. В соответствии со сказанным выше

$$S_{\text{в. с}} = S - S_{\text{м. с}} = N\sigma_y^2 - N\sigma^2\{\hat{y}_x - \hat{y}\};$$

это можно переписать в виде

$$S_{\text{ост}} = S_{\text{в. с}} = N\sigma_y^2\left(1 - \frac{\sigma^2\{\hat{y}_x - \hat{y}\}}{\sigma_y^2}\right) = N\sigma_y^2(1 - \eta_{y/x}^2).$$

Очевидно, число степеней свободы этой величины равно

$$f_{\text{ост}} = f - f_{\text{м. о}} = (N-1) - (k_x - 1) = N - k_x;$$

поэтому остаточная дисперсия будет

$$s^2_{\text{ост}} = \frac{S_{\text{ост}}}{f_{\text{ост}}} = \frac{N\sigma^2_y (1 - \eta^2_{y/x})}{N - k_x}. \qquad (***)$$

Таким образом, оценка значимости нелинейного характера корреляции сводится к оценке значимости величины

$$F = \frac{s^2_{\text{нелин}}}{s^2_{\text{ост}}} = \frac{\eta^2_{y/x} - \rho^2}{k_x - 2} : \frac{1 - \eta^2_{y/x}}{N - k_x}$$

(после сокращения на $N\sigma^2_y$). Если заменить N на объем выборки n, а $\eta_{y/x}$ и ρ — их выборочными оценками, то окончательно получится критерий нелинейности корреляции

$$F = \frac{e^2_{y/x} - r^2}{1 - e^2_{y/x}} \quad \frac{n - k_x}{k_x - 2}. \qquad (8.30)$$

Надо только помнить, что при сравнении с критическим значением $F_\alpha\,(f_1; f_2)$ следует считать $f_1 = k_x - 2$, $f_2 = n - k_x$.

Аналогично сравниваются $e_{x/y}$ и r. Корреляция нелинейна, если условие $F > F_\alpha$ выполняется для б о л ь ш е г о из корреляционных отношений.

П р и м е р 11. Проверим линейность корреляции между высотой сосны и длиной ее вершинного побега (из табл. 109).

Используя найденные выше значения $e^2_{x/y} = 0{,}633$, $e^2_{y/x} = 0{,}639$, $r = 0{,}638$ и учитывая, что $n = 330$, $k_x = 10$, получаем по формуле (8.30):

$$F = \frac{0{,}639 - 0{,}638}{1 - 0{,}639} \cdot \frac{320}{8} = 0{,}11.$$

Это много меньше, чем $F_{0{,}5}\,(8; 320) = 1{,}97$ (и даже меньше единицы). Поэтому гипотеза о линейности корреляции не опровергается и таким образом можно пользоваться уравнением линейной регрессии.

То, что r оказалось больше, чем $e_{x/y}$, не должно нас смущать: ведь r и $e_{x/y}$ — не действительные значения параметров, а их выборочные оценки; значения параметров лежат внутри соответствующих доверительных интервалов, так что противоречия с необходимым условием $\rho \leqslant \eta_{x/y}$ не будет.

Пример 12. Для данных из табл. 108 (корреляция между диаметром x и высотой y сосны) в примере 5 были получены значения: $e^2_{x/y} = 0{,}529$, $e^2_{y/x} = 0{,}482$. Чтобы найти r^2, вычислим сначала

$$(XY) = \sum_x Y_x x = (-19)(-2) + (-64)(-1) + (-2)\cdot 0 + $$
$$+ 37\cdot 1 + 31\cdot 2 + 19\cdot 3 + 9\cdot 4 = 38 + 64 + 0 + 37 + 62 + $$
$$+ 36 = 294;$$
$$(XY) = \sum_y X_y y = (-1)(-4) + (-5)(-3) + (-26)(-2) +$$
$$+ (-33)(-1) + (-19)\cdot 0 + 10\cdot 1 + 23\cdot 2 + 22\cdot 3 + 12\cdot 4 +$$
$$+ 4\cdot 5 = 4 + 15 + 52 + 33 + 10 + 46 + 66 + 48 + 20 = 294$$

(совпадение обоих значений показывает, что вычисления произведены правильно); затем найдем

$$\Sigma_{xy} = (XY) - \frac{X_{(1)}Y_{(1)}}{n} = 294 - \frac{(-13)\cdot 11}{250} = 294{,}6.$$

Значения $\Sigma_{xx} = 312{,}3$ и $\Sigma_{yy} = 568{,}5$ также были получены в примере 5. Теперь имеем

$$r^2 = \frac{(294{,}6)^2}{312{,}3\cdot 568{,}5} = 0{,}489.$$

Таблица 115

x \ y	0	3	6	9	12	15	18	21	24	27	30	n_x	$\bar{y}_x$
0								1	1			2	22,5
5						1	1	3	3	1		9	21,7
10				1	2	7	7	7	4	1		29	18,4
15			1	3	13	17	15	8	2	1		60	16,0
20		1	6	16	21	21	17	9	3	1		95	14,1
25	2	7	15	20	23	18	15	5	2	1	2	110	12,2
30		1	6	16	21	21	17	9	3	1		95	14,1
35			1	3	13	17	15	8	2	1		60	16,0
40				1	2	7	7	7	4	1		29	18,4
45						1	1	3	3	1		9	21,7
50								1	1			2	22,5
n_y	2	9	29	60	95	110	95	60	29	9	2	500	
$\bar{x}_y$	25	25	25	25	25	25	25	25	25	25	25		

Так как $e_{x/y} > e_{y/x}$, то мы сравниваем $e_{x/y}$ и r. В данном случае $e^2_{x/y} - r^2 = 0{,}040$; $1 - e^2_{x/y} = 0{,}471$; $k_y - 2 = 10 - 2 = 8$; $n - k_y = 250 - 10 = 240$. Поэтому

$$F = \frac{0{,}040 \cdot 240}{0{,}471 \cdot 8} = 2{,}55.$$

Согласно табл. XIV Приложений $F_{05}(8; 240) = 1{,}98$ и $F_{01}(8; 240) = 2{,}59$; следовательно, нелинейность можно считать значимой (F довольно близко к F_{01}).

Заметим, что в биологических совокупностях нелинейная корреляция встречается довольно часто.

Надо сказать, что при наличии некоторого опыта можно делать качественное заключение о линейности или нелинейности связи просто по виду корреляционной решетки; это видно из сравнения табл. 109 и 110 (выделение «нулевого креста» $x = 0$, $y = 0$ вносит бо́льшую ясность). В то же время следует избегать выносить суждение по этому вопросу, основываясь только на характере распределений n_x и n_y. Схематизированная табл. 115 демонстрирует это с большой убедительностью: хотя распределение по каждому из признаков симметрично (и даже похоже на нормальное), связь отнюдь не линейна (рис. 50).

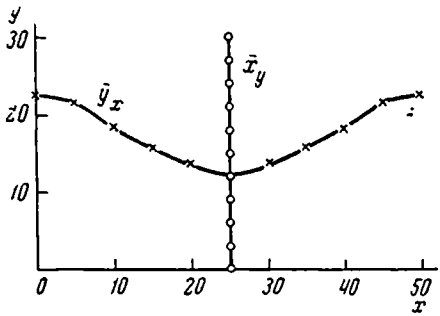

Рис. 50

И еще одно замечание: вывод об отсутствии значимой нелинейности регрессии относится только к непосредственно исследованному диапазону значений x и y; экстраполировать этот вывод за пределы указанной области нельзя.

§ 7. Доверительная зона регрессии

Выборочное значение коэффициента регрессии $b_{y/x}$ является оценкой соответствующего генерального коэффициента $\beta_{y/x}$ и отличается от него в среднем на $\sigma_{b_{y/x}}$. Это значит, что «истинная» линия регрессии заключена (при больших объемах выборки с вероятностью 68,3%) внутри пары вертикальных углов, образованных пересечением в точке $(\bar{x}, \bar{y})$ двух прямых PP и QQ с наклонами $b_{y/x} - \sigma_{b_{y/x}}$ и $b_{y/x} + \sigma_{b_{y/x}}$ (заштрихованная часть на рис. 51). Поэтому для каждого значения x «истинное» значение выравнен-

ного условного среднего заключено (с вероятностью 68,3%) в интервале $y_x \pm \sigma_{b_{y/x}}(x - \bar{x})$; на рис. 51 этот интервал изображен отрезком AB[1].

Однако имеется некоторая неопределенность не только в наклоне «истинной» линии регрессии $\hat{y}_x$, 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

$$\sqrt{\frac{1}{n}\sum_x\sum_y (y - \hat{y}_x)^2} = \sigma\{y - \hat{y}_x\}, \qquad (*)$$

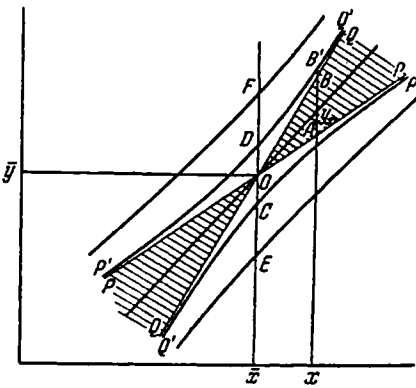

Рис. 51

а поэтому стандартная ошибка в расположении линии регрессии (в смысле смещения по вертикали) будет

$$\sigma_{y_x}^{(\text{верт})} = \frac{\sigma\{y - \hat{y}_x\}}{\sqrt{n}}; \qquad (**)$$

интервал $y \pm \sigma_{y_x}^{(\text{верт})}$ изображается отрезком CD на рис. 51.

Величина $\sigma_{y_x}^{(\text{верт})}$ имеет размерность y; например, при изучении регрессии веса телят на их возраст $\sigma_{y_x}^{(\text{верт})}$ имеет размерность веса. Что касается величины $\sigma_{b_{y/x}}$, то она должна иметь ту же размерность, что и $b_{y/x}$, т. е. размерность y/x. Величина с такой размерностью получится, если мы разделим $\sigma_{y_x}^{(\text{верт})}$ на σ_x, т. е. примем

$$\sigma_{b_{y/x}} = \frac{\sigma_{y_x}^{(\text{верт})}}{\sigma_x}. \qquad (8.31)$$

Очевидно, полная стандартная ошибка истинного условного среднего должна вычисляться как

$$\sigma_{y_x} = \sqrt{[\sigma_{b_{y/x}}(x - \bar{x})]^2 + [\sigma_{y_x}^{(\text{верт})}]^2} \qquad (***)$$

[1] Разумеется, все сказанное здесь и в дальнейшем о регрессии y на x относится в равной мере к регрессии x на y. В последнем случае все отрезки будут откладываться по горизонтали, а в формулах надо везде заменять y на x, а x на y.

или, подставив значение $\sigma_{b_{y/x}}$ из (8.31),

$$\sigma_{y_x} = \sigma_{y_x}^{(\text{верт})} \sqrt{\left(\frac{x-\bar{x}}{\sigma_x}\right)^2 + 1}. \qquad (8.32)$$

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. 51. 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).

Чтобы практически построить доверительную зону регрессии, надо найти оценку для величины $\sigma_{y_x}^{(\text{верт})}$; из формулы (**) видно, что дело сводится к нахождению оценки для $\sigma\{y - \hat{y}_x\}$. В принципе это можно сделать, пользуясь формулой (*), но это довольно громоздко в вычислительном отношении. Поэтому мы изберем другой путь.

Из схемы на стр. 296 видно, что величина $\sigma^2\{y - \hat{y}_x\}$, которая представляет собой дисперсию значений y внутри строк, оценивается величиной

$$s^2_{\text{ост}} = \frac{n\sigma_y^2(1-\eta^2_{y/x})}{n-k_x} = \frac{(1-\eta^2_{y/x})\,\Sigma_{yy}}{n-k_x}$$

Однако эта оценка может быть принята лишь в том случае, если регрессия существенно нелинейна, т. е. если величина

$$s^2_{\text{нелин}} = \frac{n\sigma_y^2(\eta^2_{y/x}-\rho^2)}{k_x-2} = \frac{(\eta^2_{y/x}-\rho^2)\,\Sigma_{yy}}{k_x-2}$$

значимо превышает $s^2_{\text{ост}}$ (см. предыдущий параграф). Так как здесь мы рассматриваем случай линейной регрессии, при которой заведомо $s^2_{\text{нелин}}$ не превышает значимо $s^2_{\text{ост}}$, то обе эти величины должны считаться оценками одной и той же дисперсии $\sigma^2\{y-\hat{y}_x\}$. Очевидно, лучшей оценкой будет результат усреднения этих двух оценок, произведенного взвешиванием по числу степеней свободы, т. е. величина

$$s^2 = \frac{(1-\eta^2_{y/x})\,\Sigma_{yy} + (\eta^2_{y/x}-\rho^2)\,\Sigma_{yy}}{(n-k_x)-(k_x-2)},$$

или, если привести подобные члены и заменить ρ^2 его оценкой r^2,

$$s^2 = \frac{(1 - r^2)\,\Sigma_{yy}}{n - 2}. \tag{8.33}$$

Следовательно, оценкой $\sigma_{y_x}^{(\text{верт})}$ будет

$$s_{y_x}^{(\text{верт})} = \sqrt{\frac{(1 - r^2)\,\Sigma_{yy}}{n\,(n - 2)}}. \tag{8.34}$$

В соответствии с формулами (8.31) и (**)

$$\sigma^2_{b_{y/x}} = \frac{\sigma^2\{y - \hat{y}_x\}}{n\sigma_x^2} = \frac{\sigma^2\{y - \hat{y}_x\}}{\Sigma_{xx}}. \tag{8.35}$$

Если использовать в качестве оценки величины $\sigma^2\{y - \hat{y}_x\}$ величину s^2 из (8.33), то получится оценка

$$s_{b_{y/x}} = \sqrt{\frac{1 - r^2}{n - 2} \cdot \frac{\Sigma_{yy}}{\Sigma_{xx}}}. \tag{8.36}$$

При помощи этой величины можно проверять значимость регрессии, т. е. правильность гипотезы $\beta_{y/x} = 0$. Эта проверка основывается на том, что величина

$$t = \frac{b_{y/x} - \beta_{y/x}}{s_{b_{y/x}}},$$

где $s_{b_{y/x}}$ дается формулой (8.36), имеет распределение Стьюдента с $n - 2$ степенями свободы.

Аналогично формулам (8.34) и (8.36) имеем:

$$s_{x_y}^{(\text{гориз})} = \sqrt{\frac{(1 - r^2)\,\Sigma_{xx}}{n\,(n - 2)}}; \tag{8.37}$$

$$s_{b_{x/y}} = \sqrt{\frac{1 - r^2}{n - 2} \cdot \frac{\Sigma_{xx}}{\Sigma_{yy}}}. \tag{8.38}$$

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.

Форма доверительной зоны для $y(x)$ существенно зависит от статистических свойств аргумента x. Дело в том, что величина x может быть либо случайно варьирующей, либо наперед задаваемой. Так, в примерах предыдущих параграфов значения x варьировали независимо от воли исследователя: в примере 1 (стр. 267) деревья просматривались (с измерением их диаметров и высот), по-видимому, в порядке их расположения; в примере 6 (стр. 288) урожайности как картофеля (y), так и пшеницы (x) располагались в порядке следования годов. Наряду с этим, однако, возможна другая постановка регрессионной и корреляционной задачи. 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.

Иногда оба рассмотренных случая различают терминологически. Именно, названия «регрессия», «регрессионный анализ» сохраняют лишь за тем случаем, когда величина x принимает заранее заданные определенные значения, т. е. случайные вариации этой величины отсутствуют. Случай же, когда x есть, как и y, случайно варьирующая величина, обозначают термином «конфлуэнция» и соответственно говорят о «конфлуэнтном анализе». Мы не будем здесь проводить этого различия, поскольку построение линии регрессии и вычисление показателей корреляции производится в обоих случаях одинаково. Однако доверительные зоны регрессии (для результатов отдельных измерений, но не для условных средних) имеют различную форму.

Мы начнем со случая строго задаваемых значений x как более простого, причем, как и ранее, ограничимся линейной регрессией.

При нахождении доверительной зоны для $y(x)$ нужно в выражении (***) этого параграфа учесть также варьирование отдельных значений y около условных средних $\bar{y}_x$. Это варьирование отражает величина $\sigma\{y - \bar{y}_x\}$, равная, согласно (**), $\sqrt{n}\,\sigma_{v_x}^{(\text{верт})}$.

Тогда вместо (8.32) будем иметь

$$\sigma_{y(x)} = \sigma_{y_x}^{\text{верт}} \sqrt{\left(\frac{x-\bar{x}}{\sigma_x}\right)^2 + 1 + n}, \tag{8.39}$$

так что границы доверительной зоны для $y(x)$ будут пересекать вертикаль $x = \bar{x}$ в точках (E и F на рис. 51), отстоящих от $\bar{y}$ на $\pm t_P \sigma_{y_x}^{(\text{верт})} \sqrt{n+1}$; используя оценку $\sigma_{y_x}^{(\text{верт})}$ из (8.33), имеем

$$[t_P s_{y(x)}]_{x=\bar{x}} = t_P s_y \sqrt{\frac{n+1}{n-2}(1-r^2)}. \tag{8.40}$$

Пример 13. В табл. 116 записаны результаты серии 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.

Таблица 116

Доза, 10^3 рентген (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	60	61	63		

Здесь численности, отвечающие отдельным значениям аргумента x, целиком 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 $y(x)$.

Чтобы получить корреляционную решетку обычного вида, надо произвести группировку вариант. Результат этой группировки представлен в табл. 117, там же даны все промежуточные расчеты.

Таблица 117

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	$Y_{(2)}$					

Теперь находим, как обычно, по формулам § 4 настоящей главы

$$\sum_{xy} = -292 - \frac{(-4)(-10)}{61} = -292 - 0{,}7 = -292{,}7;$$

$$\sum_{xx} = 248 - \frac{(-4)^2}{61} = 248 - 0{,}3 = 247{,}7;$$

$$\sum_{yy} = 360 - \frac{(-10)^2}{61} = 360 - 1{,}6 = 358{,}4.$$

Тогда

$$r^2 = \frac{(-292{,}7)^2}{247{,}7 \cdot 358{,}4} = 0{,}965.$$

Согласно формуле (8.34), 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 точность могла пострадать из-за группировки вариант — в данном случае довольно грубой. Поэтому лучше провести расчет, не делая группировки. Такой расчет показан в табл. 118. 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.

Таблица 118

Доза 10^3 рентген (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
	$Y_{(2)} = \Sigma y^2 = 9190$	−166 $Y_{(1)}$	−2102 (XY)	61 n		240 $X_{(1)}$	1192 $X_{(2)}$

Используя полученные в табл. 118 значения, находим:

$$\sum_{xy} = -2102 - \frac{(-166)\,240}{61} = -2102 + 653{,}1 = -1448{,}9;$$

$$\sum_{xx} = 1192 - \frac{240^2}{61} = 1192 - 944{,}3 = 247{,}7; \quad s_x^2 = \frac{247{,}7}{60} = 4{,}13;$$

$$\sum_{yy} = 9190 - \frac{(-166)^2}{61} = 9190 - 451{,}7 = 8738{,}3.$$

Поэтому

$$b_{y/x} = \frac{-1448{,}9}{247{,}7} = -5{,}85; \quad r^2 = \frac{(-1148{,}9)^2}{247{,}7 \cdot 8738{,}3} = 0{,}9699.$$

Кроме того,

$$\bar{x} = \frac{240}{61} = 3{,}93; \quad \bar{y} = 80 + \frac{(-166)}{61} = 77{,}28.$$

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. 52 эта точка обведена кружком). Затем от какой-нибудь точки в левом верхнем углу графика (на рис. 52 это точка I) откладываем вправо 10

единиц в масштабе x и от найденной точки II откладываем вниз (так как $b_{y/x} < 0$) отрезок длиной $|b_{y/x}| \cdot 10 = 5{,}85 \cdot 10 = 58{,}5$ единиц в масштабе y (точка III). Наконец, проводим через точку $(\bar{x}, \bar{y})$ прямую, параллельную прямой, соединяющей точки I и III. Крестиками на рис. 52 изображены эмпирические условные средние $\bar{y}_x$, взятые из табл. 118.

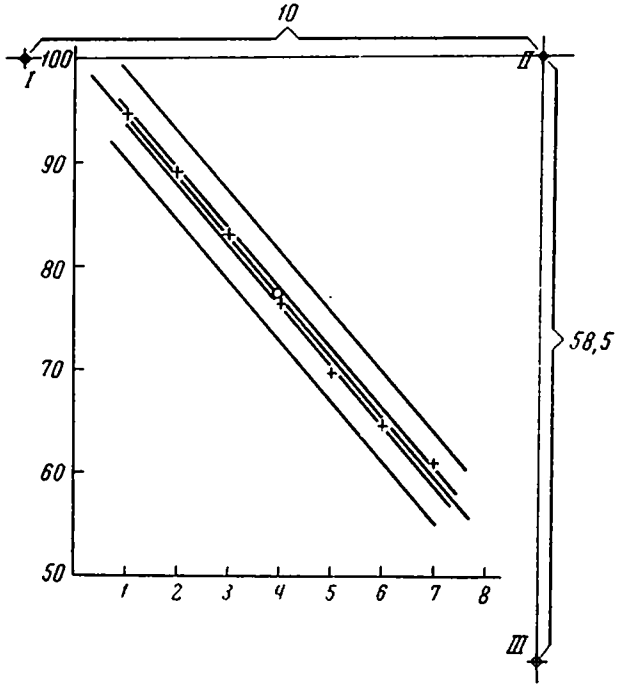

Рис. 52

Теперь построим доверительные зоны для $\hat{y}_x$ и $y(x)$. По формуле (8.34) находим

$$s_{y_x}^{(\text{верт})} = \sqrt{\frac{0{,}0301 \cdot 8738{,}3}{61 \cdot 59}} = \sqrt{0{,}0731} = 0{,}270.$$

Для максимального в данной задаче значения $|x - \bar{x}| \approx 3$ имеем

$$\frac{(x-\bar{x})^2}{s_x^2} = \frac{9}{4{,}13} = 2{,}18.$$

Если выбрать $P = 95\%$, то $t_P = 2{,}00$ (при $n - 2 = 59$). Поэтому

$$[t_P s_{y_x}]_{x-\bar{x}=0} = 2{,}00 \cdot 0{,}270 = 0{,}54;$$

$$[t_P s_{\hat{y}_x}]_{x-\bar{x}=3} = 2{,}00 \cdot 0{,}270 \sqrt{2{,}18 + 1} = 0{,}96;$$

$$[t_P s_{y(x)}]_{x-\bar{x}=0} = 2{,}00 \cdot 0{,}270 \sqrt{61 + 1} = 4{,}25;$$

$$[t_P s_{y(x)}]_{x-\bar{x}=3} = 2{,}00 \cdot 0{,}270 \sqrt{61 + 1 + 2{,}18} = 4{,}33.$$

На вертикали $x = 3{,}94$ откладываем точки с ординатами $77{,}28 - 0{,}54 = 76{,}74$; $77{,}28 + 0{,}54 = 77{,}82$ для зоны $\hat{y}_x$ и точки с ординатами $77{,}28 - 4{,}25 = 73{,}03$; $77{,}28 + 4{,}25 = 81{,}53$ 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 зон (см. рис. 52). В данном случае это почти прямые, параллельные линии регрессии.

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].

Сделаем следующее замечание. Согласно формуле (8.31), величина $\sigma^2_{b_{x/y}}$ обратно пропорциональна дисперсии σ^2_x. Поэтому наклон прямой регрессии будет найден с тем меньшей неопределенностью, чем шире будет интервал значений x. Это обстоятельство может быть использовано для снижения $\sigma^2_{b_{y/x}}$, когда x не случайная, а задаваемая в эксперименте величина.

Если линейность регрессии не вызывает сомнения, то величина σ^2_x 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 выгодней (в смысле получения наименьшего значения $\sigma_{b_{y/x}}$) выбрать не точки 21, 24, 27, 30, 33, 36° С, а точки 21 и 36°С, производя по три измерения при каждой из этих температур. Однако, если линейность регрессии нуждается в проверке, то следует использовать равноотстоящие точки по всему интервалу температур.

---

[1] Но доверительные зоны для условных средних $\hat{y}_x$ и $\hat{x}_y$, как и в предыдущем случае, ограничены парами гипербол, которые строятся по формуле (8.32).

[2] См. Я. М. Лукомский (1961, гл. 12, § 3); Ю. В. Линник (1962).

§ 8. Сравнение двух линий регрессии

На практике иногда возникает задача сравнения двух линий регрессии, например, зависимость урожайности от количества удобрения для двух сортов или ход размножения бактерий в двух средах и т. д.

Уравнения сравниваемых линий регрессии запишем в обычном виде:

$$\tilde{y}_{x(1)} - \bar{y}_{(1)} = b_{(1)}\,(x - \bar{x}_{(1)});$$
$$\tilde{y}_{x(2)} - \bar{y}_{(2)} = b_{(2)}\,(x - \bar{x}_{(2)}).$$

Индекс y/x у коэффициентов регрессии b для простоты пропущен; 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

$$t = \frac{|\,b_{(1)} - b_{(2)}\,|}{s_{b_{(1)}-b_{(2)}}} \tag{8.41}$$

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 $\sigma^2_{(1)}$ и $\sigma^2_{(2)}$ находятся по формуле (8.33), а числа степеней свободы равны соответственно $n_{(1)} - 2$ и $n_{(2)} - 2$.

Если различие между $s^2_{(1)} = (1 - r^2_{(1)})\,\Sigma_{yy(1)}/(n_{(1)} - 2)$ и $s^2_{(2)} = (1 - r^2_{(2)})\,\Sigma_{yy(2)}/(n_{(2)} - 2)$ оказалось незначимым, можно приступить к вычислению t по формуле (8.41).

Так как, согласно (8.35), $\sigma^2_b = \sigma^2\{y - \hat{y}_x\}/\Sigma_{xx}$, то оценкой $\sigma^2_{b_{(1)}-b_{(2)}}$ будет

$$s^2_{b_{(1)}-b_{(2)}} = s^2_{b(1)} + s^2_{b(2)} = s^2\left(\frac{1}{\Sigma_{xx_{(1)}}} + \frac{1}{\Sigma_{xx_{(2)}}}\right),$$

или

$$s^2_{b_{(1)}-b_{(2)}} = s^2\,\frac{\Sigma_{xx(1)} + \Sigma_{xx(2)}}{\Sigma_{xx_{(1)}} \cdot \Sigma_{xx_{(2)}}}, \tag{8.42}$$

[1] Если распределение исходных величин не очень отклоняется от нормального.

где s^2 получается объединением соответствующих оценок для обеих прямых (коль скоро различие между ними незначимо). Объединение оценок дисперсии производится как обычно — взвешиванием по числу степеней свободы:

$$s^2 = \frac{(n_{(1)} - 2)\, s^2_{(1)} + (n_{(2)} - 2)\, s^2_{(2)}}{n_{(1)} + n_{(2)} - 4}. \qquad (8.43)$$

Последовательное применение формул (8.43), (8.42) и (8.41) дает значение t, которое сравнивается 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 § 4 гл. 4.

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, как в формуле (3.22) из § 5 гл. 3, обратные значения дисперсий $s^2_{b(1)}$ и $s^2_{b(2)}$; после сокращения на общую величину $\sigma^2\{y - \hat{y}_x\}$ получается

$$\bar{b} = \frac{\Sigma_{xx(1)} b_{(1)} + \Sigma_{xx(2)} b_{(2)}}{\Sigma_{xx(1)} + \Sigma_{xx(2)}} \qquad (8.44)$$

Оценка дисперсии этой величины (которая нужна для построения доверительного интервала для β) равна

$$s^2_{\bar{b}} = \frac{s^2}{\Sigma_{xx(1)} + \Sigma_{xx(2)}}, \qquad (8.45)$$

причем s^2 берется из (8.43).

Если линии регрессии оказались параллельными (т. е. имеют общий угловой коэффициент $\bar{b}$), то их уравнения примут вид:

$$\tilde{y}_{x(1)} - \bar{y}_{(1)} = \bar{b}\,(x - \bar{x}_{(1)});$$
$$\tilde{y}_{x(2)} - \bar{y}_{(2)} = \bar{b}\,(x - \bar{x}_{(2)}),$$

или, после некоторой перестановки,

$$\tilde{y}_{x(1)} = (\bar{y}_{(1)} - \bar{b}\bar{x}_{(1)}) + \bar{b}x;$$
$$\tilde{y}_{x(2)} = (\bar{y}_{(2)} - \bar{b}\bar{x}_{(2)}) + \bar{b}x.$$

Чтобы эти линии были не только параллельны, 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

$$\frac{\bar{y}_{(1)} - \bar{y}_{(2)}}{\bar{x}_{(1)} - \bar{x}_{(2)}} = b^*, \tag{8.46}$$

полученная подстановкой в (8.46) эмпирических значений $\bar{x}_{(1)}$, $\bar{x}_{(2)}$, $\bar{y}_{(1)}$, $\bar{y}_{(2)}$, не равна величине $\bar{b}$, полученной из (8.44), то прямые не совпадают. Поэтому вопрос о совпадении двух параллельных прямых регрессии можно решать при помощи t-критерия, сравнивая с табличным t_α значение

$$t - \frac{b^* - \bar{b}}{s_{b^* - \bar{b}}} \tag{8.47}$$

Величину $s_{b^* - \bar{b}}$ находим из равенства

$$s^2_{b^* - \bar{b}} = s^2_{b^*} + s^2_{\bar{b}}, \tag{8.48}$$

причем

$$s^2_{b^*} = \frac{s^2}{(\bar{x}_{(1)} - \bar{x}_{(2)})^2}\left(\frac{1}{n_{(1)}} + \frac{1}{n_{(2)}}\right), \tag{8.49}$$

а $s^2_{\bar{b}}$ вычисляется по формуле (8.45).

Если различие окажется значимым, то можно вычислить разность

$$\begin{aligned} d_{(1)-(2)} &= [\bar{y}_{(1)} - \bar{b}\,(x - \bar{x}_{(1)})] - [\bar{y}_{(2)} - \bar{b}\,(x - \bar{x}_{(2)})] = \\ &= (\bar{y}_{(1)} - \bar{y}_{(2)}) - \bar{b}\,(\bar{x}_{(1)} - \bar{x}_{(2)}), \end{aligned} \tag{8.50}$$

которую удобней представить в эквивалентной форме

$$d_{(1)-(2)} = (\bar{x}_{(1)} - \bar{x}_{(2)})\,(b^* - \bar{b}). \tag{8.51}$$

Оценка дисперсии этой величины равна

$$s_d^2 = s^2\left[\frac{1}{n_{(1)}} + \frac{1}{n_{(2)}} + \frac{(\bar{x}_{(1)} - \bar{x}_{(2)})^2}{s^2_{x(1)} + s^2_{x(2)}}\right]. \qquad (8.52)$$

Пример 14. В примере 13 предыдущего параграфа 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 представлена в табл. 119.

Таблица 119

Доза, 10^3 рентген (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 табл. 119 и 116 не дает оснований считать, что штаммы различны. Однако окончательный вывод можно будет сделать лишь после количественного регрессионного анализа.

Таблица 120

Доза, 10^3 рентген (x)	Остаточный уровень размножения, % (y)	Y_x	$Y_x x$	n_x	$\bar{y}_x$	$n_x \bar{y}_x x$	$n_x x^2$
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 $X_{(2)}$

Основные расчеты для этого анализа выполнены в табл. 120, которая совершенно аналогична табл. 118. 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

$$b_{y/x} = \frac{-605{,}7}{108{,}3} = -5{,}593; \quad r^2 = \frac{(-605{,}7)^2}{108{,}3 \cdot 3514{,}1} = 0{,}9640.$$

Теперь найдем

$$s^2 = \frac{\sum_{yy}(1 - r^2)}{n - 2} = \frac{3514{,}1 \cdot 0{,}0360}{40} = 3{,}163.$$

Наконец,

$$\bar{x} = \frac{144}{42} = 3{,}429; \quad \bar{y} = 80 + \frac{53}{42} = 81{,}262.$$

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 $b^* - \bar{b}$, то значения $\bar{x}_{(1)}$, $\bar{y}_{(1)}$ и $b_{y/x}$ из предыдущего параграфа на всякий случай вычислены заново более точно — с сохранением еще одного десятичного знака; величина $s^2_{(1)}$ получена по формуле (8.38).

Анализ начинаем со сравнения $s^2_{(1)}$ и $s^2_{(2)}$. Так как

$$F = \frac{s^2_{(1)}}{s^2_{(2)}} = \frac{4{,}458}{3{,}163} = 1{,}41 < F_{05}(40; 59) = 1{,}64,$$

то различие между дисперсиями $s^2_{(1)}$ и $s^2_{(2)}$ незначимо, что позволяет пользоваться t-критерием для сравнения $b_{(1)}$ и $b_{(2)}$. Поскольку $s^2_{(1)}$ 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 (8.43):

$$s^2 = \frac{59 \cdot 4,458 + 40 \cdot 3,163}{59 + 40} = \frac{389,5}{99} = 3,934.$$

Теперь по формуле (8.42) 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 (8.50),

$$t_{b_{(1)}-b_{(2)}} = \frac{5,849 - 5,593}{0,228} = \frac{0,256}{0,228} = 1,12.$$

Это меньше, чем $t_{05}\,(99) = 1,98$, 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 (8.44), 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 $\sigma^2_{\bar{b}}$ получаем по формуле (8.45)

$$s^2_{\bar{b}} = \frac{3,934}{356,0} = 0,01105.$$

Проверим теперь, не совпадают ли обе линии регрессии. Для этого по формуле (8.46) вычисляем

$$b^* = \frac{77,279 - 81,262}{3,934 - 3,429} = \frac{3,983}{0,505} = -7,887.$$

Далее по формулам (8.49) и (8.48) находим

$$s^2_{b^*} = \frac{3,934}{0,505^2} \cdot \frac{61 + 42}{61 \cdot 42} = 0,6202;$$

$$s^2_{b^*-\bar{b}} = 0,6202 + 0,0110 = 0,6312;$$

$$s_{b^*-\bar{b}} = \sqrt{0,6312} = 0,795.$$

Поэтому

$$t_{b^*-\bar{b}} = \frac{7,89 - 5,75}{0,795} = \frac{2,14}{0,795} = 2,69.$$

Это превышает $t_{01}(99) = 2{,}63$, так что приходится отвергнуть гипотезу о совпадении обеих прямых. На рис. 53 изображены крестиками и кружками точки $(x, \bar{y}_x)$ соответственно для первой и второй регрессий, а сплошными линиями — их линии регрессии. Пунктирная линия изображает прямую с угловым коэффициентом b^*, проходящую через центры обеих регрессий (жирные точки).

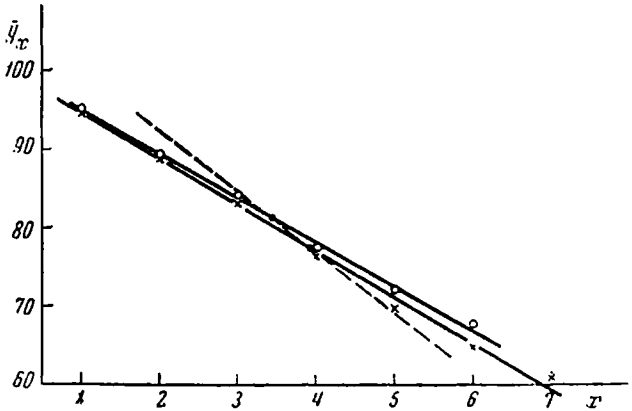

Рис. 53

Поскольку две прямые регрессии значимо не совпадают, мы вычислим по формуле (8.51) 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

$$s_d^2 = 3{,}934\left(\frac{1}{61} + \frac{1}{42} + \frac{0{,}505^2}{356{,}0}\right) = 3{,}934\,(0{,}0164 + 0{,}0238 + 0{,}0007) = 3{,}934 \cdot 0{,}0409 = 0{,}1609,$$

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 книге А. Хальда (1956, стр. 495).

§ 9. Множественная корреляция

Характер корреляции между двумя признаками существенно зависит от того, по какому третьему признаку мы определяем принадлежность элементов к изучаемой совокупности. Например, если предметом изучения является совокупность всех детей определенного возраста (рост которых может быть различным), то корреляция между длиной туловища и длиной ног будет положительной; если же мы станем изучать совокупность всех детей определенного роста (при любом возрасте), то корреляция между длиной туловища и длиной ноги окажется отрицательной (ибо чем длиннее туловище, тем короче должны быть ноги, чтобы рост остался на заданном уровне).

Конечно, далеко не всегда проблема носит столь примитивный характер, как в данном примере, выбранном нарочито простым для большей наглядности. Например, вычисленный обычным образом коэффициент корреляции между урожаем и количеством осадков не дает правильного представления о связи между этими двумя признаками, так как имеется заметная корреляция между количеством осадков и температурой воздуха, а последняя в свою очередь существенно влияет на урожай; поэтому для получения правильного заключения о влиянии на урожай именно интересующего нас фактора (количества осадков) нужно изучить корреляцию между урожаем и количеством осадков при п о с т о я н н о й температуре.

В общем виде задача формулируется следующим образом: помимо (или в отличие от) коэффициента корреляции между двумя признаками x и y (эту величину обозначим ρ_{xy}), когда на значения какого-либо третьего признака z не налагается никаких ограничений, мы хотим найти коэффициент корреляции между x и y при условии, что признак z имеет постоянное значение; последнюю величину называют *частным* или *парциальным коэффициентом корреляции* и обозначают $\rho_{xy\,(z)}$.

Прямой путь решения задачи состоял бы в том, чтобы отбирать в совокупность только те варианты, которые соответствуют какому-либо одному уровню признака z. Однако это очень затрудняло бы составление совокупностей достаточного объема. Методы математической статистики позволяют получить нужный результат без такого ограничения эмпирического материала.

Мы приведем здесь готовую формулу, не останавливаясь на ее выводе. Для нахождения $\rho_{xy(z)}$ нужно сначала найти обычным образом попарные полные коэффициенты корреляции ρ_{xy}, ρ_{xz} и

ρ_{yz}, после чего воспользоваться формулой

$$\rho_{xy(z)} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{(1-\rho_{xz}^2)(1-\rho_{yz}^2)}}. \qquad (8.53)$$

Пример 15. В табл. 121 приведены схематизированные (для наглядности) данные о размерах и весе огурцов. Нужно найти корреляцию между длиной и толщиной при постоянном весе.

Таблица 121

Длина x, *см*	Толщина y, *см*	Вес z, *г*	Длина x, *см*	Толщина y, *см*	Вес z, *г*
12	6	150	8	5	50
10	5	50	12	7	250
10	3	50	12	4	150
14	7	250	10	6	150
12	4	50	14	5	250
16	5	250	12	3	50
12	6	250	16	6	250
8	4	50	14	4	150
10	5	150	14	5	150

Обычный расчет дает

$$\rho_{xy} = 0{,}25;\ \rho_{xz} = 0{,}71;\ \rho_{yz} = 0{,}71,$$

т. е. при увеличении веса длина и толщина увеличиваются одинаково. Теперь по формуле (8.53) находим

$$\rho_{xy(z)} = \frac{0{,}25 - 0{,}71 \cdot 0{,}71}{\sqrt{(1-0{,}71)^2(1-0{,}71^2)}} = \frac{0{,}25 - 0{,}50}{\sqrt{(1-0{,}5)(1-0{,}5)}} = -0{,}50,$$

т. е. при постоянном весе корреляция между длиной и толщиной отрицательна (что вполне понятно); неполноту этой корреляции (т. е. то, что $\rho \neq -1$) следует отнести в основном за счет слишком грубой группировки по весу (значения 50, 150, 250 *г* представляют собой, разумеется, середины разрядов группировки, имеющих в данном случае очень большую ширину — 100 *г*).

Этот же результат можно получить и прямым путем, обрабатывая совокупности, для которых вес одинаков. В нашем примере можно образовать три таких совокупности—при весе 50, 150 и 250 *г*. Соответствующие корреляционные решетки приведены в табл. 122.

Таблица 122

$z = 50$ *в*

x \ y	3	4	5	6	7
8		1	1		
10	1		1		
12	1	1			
14					
16					

$z = 150$ *в*

3	4	5	6	7
		1	1	
	1		1	
	1	1		

$z = 250$ *г*

3	4	5	6	7
			1	1
		1		1
		1	1	

Расчет для каждой из них дает одинаковый результат $\rho_{xy(z)} = -0{,}5$, совпадающий с полученным ранее. Конечно, в реальном, а не схематизированном примере полученные таким образом значения $\rho_{xy(z)}$ несколько различались бы между собой, и мы должны были бы взять средневзвешенное значение — последнее и совпадало бы с вычисленным по формуле (8. 53).

Превращение отрицательной корреляции $\rho_{xy(z)}$ в положительную корреляцию ρ_{xy} иллюстрируется сравнением табл. 122 со сводной табл. 123, которая представляет собой корреляционную решетку для всех вариант из первых двух столбцов табл. 121.

Таблица 123

x \ y	3	4	5	6	7
8		1	1		
10	1		2	1	
12	1	2		2	1
14		1	2		1
16			1	1	

На рис. 54 показано схематически, как из сложения положительных частных корреляционных эллипсов может получиться отрицательный общий корреляционный эллипс. Но на том же рисунке видно, что полная корреляция может оставаться положительной или же стать равной нулю. Очевидно, реализация того или

иного случая зависит просто от того, насколько вообще велик интервал значений признаков x и y при всех встречающихся значениях признака z; поэтому сказать заранее, без расчета, одинаков ли знак $\rho_{xy(z)}$ и ρ_{xy} или нет, очень трудно. Правда, из (8.53) видно, что если $\rho_{xy} > 0$, а ρ_{xz} и ρ_{yz} имеют разные знаки, то и $\rho_{xy(z)} > 0$; если $\rho_{xy} < 0$, а ρ_{xz} и ρ_{yz} имеют одинаковые знаки, то $\rho_{xy(z)} < 0$. Некоторые другие случаи, помимо изображенных на рис. 54, показаны на рис. 55.

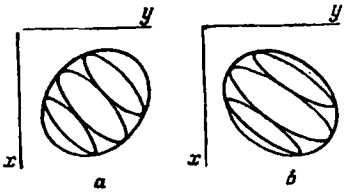

Рис. 54

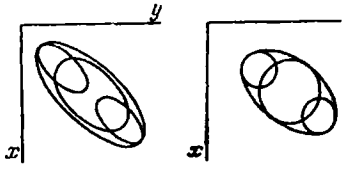

Рис. 55

Иногда возникает необходимость вычислить полные коэффициенты корреляции по известным частным коэффициентам. Тогда используется формула

$$\rho_{xy} = \frac{\rho_{xy\,(z)} + \rho_{xz\,(y)}\,\rho_{yz\,(x)}}{\sqrt{(1-\rho^2_{xz\,(y)})\,(1-\rho^2_{yz\,(x)})}}; \qquad (8.54)$$

аналогичные формулы можно написать для ρ_{xz} и ρ_{yz}.

Подобно тому как мы находим частный коэффициент корреляции между признаками x и y при постоянном значении признака z, можно определять корреляцию между x и y при постоянных значениях признаков z и u, при постоянных значениях z, u и w и т. д. Соответствующие частные коэффициенты корреляции $\rho_{xy(zu)}$, $\rho_{xy(zuw)}$ и т. д. могут вычисляться по формулам, непосредственно обобщающим формулу (8.53). Например,

$$\rho_{xy\,(zu)} = \frac{\rho_{xy\,(u)} - \rho_{xz\,(u)}\,\rho_{yz\,(u)}}{\sqrt{(1-\rho^2_{xz\,(u)})\,(1-\rho^2_{yz\,(u)})}}, \qquad (8.55)$$

т. е. к индексам всех величин, входящих в формулу (8.53), просто добавлено (u), причем эти новые величины $\rho_{xy(u)}$, $\rho_{xz(u)}$ и $\rho_{yz(u)}$ в свою очередь получаются по формуле (8.53). Можно было бы выразить $\rho_{xy(zu)}$ и через $\rho_{xy(z)}$, $\rho_{xu(z)}$ и $\rho_{yu(z)}$, поскольку $\rho_{xy(zu)} = \rho_{xy(uz)}$.

Надо, однако, иметь в виду, что чем больше признаков принято во внимание, тем строже должна соблюдаться линейность всех корреляций — иначе формулы для перехода от полных коэффициентов корреляции к частным коэффициентам и обратно дают слишком неточные результаты. Поскольку в реальных совокупностях корреляция никогда не бывает строго линейной, то нахождение частных коэффициентов корреляции при исключении более чем одного признака является в значительной мере бессмысленным (если не считать случаев многотысячных совокупностей, где можно достаточно надежно выявлять линейность корреляции).

Наряду с определением частных коэффициентов корреляции может возникнуть необходимость в нахождении частных коэффициентов регрессии: y на x при постоянном z, или z на x при постоянном y, или x на y при постоянном z и т. д. — всего шесть разных частных коэффициентов регрессии:

$$\beta_{(y/x)z},\ \beta_{(x/y)z},\ \beta_{(x/z)y},\ \beta_{(z/x)y},\ \beta_{(y/z)x},\ \beta_{(z/y)x}.$$

Для их вычисления существуют формулы

$$\beta_{(x/y)z} = \frac{\beta_{x/y} - \beta_{x/z}\beta_{z/y}}{1 - \beta_{y/z}\beta_{z/y}} \tag{8.56}$$

и аналогично для других пяти коэффициентов. Обратные формулы имеют вид

$$\beta_{x/y} = \frac{\beta_{(x/y)z} + \beta_{(x/z)y}\beta_{(z/y)x}}{1 - \beta_{(x/z)y}\beta_{(z/x)y}}; \tag{8.57}$$

аналогичный вид имеют формулы и для остальных пяти коэффициентов.

Если подставить в выражение (8.56) значения оценок для $\beta_{x/y}$, $\beta_{x/z}$, $\beta_{z/y}$, $\beta_{y/z}$ согласно (8.6), то получится формула, по которой можно сразу вычислять оценку для $\beta_{(x/y)z}$:

$$b_{(x/y)z} = \frac{\Sigma_{xy}\Sigma_{zz} - \Sigma_{xz}\Sigma_{yz}}{\Sigma_{yy}\Sigma_{zz} - (\Sigma_{yz})^2};$$

эта формула имеет очень симметричный вид. Остальные коэффициенты получаются простой перестановкой индексов.

Рассмотрение вопросов, связанных с частной корреляцией, показывает лишний раз, что интерпретация результатов корреляционного анализа требует большой осмотрительности. В табл. 124 приведены данные о корреляции между индексом плотности колоса пшеницы (y) и числом зерен в колосе (x). Измерения проводились отдельно для трех чистых линий пшеницы; числа, относящиеся к разным линиям, даны различным шрифтом. Мы видим, что в чистых линиях корреляция отрицательна, но при обработке смешанного материала получилась бы положительная корреляция.

Таблица 124

x \ y	15	16	17	18	19	20	21	22	23	24	25	26	27	n_x
18					1									1
21			1	1		1	1	1						5
24		1	3	6	3	1	3	2						19
27	1	2	4	3	1	4	4	3		1	1	1		25
30	1	1	2	1	5	7	4	2	1	2	1		1	28
33		1	1	1	3	3	1			2	3	1		16
36				1	2	2	1	1	1		1			9
39					1	1	1	1	3	1				8
42				1				1	2	1		1		6
45						1	1			1				3
48								1	1					2
n_y	2	5	11	14	16	20	16	12	8	8	6	3	1	122

В некоторых случаях корреляция может вообще быть ложной. Пусть, например, имеются три независимо варьирующие величины x, y, z. Если по ходу анализа будут введены индексы x/z и y/z (прием, часто применяемый в разного рода задачах), то эти индексы окажутся коррелированными. Причина ясна — они содержат общую величину z.

ВЫРАВНИВАНИЕ РЯДОВ

§ 1. Постановка задачи

В предыдущей главе рассматривались линии регрессии, устанавливающие связь между двумя варьирующими величинами x и y. Существенно, что при этом варьирование обеих величин x и y находится вне произвола исследователя. Например, при изучении связи между толщиной ствола и высотой деревьев в посадке мы можем выбирать произвольно лишь порядок просмотра деревьев (вдоль полосы, поперек, в шахматном порядке и т. д.), обе же интересующие нас величины—толщина ствола x и высота дерева y — варьируют «сами по себе».

Между тем 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. Так, можно изучать зависимость веса животного от возраста, взвешивая животное в конце каждого года (или месяца) жизни, или зависимость урожая картофеля от количества удобрений, внося на каждую из ряда делянок вполне определенные количества удобрений. В этих случаях, очевидно, агрумент x не является статистически варьирующей величиной, а это значит, что имеет смысл говорить лишь о регрессии y на x, но не о регрессии x на y. Чаще всего в такого рода задачах независимой переменной является время; тогда говорят о *рядах динамики*.

В табл. 113 были приведены данные об урожайности картофеля по годам:

Годы	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
Урожайность, *т/га* .	7,2	7,1	7,4	6,1	6,0	7,3	9,4	9,2	8,8	10,4	8,0	9,7

Для этого примера характерно то, что каждому значению аргумента x соответствует не ряд значений y, а лишь одно такое значение. Конечно, если повторить опыт для каждого значения x по

несколько раз, то получим несколько несовпадающих значений y_x, соответствующих каждому значению x. Ниже для простоты ограничимся случаем, когда каждому значению x соответствует лишь одно значение y; впрочем, это значение y может быть средним из нескольких значений y_x для данного x, но при этом число значений y_x должно быть для всех x одинаково (с тем, чтобы все эмпирические точки имели одинаковый «вес»).

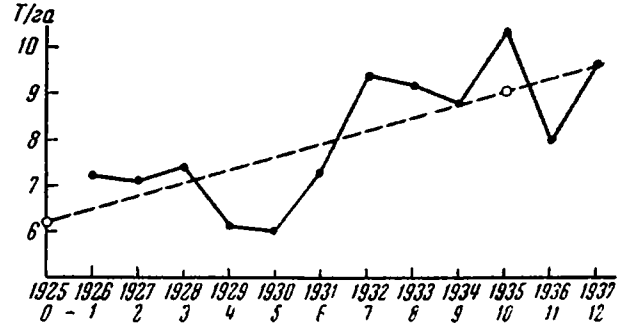

Рис. 56

При изучении регрессии между двумя варьирующими величинами мы ранее ограничились случаем линейной регрессии (точнее — линейным приближением к истинной зависимости). Это объяснялось тем, что только в таком случае расчеты являются относительно простыми; при нелинейной регрессии они настолько усложняются, что уточнение результата редко оправдывает усилия, затраченные на его получение.

В данной главе дело обстоит иначе. Это связано с тем, что мы с самого начала упрощаем задачу в ряде отношений, считая, что а) величина x не варьирует, а принимает определенные, заранее заданные значения и что б) каждому значению x соответствует лишь одно значение y; кроме того, мы будем считать, что в) значения x являются, как правило, равноотстоящими. Такое существенное упрощение задачи позволяет нам не ограничиваться линейным приближением связи, а рассмотреть и другие, более сложные зависимости.

Задача, которую мы ставим себе здесь, состоит в том, чтобы, имея некоторый эмпирический ряд точек x, y, выявить ту основную тенденцию, которая характеризует этот ряд и на которую накладываются статистические вариации. Так, в написанном выше ряде указанная тенденция заключалась в том, что урожайность картофеля в общем возрастает (рис. 56); то обстоятельство, что эмпирические точки не ложатся на одну прямую, мы рассматриваем как

результат статистических вариаций, несколько смещающих «истинные» значения y вверх или вниз (как следствие случайного сочетания воздействий многих небольших побочных факторов).

Сформулированную выше задачу принято называть задачей о *выравнивании* рядов. Такое название отражает тот факт, что мы как бы в ы р а в н и в а е м (приводим к одному уровню) действие побочных факторов при каждом значении аргумента.

Во многих случаях выравнивание ряда можно произвести графически, проведя на глаз более или менее плавную линию, достаточно близкую к эмпирическим точкам. Если желательно получить более точный результат и одновременно избежать произвола, всегда сопровождающего графическое выравнивание, то лучше применять аналитические методы, к описанию которых мы и переходим.

§ 2. Метод наименьших квадратов

Весьма часто исследуемая зависимость принадлежит к хорошо изученному типу, и ее аналитическое (алгебраическое) выражение точно известно; при этом целью исследования является определение числовых 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 ответствен за активность почвы).

Другой пример: есть все основания считать, что привес животных вообще пропорционален количеству корма:

$$\tilde{y} = ax, \qquad (**)$$

но мы хотим найти коэффициент пропорциональности a для интересующего нас вида корма (через $\tilde{y}$ мы будем впредь обозначать выравненные значения y).

Задачи, подобные вышеуказанным, можно решать при помощи метода наименьших квадратов, о котором уже говорилось в гл. 8, § 2. Напомним сущность этого метода: правильными будут считаться такие значения параметров уравнений типа (*) или (**), при которых сумма квадратов отклонений эмпирических значений y от расчетных значений $\tilde{y}$ окажется наименьшей.

Учет условия 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

$$\tilde{y} = a_0 + a_1 x + a_2 x^2 + \ldots + a_h x^h, \tag{9.1}$$

то дело упрощается. Это связано с тем, что в выражении (9.1) $\tilde{y}$ зависит от параметров $a_0, a_1, a_2, \ldots, a_h$ линейно, а поэтому упомянутая выше система уравнений получается также линейной. Система же линейных уравнений может быть решена совершенно элементарными приемами, например, способом подстановки или способом исключения неизвестных.

Пусть x_i $(i = 1, 2, \ldots, n)$ — заданные значения x, а y_i — отвечающие им эмпирические значения y. В соответствии с (9.1) отклонения $y_i - \tilde{y}_i$ равны

$$y_i - \tilde{y}_i = y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_h x_i^h. \tag{9.2}$$

Можно показать, что сумма $\sum_i (y_i - \tilde{y}_i)^2$ имеет минимальное значение, если одновременно выполняются равенства:

$$\left.\begin{aligned}
\Sigma (y_i - \tilde{y}_i)\, x_i^0 &= \Sigma (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots \\
&\qquad - a_h x_i^h)\, x_i^0 = 0; \\
\Sigma (y_i - \tilde{y}_i)\, x_i^1 &= \Sigma (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots \\
&\qquad - a_h x_i^h)\, x_i^1 = 0; \\
\Sigma (y_i - \tilde{y}_i)\, x_i^2 &= \Sigma (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots \\
&\qquad - a_h x_i^h)\, x_i^2 = 0; \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
\Sigma (y_i - \tilde{y}_i)\, x_i^h &= \Sigma (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots \\
&\qquad - a_h x_i^h)\, x_i^h = 0.
\end{aligned}\right\} \tag{9.3}$$

Каждое j-е уравнение этой системы получено умножением $(y_i - \tilde{y}_i)$ на x_i в j-й степени с последующим суммированием по всем i (при этом $x_i^0 = 1$, $x_i^1 = x_i$). Неизвестными здесь являются параметры $a_0, a_1, a_2, \ldots, a_h$, а величины x_i^0 $(= 1)$, x_i, x_i^2, $\ldots$, x_i^h являются коэффициентами при этих неизвестных в выражении (9.2). Число уравнений системы (9.3) равно $h + 1$, как и число неизвестных параметров $a_0, a_1, a_2, \ldots, a_h$.

Систему (9.3) можно переписать в виде:

$$\left.\begin{aligned} a_0 n + a_1\sum x_i + a_2\sum x_i^2 + \quad & + a_h\sum x_i^h = \sum y_i; \\ a_0\sum x_i + a_1\sum x_i^2 + a_2\sum x_i^3 + \quad & + a_h\sum x_i^{h+1} = \sum y_i x_i; \\ a_0\sum x_i^2 + a_1\sum x_i^3 + a_2\sum x_i^4 + \quad & + a_h\sum x_i^{h+2} = \sum y_i x_i^2; \\ & \ldots \\ a_0\sum x_i^h + a_1\sum x_i^{h+1} + a_2\sum x_i^{h+2} + \quad & + a_h\sum x_i^{2h} = \sum y_i x_i^h. \end{aligned}\right\} \qquad (9.4)$$

Входящие сюда $\sum x_i, \sum x_i^2, \ldots, \sum x_i^{2h}$ и $\sum y_i, \sum y_i x_i, \ldots, \sum y_i x_i^h$ вычисляются по эмпирическим данным $\left(n = \sum_{i=1}^{n} 1\right)$. Систему (9.4) называют *системой нормальных уравнений.*

§ 3. Линейная зависимость

Пусть предполагаемая зависимость между x и y является линейной, т. е. имеет вид

$$\tilde{y} = a_0 + a_1 x. \qquad (9.5)$$

Здесь $h = 1$. так что нормальных уравнений будет $h + 1 = 2$. В первом уравнении наивысшая степень x_i будет $h = 1$, а во втором уравнении $h + 1 = 2$. Таким образом, имеем систему:

$$\left.\begin{aligned} a_0 n + a_1\sum x_i &= \sum y_i; \\ a_0\sum x_i + a_1\sum x_i^2 &= \sum y_i x_i. \end{aligned}\right\} \qquad (9.6)$$

Пример 1. Решим задачу для ряда из § 1 этой главы. Составив табл. 125, находим:

$$\sum x_i = 78; \quad \sum x_i^2 = 650; \quad \sum y_i = 96{,}6; \quad \sum y_i x_i = 668{,}8$$

(в качестве начала отсчета величин x_i мы для удобства взяли число 1925). Теперь имеем систему уравнений:

$$12a_0 + 78a_1 = 96{,}6;$$
$$78a_0 + 650a_1 = 668{,}8.$$

Чтобы решить эту систему, умножаем первое равенство на $78 : 12 = 6{,}5$ и результат

$$78a_0 + 507a_1 = 627{,}9$$

Таблица 125

	x_i^2	y_i	$y_i x_i$
1	1	7,2	7,2
2	4	7,1	14,2
3	9	7,4	22,2
4	16	6,1	24,4
5	25	6,0	30,0
6	36	7,3	43,8
7	49	9,4	65,8
8	64	9,2	73,3
9	81	8,8	79,2
10	100	10,4	104,0
11	121	8,0	88,0
12	141	9,7	116,4
78	650	96,6	668,8

вычитаем из второго равенства. Это дает

$$143a_1 = 40{,}9,$$

откуда

$$a_1 = \frac{40{,}9}{143} = 0{,}286.$$

Теперь из первого уравнения нормальной системы находим

$$a_0 = \frac{96{,}6 - 78 \cdot 0{,}286}{12} = 6{,}19.$$

Значит, уравнение регрессии имеет вид

$$\tilde{y} = 6{,}19 + 0{,}286x.$$

Эта прямая изображена на рис. 56 (она проведена через точки $\tilde{y}_0 = 6{,}19 + 0{,}286 \cdot 0 = 6{,}19$ и $\tilde{y}_{10} = 6{,}19 + 0{,}286 \cdot 10 = 6{,}19 + 2{,}86 = 9{,}05$).

Если решить систему уравнений (9.6) в общем виде, то получим готовые формулы для вычисления a_0 и a_1. Решение прове-

дем следующим образом. Разделив первое уравнение на n, получаем

$$a_0 + a_1\hat{x} = \hat{y}; \qquad (***)$$

если вычесть это из (9.5), то получится

$$\tilde{y} - \bar{y} = a_0 + a_1 x - a_0 - a_1\bar{x} = a_1 (x - \bar{x});$$

тогда сравнение с (8.1) показывает, что a_1 есть не что иное, как коэффициент регрессии $b_{y/x}$. Чтобы найти величину a_1, умножим первое уравнение системы (9.6) на $\Sigma x_i/n$ и вычтем получившееся равенство из второго уравнения этой системы:

$$[a_0\Sigma x_i + a_1\Sigma x_i^2] - [a_0\Sigma x_i + a_1 \frac{1}{n}(\Sigma x_i)^2] = \Sigma x_i y_i - \frac{1}{n}\Sigma x_i y_i,$$

отсюда

$$a_1 = \frac{\Sigma x_i y_i - \Sigma x_i \Sigma y_i/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n}$$

Если применить обозначения (5.6), то

$$a_1 = \frac{(XY) - X_{(1)} Y_{(1)}/n}{X_{(2)} - X_{(1)}^2/n}, \qquad (9.7)$$

т. е. получаем формулу, которую имели и ранее для коэффициента регрессии $b_{y/x}$. Из (***) теперь находим

$$a_0 = (Y_{(1)} - a_1 X_{(1)})/n, \qquad (9.8)$$

где a_1 вычисляется по формуле (9.7).

Пример 2. Применив формулы (9.7) и (9.8) к примеру 1, получаем сразу:

$$a_1 = \frac{668{,}8 - 78 \cdot 96{,}6/12}{650 - 78^2/12} = \frac{668{,}8 - 627{,}9}{650 - 507} = \frac{40{,}9}{143} = 0{,}286;$$

$$a_0 = \frac{96{,}6 - 0{,}286 \cdot 78}{12} = \frac{96{,}6 - 22{,}3}{12} = \frac{74{,}3}{12} = 6{,}19.$$

Если значения x_i являются равноотстоящими, то систему нормальных уравнений можно сильно упростить. При нечетном числе значений x_i это достигается использованием условной шкалы с началом отсчета в середине ряда (так, ряд 46, 49, 52, 55, 58, 61, 64, 67, 70 заменяется рядом -4, -3, -2, -1, 0, 1, 2, 3, 4). Тогда, очевидно, суммы нечетных степеней x_i будут равны нулю.

В частности, вместо системы (9.6) получим два уравнения:

$$a_0 n = \sum y_i; \quad a_1 \sum x_i^2 = \sum y_i x_i,$$

откуда

$$a_0 = \frac{\sum y_i}{n}; \quad a_1 = \frac{\sum y_i x_i}{\sum x_i^2}. \tag{9.9}$$

Если число значений x_i четно, то начало отсчета переносится в точку, лежащую между двумя срединными значениями (например, при $n = 10$ это будет 5,5), а единица измерения x уменьшается вдвое. Тогда новые значения x_i будут— 1, —3, —5 и т. д. влево и 1, 3, 5 и т. д. вправо от середины ряда. Например:

Старая шкала	1	2	3	4	5	6	7	8	9	10
Новая шкала	—9	—7	—5	—3	—1	1	3	5	7	9

Здесь опять суммы нечетных степеней x_i равны нулю.

Пример 3. Для ряда из примера 1 новая расчетная таблица будет иметь вид табл. 126.

Таблица 126

x_i	x_i^2	y_i	$y_i x_i$
—11	121	7,2	—79,2
— 9	81	7,1	—63,9
— 7	49	7,4	—51,8
— 5	25	6,1	—30,5
— 3	9	6,0	—18,0
— 1	1	7,3	—7,3
1	9	9,4	9,4
3	1	9,2	27,6
5	25	8,8	44,0
7	49	10,4	72,8
9	81	8,0	72,0
11	121	9,7	106,7
Сумма.	572	96,6	81,8

Отсюда по формулам (9.9) получаем:

$$a = \frac{96{,}6}{12} = 8{,}05; \quad a_1 = \frac{81{,}8}{572} = 0{,}143,$$

так что уравнение регрессии будет

$$\tilde{y} = 8{,}05 + 0{,}143\, x';$$

если подставить сюда $x' = 2\,(x - 6{,}5)$ — начало отсчета было сдвинуто на 6,5 единиц, а масштаб был уменьшен вдвое, — то получится, как и ранее,

$$\tilde{y} = 6{,}19 + 0{,}286x.$$

В некоторых случаях предполагаемая зависимость между y и x не является линейной, но может быть приведена к таковой надлежащим преобразованием координат.

Пример 4. При облучении гамма-лучами фермента наблюдается падение его активности. В табл. 127 приведены значения активности A (в процентах к начальной) при разных дозах

Таблица 127

$x = D$, тыс. рентген	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
$X_{(1)}$		$Y_{(1)}$	$X_{(2)}$	(XY)

облучения D. Известно, что активность убывает при увеличении дозы облучения по показательному закону

$$A = A_0 e^{-\alpha D}$$

(e — основание натуральных логарифмов). Требуется найти коэффициент α.

Данную нелинейную зависимость можно привести к линейному виду, если выполнить преобразование

$$\lg A = \lg A_0 - \alpha D \lg e.$$

Теперь мы ищем регрессию между $D = x$ и $\lg A = y$. Проведя обычным образом вычисления (они показаны в табл. 127), находим:

$$\sum_{xy} = (XY) - \frac{X_{(1)} Y_{(1)}}{n} = 169{,}2 - \frac{160{,}5 \cdot 10{,}42}{7} = -69{,}8;$$

$$\sum_{xx} = X_{(2)} - \frac{X_{(1)}^2}{n} = 6815 - \frac{160{,}5^2}{7} = 3135;$$

$$b_{y/x} = \frac{-69{,}8}{3135} = -0{,}0223.$$

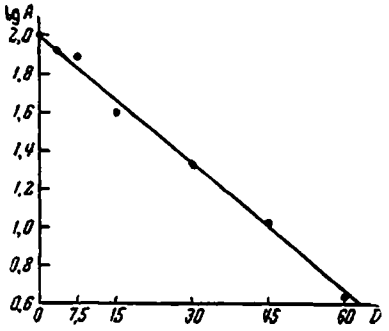

Рис. 57

Мы получили величину $-\alpha \lg e$; так как $\lg e = 0{,}4343$, то

$$\alpha = \frac{-0{,}0223}{-0{,}4343} = 0{,}0514.$$

Зависимость между $\lg A$ и D изображена на рис. 57. Там же показана линия регрессии, проведенная по найденному значению $b_{y/x}$.

Аналогичным образом поступаем, если предполагаемая функциональная зависимость имеет другой вид, — стараемся привести ее к линейной подходящим преобразованием. Несколько примеров таких преобразований дает табл. 128[1].

Таблица 128

Исходная функция	Искомые параметры	Преобразованное уравнение	Вспомогательные величины
$y = ax^s$	s	$\lg y = \lg a + s \lg x$	$y' = \lg y,\ a' = \lg a,\ x' = \lg x$
$y = b \lg ax$	$a,\ b$	$y = b \lg a + b \lg x$	$a' = b \lg a,\ x' = \lg x$
$y = a + \frac{b}{x}$	$a,\ b$	$y = a + b\frac{1}{x}$	$x' = \frac{1}{x}$
$y = ax + bx^3$	$a,\ b$	$\frac{y}{x} = a + bx^2$	$y' = \frac{y}{x},\ x' = x^2$
$y = ae^{-b/x}$	$a,\ b$	$\lg y = \lg a - \frac{b \lg e}{x}$	$y' = \lg y,\ x' = \frac{1}{x},$ $a' = \lg a,\ b' = b \lg e$

[1] Более подробно этот вопрос рассмотрен в книге А. Хальда (1956, гл. XVII, § 7).

При пользовании всеми этими преобразованиями надо 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.

§ 4. Квадратичная зависимость

Если предполагаемая зависимость между y и x квадратична

$$\tilde{y} = a_0 + a_1 x + a_2 x^2, \tag{9.10}$$

то система нормальных уравнений имеет вид:

$$\left.\begin{aligned} a_0 n + a_1 \sum x_i + a_2 \sum x_i^2 &= \sum y_i; \\ a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 &= \sum y_i x_i; \\ a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 &= \sum y_i x_i^2. \end{aligned}\right\} \tag{9.11}$$

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

$$\tilde{y} = b_0 + b_1 (x - \bar{x}) + b_2 [(x - \bar{x})^2 - \gamma], \tag{9.12}$$

где

$$\bar{x} = \frac{1}{n} \sum x_i; \quad \gamma = \frac{1}{n} \sum (x_i - \bar{x})^2. \tag{9.13}$$

Тогда

$$\left.\begin{aligned} b_0 = \bar{y} = \frac{1}{n} \sum y_i; \quad b_1 &= \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}; \\ b_2 &= \frac{\sum (x_i - \bar{x})^2 y_i - n\gamma\bar{y}}{\sum (x - \bar{x})^4 - n\gamma^2}. \end{aligned}\right\} \tag{9.14}$$

Пример 5. В табл. 129 приведены данные об удоях коров за первую, вторую и т. д. лактации. Значения y_i представляют

Таблица 129

x (номер лактации)	1	2	3	4	5	6	7	8	9
y (удой в ц за лактацию)	30,7	32,9	35,8	38,2	38,7	39,2	39,9	38,4	37,1

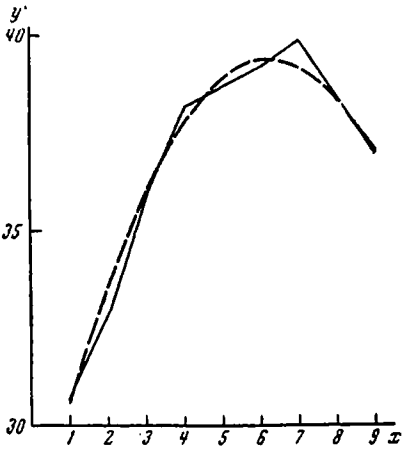

Рис. 58

собой усредненные данные для группы коров.

Вид графика на рис. 58 (ломаная линия) подсказывает, что подходящим аналитическим выражением для этой зависимости может служить квадратичная парабола типа (9.10).

Заменив для сокращения вычислений y_i на $y_i' = y_i - 30$, составляем вспомогательную табл. 130; в ней мы производим все промежуточные расчеты, необходимые для получения сумм, входящих в формулы (9.14).

Используя табличные данные, имеем:

$$\bar{x} = \frac{45}{9} = 5;\quad \gamma = \frac{60}{9} = \frac{20}{3} = 6{,}667;\quad b_0 = \frac{60{,}9}{9} = 6{,}767;$$

$$b_1 = \frac{51{,}3}{60} = 0{,}855;\quad b_2 = \frac{306{,}7 - 9\cdot\frac{20}{3}\cdot\frac{60{,}9}{9}}{708 - 9\cdot\frac{400}{9}} = \frac{-99{,}3}{308} = -0{,}3224.$$

Поэтому

$$\tilde{y} = 30 + 6{,}77 + 0{,}855(x-5) - 0{,}3224[(x-5)^2 - 6{,}667]. \qquad (*)$$

Таблица 130

x_i	$x_i - \bar{x}$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^4$	y_i'	$(x_i-\bar{x})y_i'$	$(x_i-\bar{x})^2 y_i'$	$\tilde{y}_i$
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	9,2	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	

Раскрыв скобки в (*), придем к форме (9.10). Значения $\tilde{y}_i$, вычисленные по формуле (*), проставлены в последнем столбце, табл. 130. Кривая, изображающая функцию (*), нанесена пунктиром на рис. 58.

§ 5. Выбор степени полинома

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 пособ 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:

$$x \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \text{ и т. д.;}$$

$$y \quad 5 \quad 7 \quad 9 \quad 11 \quad 13 \text{ и т. д.}$$

Разности соседних значений y (или приращения Δy) все равны 2: $y_2 - y_1 = 7 - 5 = 2$, $y_3 - y_2 = 9 - 7 = 2$, $y_4 - y_3 = 11 - 9 = 2$ и т. д. Если функция $y = f(x)$ квадратична, то приращения Δy неодинаковы, но зато одинаковы приращения этих приращений (обозначим их $\Delta^2 y$). Например, если

$$y = x^2,$$

то ряд будет:

$$x \quad 1 \quad 2 \quad 3 \quad 4 \quad 5$$

$$y \quad 1 \quad 4 \quad 9 \quad 16 \quad 25\,.$$

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 таблицу (табл. 131).

Таблица 131

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	6
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 систематического характера, а имели бы вид случайных вариаций. Составим, например, таблицу приращений (табл. 132) для ряда из табл. 129.

Здесь изменения Δy при переходе от низших 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

$$\tilde{y} = a_0 + a_1 x + a_2 x^2.$$

Для ряда из примера 1 случайно варьируют уже приращения первого порядка

x	1	2	3	4	5	6	7	8	9	10	11	12
y	7,1	7,2	7,4	6,1	6,0	7,3	9,4	9,2	8,8	10,4	8,0	9,7
Δy		—0,1	0,3	—1,3	—0,1	1,3	2,1	—0,2	—0,4	1,6	—2,4	1,7

поэтому выравнивание может производиться при помощи линейной функции $\tilde{y} = a_0 + a_1 x$, как это и было сделано выше.

Таблица 132

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	—1,4
9	37,1	—2,4	—
10	34,7	—	—

Если значения x_i не равноотстоящие, то для проверки линейности полинома нужно составить ряд относительных приращений $\Delta y_i / \Delta x_i$.

§ 6. Способ скользящего среднего

Выравнивание эмпирического ряда при помощи полинома

$$\tilde{y} = a_0 + a_1 x + a_2 x^2 + \ldots + a_h x^h \qquad (9.1)$$

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 § 3

этой главы). Если же таких оснований нет, то полином (9. 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 удобней 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 этой 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,

$$\begin{array}{ccc} x_4 & x_5 & x_6 \\ y_4 & y_5 & y_6. \end{array}$$

Прямая линия регрессии, которую можно провести по этим трем точкам, проходит, как мы знаем, через точку с координатами

$$\bar{x} = \frac{x_4 + x_5 + x_6}{3}; \quad \bar{y} = \frac{y_4 + y_5 + y_6}{3}.$$

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 $b_{y/x}$ 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 вычислитель-

ную работу, связанную с определением b_x. Поэтому обычно ограничиваются нахождением выравненных точек $(\widetilde{x}_i, \widetilde{y}_i)$.

Выравненные точки $(\widetilde{x}_6, \widetilde{y}_6, x_7, \widetilde{y}_7)$ и т. д. найдем из равенств:

$$\widetilde{x}_6 = \frac{x_5 + x_6 + x_7}{3} \quad \widetilde{y}_6 = \frac{y_5 + y_6 + y_7}{3};$$

$$\widetilde{x}_7 = \frac{x_6 + x_7 + x_8}{3}; \quad \widetilde{y}_7 = \frac{y_6 + y_7 + y_8}{3}$$

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 $\widetilde{x}_i$, $\widetilde{y}_i$ называют *скользящими*; отсюда и название разбираемого метода выравнивания.

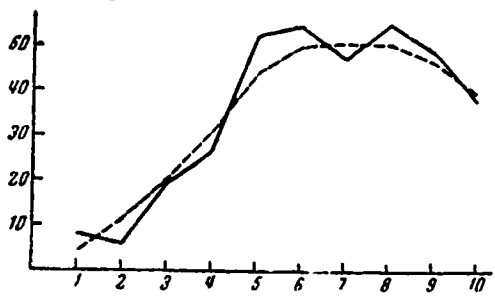

Рис. 59

Процедура выравнивания сильно упрощается, если значения x_i являются равноотстоящими. Тогда $\widetilde{x}_5 = x_5$, $\widetilde{x}_6 = x_6$ и т. д. (т. е. вообще $\widetilde{x}_i = x_i$), так что остается только находить значения

$$\widetilde{y}_i = (y_{i-1} + y_i + y_{i+1}) : 3. \qquad (9.15)$$

Таблица 133

x	1	2	3	4	5	6	7	8	9	10
y	8	6	20	27	53	55	48	56	50	39
$\widetilde{y}$ (3)	4,4	11,3	17,7	33,3	45,0	52,0	53,0	51,3	48,3	40,0
$\widetilde{y}$ (5)	—	—	22,8	32,2	40,6	47,8	52,4	49,6	—	—
$\widetilde{y}$ (взв)	3,9	10,1	20,7	31,5	44,4	50,5	51,7	51,4	47,1	41,9

Пример 6. В первых двух строках табл. 133 записан ряд, изображение которого дано на рис. 59 (сплошная линия).

Пользуясь формулой (9.15), имеем: $\widetilde{y}_2 = (8 + 6 + 20) : 3 = 11{,}3$; $\widetilde{y}_3 = (6 + 20 + 27)\ 3 = 17{,}7$ и т. д. (эти значения записаны в третьей строке табл. 133).

Формула (9.15) не годится, конечно, для вычисления выравненных значений первого и последнего y_i (т. е. $\widetilde{y}_1$ и $\widetilde{y}_n$), так как для $\widetilde{y}_1$ у нас нет значения $y_{i-1} = y_{1-1} = y_0$, а для $\widetilde{y}_n$ нет значения $y_{i+1} = y_{n+1}$. Чтобы обойти это затруднение и в то же время не потерять двух точек ряда, поступим следующим образом. Найдя прямую линию регрессии, соответствующую трем первым точкам (эту прямую мы выше использовали для нахождения $\widetilde{y}_2$), продолжим ее влево до пересечения с вертикалью $x_0 = 0$; это даст некоторое условное значение y_0, представляющее собой результат линейной экстраполяции за пределы ряда. Полученное таким образом значение y_0 затем используем для нахождения $\widetilde{y}_1$ по формуле (9.15):

$$\widetilde{y}_1 = (y_0 + y_1 + y_2) : 3. \qquad (*)$$

Величину y_0 найдем при помощи уравнения линейной регрессии

$$y_x = \bar{y} + b_{y/x}\,(x - \bar{x}), \qquad (8.10)$$

где согласно (8.6) и (8.8)

$$b_{y/x} = \frac{\Sigma x_i y_i - \Sigma x_i \Sigma y_i / n}{\Sigma x_i^2 - (\Sigma x_i)^2/n} \qquad (**)$$

В нашем случае $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, так что

$$\Sigma x_i = 1 + 2 + 3 = 6;\quad \Sigma x_i^2 = 1 + 4 + 9 = 14;$$
$$\Sigma x_i y_i = y_1 + 2y_2 + 3y_3;$$
$$n = 3;\quad \bar{x} = \Sigma x_i / n = 6/3 = 2;\quad \bar{y} = (y_1 + y_2 + y_3)/3.$$

Подставляя эти значения в (**) и (8.10) и учитывая, что $x_0 = 0$, получаем

$$y_0 = (4y_1 + y_2 - 2y_3) : 3.$$

Если подставить это в (*), то получится

$$\widetilde{y}_1 = (7y_1 + 4y_2 - 2y_3) : 9. \qquad (9.16)$$

По этой же формуле можно вычислить и $\widetilde{y}_n$, если нумерацию точек вести с конца. Для ряда из табл. 133 имеем:

$$\widetilde{y}_1 = (7\cdot 8 + 4\cdot 6 - 2\cdot 20) : 9 = 40 : 9 \approx 4{,}4;$$

$$\widetilde{y}_{10} = (7\cdot 39 + 4\cdot 50 - 2\cdot 56) : 9 = 361 : 9 \approx 40{,}0.$$

Эти числа также записаны в третьей строке табл. 133.

При выравнивании необязательно пользоваться именно тройками чисел. Однако удобней, чтобы число точек, по которым производится усреднение, было нечетным, ибо иначе не будет выполняться условие $\widetilde{x}_i = x_i$. 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. С другой стороны, 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:

$$\widetilde{y}_i = (y_{i-2} + y_{i-1} + y_i + y_{i+1} + y_{i+2}) : 5.$$

Соответствующие значения для ряда из примера 5 приведены в четвертой строке табл. 133.

§ 7. Взвешенное скользящее среднее

Усреднение по пяти точкам даст достаточно точные результаты и при большой кривизне эмпирического ряда, если выравнивать его не отрезками прямой, а отрезками какой-нибудь простой кривой, например, квадратичной параболы. Как и в случае прямолинейного приближения, здесь не потребуется определения уравнений регрессии для каждого отрезка. Если ограничиться только нахождением выравненных точек (не интересуясь направлением и кривизной выравненной кривой в этих точках), то дело опять сводится просто к вычислению средних $\widetilde{y}_i$. Разница только в том, что в этом случае $\widetilde{y}_i$ есть не просто среднее арифметическое, а среднее взвешенное, т. е. значения y_{i-2}, y_{i-1}, y_i, y_{i+1} и y_{i+2} перед суммированием умножаются на определенные весовые множители. Расчет показывает, что эти множители (веса):

$$-\frac{1}{12},\quad \frac{4}{12},\quad \frac{6}{12},\quad \frac{4}{12},\quad -\frac{1}{12};$$

сумма этих весов равна, конечно, единице [1]. Для ряда из табл. 133 получим, например:

$$\tilde{y}_4 = (-1\cdot 6 + 4\cdot 20 + 6\cdot 27 + 4\cdot 53 - 1\cdot 55) : 12 = 32{,}8$$

(вместо 32,2 при прямолинейном выравнивании).

Вместо квадратичной параболы можно использовать для выравнивания отдельных отрезков какую-нибудь другую кривую. При этом не столь важно, чтобы эта кривая описывалась достаточно простым алгебраическим выражением — для практики выравнивания гораздо важней, чтобы простым получался набор весовых множителей. Например, можно употребить такой способ выравнивания. Сначала мы получаем ряд скользящих средних, усредняя прямолинейно (т. е. с одинаковыми весами) по три значения y_i; это дает нам

$$\tilde{y}_2 = \frac{1}{3}(y_1 + y_2 + y_3); \quad \tilde{y}_3 = \frac{1}{3}(y_2 + y_3 + y_4);$$
$$\tilde{y}_4 = \frac{1}{3}(y_3 + y_4 + y_5) \text{ и т. д.}$$

Затем таким же образом выравниваем новый ряд $\tilde{y}_2, \tilde{y}_3, \ldots$, т. е. вычисляем

$$\tilde{\tilde{y}}_3 = \frac{1}{2}(\tilde{y}_2 + \tilde{y}_3 + \tilde{y}_4), \quad \tilde{\tilde{y}}_4 = \frac{1}{3}(\tilde{y}_3 + \tilde{y}_4 + \tilde{y}_5) \text{ и т. д.}$$

Но

$$\tilde{\tilde{y}}_3 = \frac{1}{3}\left[\frac{1}{3}(y_1 + y_2 + y_3) + \frac{1}{3}(y_2 + y_3 + y_4) + \frac{1}{3}(y_3 + y_4 + y_5)\right] =$$
$$= \frac{1}{9}(y_1 + 2y_2 + 3y_3 + 2y_4 + y_5),$$

так что описанный способ двухступенчатого выравнивания сводится по существу к одноступенчатому выравниванию по пяти точкам с весовыми множителями 1, 2, 3, 2, 1 (это — числители весов; сумма их равна, конечно, общему знаменателю этих весов: $1 + 2 + 3 + 2 + 1 = 9$).

Аналогично можно провести, например, четырехступенчатое прямолинейное выравнивание по двум точкам, что также сведется к одноступенчатому криволинейному (т. е. с неодинаковыми весами) выравниванию по пяти точкам. Так, последовательные ступени дают:

1) $\frac{1}{2}(y_1 + y_2)$, $\frac{1}{2}(y_2 + y_3)$, $\frac{1}{2}(y_3 + y_4)$,

[1] Точнее, веса равны: $-\frac{3}{35}, \frac{12}{35}, \frac{17}{35}, \frac{12}{35}, -\frac{3}{35}$.

$$2)\ \frac{1}{2}\left[\frac{1}{2}(y_1+y_2)+\frac{1}{2}(y_2+y_3)\right]=\frac{1}{4}(y_1+2y_2+y_3),$$

$$3)\ \frac{1}{2}\left[\frac{1}{4}(y_1+2y_2+y_3)+\frac{1}{4}(y_2+2y_3+y_4)\right]=$$

$$=\frac{1}{8}(y_1+3y_2+3y_3+y_4),$$

$$4)\ \frac{1}{2}\left[\frac{1}{8}(y_1+3y_2+3y_3+y_4)+\frac{1}{8}(y_2+3y_3+3y_4+y_5)\right]=$$

$$=\frac{1}{16}(y_1+4y_2+6y_3+4y_4+y_5),$$

Таким образом, здесь получаются весовые множители 1, 4, 6, 4, 1 (легко видеть, что они представляют собой биномиальные коэффициенты — так же, как и на предыдущих ступенях этого выравнивания).

Итак, мы имеем уже по крайней мере три набора весовых множителей для криволинейного выравнивания:

Числители					Знаменатели
−1,	4,	6,	4,	−1	12
1,	2,	3,	2,	1	9
1,	4,	6,	4,	1	16

Нетрудно сконструировать и другие наборы весов.

Можно ли выделить один какой-нибудь определенный набор весов как наиболее «правильный»? Это невозможно по следующей причине. Каждый набор весов отвечает определенной кривой, применяемой для моделирования фактической зависимости. Но последняя ведь неизвестна — она как раз и составляет предмет поисков. Поэтому заранее нельзя сказать, какая кривая будет лучше всего моделировать фактическую зависимость. Более того, эта зависимость может быть такова, что на разных участках она лучше всего моделируется разными кривыми.

Поэтому достаточно разумный критерий для предпочтения определенного набора весов может состоять в том, чтобы этот набор приводил к возможно более простым вычислениям. Очевидно, таким будет набор

$$1,\ 2,\ 4,\ 2,\ 1,$$

так как знаменатель здесь равен 10 $(1+2+4+2+1=10)$; тем самым мы избавляемся от операции деления для каждой выравненной точки, заменяя эту операцию простым смещением запятой. Для рассмотренного выше примера имеем

$$\widetilde{y}_4=(1\cdot 6+2\cdot 20+4\cdot 27+2\cdot 53+1\cdot 55):10=315:10=31{,}5,$$

а в общем виде

$$\tilde{y}_i = (y_{i-2} + 2y_{i-1} + 4y_i + 2y_{i+1} + y_{i+2}) : 10. \qquad (9.17)$$

Чтобы не потерять четыре точки ряда (по две с каждой стороны), для которых формула (9.17) неприменима, поступаем так же, как и при прямолинейном выравнивании,— сначала производим экстраполяцию за пределы ряда, находя значения y_{n+1} и y_{n+2} по эмпирическим значениям y_n, y_{n-1}, y_{n-2} и y_{n-3}, а затем получившиеся значения y_{n+1} и y_{n+2}; используем в формуле (9.17) для вычисления выравненных значений $\tilde{y}_{n-1}$ и $\tilde{y}_n$; аналогичным образом получаем выравненные значения $\tilde{y}_1$ и $\tilde{y}_2$. Разница только в том, что для экстраполяции мы здесь применяем не прямую, а соответствующую кривую. Не приводя расчетов, дадим сразу готовые формулы:

$$\tilde{y}_1 = (7y_1 + 5y_2 - y_3 - y_4) : 10; \qquad (9.18)$$

$$\tilde{y}_2 = (3y_1 + 5y_2 + y_3 + y_4) \quad 10. \qquad (9.19)$$

При вычислении $\tilde{y}_n$ и $\tilde{y}_{n-1}$ нужно нумерацию точек вести с конца.

Значения $\tilde{y}_i$, вычисленные для нашего примера по формулам (9.17) — (9.19), записаны в последней строке табл. 133. Так,

$$\tilde{y}_1 = (7 \cdot 8 + 5 \cdot 6 - 20 - 27) \quad 10 = 3{,}9;$$
$$\tilde{y}_2 = (3 \cdot 8 + 5 \cdot 6 + 20 + 27) : 10 = 10{,}1;$$
$$\tilde{y}_3 = (8 + 2 \cdot 6 + 4 \cdot 20 + 2 \cdot 27 + 53) \quad 10 = 20{,}7;$$
$$\cdots$$
$$\tilde{y}_{10} = (7 \cdot 39 + 5 \cdot 50 - 56 - 48) : 10 = 41{,}9.$$

Выравненный таким способом ряд изображен на рис. 59 (пунктирная линия).

Для расчетов по формуле (9.17) удобно составить на крае листа бумаги трафаретку с числами 1, 2, 4, 2, 1, которую перемещают вдоль выравниваемого ряда.

КОРРЕЛЯЦИЯ ПРИ ПОРЯДКОВЫХ И КАЧЕСТВЕННЫХ ПРИЗНАКАХ

§ 1. Корреляция рангов

Изложенный в гл. 8, § 4 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 1. Бегуны, ранги которых при построении по росту были 1, 2,... 10, заняли на состязаниях 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

$$\rho = \frac{\Sigma_{xy}}{\sqrt{\Sigma_{xx}\Sigma_{yy}}} = \frac{\Sigma\Sigma\,(x-\hat{x})\,(y-\hat{y})}{\sqrt{\Sigma\,(x-\hat{x})^2\,\Sigma\,(y-\hat{y})^2}} \qquad (8.20)$$

Однако мы знаем, что 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).

Расчет по формуле (8. 20) показан в табл. 134. Корреляция между рангами оказалась сравнительно небольшой ($\rho^S \approx 0{,}45$).

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

Таблица 134

Ранг роста x	Ранг в беге y	x^2	y^2	xy
1	6	1	36	6
2	5	4	25	10
3	1	9	1	3
4	4	16	16	16
5	2	25	4	10
6	7	36	49	42
7	8	49	64	56
8	10	64	100	80
9	3	81	9	27
10	9	100	81	90
55	55	385	385	340
$X_{(1)}$	$Y_{(1)}$	$X_{(2)}$	$Y_{(2)}$	(XY)

$$\Sigma_{xx} = 385 - \frac{55^2}{10} = 385 - 302{,}5 = 82{,}5 = \Sigma_{yy};$$

$$\Sigma_{xy} = 340 - \frac{55 \cdot 55}{10} = 340 - 302{,}5 = 37{,}5;$$

$$\rho^S = \frac{37{,}5}{82{,}5} \approx 0{,}45.$$

рангом признака y и соответствующим ему рангом признака x:

$$d = y - x;$$

в нашем примере

$$d_1 = y_1 - x_1 = 6 - 1 = 5;$$
$$d_2 = y_2 - x_2 = 5 - 2 = 3;$$
$$d_3 = y_3 - x_3 = 1 - 3 = -2 \text{ и т. д.}$$

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 $d_{\text{ср}}$. Однако это не должна быть средняя арифметическая

$$\hat{d} = \frac{\Sigma d}{N},$$

ибо 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 квадратичную, 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 (8.20). В данном случае

$$X_{(1)} = Y_{(1)} = \Sigma x = \Sigma y = 1 + 2 + \ldots + N;$$
$$X_{(2)} = Y_{(2)} = \Sigma x^2 = \Sigma y^2 = 1^2 + 2^2 + \ldots + N^2.$$

Из алгебры известно, что

$$1 + 2 + \ldots + N = \frac{N(N+1)}{2},$$
$$1^2 + 2^2 + \ldots + N^2 = \frac{N(N+1)(2N+1)}{6};$$

поэтому

$$\Sigma_{xx} = \Sigma_{yy} = X_{(2)} - \frac{X_{(1)}^2}{N} = Y_{(2)} - \frac{Y_{(1)}^2}{N} =$$
$$= \frac{N(N+1)(2N+1)}{6} - \frac{N^2(N+1)^2}{4N}.$$

После элементарных преобразований получаем

$$\frac{N(N+1)}{12}[2(2N+1) - 3(N+1)] = \frac{N(N+1)(N-1)}{12} = \frac{N(N^2-1)}{12}.$$

Далее,

$$\Sigma d^2 = \Sigma (y - x)^2 = \Sigma [(y - \hat{y}) - (x - \hat{x})]^2 =$$
$$= \Sigma (y - \hat{y})^2 + \Sigma (x - \hat{x})^2 - 2\Sigma (x - \hat{x})(y - \hat{y}),$$

так что

$$\Sigma_{xy} = \Sigma (x - \hat{x})(y - \hat{y}) = \frac{1}{2}(\Sigma_{yy} + \Sigma_{xx} - \Sigma d^2) = \frac{N(N^2-1)}{12} - \frac{1}{2}\Sigma d^2.$$

Поэтому окончательно

$$\rho^S \quad \frac{\Sigma_{xy}}{\sqrt{\Sigma_{xx}\Sigma_{yy}}} = \frac{\Sigma_{xy}}{\Sigma_{xx}} = 1 - \frac{6\Sigma d^2}{N(N^2-1)} \tag{10.1}$$

(формула Спирмена).

Применяя эту формулу к нашему примеру (см. табл. 135), имеем:

$$\Sigma d^2 = 25 + 9 + 4 + 0 + 9 + \quad + 36 + 1 = 90,$$

$$N(N^2 - 1) = 10(100 - 1) = 990, \quad \rho^S = 1 - \frac{6 \cdot 90}{990} \approx 0{,}45,$$

как и ранее.

Таблица 135

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 2. 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. 136, имеем:

$$\Sigma d^2 = 118, \quad N(N^2 - 1) = 15(225 - 1) = 3360;$$

$$\rho^S = 1 - \frac{6 \cdot 118}{3360} \approx 1 - 0{,}21 = 0{,}79.$$

Таблица 136

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 3. Произведем приближенную оценку коэффициента корреляции между урожайностью пшеницы и картофеля на соседних полях по данным табл. 113. Составив табл. 137, имеем

$$\rho^S = 1 - \frac{6 \cdot 16}{12 \cdot 143} = 1 - 0{,}056 = 0{,}944,$$

что почти точно совпадает с вычисленным ранее коэффициентом корреляции $\rho = 0{,}941$.

Таблица 137

Годы	Пшеница, u	Картофель, m	Ранг x	Ранг y	d	d^2
1926	20,1	7,2	3	4	1	1
1927	23,6	7,1	5	3	—2	4
1928	26,3	7,4	6	6	0	0
1929	19,9	6,1	2	2	0	0
1930	16,7	6,0	1	1	0	0
1931	23,2	7,3	4	5	1	1
1932	31,4	9,4	10	10	0	0
1933	33,5	9,2	11	9	—2	4
1934	28,2	8,8	7	8	1	1
1935	35,3	10,4	12	12	0	0
1936	29,3	8,0	8	7	—1	1
1937	30,5	9,7	9	11	2	4
$N = 12$					0	16

Конечно, не всегда совпадение бывает таким хорошим, но в большинстве случаев значения ρ^S и ρ 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 ρ^S, то она очевидна из сравнения примеров 3 из этой главы и 6 из главы 8.

В § 5 гл. 8 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.

Аналогичным образом можно указать минимальные доверительные значения выборочного показателя корреляции рангов r^S. Анализ показывает, что при $n > 10$ хорошую оценку величины r_α^S дает выражение

$$r_\alpha^S(n) = \frac{u_\alpha}{\sqrt{n-1}}\left[1 - \frac{0{,}19}{n-1}(u_\alpha^2 - 3)\right]. \qquad (10.2)$$

Учитывая, что $u_{05} = 1{,}96$, $u_{01} = 2{,}58$, получаем:

$$r_{05}^S(n) = \frac{1{,}96}{\sqrt{n-1}}\left(1 - \frac{0{,}16}{n-1}\right); \quad r_{01}^S(n) = \frac{2{,}58}{\sqrt{n-1}}\left(1 - \frac{0{,}69}{n-1}\right).$$

Значения $r_{05}^S(n)$ и $r_{01}^S(n)$ приведены в табл. XXVIII Приложений.

§ 2. Связь между признаками с качественной группировкой

При группировке по качественному признаку единственными числами, имеющимися в нашем распоряжении, являются частоты распределения; поэтому задача построения количественной характеристики связи между двумя качественными признаками (в этом случае связь обычно называют *сопряженностью*) сводится к построению такого выражения, которое включало бы только частоты (численности).

Пример 4. В табл. 138 представлены данные об окраске цветов и характере поверхности плодов у гибридов дурмана. Имеется ли сопряженность между этими двумя признаками?

На первый взгляд ответ на поставленный вопрос может быть дан без всяких вычислений. Действительно, среди «лиловых» растений больше «колючих» (47), чем «гладких» (12); в то же время среди «колючих» растений больше «лиловых» (47), чем «белых» (21). Поэтому сопряженность лиловой окраски цветов с колючей поверхностью плодов представляется очевидной.

Однако спросим себя, какого распределения частот следовало бы ожидать, если бы такая сопряженность заведомо отсутствовала; можно ли утверждать, что в этом случае число растений с колючими и гладкими плодами среди имеющих лиловые цветы было бы одинаковым? Очевидно, нет, ибо растений с колючими плодами в исследованной совокупности вообще больше, чем растений с гладкими плодами. Стало быть, тот факт, что среди «лиловых» растений больше «колючих», чем «гладких», может быть просто следствием неравномерности общего распределения «колючих» и «гладких» растений (68 и 15).

Таблица 138

Окраска цветов	Поверхность плодов		Сумма
	колючая	гладкая	
Лиловая .	47_{48}	12_{11}	59
Белая .	21_{20}	3_4	24
Сумма .	68	15	83

Поэтому сопряженность между лиловой окраской цветов и колючей поверхностью плодов следует считать действительно существующей не тогда, когда сочетание этих вариант появляется чаще, чем какое-либо другое сочетание, а тогда, когда оно появляется чаще, чем это должно было быть при независимости обоих признаков (окраски цветов и поверхности плодов).

Пусть мы имеем распределение цветов по двум качественным признакам: по окраске A с оттенками $A_1, A_2, \ldots, A_i, \ldots, A_{k_A}$ и по форме B с видами формы $B_1, B_2, \ldots, B_j, \ldots, B_{k_B}$. Цветы, имеющие окраску A_i и форму B_j, обозначим A_iB_j, а число таких цветов в исследуемой совокупности обозначим через n_{ij}. Тогда мы получим распределение, показанное в табл. 139. Очевидно,

$$n_{i1} + n_{i2} + \quad + n_{ij} + \ldots + n_{ik_B} = n_{i\cdot}$$

есть число всех цветов с окраской A_i (со всеми формами); аналогично

$$n_{1j} + n_{2j} + \ldots + n_{ij} + \ldots + n_{k_Aj} = n_{\cdot j}$$

есть число всех цветов с формой B_j (со всеми оттенками окраски). Далее,

$$n_{1\cdot} + n_{2\cdot} + \ldots + n_{i\cdot} + \ldots + n_{k_A\cdot} = N$$

есть число всех цветов в совокупности, т. е. ее объем N, причем также

$$n_{\cdot 1} + n_{\cdot 2} + \quad + n_{\cdot j} + \ldots + n_{\cdot k_B} = N.$$

Если признаки A и B взаимно независимы, то цветы с окраской A_i делятся между формами $B_1, B_2, \ldots, B_{k_B}$ в той же пропорции, в какой все цветы делятся между этими формами, т. е. численности n_{i1}, n_{i2} и т. д. пропорциональны суммарным численностям

Таблица 139

A \ B	B_1	B_2		B_j		B_{k_B}	
A_1	n_{11}	n_{12}		n_{1j}		n_{1k_B}	$n_{1\bullet}$
A_2	n_{21}	n_{22}		n_{2j}		n_{2k_B}	$n_{2\bullet}$
A_i	n_{i1}	n_{i2}		n_{ij}		n_{ik_B}	$n_{i\bullet}$
. . .	. . .	. . .		. . .		. . .	. . .
A_{k_A}	n_{k_A1}	n_{k_A2}		n_{k_Aj}		$n_{k_Ak_B}$	$n_{k_A\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet j}$		$n_{\bullet k_B}$	N

n_{*1}, n_{*2} и т. д. Но это значит, что доля цветов с формой B_j и окраской A_i среди всех цветов с окраской A_i такова же, как доля всех цветов с формой B_j среди всех цветов совокупности.

Алгебраически это запишется в виде

$$\frac{n_{ij}}{n_{i_\bullet}} = \frac{n_{\bullet j}}{N}$$

Это равенство можно переписать так:

$$n_{ij} = \frac{n_{i_\bullet} n_{\bullet j}}{N}, \tag{10.3}$$

или

$$\frac{n_{ij}}{N} = \frac{n_{i_\bullet}}{N} \cdot \frac{n_{\bullet j}}{N}, \tag{10.4}$$

т. е. при независимости признаков A и B доля n_{ij} в совокупности равна произведению долей $n_{i_\bullet}$ и $n_{\bullet j}$.

Применяя условие (10.3) к данным из табл. 138, находим, что при независимости окраски цветов и характера поверхности плодов число растений с лиловыми цветами и колючими плодами должно было бы составлять

$$\frac{59 \cdot 68}{83} \approx 48,$$

т. е. даже несколько больше, чем фактическое число 47; значит, нет никаких оснований утверждать, что имеется положительная корреляция между лиловой окраской цветов и колючей поверхностью плодов.

Если через $\hat{n}_{ij}$ обозначить «теоретическую» численность варианты A_iB, т. е. численность, отвечающую условию независимости признаков (10.3), то в данном случае будет

$$n_{11} < \hat{n}_{11},$$

что говорит скорей об отрицательной корреляции.

Величину и знак корреляции между классами A_i и B_j естественно характеризовать величиной отклонения эмпирической численности от «теоретической»:

$$\xi_{ij} = n_{ij} - \hat{n}_{ij}.$$

Однако нашей целью является изучение корреляции не для каждой клетки корреляционной решетки, а для всей решетки в целом, т. е. вообще для признаков A и B. Поэтому нужно построить некий обобщенный показатель, представляющий собой известное усреднение ξ_{ij}. При построении этого показателя должно быть учтено следующее.

А. При одновременном увеличении всех численностей в l раз во столько же раз увеличатся и все ξ_{ij}, но искомый показатель (коэффициент) сопряженности не должен при этом и измениться; поэтому следует усреднять не абсолютные отклонения ξ_{ij}, а относительные отклонения

$$\delta_{ij} = \frac{\xi_{ij}}{\hat{n}_{ij}} = \frac{n_{ij} - \hat{n}_{ij}}{\hat{n}_{ij}}. \qquad (*)$$

Отнесение отклонений ξ_{ij} к «теоретическим» численностям $\hat{n}_{ij}$, а не к фактическим n_{ij} удобней потому, что некоторые n_{ij} могут быть равны нулю; кроме того, поскольку отклонения вообще отсчитываются от «теоретических» численностей, то тем самым последние как бы рассматриваются в качестве неких «базовых» величин.

Б. Хотя сумма относительных отклонений δ_{ij}, в отличие от суммы абсолютных отклонений ξ_{ij}, в общем случае не равна нулю, все же при суммировании δ_{ij} будет происходить частичная компенсация положительных и отрицательных δ_{ij}; в некоторых случаях компенсация может оказаться даже полной, несмотря на явную сопряженность признаков. Так, для табл. 140 все $\hat{n}_{ij}$ равны

$$\frac{10 \cdot 10}{20} = 5,$$

так что

$$\delta_{11} = \frac{10-5}{5} = 1; \quad \delta_{12} = \frac{0-5}{5} = -1;$$

$$\delta_{21} = \frac{0-5}{5} = -1; \quad \delta_{22} = \frac{10-5}{5} = 1;$$

сумма всех δ_{ij} равна нулю, в то время как сопряженность является полной.

Таблица 140

	B_1	B_2	B
A_1	10	0	10
A_2	0	10	10
A	10	10	20

Чтобы избежать такой компенсации, следует, очевидно, усреднять не сами δ_{ij}, а их квадраты δ_{ij}^2.

В. Целесообразно, чтобы варианты, представленные в совокупности в большем числе, оказывали большее влияние на результат усреднения. Это значит, что усреднение должно производиться с учетом «веса» каждой варианты. При этом в качестве «весов» будем брать не фактические, а теоретические численности, поскольку именно последним отведена роль «базовых» величин; кроме того, это удобней и в вычислительном отношении.

Таким образом, мы приходим к величине

$$\varphi = \sqrt{\frac{1}{N}\sum_{i,j}\hat{n}_{ij}\delta_{ij}^2}\,, \qquad (**)$$

которая представляет собой, очевидно, среднее квадратичное относительное отклонение. Подставляя сюда значение δ_{ij} из (*), получаем

$$\varphi = \sqrt{\frac{1}{N}\sum_{i,j}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}}$$

Так как

$$\sum_{i,j}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} = \chi^2 \qquad (10.5)$$

(см. гл. 6), то

$$\varphi = \sqrt{\frac{\chi^2}{N}}. \qquad (10.6)$$

Г. Целесообразно подобрать характеристику сопряженности так, чтобы при полной сопряженности она равнялась единице.

Рассмотрим пример такой полной сопряженности, причем для простоты возьмем тот случай, когда 1) оба признака A и B разделяются на одинаковое число k классов, так что корреляционная решетка является квадратной, и 2) численность всех классов одинакова. Тогда корреляционная решетка будет иметь примерный вид табл. 141.

Таблица 141

A \ B	B_1	B_2	B_3	B_4		B_k	
A_1		ν					ν
A_2						ν	ν
A_3			ν				ν
.							
A_{k-1}	ν						ν
A_k				ν			ν
$n_{\cdot j}$	ν	ν	ν	ν		ν	N

Но качественные признаки характеризуются тем, что классы группировки можно располагать в любой последовательности, так что нумерация классов является произвольной. Поэтому мы можем так изменить эту нумерацию, чтобы численности ν расположились по диагонали. Если, например, приписать классу B_2 номер B_1, классу B_k номер B_2 и т. д., то получится табл. 142. В этом случае $n_{ij} = \nu$ при $i = j$ (т. е. для диагональных клеток) и $n_{ij} = 0$ при $i \neq j$.

Таблица 142

A \ B	B_1	B_2		B_k	$n_{i\cdot}$
A_1	ν				ν
A_2		ν			ν
.					.
A_k				ν	ν
$n_{\cdot j}$	ν	ν		ν	N

Однако мы могли бы с равным правом приписать классу B_2 номер B_k, классу B_k номер B_{k-1} и т. д., и тогда численности

расположились бы по отрицательной диагонали. Отсюда ясно, что при качественной классификации нет оснований различать положительную и отрицательную сопряженность. Поэтому будем всегда считать квадратный корень в (**) положительным.]

Итак, мы получили табл. 142. Поскольку все $n_{i\cdot}$ и $n_{\cdot j}$ одинаковы и равны ν, а $N = k\nu$, то все «теоретические» численности будут

$$\hat{n}_{ij} \equiv \nu_0 = \frac{\nu^2}{N} = \frac{\nu}{k}.$$

Тогда для диагональных клеток

$$\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \frac{(\nu - \nu_0)^2}{\nu_0} = \frac{(\nu - \nu/k)^2}{\nu/k} = \frac{\nu(k-1)^2}{k},$$

а для остальных клеток

$$\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \frac{(0 - \nu_0)^2}{\nu_0} = \nu_0 = \frac{\nu}{k}.$$

Поэтому вклад в χ^2 от диагональных клеток, число которых равно k, будет

$$k\,\frac{\nu(k-1)^2}{k} = \nu(k-1)^2,$$

а вклад от недиагональных клеток, число которых есть $k^2 - k = k(k-1)$, будет

$$k(k-1)\,\frac{\nu}{k} = \nu(k-1),$$

так что для рассматриваемого случая полной сопряженности получится

$$\chi^2 = \nu(k-1)^2 + \nu(k-1) = \nu(k-1)[(k-1)+1] = k\nu(k-1) = N(k-1)$$

и

$$\varphi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{N(k-1)}{N}} = \sqrt{k-1}.$$

Отсюда следует, что в данном случае

$$\frac{\varphi}{\sqrt{k-1}} = 1.$$

Значит, при полной корреляции равна единице не величина (10.6), а величина

$$K = \sqrt{\frac{\chi^2}{N(k-1)}}.$$

В общем случае, когда число классов в признаках A и B неодинаково ($k_A \neq k_B$), величина $k-1$ заменяется средним геометрическим из $k_A - 1$ и $k_B - 1$, т. е. величиной

$$\sqrt{(k_A - 1)(k_B - 1)};$$

тогда

$$K = \sqrt{\frac{\chi^2}{N\sqrt{(k_A - 1)(k_B - 1)}}}. \qquad (10.7)$$

Эта величина (предложенная А. А. Чупровым) называется *коэффициентом взаимной сопряженности.*

Таблица 143

Цвет глаз	Цвет волос				Сумма
	светлый B_1	русый B_2	черный B_3	рыжий B_4	
Голубой A_1	177 117 60	71 96 −25	17 47 −30	14 19 −5	279
Серый A_2	95 131 −36	119 108 11	75 53 22	25 22 3	314
Карий A_3	12 36 −24	44 30 14	23 15 8	8 6 2	87
Сумма	284	234	115	47	680

Пример 5. Табл. 143 содержит данные о группировке 680 человек по двум признакам — по цвету волос и по цвету глаз. В каждой клетке записано три числа. Первое означает фактическую численность, например, 75 человек из 680 имеют серые глаза и черные волосы. Второе число представляет собой «теоретическую» частоту (т. е. вычисленную в предположении независимости признаков); эти числа получаются по формуле (10.3);

например,

$$\frac{314 \cdot 115}{680} \approx 53.$$

Третье число в каждой клетке есть

$$\xi_{ij} = n_{ij} - \hat{n}_{ij};$$

например,

$$75 - 53 = 22.$$

Если все вычисления произведены правильно, то сумма отклонений для каждой строки и каждого столбца должна равняться нулю; например, для первой строки

$$60 + (-25) + (-30) + (-5) = 0,$$

для третьего столбца

$$-30 + 22 + 8 = 0$$

и т. д.

Теперь находим χ^2:

$$\begin{array}{ll} 60^2 : 117 = 30{,}8 & 22^2 : 53 = 9{,}1 \\ 36^2 : 131 = 9{,}9 & 8^2 : 15 = 4{,}3 \\ 24^2 : 36 = 16{,}0 & 5^2 : 19 = 1{,}3 \\ 25^2 : 96 = 6{,}5 & 3^2 : 22 = 0{,}4 \\ 11^2 : 108 = 1{,}1 & 2^2 : 6 = 0{,}7 \\ 14^2 : 30 = 6{,}5 & \hline \\ 30^2 : 47 = 19{,}2 & \chi^2 = 105{,}8 \approx 106 \end{array}$$

Поскольку в данном случае значение χ^2 велико, можно пользоваться упрощенной формулой (6.2). Для нашего примера получим:

$$\begin{array}{ll} 177^2 : 117 = 268 & 75^2 : 53 = 106 \\ 95^2 : 131 = 69 & 23^2 : 15 = 35 \\ 12^2 : 36 = 4 & 14^2 : 19 = 10 \\ 71^2 : 96 = 53 & 25^2 : 22 = 28 \\ 119^2 : 108 = 131 & 8^2 : 6 = 11 \\ 44^2 : 30 = 65 & \hline \\ 17^2 \quad 47 = 6 & 786 \end{array}$$

так что

$$\chi^2 = 786 - 680 = 106,$$

как и ранее,

Принятая здесь точность расчетов вполне достаточна; увеличение числа значащих цифр практически не сказалось бы на окончательном результате. Напомним, однако, что значение χ^2 может содержать ошибку, если имеются очень малые «теоретические» частоты: слагаемые χ^2, в которые эти частоты входят делителями, особенно чувствительны к неточностям в этих делителях именно при малых значениях последних. Практически принимается, что ни одно из значений $\hat{n}_{ij}$ не должно быть меньше 3; в противном случае приходится прибегать к объединению групп. Например, если бы в табл. 143 получилось для $\hat{n}_{34}$ не 6, а, скажем, 2, то пришлось бы вообще ликвидировать эту варианту как самостоятельный элемент группировки. Практически это означает, что нужно было бы либо включить строку A_3 в какую-нибудь другую строку, либо включить столбец B_4 в какой-нибудь другой столбец. Так как в данном случае численность столбца B_4 меньше, чем численность строки A_3 ($47 < 87$), то было бы целесообразно ликвидировать именно столбец B_4. Вопрос о том, с каким именно столбцом слить этот столбец, должен решаться, разумеется, в соответствии с тем, к какому он ближе по смыслу. Вероятно, в данном случае было бы резонно слить его со столбцом B_1 (светлые волосы).

Итак, мы получили $\chi^2 = 106$. Тогда

$$K = \sqrt{\frac{106}{680\sqrt{2\cdot 3}}} \approx \sqrt{0{,}064} \approx 0{,}25.$$

Мы видим, что сопряженность между цветом волос и цветом глаз в общем не особенно велика.

Отметим в заключение, что если в исходной таблице имеются клетки с нулевой численностью, то эти клетки также должны входить в счет при вычислении χ^2. Очевидно, вклад каждой такой клетки равен просто $\hat{n}_{ij}$:

$$\frac{(0-\hat{n}_{ij})^2}{\hat{n}_{ij}} = \hat{n}_{ij}.$$

§ 3. Двумерное альтернативное распределение

Весьма важным частным случаем является четырехклеточная группировка типа табл. 95. В этом случае вычисление коэффициента взаимной сопряженности сильно упрощается. Прежде всего, при этом $k_A = k_B = 2$, так что

$$\sqrt{(k_A - 1)(k_B - 1)} = 1.$$

Далее, как указывалось в § 4 гл. 6, при $k_A = k_B = 2$ формула (10.5) принимает вид

$$\chi^2 = \frac{|n_{11}n_{22} - n_{12}n_{21}|^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}} N; \tag{10.8}$$

поэтому

$$K = \frac{|n_{11}n_{22} - n_{12}n_{21}|}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}. \tag{10.9}$$

Здесь отпадает необходимость в вычислении ожидаемых численностей. Если ввести поправку на группировку (см. § 4 гл. 6), то

$$\chi^2 = \frac{(|n_{11}n_{22} - n_{12}n_{21}| - N/2)^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}} N \tag{10.10}$$

и

$$K = \frac{|n_{11}n_{22} - n_{12}n_{21}| - N/2}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{2\cdot}}}. \tag{10.11}$$

Пример 6. Вычислим коэффициент взаимной сопряженности для опыта с иммунизацией телят от туберкулеза (см. пример 16 из гл. 4 и пример 8 из гл. 6).

Чтобы воспользоваться общей формулой (10.7), нужно сначала вычислить «теоретические» численности $\hat{n}_{ij}$ и отклонения ξ_{ij} во всех четырех клетках; соответствующие значения записаны в табл. 144. Теперь находим

$$\chi^2 = 5{,}29^2 \left(\frac{1}{11{,}29} + \frac{1}{10{,}71} + \frac{1}{8{,}71} + \frac{1}{8{,}29}\right) =$$
$$= 28{,}0\ (0{,}089 + 0{,}093 + 0{,}115 + 0{,}121) = 28{,}0 \cdot 0{,}418,$$

Таблица 144

	Заболевшие	Незаболевшие	Сумма
С прививкой	6 11,29 −5,29	14 8,71 5,29	20
Без прививки	16 10,71 5,29	3 8,29 −5,29	19
Сумма	22	17	39

так что

$$K = \sqrt{\frac{28{,}0 \cdot 0{,}418}{38}} \approx \sqrt{0{,}30} \approx 0{,}55.$$

Если же воспользоваться формулой (10.9), то получим сразу

$$K = \frac{|6 \cdot 3 - 14 \cdot 16|}{\sqrt{20 \cdot 19 \cdot 22 \cdot 17}} \approx \frac{|18 - 224|}{\sqrt{14{,}2 \cdot 10^4}} \approx \frac{206}{377} \approx 0{,}55.$$

Более правильное значение получится по формуле (10.11):

$$K = \frac{|6 \cdot 3 - 14 \cdot 16| - 39/2}{\sqrt{20 \cdot 19 \cdot 22 \cdot 17}} \approx \frac{186{,}5}{377} \approx 0{,}495.$$

Как и всякий статистический параметр, характеризующий совокупности с качественной классификацией, коэффициент взаимной сопряженности может быть применен и к совокупностям

Таблица 145

	B_1	B_2	n_x	x	$n_x x$	$n_x x^2$
A_1	6	14	20	−1	−20	20
A_2	16	3	19	+1	+19	19
n_y	22	17	39		−1	39
y	−1	1			$X_{(1)}$	$X_{(2)}$
$n_y y$	−22	17	−5	$Y_{(1)}$	+	9
$n_y y^2$	22	17	39	$Y_{(2)}$	−	30
					(XY)	−21

$$\Sigma_{xx} = 39 - \frac{1}{39} \approx 38{,}98;$$

$$\Sigma_{yy} = 39 - \frac{25}{39} \approx 38{,}36;$$

$$\Sigma_{xy} = -21 - \frac{5}{39} \approx -21{,}13;$$

$$\rho = \frac{|-21{,}13|}{\sqrt{38{,}98 \cdot 38{,}36}} \approx 0{,}55.$$

с количественной классификацией (считая количественные разряды качественными классами). Но при этом значение K в общем случае не будет совпадать со значением коэффициента корреляции ρ.

Однако для четырехклеточных таблиц K и ρ совпадают, в чем можно убедиться как общими выкладками, так и численной проверкой (табл. 145).

§ 4. Оценка значимости коэффициента взаимной сопряженности

Согласно формуле (10.7) выборочный коэффициент взаимной сопряженности между признаками с качественной группировкой равен

$$K = \sqrt{\frac{\chi^2}{n\sqrt{(k_A-1)(k_B-1)}}},$$

где n — объем выборки, а k_A и k_B — число строк и столбцов корреляционной решетки. Естественно считать величину K значимо отличной от нуля, если значима входящая в K величина χ^2. Число степеней свободы равно здесь

$$f = (k_A - 1)(k_B - 1).$$

Проверим значимость рассмотренной в § 2 этой главы сопряженности между цветом волос и цветом глаз (см. табл. 143). При $k_A = 3$, $k_B = 4$ и $n = 680$ было получено $\chi^2 \approx 106$, $K \approx 0{,}25$. Поскольку при $f = 2 \cdot 3 = 6$

$$\chi^2_{01}(6) = 16{,}8,$$

то ясно, что полученная величина K значимо отлична от нуля $(\chi^2 > \chi^2_{01})$.

Если имеются данные о сопряженности по нескольким выборкам из генеральной совокупности, то для нахождения обобщенной оценки значимости этой сопряженности нет нужды объединять эти выборки в одну общую выборку с последующим вычислением коэффициента сопряженности на основании этой общей выборки. Достаточно просто сложить все частные значения χ^2 в одну общую величину $\chi^2_{\text{общ}}$, приписав ей число степеней свободы $f_{\text{общ}}$, равное сумме частных значений f.

Когда количество классов группировки велико, число степеней свободы $f = (k_A - 1)(k_B - 1)$ может оказаться больше предусмотренного в табл. XIX Приложений. В таких случаях можно

воспользоваться тем, что при большом числе степеней свободы f величина $\sqrt{2\chi^2}$ распределена приближенно нормально со средним значением $\sqrt{2f-1}$ и стандартным отклонением 1. Поэтому значимость величины χ^2 можно оценить, сравнив

$$u = \sqrt{2\chi^2} - \sqrt{2f-1} \qquad (10.12)$$

с u_α. В данном случае критерий должен быть односторонним, ибо 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

$$\Phi(u_\alpha) = 1 - \alpha \qquad (*)$$

(Φ — интеграл вероятностей). Так как

$$\Phi(u) = \frac{1}{2}\theta(u) + \frac{1}{2}$$

[см. формулу (2.5) в гл. 2, § 3], то условие (*) можно переписать в виде

$$\theta(u_\alpha) = 1 - 2\alpha.$$

Поэтому для одностороннего u-критерия

$$u_{05} = 1{,}64, \quad u_{01} = 2{,}33$$

(см. табл. I Приложений). Эти значения можно получить также непосредственно из табл. II Приложений, так как согласно (*)

$$u_\alpha = \Psi(1 - \alpha),$$

тогда

$$u_{05} = \Psi(1 - 0{,}05) = \Psi(0{,}95) = 1{,}64;$$
$$u_{01} = \Psi(1 - 0{,}01) = \Psi(0{,}99) = 2{,}33.$$

Пример 7. Рассмотрение установленных последовательностей аминокислотных остатков в белковых молекулах позволяет составить таблицу сопряженности, каждая клетка которой показывает, сколь часто после остатка, стоящего в названии строки, следует остаток, стоящий в названии столбца. Выяснение того, является ли сопряженность между последующими и предшествующими остатками значимой, имеет большое значение для раскрытия закономерностей синтеза белков.

Расчет дает для данной таблицы $\chi^2 = 392$. Так как число различных аминокислот, входящих в природные белки, равно 20, то таблица сопряженности имеет 20 строк и 20 столбцов, а число степеней свободы равно, следовательно,

$$f = (20 - 1)(20 - 1) = 361.$$

Тогда формула (10.12) дает

$$u = \sqrt{2 \cdot 392} - \sqrt{2 \cdot 361 - 1} = 28{,}00 - 26{,}85 = 1{,}15.$$

Так как это меньше, чем $u_{05} = 1{,}64$, то мы делаем вывод, что нулевая гипотеза об отсутствии корреляции между соседними аминокислотными остатками в белках не опровергается.

Некоторым неудобством формулы (10.12) является то, что значение u получается как малая разность двух больших величин, а это предъявляет повышенные требования к точности вычислений. Так, в примере 10 пришлось вычислять квадратные корни с четырьмя значащими цифрами; это не позволяет использовать логарифмическую линейку. 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 со средним значением f и стандартным отклонением $\sqrt{2f}$. Поэтому можно принять

$$u = \frac{\chi^2 - f}{\sqrt{2f}}. \qquad (10.13)$$

Для примера 7 имеем

$$u = \frac{392 - 361}{\sqrt{2 \cdot 361}} = \frac{31}{26{,}9} = 1{,}15;$$

здесь оказалось достаточно точности до трех значащих цифр.

При альтернативном распределении (для 4-клеточных таблиц типа 2×2) значение K получается таким же, как вычисленный обычным образом коэффициент корреляции; поэтому значимость K можно проверять в этих случаях по табл. XXVII Приложений. Кроме того, можно проверить значимость K и по χ^2-критерию. Именно, согласно (10.7) имеем здесь

$$\chi^2 = nK^2 \qquad (10.14)$$

при

$$f = (2 - 1)(2 - 1) = 1.$$

В примере 6 мы получили $K = 0{,}495$ при $n = 39$. Тогда

$$\chi^2 = 39 \cdot 0{,}495^2 = 9{,}55,$$

а так как $\chi^2_{01}(1) = 6{,}63$, то найденная сопряженность значима. Тот же результат получим по табл. XXVII Приложений:

$$r_{01}(39) \approx r_{01}(40) = 0{,}407, \text{ т. е. } r > r_{01}.$$

ОБЩАЯ СХЕМА СТАТИСТИЧЕСКОГО АНАЛИЗА

Укажем теперь в общих чертах последовательные этапы выполнения статистического анализа. Будем следовать в основном А. Хальду (1956), излагающему относящиеся к этому вопросу идеи Р. Фишера.

1. Выбор математико-статистической модели

Речь идет прежде всего о выборе типа распределения вариант в генеральной совокупности. Этот выбор определяется имеющейся у исследователя предварительной информацией о свойствах объекта, а также характером задачи.

При решении вопросов, связанных с измерениями размеров, весов и т. п. свойств живых объектов, распределение вариант можно предполагать нормальным (§ 3 гл. 2); при изучении редких явлений (например, появление бактерий в ячейках цитометра) чаще всего употребляется в качестве модели распределение Пуассона (§ 6 гл. 2); если интересуются числом доминантных и рецессивных форм при скрещивании, то подходящей моделью будет альтернативное распределение (§ 8 гл. 2) и т. д. Вообще различным статистическим моделям для одномерных совокупностей посвящена гл. 2.

При исследовании связи между признаками моделью могут служить линейная или нелинейная регрессия (§ 2 гл. 8), разные виды выравнивания (гл. 9), сопряженность ранжированных (§ 1 гл. 10) или качественных признаков (§ 2 и 3 гл. 10).

Решению вопроса о выборе модели очень помогает предварительный графический анализ — построение гистограммы или полигона частот (§ 2 гл. 1), использование вероятностной бумаги (§ 3 гл. 2), составление корреляционного поля (§ 1 гл. 8), различные преобразования координат (§ 4 гл. 2 и § 3 гл. 9) и т. д.; в частности, анализ S-образных кривых производится чаще всего при помощи пробит-метода (§ 5 гл. 2).

Конечно, на первом плане всегда должны стоять соображения, связанные с существом дела. Но обычно известное значение имеет также требование математической простоты модели.

2. Планирование выборочного исследования

Здесь имеется в виду, конечно, не выбор биологического объекта исследования, экспериментальной методики и т. п., а лишь статистическая сторона дела. Поскольку при биологических исследованиях почти всегда имеют дело с выборками, а не с генеральными совокупностями, возникают следующие две задачи:

а) выбор метода получения выборок. Это может быть случайная, зональная, механическая или какая-либо другая выборка — в зависимости от известных или предполагаемых свойств генеральной совокупности. Эти вопросы обсуждаются в § 1 гл. 3;

б) планирование объема выборки, т. е. числа наблюдений. Как правило, это требует наличия некоторой информации о дисперсии генеральной совокупности. Иногда эта информация имеется из предыдущих аналогичных исследований. В общем же случае примерная оценка σ получается из того предварительного исследования, которое проводится с малой выборкой для выбора статистической модели. После того как получена оценка σ, вычисляется необходимый объем окончательной выборки — по формуле, соответствующей выбранной модели и поставленной задаче. Именно, если речь идет о нахождении доверительного интервала для параметра, то используются формулы: (3.38) или (3.39) для доли вариант и (3.20) — для среднего значения в нормальном распределении. Требуемый объем выборки будет зависеть, помимо всего прочего, от принятого доверительного уровня, который определяется целями исследования.

3. Оценка параметров модели

После того как эмпирический материал получен, можно найти выборочные оценки и доверительные интервалы для параметров принятой математико-статистической модели. Прежде всего решается вопрос об исключении из выборки сильно отклоняющихся вариант — при помощи критериев, описанных в § 2 гл. 4. Затем вычисляют значения параметров.

Чаще всего интересуются средним значением (§ 3 и 4 гл. 1) и дисперсией (§ 6 гл. 1). Иногда возникает необходимость в нахождении других показателей: медианы и моды (§ 5 гл. 1), а также коэффициентов асимметрии (§ 7 гл. 1) и эксцесса (§ 4 гл. 2) и других статистических моментов (§ 8 гл. 1).

Характеристиками двумерных совокупностей служат коэффициенты регрессии (§ 2 гл. 8) и корреляции (§ 4 и 9 гл. 8), корреляционные отношения (§ 3 гл. 8), показатель корреляции рангов (§ 1 гл. 10) и коэффициент взаимной сопряженности качественных признаков (§ 2 и 3 гл. 10).

Нахождению выборочных оценок и доверительных интервалов для параметров одномерных совокупностей посвящена гл. 3: для нормального распределения — §§ 2—7, для распределения Пуассона — § 8, для альтернативного распределения — § 9. Построение доверительного интервала для коэффициента корреляции описано в § 5 гл. 8, а доверительной зоны регрессии — в § 7 гл. 8.

4. Проверка пригодности модели

Вычисленные параметры могут служить для адекватного описания изучаемого явления только в том случае, если статистическая модель выбрана правильно; например, коэффициент корреляции вообще не имеет реального смысла, если зависимость между признаками нелинейна. Проверка того, что подобранная модель соответствует (вернее, не противоречит) эмпирическим данным, производится при помощи определенных критериев.

Нормальность распределения можно проверить способами, описанными в § 4 гл. 2 и § 3 гл. 4; второй способ, являясь количественным, более надежен, но зато несколько сложнее. Другие типы распределения удобней проверять по критерию χ^2 (§ 1 гл. 6); для проверки нормальности распределения этот критерий чаще всего оказывается слишком громоздким.

Линейность регрессии проверяется при помощи процедуры, описанной в § 6 гл. 8. Если предположение о линейности не оправдалось, имеет смысл испытать различные преобразования координат, приводящие к линейной зависимости (§ 3 гл. 9). При малом числе наблюдений можно получить указание о характере зависимости по способу, описанному в § 5 гл. 9. Если ни линейная (или приводимая к ней), ни квадратичная модель (§ 4 гл. 9) не подходят, целесообразно использовать выравнивание с помощью скользящего среднего — простого (§ 6 гл. 9) или взвешенного (§ 7 гл. 9).

5. Проверка гипотез при помощи критериев значимости

В биологических приложениях математической статистики этот вопрос является одним из основных. Поэтому в книге ему уделено больше всего внимания. В соответствии с многообразием относящихся сюда задач существует большое число различных критериев и методов. Им целиком посвящены четыре главы — 4—7, а также отдельные параграфы в гл. 8.

Если требуется проверить однотипность двух эмпирических распределений, то при больших объемах выборок удобней критерий χ^2 (§ 2 и 3 гл. 6), а при малых объемах — серийный критерий

(§ 4 гл. 7) и критерий Колмогорова — Смирнова (§ 5 гл. 7), причем второй из них более мощный («чувствительный»), но зато несколько сложнее.

Если интересуются только сравнением «центральных тенденций», то пользуются критериями Вилкоксона (§ 2 гл. 7) или ван дер Вардена (§ 3 гл. 7), причем опять второй — более мощный, но более громоздкий.

Часто желательно не только получить заключение о различии двух средних значений, но и построить доверительный интервал для их разности; тогда пользуются критерием Стьюдента (§ 4 гл. 4).

Если выборочный комплекс представляет собой набор сопряженных пар, то критерии несколько модифицируются (§ 5 гл. 4 и § 6 и 7 гл. 7). Более сложные комплексы, в которых сравниваются сразу несколько выборок, исследуются при помощи дисперсионного анализа (гл. 5); при этом комплекс может быть различным по составу (бесповторный, равномерный, неравномерный) и анализироваться по одному, двум или большему числу факторов (см. оглавление).

Сравнение дисперсий производится по F-критерию (§ 7 гл. 4), а сравнение долей вариант в альтернативных распределениях — по критерию χ^2 (§ 4 гл. 6) или по параметрическому критерию (§ 8 гл. 4).

Если в формулировке нулевой и альтернативной гипотез участвуют два разных численных значения параметра, то целесообразно применять последовательный (секвенциальный) анализ (§ 6 гл. 4).

В § 5 гл. 8 описано сравнение коэффициентов корреляции, в § 8 гл. 8 — сравнение двух линий регрессии и в § 4 гл. 10 — проверка значимости коэффициента взаимной сопряженности качественных признаков.

В заключение подчеркнем еще раз, что если примененный критерий не отвергает нулевую гипотезу, то это вовсе не означает, что справедливость этой гипотезы доказана. Когда неопровергнутая нулевая гипотеза принимается в качестве рабочей гипотезы, имеется некоторая вероятность β ошибиться (вероятность ошибки II рода — см. § 1 гл. 4). Получение более полных данных может в дальнейшем привести к тому, что эту гипотезу придется отвергнуть.

ПРИЛОЖЕНИЯ

ВСПОМОГАТЕЛЬНЫЕ ТАБЛИЦЫ (I—XXIX)

(В квадратных скобках указаны глава и параграф, в котором дается определение табулированной функции)

Таблица I

Значения $\theta(u)$ — площади под нормальной кривой в пределах от $\hat{x} - u\sigma$ до $\hat{x} + u\sigma$ [гл. 2, §3]

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	2586	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	8385	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	9476	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	9715
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	9981	9986	9990	9993	9995	9997	9998	9999	9999

Таблица II

Значения $\Psi(p) = u(\Phi)$ — функции, обратной к интегралу вероятностей [гл. 2, § 5]

$p \rightarrow$ / $\downarrow$	0	1	2	3	4	5	6	7	8	9
0,00	$-\infty$	−3,09	−2,88	−2,75	−2,65	−2,58	−2,51	−2,46	−2,41	−2,37
0,01	−2,33	−2,29	−2,26	−2,23	−2,20	−2,17	−2,14	−2,12	−2,10	−2,07
0,02	−2,05	−2,03	−2,01	−2,00	−1,98	−1,96	−1,94	−1,93	−1,91	−1,90
0,03	−1,88	−1,87	−1,85	−1,84	−1,83	−1,81	−1,80	−1,79	−1,77	−1,76
0,04	−1,75	−1,74	−1,73	−1,72	−1,71	−1,70	−1,68	−1,67	−1,66	−1,65
0,05	−1,64	−1,64	−1,63	−1,62	−1,61	−1,60	−1,59	−1,58	−1,57	−1,56
0,06	−1,55	−1,55	−1,54	−1,53	−1,52	−1,51	−1,51	−1,50	−1,49	−1,48
0,07	−1,48	−1,47	−1,46	−1,45	−1,45	−1,44	−1,43	−1,43	−1,42	−1,41
0,08	−1,41	−1,40	−1,39	−1,39	−1,38	−1,37	−1,37	−1,36	−1,35	−1,35
0,09	−1,34	−1,33	−1,33	−1,32	−1,32	−1,31	−1,30	−1,30	−1,29	−1,29
0,10	−1,28	−1,28	−1,27	−1,26	−1,26	−1,25	−1,25	−1,24	−1,24	−1,23
0,11	−1,23	−1,22	−1,22	−1,21	−1,21	−1,20	−1,20	−1,19	−1,19	−1,18
0,12	−1,18	−1,17	−1,17	−1,16	−1,16	−1,15	−1,15	−1,14	−1,14	−1,13
0,13	−1,13	−1,12	−1,12	−1,11	−1,11	−1,10	−1,10	−1,09	−1,09	−1,09
0,14	−1,08	−1,08	−1,07	−1,07	−1,06	−1,06	−1,05	−1,05	−1,05	−1,04
0,15	−1,04	−1,03	−1,03	−1,02	−1,02	−1,02	−1,01	−1,01	−1,00	−1,00
0,16	−0,99	−0,99	−0,99	−0,98	−0,98	−0,97	−0,97	−0,97	−0,96	−0,96
0,17	−0,95	−0,95	−0,95	−0,94	−0,94	−0,93	−0,93	−0,93	−0,92	−0,92
0,18	−0,92	−0,91	−0,91	−0,90	−0,90	−0,90	−0,89	−0,89	−0,89	−0,88
0,19	−0,88	−0,87	−0,87	−0,87	−0,86	−0,86	−0,86	−0,85	−0,85	−0,85
0,20	−0,84	−0,84	−0,83	−0,83	−0,83	−0,82	−0,82	−0,82	−0,81	−0,81
0,21	−0,81	−0,80	−0,80	−0,80	−0,79	−0,79	−0,79	−0,78	−0,78	−0,78
0,22	−0,77	−0,77	−0,77	−0,76	−0,76	−0,76	−0,75	−0,75	−0,75	−0,74
0,23	−0,74	−0,74	−0,73	−0,73	−0,73	−0,72	−0,72	−0,72	−0,71	−0,71
0,24	−0,71	−0,70	−0,70	−0,70	−0,69	−0,69	−0,69	−0,68	−0,68	−0,68
0,25	−0,67	−0,67	−0,67	−0,67	−0,66	−0,66	−0,66	−0,65	−0,65	−0,65
0,26	−0,64	−0,64	−0,64	−0,63	−0,63	−0,63	−0,63	−0,62	−0,62	−0,62
0,27	−0,61	−0,61	−0,61	−0,60	−0,60	−0,60	−0,59	−0,59	−0,59	−0,59
0,28	−0,58	−0,58	−0,58	−0,57	−0,57	−0,57	−0,57	−0,56	−0,56	−0,56
0,29	−0,55	−0,55	−0,55	−0,54	−0,54	−0,54	−0,54	−0,53	−0,53	−0,53
0,30	−0,52	−0,52	−0,52	−0,52	−0,51	−0,51	−0,51	−0,50	−0,50	−0,50
0,31	−0,50	−0,49	−0,49	−0,49	−0,48	−0,48	−0,48	−0,48	−0,47	−0,47
0,32	−0,47	−0,46	−0,46	−0,46	−0,46	−0,45	−0,45	−0,45	−0,45	−0,44
0,33	−0,44	−0,44	−0,43	−0,43	−0,43	−0,43	−0,42	−0,42	−0,42	−0,42
0,34	−0,41	−0,41	−0,41	−0,40	−0,40	−0,40	−0,40	−0,39	−0,39	−0,39
0,35	−0,39	−0,38	−0,38	−0,38	−0,37	−0,37	−0,37	−0,37	−0,36	−0,36
0,36	−0,36	−0,36	−0,35	−0,35	−0,35	−0,35	−0,34	−0,34	−0,34	−0,33
0,37	−0,33	−0,33	−0,33	−0,32	−0,32	−0,32	−0,32	−0,31	−0,31	−0,31
0,38	−0,31	−0,30	−0,30	−0,30	−0,30	−0,29	−0,29	−0,29	−0,28	−0,28

Таблица II (продолжение)

p↓	0	1	2	3	4	5	6	7	8	9
0,39	—0,28	—0,28	—0,27	—0,27	—0,27	—0,27	—0,26	—0,26	—0,26	—0,26
0,40	—0,25	—0,25	—0,25	—0,25	—0,24	—0,24	—0,24	—0,24	—0,23	—0,23
0,41	—0,23	—0,23	—0,22	—0,22	—0,22	—0,21	—0,21	—0,21	—0,21	—0,20
0,42	—0,20	—0,20	—0,20	—0,19	—0,19	—0,19	—0,19	—0,18	—0,18	—0,18
0,43	—0,18	—0,17	—0,17	—0,17	—0,17	—0,16	—0,16	—0,16	—0,16	—0,15
0,44	—0,15	—0,15	—0,15	—0,14	—0,14	—0,14	—0,14	—0,13	—0,13	—0,13
0,45	—0,13	—0,12	—0,12	—0,12	—0,12	—0,11	—0,11	—0,11	—0,11	—0,10
0,46	—0,10	—0,10	—0,10	—0,09	—0,09	—0,09	—0,09	—0,08	—0,08	—0,08
0,47	—0,08	—0,07	—0,07	—0,07	—0,07	—0,06	—0,06	—0,06	—0,06	—0,05
0,48	—0,05	—0,05	—0,05	—0,04	—0,04	—0,04	—0,04	—0,03	—0,03	—0,03
0,49	—0,03	—0,02	—0,02	—0,02	—0,02	—0,01	—0,01	—0,01	—0,01	—0,00
0,50	0,00	0,00	0,01	0,01	0,01	0,01	0,02	0,02	0,02	0,02
0,51	0,03	0,03	0,03	0,03	0,04	0,04	0,04	0,04	0,05	0,05
0,52	0,05	0,05	0,06	0,06	0,06	0,06	0,07	0,07	0,07	0,07
0,53	0,08	0,08	0,08	0,08	0,09	0,09	0,09	0,09	0,10	0,10
0,54	0,10	0,10	0,11	0,11	0,11	0,11	0,12	0,12	0,12	0,12
0,55	0,13	0,13	0,13	0,13	0,14	0,14	0,14	0,14	0,15	0,15
0,56	0,15	0,15	0,16	0,16	0,16	0,16	0,17	0,17	0,17	0,17
0,57	0,18	0,18	0,18	0,18	0,19	0,19	0,19	0,19	0,20	0,20
0,58	0,20	0,20	0,21	0,21	0,21	0,21	0,22	0,22	0,22	0,23
0,59	0,23	0,23	0,23	0,24	0,24	0,24	0,24	0,25	0,25	0,25
0,60	0,25	0,26	0,26	0,26	0,26	0,27	0,27	0,27	0,27	0,28
0,61	0,28	0,28	0,28	0,29	0,29	0,29	0,30	0,30	0,30	0,30
0,62	0,31	0,31	0,31	0,31	0,32	0,32	0,32	0,32	0,33	0,33
0,63	0,33	0,33	0,34	0,34	0,34	0,35	0,35	0,35	0,35	0,36
0,64	0,36	0,36	0,36	0,37	0,37	0,37	0,37	0,38	0,38	0,38
0,65	0,39	0,39	0,39	0,39	0,40	0,40	0,40	0,40	0,41	0,41
0,66	0,41	0,42	0,42	0,42	0,42	0,43	0,43	0,43	0,43	0,44
0,67	0,44	0,44	0,45	0,45	0,45	0,45	0,46	0,46	0,46	0,46
0,68	0,47	0,47	0,47	0,48	0,48	0,48	0,48	0,49	0,49	0,49
0,69	0,50	0,50	0,50	0,50	0,51	0,51	0,51	0,52	0,52	0,52
0,70	0,52	0,53	0,53	0,53	0,54	0,54	0,54	0,54	0,55	0,55
0,71	0,55	0,56	0,56	0,56	0,57	0,57	0,57	0,57	0,58	0,58
0,72	0,58	0,59	0,59	0,59	0,59	0,60	0,60	0,60	0,61	0,61
0,73	0,61	0,62	0,62	0,62	0,63	0,63	0,63	0,63	0,64	0,64
0,74	0,64	0,65	0,65	0,65	0,66	0,66	0,66	0,67	0,67	0,67
0,75	0,67	0,68	0,68	0,68	0,69	0,69	0,69	0,70	0,70	0,70
0,76	0,71	0,71	0,71	0,72	0,72	0,72	0,73	0,73	0,73	0,74
0,77	0,74	0,74	0,75	0,75	0,75	0,76	0,76	0,76	0,77	0,77
0,78	0,77	0,78	0,78	0,78	0,79	0,79	0,79	0,80	0,80	0,80

Таблица II (окончание)

p→ ↓	0	1		3	4	5	6	7	8	9
0,79	0,81	0,81	0,81	0,82	0,82	0,82	0,83	0,83	0,83	0,84
0,80	0,84	0,85	0,85	0,85	0,86	0,86	0,86	0,87	0,87	0,87
0,81	0,88	0,88	0,89	0,89	0,89	0,90	0,90	0,90	0,91	0,91
0,82	0,92	0,92	0,92	0,93	0,93	0,93	0,94	0,94	0,95	0,95
0,83	0,95	0,96	0,96	0,97	0,97	0,97	0,98	0,98	0,99	0,99
0,84	0,99	1,00	1,00	1,01	1,01	1,02	1,02	1,02	1,03	1,03
0,85	1,04	1,04	1,05	1,05	1,05	1,06	1,06	1,07	1,07	1,08
0,86	1,08	1,09	1,09	1,09	1,10	1,10	1,11	1,11	1,12	1,12
0,87	1,13	1,13	1,14	1,14	1,15	1,15	1,16	1,16	1,17	1,17
0,88	1,18	1,18	1,19	1,19	1,20	1,20	1,21	1,21	1,22	1,22
0,89	1,23	1,23	1,24	1,24	1,25	1,25	1,26	1,26	1,27	1,28
0,90	1,28	1,29	1,29	1,30	1,30	1,31	1,32	1,32	1,33	1,33
0,91	1,34	1,35	1,35	1,36	1,37	1,37	1,38	1,39	1,39	1,40
0,92	1,41	1,41	1,42	1,43	1,43	1,44	1,45	1,45	1,46	1,47
0,93	1,48	1,48	0,49	1,50	1,51	1,51	1,52	1,53	1,54	1,55
0,94	1,55	1,56	1,57	1,58	1,59	1,60	1,61	1,62	1,63	1,64
0,95	1,64	1,65	1,66	1,67	1,68	1,70	1,71	1,72	1,73	1,74
0,96	1,75	1,76	1,77	1,79	1,80	1,81	1,83	1,84	1,85	1,87
0,97	1,88	1,90	1,91	1,93	1,94	1,96	1,98	2,00	2,01	2,03
0,98	2,05	2,07	2,10	2,12	2,14	2,17	2,20	2,23	2,26	2,29
0,99	23,3	2,37	2,41	2,46	2,51	2,58	2,65	2,75	2,88	3,09

Таблица III

Значения $\gamma = v\sqrt{0{,}5 + (v/100)^2}$ для определения $s_v = \gamma/\sqrt{n}$ [гл. 3, §3]

v	γ		γ	v	γ		γ	v	γ
1	0,71	11	7,87	21	15,49	31	23,93	41	33,51
2	1,41	12	8,61	22	16,29	32	24,84	42	34,54
3	2,12	13	9,35	23	17,10	33	25,75	43	35,59
4	2,83	14	10,09	24	17,92	34	26,68	44	36,64
5	3,54	15	10,84	25	18,75	35	27,61	45	37,72
6	4,25	16	11,60	26	19,59	36	28,57	46	38,80
7	4,97	17	12,36	27	20,44	37	29,53	47	39,91
8	5,69	18	13,13	28	21,29	38	30,50	48	41,02
9	6,41	19	13,91	29	22,16	39	31,49	49	42,15
10	7,14	20	14,70	30	23,04	40	32,50	50	43,30

Таблица IV

Критические значения t_P и t_α (критерия Стьюдента)

[гл. 3, § 4; гл. 4, § 4]

f	Доверительные уровни			f	Доверительные уровни		
	95%	99%	99,9%		95%	99%	99,9%
1	12,71	63,60		21	2,08	2,83	3,82
2	4,30	9,93	31,60	22	2,07	2,82	3,79
3	3,18	5,84	12,94	23	2,07	2,81	3,77
4	2,78	4,60	8,61	24	2,06	2,80	3,75
5	2,57	4,03	6,86	25	2,06	2,79	3,73
6	2,45	3,71	5,96	26	2,06	2,78	3,71
7	2,37	3,50	5,41	27	2,05	2,77	3,69
8	2,31	3,36	5,04	28	2,05	2,76	3,67
9	2,26	3,25	4,78	29	2,04	2,76	3,66
10	2,23	3,17	4,59	30	2,04	2,75	3,65
11	2,20	3,11	4,44	40	2,02	2,70	3,55
12	2,18	3,06	4,32	50	2,01	2,68	3,50
13	2,16	3,01	4,22	60	2,00	2,66	3,46
14	2,15	2,98	4,14	80	1,99	2,64	3,42
15	2,13	2,95	4,07	100	1,98	2,63	3,39
16	2,12	2,92	4,02	120	1,98	2,62	3,37
17	2,11	2,90	3,97	200	1,97	2,60	3,34
18	2,10	2,88	3,92	500	1,96	2,59	3,31
19	2,09	2,86	3,88	∞	1,96	2,58	3,29
20	2,09	2,85	3,85				
f	5%	1%	0,1%	f	5%	1%	0,1%
	Уровни значимости				Уровни значимости		

Нулевая гипотеза принимается при $t \leqslant t_{05}$ и отвергается при $t > t_{01}$.

Таблица V

Случайные числа [гл. 3, § 1]

5489	5583	3156	0835	1988	3912	0938	7460	0869	4420
3522	0935	7877	5665	7020	9555	7379	7124	7878	5544
7555	7579	2550	2487	9477	0864	2349	1012	8250	2633
5759	3554	5080	9074	7001	6249	3224	6368	9102	2672
6303	6895	3371	3196	7231	2918	7380	0438	7547	2644
7351	5634	5323	2623	7803	8374	2191	0464	0696	9529
7068	7803	8832	5119	6350	0120	5026	3684	5657	0304
3613	1428	1796	8147	0503	5654	3254	7336	9536	1944
5148	4534	2105	0368	7890	2473	4240	8652	9435	1422
9815	5144	7649	8638	6137	8070	5345	4865	2456	5708
5780	1277	6316	1013	2867	9938	3930	3203	5696	1769
1187	0951	5991	5245	5700	5564	7352	0891	6249	6568
4184	2179	4554	9088	2254	2435	2965	5154	1209	7069
2916	2972	9885	0275	0144	8034	8122	3213	7666	0230
5524	1341	9860	6565	6981	9842	0171	2284	2707	3008
0146	5291	2354	5694	0377	5336	6460	9585	3415	2358
4920	2826	5238	5402	7937	1993	4332	2327	6875	5230
7978	1947	6380	3425	7267	7285	1130	7722	0164	8573
7453	0653	3645	7497	5969	8682	4191	2976	0361	9334
1473	6938	4899	5348	1641	3652	0852	5296	4538	4456
8162	8797	8000	4707	1880	9660	8446	1883	9768	0881
5645	4219	0807	3301	4279	4168	4305	9937	3120	5547
2042	1192	1175	8851	6432	4635	5757	6656	1660	5389
5470	7702	6958	9080	5925	8519	0127	9233	2452	7341
4045	1730	6005	1704	0345	3275	4738	4862	2556	8333
5880	1257	6163	4439	7276	6353	6912	0731	9033	5294
9083	4260	5277	4998	4298	5204	3965	4028	8936	5148
1762	8713	1189	1090	8989	7273	3213	1935	9321	4820
2023	2589	1740	0424	8924	0005	1969	1636	7237	1227
7965	3855	4765	0703	1678	0841	7543	0308	9732	1289
7690	0480	8098	9629	4819	7219	7241	5128	3853	1921
9292	0426	9573	4903	5916	6576	8368	3270	6641	0033
0867	1656	7016	4220	2533	6345	8227	1904	5138	2537
0505	2127	8255	5276	2233	3956	4118	8199	6380	6340
6295	9795	1112	5761	2575	6837	3336	9232	7403	8345
6323	2615	3410	3365	1117	2417	3176	2434	5240	5455
8672	8536	2966	5773	5412	8114	0930	4697	6919	4569
1422	5507	7596	0670	3013	1351	3886	3268	9469	2584
2653	1472	5113	5735	1469	9545	9331	5303	9914	6394
0438	4376	3328	8649	8327	0110	4549	7955	5275	2890

Таблица V (окончание)

2851	2157	0047	7085	1129	0460	6821	8323	2572	8962
7962	2753	3077	8718	7418	8004	1425	3706	8822	1494
3837	4098	0220	1217	4732	0150	1637	1097	1040	7372
8542	4126	9274	2251	0607	4301	8730	7690	6235	3477
0139	0765	8039	9484	2577	7859	1976	0623	1418	6685
6687	1943	4307	0579	8171	8224	8641	7034	3595	3875
6242	5582	5872	3197	4919	2792	5991	4058	9769	1918
6859	9606	0522	4993	0345	8958	1289	8825	6941	7685
6590	1932	6043	3623	1973	4112	1795	8465	2110	8045
3482	0478	0221	6738	7323	5643	4767	0106	2272	9862

Таблица VI

Значения $q'_P(f)$ и $q''_P(f)$ для построения доверительного интервала для стандартного отклонения [гл. 3, § 4]

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	0,519	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

Таблица VII

Значения $\varphi = 2 \arcsin \sqrt{p}$ [гл. 3, § 9]

%	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

Таблица VII (продолжение)

%	0	1	2	3	4	5	6	7	8	9
31	1,182	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
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	0,693	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	1,963	1,965	1,967	1,969	1,971	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

Таблица VII (окончание)

%	0	1	2	3	4	5	6	7	8	9
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	2,111	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
80	2,214	2,217	2,219	2,222	2,224	2,227	2,229	2,231	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	2,294	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	2,664	2,668	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	3,983	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	3,004	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									

Таблица VIII

Достаточно большие объемы выборок при доверительной вероятности $P=0{,}99$ [гл. 3, § 4]

v, %	$\varepsilon=0{,}05$	$\varepsilon=0{,}03$	v, %	$\varepsilon=0{,}05$	$\varepsilon=0{,}03$	v, %	$\varepsilon=0{,}05$	$\varepsilon=0{,}03$
1		1	18	87	240	35	326	910
2	1	3	19	97	267	36	345	960
3	3	7	20	107	296	37	365	1015
4	5	12	21	118	327	38	385	1070
5	7	19	22	129	358	39	405	1125
6	10	27	23	141	392	40	426	1185
7	13	37	24	153	426	41	448	1245
8	17	48	25	166	463	42	470	1305
9	22	60	26	180	500	43	493	1370
10	27	75	27	194	540	44	515	1430
11	33	90	28	208	580	45	540	1500
12	39	107	29	224	625	46	564	1570
13	45	125	30	240	665	47	588	1640
14	53	145	31	256	710	48	613	1710
15	60	167	32	273	760	49	640	1780
16	68	190	33	290	810	50	665	1850
17	77	214	34	307	860			

Таблица IX

Значения $\rho^{(n')}=s/\overline{R}^{(n')}$ для оценки дисперсии по среднему размаху варьирования $\overline{R}^{(n')}$ [гл. 3, § 7]

n'	$\rho^{(n')}$	n'	$\rho^{(n')}$	n'	$\rho^{(n')}$
2	0,886	6	0,395	10	0,325
3	0,591	7	0,370	11	0,315
4	0,486	8	0,351	12	0,307
5	0,430	9	0,337	14	0,294

Таблица X

Доверительные значения $d_P^{(n)}$ для построения доверительных интервалов по размаху варьирования [гл. 3, § 7]

Объем выборки	Доверительные уровни		Объем выборки	Доверительные уровни	
	95%	99%		95%	99%
3	0,668	0,879	12	0,174	0,229
4	0,476	0,626	13	0,163	0,214
5	0,377	0,496	14	0,154	0,202
6	0,316	0,416	15	0,145	0,191
7	0,274	0,361	16	0,139	0,183
8	0,243	0,320	17	0,132	0,174
9	0,220	0,289	18	0,127	0,167
10	0,201	0,265	19	0,122	0,160
11	0,186	0,245	20	0,117	0,154

Таблица XI

Критические значения $\tau' = \dfrac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}$ и $\tau'' = \dfrac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$ для $\alpha = 0{,}01$ и $0{,}05$ [гл. 4, § 2]

n	τ		n	τ		n	τ	
	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	0,391	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

Если $\tau > \tau_{0,01}$, то варианта отбрасывается; если $\tau \leqslant \tau_{0,05}$, то варианта оставляется.

Таблица XII

Критерий для отбрасывания крайних вариант [гл. 4, § 2]

n	τ_α		n	τ_α		n	τ_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

Таблица XIII

Доверительные границы для параметра в распределении Пуассона [гл. 3, § 8]

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					

Критические значения F_α (при F_{05} набраны обычным

f_2 — степени свободы для меньшей дисперсии	f_2	f_1 — степени свободы											
		1	2	3	4	5	6	7	8	9	10	11	12
	1	161 4 052	200 4 999	216 5 403	225 5 625	230 5 764	234 5 889	237 5 928	239 5 981	241 6 022	242 6 056	243 6 082	244 6 106
	2	18,51 98,49	19,00 99,01	19,16 99,17	19,25 99,25	19,30 99,30	19,33 99,33	19,36 99,34	19,37 99,36	19,38 99,38	19,39 99,40	19,40 99,41	19,41 99,42
	3	10,13 34,12	9,55 30,81	9,28 29,46	9,12 28,71	9,01 28,24	8,94 27,91	8,88 27,67	8,84 27,49	8,81 27,34	8,78 27,23	8,76 27,13	8,74 27,05
	4	7,71 21,20	6,94 18,00	6,59 16,69	6,39 15,98	6,26 15,52	6,16 15,21	6,09 14,98	6,04 14,80	6,00 14,66	5,96 14,54	5,93 14,45	5,91 14,37
	5	6,61 16,26	5,79 13,27	5,41 12,06	5,19 11,39	5,05 10,97	4,95 10,67	4,88 10,45	4,82 10,27	4,78 10,15	4,74 10,05	4,70 9,96	4,68 9,89
	6	5,99 13,74	5,14 10,92	4,76 9,78	4,53 9,15	4,39 8,75	4,28 8,47	4,21 8,26	4,15 8,10	4,10 7,98	4,06 7,87	4,03 7,79	4,00 7,72
	7	5,59 12,25	4,74 9,55	4,35 8,45	4,12 7,85	3,97 7,46	3,87 7,19	3,79 7,00	3,73 6,84	3,68 6,71	3,63 6,62	3,60 6,54	3,57 6,47
	8	5,32 11,26	4,46 8,65	4,07 7,59	3,84 7,01	3,69 6,63	3,58 6,37	3,50 6,19	3,44 6,03	3,39 5,91	3,34 5,82	3,31 5,74	3,28 5,67
	9	5,12 10,56	4,26 8,02	3,86 6,99	3,63 6,42	3,48 6,06	3,37 5,80	3,29 5,62	3,23 5,47	3,18 5,35	3,13 5,26	3,10 5,18	3,07 5,11
	10	4,96 10,04	4,10 7,56	3,71 6,55	3,48 5,99	3,33 5,64	3,22 5,39	3,14 5,21	3,07 5,06	3,02 4,95	2,97 4,85	2,94 4,78	2,91 4,71
	11	4,84 9,85	3,98 7,20	3,59 6,22	3,36 5,67	3,20 5,32	3,09 5,07	3,01 4,88	2,95 4,74	2,90 4,63	2,86 4,54	2,82 4,46	2,79 4,40
	12	4,75 9,33	3,88 6,93	3,49 5,95	3,26 5,41	3,11 5,06	3,00 4,82	2,92 4,65	2,85 4,50	2,80 4,39	2,76 4,30	2,72 4,22	2,69 4,16
	13	4,67 9,07	3,80 6,70	3,41 5,74	3,18 5,20	3,02 4,86	2,92 4,62	2,84 4,44	2,77 4,30	2,72 4,19	2,67 4,10	2,63 4,02	2,60 3,96
	14	4,60 8,86	3,74 6,51	3,34 5,56	3,11 5,03	2,96 4,69	2,85 4,46	2,77 4,28	2,70 4,14	2,65 4,03	2,60 3,94	2,56 3,86	2,53 3,80
	15	4,54 8,68	3,68 6,36	3,29 5,42	3,06 4,89	2,90 4,56	2,79 4,32	2,70 4,14	2,64 4,00	2,59 3,89	2,55 3,80	2,51 3,73	2,48 3,67
	16	4,49 8,53	3,63 6,23	3,24 5,29	3,01 4,77	2,85 4,44	2,74 4,20	2,66 4,03	2,59 3,89	2,54 3,78	2,49 3,69	2,45 3,61	2,42 3,55
	17	4,45 8,40	3,59 6,11	3,20 5,18	2,96 4,67	2,81 4,34	2,70 4,10	2,62 3,93	2,55 3,79	2,50 3,68	2,45 3,59	2,41 3,52	2,38 3,45

Критические значения F_α (кри
F_{05} избраны обычным

f_3 — степени свободы для меньшей дисперсии	f_3	f_1 — степени свободы											
		1	2	3	4	5	6	7	8	9	10	11	12
	1	161 4 052	200 4 999	216 5 403	225 5 625	230 5 764	234 5 889	237 5 928	239 5 981	241 6 022	242 6 056	243 6 082	244 6 106
	2	18,51 98,49	19,00 99,01	19,16 99,17	19,25 99,25	19,30 99,30	19,33 99,33	19,36 99,34	19,37 99,36	19,38 99,38	19,39 99,40	19,40 99,41	19,41 99,42
	3	10,13 34,12	9,55 30,81	9,28 29,46	9,12 28,71	9,01 28,24	8,94 27,91	8,88 27,67	8,84 27,49	8,81 27,34	8,78 27,23	8,76 27,13	8,74 27,05
	4	7,71 21,20	6,94 18,00	6,59 16,69	6,39 15,98	6,26 15,52	6,16 15,21	6,09 14,98	6,04 14,80	6,00 14,66	5,96 14,54	5,93 14,45	5,91 14,37
	5	6,61 16,26	5,79 13,27	5,41 12,06	5,19 11,39	5,05 10,97	4,95 10,67	4,88 10,45	4,82 10,27	4,78 10,15	4,74 10,05	4,70 9,96	4,68 9,89
	6	5,99 13,74	5,14 10,92	4,76 9,78	4,53 9,15	4,39 8,75	4,28 8,47	4,21 8,26	4,15 8,10	4,10 7,98	4,06 7,87	4,03 7,79	4,00 7,72
	7	5,59 12,25	4,74 9,55	4,35 8,45	4,12 7,85	3,97 7,46	3,87 7,19	3,79 7,00	3,73 6,84	3,68 6,71	3,63 6,62	3,60 6,54	3,57 6,47
	8	5,32 11,26	4,46 8,65	4,07 7,59	3,84 7,01	3,69 6,63	3,58 6,37	3,50 6,19	3,44 6,03	3,39 5,91	3,34 5,82	3,31 5,74	3,28 5,67
	9	5,12 10,56	4,26 8,02	3,86 6,99	3,63 6,42	3,48 6,06	3,37 5,80	3,29 5,62	3,23 5,47	3,18 5,35	3,13 5,26	3,10 5,18	3,07 5,11
	10	4,96 10,04	4,10 7,56	3,71 6,55	3,48 5,99	3,33 5,64	3,22 5,39	3,14 5,21	3,07 5,06	3,02 4,95	2,97 4,85	2,94 4,78	2,91 4,71
	11	4,84 9,86	3,98 7,20	3,59 6,22	3,36 5,67	3,20 5,32	3,09 5,07	3,01 4,88	2,95 4,74	2,90 4,63	2,86 4,54	2,82 4,46	2,79 4,40
	12	4,75 9,33	3,88 6,93	3,49 5,95	3,26 5,41	3,11 5,06	3,00 4,82	2,92 4,65	2,85 4,50	2,80 4,39	2,76 4,30	2,72 4,22	2,69 4,16
	13	4,67 9,07	3,80 6,70	3,41 5,74	3,18 5,20	3,02 4,86	2,92 4,62	2,84 4,44	2,77 4,30	2,72 4,19	2,67 4,10	2,63 4,02	2,60 3,96
	14	4,60 8,86	3,74 6,51	3,34 5,56	3,11 5,03	2,96 4,69	2,85 4,46	2,77 4,28	2,70 4,14	2,65 4,03	2,60 3,94	2,56 3,86	2,53 3,80
	15	4,54 8,68	3,68 6,36	3,29 5,42	3,06 4,89	2,90 4,56	2,79 4,32	2,70 4,14	2,64 4,00	2,59 3,89	2,55 3,80	2,51 3,73	2,48 3,67
	16	4,49 8,53	3,63 6,23	3,24 5,29	3,01 4,77	2,85 4,44	2,74 4,20	2,66 4,03	2,59 3,89	2,54 3,78	2,49 3,69	2,45 3,61	2,42 3,55
	17	4,45 8,40	3,59 6,11	3,20 5,18	2,96 4,67	2,81 4,34	2,70 4,10	2,62 3,93	2,55 3,79	2,50 3,68	2,45 3,59	2,41 3,52	2,38 3,46

терия Фишера) [гл. 4, § 7] *Таблица XIV*
шрифтом, F_{01} — жирным

для большей дисперсии												f_1
14	16	20		30	40	50	75	100	200	500	∞	
245 6 142	246 6 169	248 6 208	249 6 234	250 6 258	251 6 286	252 6 302	253 6 323	253 6 334	254 6 352	254 6 361	254 6 366	1
19,42 19,43	19,43 99,44	19,44 99,45	19,45 99,46	19,46 99,47	19,47 99,48	19,47 99,48	19,48 99,49	19,49 99,49	19,49 99,49	19,50 99,50	19,50 99,50	2
8,71 26,92	8,69 26,83	8,66 26,69	8,64 26,60	8,62 26,50	8,60 26,41	8,58 26,35	8,57 26,27	8,56 26,23	8,54 26,18	8,54 26,14	8,53 26,12	3
5,87 14,24	5,84 14,15	5,80 14,02	5,77 13,93	5,74 13,83	5,71 13,74	5,70 13,69	5,68 13,61	5,66 13,57	5,65 13,52	5,64 13,48	5,63 13,46	4
4,64 9,77	4,60 9,68	4,56 9,55	4,53 9,47	4,50 9,38	4,46 9,29	4,44 9,24	4,42 9,17	4,40 9,13	4,38 9,07	4,37 9,04	4,36 9,02	5
3,96 7,60	3,92 7,52	3,87 7,39	3,84 7,31	3,81 7,23	3,77 7,14	3,75 7,09	3,72 7,02	3,71 6,99	3,69 6,94	3,68 6,90	3,67 6,88	6
3,52 6,35	3,49 6,27	3,44 6,15	3,41 6,07	3,38 5,98	3,34 5,90	3,32 5,85	3,29 5,78	3,28 5,75	3,25 5,70	3,24 5,67	3,23 5,65	7
3,23 5,56	3,20 5,48	3,15 5,36	3,12 5,28	3,08 5,20	3,05 5,11	3,03 5,06	3,00 5,00	2,98 4,96	2,96 4,91	2,94 4,88	2,93 4,86	8
3,02 5,00	2,98 4,92	2,93 4,80	2,90 4,73	2,86 4,64	2,82 4,56	2,80 4,51	2,77 4,45	2,76 4,41	2,73 4,36	2,72 4,33	2,71 4,31	9
2,86 4,60	2,82 4,52	2,77 4,41	2,74 4,33	2,70 4,25	2,67 4,17	2,64 4,12	2,61 4,05	2,59 4,01	2,56 3,96	2,55 3,93	2,54 3,91	10
2,74 4,29	2,70 4,21	2,65 4,10	2,61 4,02	2,57 3,94	2,53 3,86	2,50 3,80	2,47 3,74	2,45 3,70	2,42 3,66	2,41 3,62	2,40 3,60	11
2,64 4,06	2,60 3,98	2,54 3,86	2,50 3,78	2,46 3,70	2,42 3,61	2,40 3,56	2,36 3,49	2,35 3,46	2,32 3,41	2,31 3,38	2,30 3,36	12
2,55 3,85	2,51 3,78	2,46 3,67	2,42 3,59	2,38 3,51	2,34 3,42	2,32 3,37	2,28 3,30	2,26 3,27	2,24 3,21	2,22 3,18	2,21 3,16	13
2,48 3,70	2,44 3,62	2,39 3,51	2,35 3,43	2,31 3,34	2,27 3,26	2,24 3,21	2,21 3,14	2,19 3,11	2,16 3,06	2,14 3,02	2,13 3,00	14
2,43 3,56	2,39 3,48	2,33 3,36	2,29 3,29	2,25 3,20	2,21 3,12	2,18 3,07	2,15 3,00	2,12 2,97	2,10 2,92	2,08 2,89	2,07 2,87	15
2,37 3,46	2,33 3,37	2,28 3,25	2,24 3,18	2,20 3,10	2,16 3,01	2,13 2,96	2,09 2,89	2,07 2,86	2,04 2,80	2,02 2,77	2,01 2,75	16
2,33 3,35	2,29 3,27	2,23 3,16	2,19 3,08	2,15 3,00	2,11 2,92	2,08 2,86	2,04 2,79	2,02 2,76	1,99 2,70	1,87 2,67	1,96 2,65	17

терия Фишера) [гл. 4, § 7] *Таблица XIV*
ирифтом, F_{01} — жирным

для большей дисперсии												
14	16	20	24	30	40	50	75	100	200	500	∞	l_1
245	246	248	249	250	251	252	253	253	254	254	254	
6 142	6 169	6 208	6 234	6 258	6 286	6 302	6 323	6 334	6 352	6 361	6 366	1
19,42	19,43	19,44	19,45	19,46	19,47	19,47	19,48	19,49	19,49	19,50	19,50	
19,43	99,44	99,45	99,46	99,47	99,48	99,48	99,49	99,49	99,49	99,50	99,50	2
8,71	8,69	8,66	8,64	8,62	8,60	8,58	8,57	8,56	8,54	8,54	8,53	
26,92	26,83	26,69	26,60	26,50	26,41	26,35	26,27	26,23	26,18	26,14	26,12	3
5,87	5,84	5,80	5,77	5,74	5,71	5,70	5,68	5,66	5,65	5,64	5,63	
14,24	14,15	14,02	13,93	13,83	13,74	13,69	13,61	13,57	13,52	13,48	13,46	4
4,64	4,60	4,56	4,53	4,50	4,46	4,44	4,42	4,40	4,38	4,37	4,36	
9,77	9,68	9,55	9,47	9,38	9,29	9,24	9,17	9,13	9,07	9,04	9,02	5
3,96	3,92	3,87	3,84	3,81	3,77	3,75	3,72	3,71	3,69	3,68	3,67	
7,60	7,52	7,39	7,31	7,23	7,14	7,09	7,02	6,99	6,94	6,90	6,88	6
3,52	3,49	3,44	3,41	3,38	3,34	3,32	3,29	3,28	3,25	3,24	3,23	
6,35	6,27	6,15	6,07	5,98	5,90	5,85	5,78	5,75	5,70	5,67	5,65	7
3,23	3,20	3,15	3,12	3,08	3,05	3,03	3,00	2,98	2,96	2,94	2,93	
5,56	5,48	5,36	5,28	5,20	5,11	5,06	5,00	4,96	4,91	4,88	4,86	8
3,02	2,98	2,93	2,90	2,86	2,82	2,80	2,77	2,76	2,73	2,72	2,71	
5,00	4,92	4,80	4,73	4,64	4,56	4,51	4,45	4,41	4,36	4,33	4,31	9
2,86	2,82	2,77	2,74	2,70	2,67	2,64	2,61	2,59	2,56	2,55	2,54	
4,60	4,52	4,41	4,33	4,25	4,17	4,12	4,05	4,01	3,96	3,93	3,91	10
2,74	2,70	2,65	2,61	2,57	2,53	2,50	2,47	2,45	2,42	2,41	2,40	
4,29	4,21	4,10	4,02	3,94	3,86	3,80	3,74	3,70	3,66	3,62	3,60	11
2,64	2,60	2,54	2,50	2,46	2,42	2,40	2,36	2,35	2,32	2,31	2,30	
4,05	3,98	3,86	3,78	3,70	3,61	3,56	3,49	3,46	3,41	3,38	3,36	12
2,55	2,51	2,46	2,42	2,38	2,34	2,32	2,28	2,26	2,24	2,22	2,21	
3,85	3,78	3,67	3,59	3,51	3,42	3,37	3,30	3,27	3,21	3,18	3,16	13
2,48	2,44	2,39	2,35	2,31	2,27	2,24	2,21	2,19	2,16	2,14	2,13	
3,70	3,62	3,51	3,43	3,34	3,26	3,21	3,14	3,11	3,06	3,02	3,00	14
2,43	2,39	2,33	2,29	2,25	2,21	2,18	2,15	2,12	2,10	2,08	2,07	
3,56	3,48	3,36	3,29	3,20	3,12	3,07	3,00	2,97	2,92	2,89	2,87	15
2,37	2,33	2,28	2,24	2,20	2,16	2,13	2,09	2,07	2,04	2,02	2,01	
3,45	3,37	3,25	3,18	3,10	3,01	2,96	2,89	2,86	2,80	2,77	2,75	16
2,33	2,29	2,23	2,19	2,15	2,11	2,08	2,04	2,02	1,99	1,87	1,96	
3,35	3,27	3,16	3,08	3,00	2,92	2,86	2,79	2,76	2,70	2,67	2,65	17

f_2 — степени свободы для меньшей дисперсии	f_1 — степени свободы											
	1		3	4	5	6	7	8	9	10	11	12
	18	4,41 8,28	3,55 6,01	3,16 5,09	2,93 4,58	2,77 4,25	2,66 4,01	2,58 3,85	2,51 3,71	2,46 3,60	2,41 3,51	2,37 3,44	2,34 3,37
	19	4,38 8,18	3,52 5,93	3,13 5,01	2,90 4,50	2,74 4,17	2,63 3,94	2,55 3,77	2,48 3,63	2,43 3,52	2,38 2,43	2,34 3,36	2,31 3,30
	20	4,35 8,10	3,49 5,85	3,10 4,94	2,87 4,43	2,71 4,10	2,60 3,87	2,52 3,71	2,45 3,56	2,40 3,45	2,35 3,37	2,31 3,30	2,28 3,23
	21	4,32 8,02	3,47 5,78	3,07 4,87	2,84 4,37	2,68 4,04	2,57 3,81	2,49 3,65	2,42 3,51	2,37 3,40	2,32 3,31	2,28 3,24	2,25 3,17
	22	4,30 7,94	3,44 5,72	3,05 4,82	2,82 4,31	2,66 3,99	2,55 3,76	2,47 3,59	2,40 3,45	2,35 3,35	2,30 3,26	2,26 3,18	2,23 3,12
	23	4,28 7,88	3,42 5,66	3,03 4,76	2,80 4,26	2,64 3,94	2,53 3,71	2,45 3,54	2,38 3,41	2,32 3,30	2,28 3,21	2,24 3,14	2,20 3,07
	24	4,26 7,82	3,40 5,61	3,01 4,72	2,78 4,22	2,62 3,90	2,51 3,67	2,43 3,50	2,36 3,36	2,30 3,25	2,26 3,17	2,22 3,09	2,18 3,03
	25	4,24 7,77	3,38 5,57	2,99 4,68	2,76 4,18	2,60 3,86	2,49 3,63	2,41 3,46	2,34 3,32	2,25 3,21	2,24 3,13	2,20 3,05	2,16 2,99
	26	4,22 7,72	3,37 5,53	2,98 4,64	2,74 4,14	2,59 3,82	2,47 3,59	2,39 3,42	2,32 3,29	2,27 3,17	2,22 3,09	2,18 3,02	2,15 2,96
	27	4,21 7,68	3,35 5,49	2,96 4,60	2,73 4,11	2,57 3,79	2,46 3,56	2,37 3,39	2,30 3,26	2,25 3,14	2,20 3,06	2,16 2,98	3,13 2,93
	28	4,20 7,64	3,34 5,45	2,95 4,57	2,71 4,07	2,56 3,76	2,44 3,53	2,36 3,36	2,29 3,23	2,24 3,11	2,19 3,03	2,15 2,95	2,12 2,90
	29	4,18 7,60	3,33 5,42	2,93 4,54	2,70 4,04	2,54 3,73	2,43 2,50	2,35 3,33	2,28 2,20	2,22 3,08	2,18 3,00	2,14 2,92	2,10 2,87
	30	4,17 7,56	3,32 5,39	2,92 4,51	2,69 4,02	2,53 3,70	2,42 3,47	2,34 3,30	2,27 3,17	2,21 3,06	2,16 2,98	2,12 2,90	2,09 2,84
	32	4,15 7,50	3,30 5,34	2,90 4,46	2,67 3,97	2,51 3,66	2,40 3,42	2,32 3,25	2,25 3,12	2,19 3,01	2,14 2,94	2,10 2,86	2,07 2,80
	34	4,13 7,44	3,28 5,29	2,88 4,42	2,65 3,93	2,49 3,61	2,38 3,38	2,30 3,21	2,23 3,08	2,17 2,97	2,12 2,89	2,08 2,82	2,05 2,76
	36	4,11 7,39	3,26 5,25	2,86 4,38	2,63 3,89	2,48 2,58	2,36 3,35	2,28 3,18	2,21 3,04	2,15 2,94	2,10 2,86	2,06 3,78	2,03 2,72
	38	4,10 7,35	3,25 5,21	2,85 4,34	2,65 3,86	2,46 3,54	2,35 3,32	2,26 3,15	2,19 3,02	2,14 2,91	2,09 2,82	2,05 2,75	2,02 2,69
	40	4,08 7,31	3,23 5,18	2,84 4,31	2,61 3,83	2,45 3,51	2,34 3,29	2,25 3,12	2,18 2,99	2,12 2,88	2,07 2,80	2,04 2,73	2,00 2,66

Таблица XIV (продолжение)

для большей дисперсии												f_1
14	16	20	24	30	40	50	75	100	200	500	∞	
2,29	2,25	2,19	2,15	2,11	2,07	2,04	2,00	1,98	1,95	1,93	1,92	18
3,27	3,19	3,07	3,00	2,91	2,83	2,78	2,71	2,68	2,62	2,59	2,57	
2,26	2,21	2,15	2,11	2,07	2,02	2,00	1,96	1,94	1,91	1,90	1,88	19
3,19	3,12	3,00	2,92	2,84	2,76	2,70	2,63	2,60	2,54	2,51	2,49	
2,23	2,18	1,12	2,08	2,04	1,99	1,96	1,92	1,90	1,87	1,85	1,84	20
3,13	3,05	2,94	2,86	2,77	2,69	2,63	2,56	2,53	2,47	2,44	2,42	
2,20	2,15	2,09	2,05	2,00	1,96	1,93	1,89	1,87	1,84	1,82	1,81	21
3,07	2,99	2,88	2,80	2,72	2,63	2,58	2,51	2,47	2,42	2,38	2,36	
2,18	2,13	2,07	2,03	1,98	1,93	1,91	1,87	1,84	1,81	1,80	1,78	22
3,02	2,94	2,83	2,75	2,67	2,58	2,53	2,46	2,42	2,37	2,33	2,31	
2,14	2,10	2,04	2,00	1,96	1,91	1,88	1,84	1,82	1,79	1,77	1,76	23
2,97	2,89	2,78	2,70	2,62	2,53	2,48	2,41	2,37	2,32	2,28	2,26	
2,13	2,09	2,02	1,98	1,94	1,89	1,86	1,82	1,80	1,76	1,74	1,73	24
2,93	2,85	2,74	2,66	2,58	2,49	2,44	2,36	2,33	2,27	2,23	2,21	
2,11	2,06	2,00	1,96	1,92	1,87	1,84	1,80	1,77	1,74	1,72	1,71	25
2,89	2,81	2,70	2,62	2,54	2,45	2,40	2,32	2,29	2,23	2,19	2,17	
2,16	2,05	1,99	1,95	1,90	1,85	1,82	1,78	1,76	1,72	1,70	1,69	26
2,86	2,77	2,66	2,58	2,50	2,41	2,36	2,28	2,25	2,19	2,15	2,13	
2,08	2,03	1,97	1,93	1,88	1,84	1,80	1,76	1,74	1,71	1,68	1,67	27
2,83	2,74	2,63	2,55	2,47	2,38	2,33	2,25	2,21	2,16	2,12	2,10	
2,07	2,02	1,96	1,91	1,87	1,81	1,78	1,75	1,72	1,69	1,67	1,65	28
2,80	2,71	2,60	2,52	2,44	2,35	2,30	2,22	2,18	2,13	2,09	2,06	
2,05	2,00	1,94	1,90	1,85	1,80	1,77	1,73	1,71	1,68	1,65	1,64	29
2,77	2,68	2,57	2,49	2,41	2,32	2,27	2,19	2,15	2,10	2,06	2,03	
2,04	1,99	1,93	1,89	1,84	1,79	1,76	1,72	1,69	1,66	1,64	1,62	30
2,74	2,66	2,55	2,47	2,38	2,29	2,24	2,16	2,13	2,07	2,03	2,01	
2,02	1,97	1,91	1,86	1,82	1,76	1,74	1,69	1,67	1,64	1,61	1,59	32
2,70	2,62	2,51	2,42	2,34	2,25	2,20	2,12	2,08	2,02	1,98	1,96	
2,00	1,95	1,89	1,84	1,80	1,74	1,71	1,67	1,64	1,61	1,59	1,57	34
2,66	2,58	2,47	2,38	2,30	2,21	2,15	2,08	2,04	1,98	1,94	1,91	
1,98	1,93	1,87	1,82	1,78	1,72	1,69	1,65	1,62	1,59	1,56	1,55	36
2,62	2,54	2,43	2,35	2,26	2,17	2,12	2,04	2,00	1,94	1,90	1,87	
1,96	1,92	1,85	1,80	1,76	1,71	1,67	1,63	1,60	1,57	1,54	1,53	38
2,59	2,51	2,40	2,32	2,22	2,14	2,08	2,00	1,97	1,90	1,86	1,84	
1,95	1,90	1,84	1,79	1,74	1,69	1,66	1,61	1,59	1,55	1,53	1,51	40
2,56	2,49	2,37	2,29	2,20	2,11	2,05	1,97	1,94	1,88	1,84	1,81	

Таблица XV

Критические значения G_α (критерия Кохрена) [гл. 4, § 7]

Значения G_α для $\alpha = 5\%$ напечатаны обычным шрифтом, для $\alpha = 1\%$ — жирным шрифтом. Ноль (целых) и запятая опущены

f \ m'	3	4	5	6	7	8	9	10
1	967 993	906 968	841 928	781 883	727 838	680 794	638 754	602 718
2	871 942	768 864	684 788	616 722	561 664	516 615	478 573	445 534
3	798 883	684 781	598 696	532 626	480 568	438 521	403 481	373 447
4	746 834	629 721	544 633	480 564	431 508	391 463	358 425	331 393
5	707 793	590 676	506 588	445 520	397 466	360 423	329 387	303 357
6	677 761	560 641	478 553	418 487	373 435	336 393	307 359	282 331
7	653 734	536 613	456 526	398 461	354 410	318 370	290 338	267 311
8	633 711	518 590	439 504	382 440	338 391	304 352	277 321	254 294
9	617 691	502 570	424 485	368 423	326 375	293 337	266 307	244 281
10	602 674	488 554	412 470	357 408	315 362	283 325	257 295	235 270
12	580 647	466 527	392 445	339 384	298 340	267 305	244 277	224 254
14	560 625	450 503	377 425	325 368	285 322	256 290	232 263	213 240
16	547 606	437 488	364 409	313 363	277 310	246 278	223 251	203 230
18	536 590	425 474	353 397	305 340	267 298	239 269	217 243	199 221
20	526 577	416 461	335 845	298 330	259 289	230 260	210 235	191 215
25	504 550	397 438	329 363	282 312	245 270	219 243	198 219	180 200
30	490 533	383 420	315 349	270 299	233 259	209 230	190 209	173 190
35	479 518	372 408	306 337	262 288	227 250	201 222	183 200	216 182
40	469 505	364 397	298 327	255 280	220 242	195 215	177 192	162 177
45	460 496	357 387	291 319	250 272	215 235	190 209	173 288	157 172

Таблица XV (окончание)

f \ w	3	4	5	6	7	8	9	10
50	455 488	350 380	288 312	246 266	210 230	186 204	170 184	152 167
60	444 474	341 368	279 300	239 256	204 220	179 198	161 177	148 160
80	429 455	329 350	268 285	227 243	195 210	170 186	154 164	140 150
100	417 441	319 338	261 277	220 235	189 202	165 180	150 160	136 143
150	401 421	305 322	250 263	211 220	180 194	160 170	145 150	130 137
200	390 410	299 310	246 255	210 215	180 190	158 165	142 148	130 135

Нулевая гипотеза принимается при $G \leqslant G_{05}$ и отвергается при $G > G_{01}$.

Таблица XVI

Критические значения $t_{\alpha}^{(R\Delta)}$ для сравнения совокупностей с попарно связанными вариантами [гл. 4, § 5]

Объем каждой выборки	Уровни значимости		Объем каждой выборки	Уровни значимости	
	5%	1%		5%	1%
3	1,272	2,093	12	0,260	0,355
4	0,813	1,237	13	0,243	0,331
5	0,613	0,896	14	0,228	0,311
6	0,499	0,714	15	0,216	0,293
7	0,426	0,600	16	0,205	0,278
8	0,373	0,521	17	0,195	0,264
9	0,334	0,464	18	0,187	0,252
10	0,304	0,419	19	0,179	0,242
11	0,280	0,384	20	0,172	0,232

Нулевая гипотеза принимается при $t^{(R\Delta)} \leqslant t_{05}^{(R\Delta)}$ и отвергается при $t^{(R\Delta)} > t_{01}^{(R\Delta)}$.

Таблица XVII

Критические значения T_α (критерия Вилкоксона) [гл. 7, § 2]

n_y \ n_x	4	5	6	7	8	9	10
4	10	11	12 10	13 10	14 11	15 11	15 12
5		17 15	18 16	20 17	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						63 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 $T \geqslant T_{05}$ и отвергается при $T < T_{01}$. Более обширная таблица (до $n = 30$) имеется в книге Снедекора (1961), стр. 126.

Таблица XVIII

Критические значения X_α (критерия ван дер Вардена) [гл. 7, § 3]

n	$n_x - n_y = 0$ или 1		$n_x - n_y = 2$ или 3		$n_x - n_y = 4$ или 5		n	$n_x - n_y = 0$ или 1		$n_x - n_y = 2$ или 3		$n_x - n_y = 4$ или 5	
	5%	1%	5%	1%	5%	1%		5%	1%	5%	1%	5%	1%
6	∞	∞	∞	∞	∞	∞	29	4,78	6,22	4,76	6,19	4,72	6,13
7	∞	∞	∞	∞	∞	∞	30	4,88	6,35	4,87	6,34	4,84	6,30
8	2,40	∞	2,30	∞	∞	∞	31	4,97	6,47	4,95	6,44	4,91	6,39
9	2,38	∞	2,20	∞	∞	∞	32	5,07	6,60	5,06	6,58	5,03	6,55
10	2,60	3,20	2,49	3,10	2,30	∞	33	5,15	6,71	5,13	6,69	5,10	6,64
11	2,72	3,40	2,58	3,40	2,40	∞	34	5,25	6,84	5,24	6,82	5,21	6,79
12	2,86	3,60	2,79	3,58	2,68	3,40	35	5,33	6,95	5,31	6,92	5,28	6,88
13	2,96	3,71	2,91	3,64	2,78	3,50	36	5,42	7,06	5,41	7,05	5,38	7,02
14	3,11	3,94	3,06	3,88	3,00	3,76	37	5,50	7,17	5,48	7,15	5,45	7,11
15	3,24	4,07	3,19	4,05	3,06	3,88	38	5,59	7,28	5,58	7,27	5,55	7,25
16	3,39	4,26	3,36	4,25	3,28	4,12	39	5,67	7,39	5,65	7,37	5,62	7,33
17	3,49	4,44	3,44	4,37	3,36	4,23	40	5,75	7,50	5,74	7,49	5,72	7,47
18	3,63	4,60	3,60	4,58	3,53	4,50	41	5,83	7,62	5,81	7,60	5,79	7,56
19	3,73	4,77	3,69	4,71	3,61	4,62	42	5,91	7,72	5,90	7,71	5,88	7,69
20	3,86	4,94	3,84	4,92	3,78	4,85	43	5,99	7,82	5,97	7,81	5,95	7,77
21	3,96	5,10	3,92	5,05	3,85	4,96	44	6,06	7,93	6,06	7,92	6,04	7,90
22	4,08	5,26	4,06	5,24	4,01	5,17	45	6,14	8,02	6,12	8,01	6,10	7,98
23	4,18	5,40	4,15	5,36	4,08	5,27	46	6,21	7,13	6,21	8,12	6,19	8,10
24	4,29	5,55	4,27	5,53	4,23	5,48	47	6,29	8,22	6,27	8,21	6,25	8,18
25	4,39	5,68	4,36	5,65	4,30	5,58	48	6,36	8,32	6,35	8,31	6,34	8,29
26	4,50	5,83	4,48	5,81	4,44	5,76	49	6,43	8,41	6,42	8,40	6,39	8,37
27	4,59	5,95	4,56	5,92	4,51	5,85	50	6,50	8,51	6,50	8,50	6,48	8,48
28	4,69	6,09	4,68	6,07	4,64	6,03							

Таблица XXII

Значения ψ для вариант, распределенных по закону Пуассона [гл. 5, § 2]

x	ψ	x	ψ	x	ψ	x	ψ
1	1,182	26	5,136	51	7,168	76	8,739
2	1,545	27	5,232	52	7,237	77	8,796
3	1,849	28	5,327	53	7,306	78	8,853
4	2,093	29	5,420	54	7,374	79	8,909
5	2,319	30	5,511	55	7,441	80	8,965
6	2,526	31	5,601	56	7,508	81	9,021
7	2,716	32	5,690	57	7,575	82	9,076
8	2,894	33	5,777	58	7,640	83	9,131
9	3,062	34	5,863	59	7,706	84	9,186
10	3,221	35	5,948	60	7,770	85	9,240
11	3,373	36	6,031	61	7,834	86	9,294
12	3,518	37	6,114	62	7,898	87	9,347
13	3,657	38	6,195	63	7,961	88	9,401
14	3,792	39	6,275	64	8,023	89	9,454
15	3,921	40	6,354	65	8,086	90	9,507
16	4,047	41	6,432	66	8,147	91	9,559
17	4,168	42	6,510	67	8,208	92	9,611
18	4,287	43	6,586	68	8,269	93	9,663
19	4,402	44	6,661	69	8,329	94	9,715
20	4,514	45	6,736	70	8,389	95	9,766
21	4,623	46	6,810	71	8,448	96	9,817
22	4,730	47	6,883	72	8,507	97	9,868
23	4,835	48	6,955	73	8,566	98	9,918
24	4,937	49	7,027	74	8,624	99	9,969
25	5,037	50	7,098	75	8,682	100	10,019

Таблица XXIII

Критические значения числа знаков (менее часто встречающихся) Z_α [гл. 7, § 6]

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		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 $Z \geqslant Z_{05}$ и отвергается при $Z < Z_{01}$.

Таблица XXIV

Критические значения числа серий S_{05} [гл. 7, § 4]

n_x \ n_y	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
6	3	3	3														
7	3	3	4	4													
8	3	3	4	4	5												
9	3	4	4	5	5	6											
10	3	4	5	5	6	6	6										
11	3	4	5	5	6	6	7	7									
12	4	4	5	6	6	7	7	8	8								
13	4	4	5	6	6	7	8	8	9	9							
14	4	5	5	6	7	7	8	8	9	9	10						
15	4	5	6	6	7	8	8	9	9	10	10	11					
16	4	5	6	6	7	8	8	9	10	10	11	11	11				
17	4	5	6	7	7	8	9	9	10	10	11	11	12	12			
18	4	5	6	7	8	8	9	10	10	11	11	12	12	13	13		
19	4	5	6	7	8	8	9	10	10	11	12	12	13	13	14	14	
20	4	5	6	7	8	9	9	10	11	11	12	12	13	13	14	14	15

Нулевая гипотеза принимается при $S \geqslant S_{05}$ и отвергается при $S < S_{05} - 2$

Таблица XXV

Значения величины $z(r) = \frac{1}{2} \ln \frac{1+r}{1-r}$ [гл. 8, § 5]

r	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0501	0,0601	0,0701	0,0802	0,0902
1	0,1003	0,1105	0,1206	0,1308	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
2	0,2027	0,2132	0,2237	0,2342	0,2448	0,2554	0,2661	0,2769	0,2877	0,2986
3	0,3095	0,3206	0,3317	0,3428	0,3541	0,3654	0,3769	0,3884	0,4001	0,4118
4	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
5	0,5493	0,5627	0,5763	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
6	0,6931	0,7089	0,7250	0,7414	0,7582	0,7753	0,7928	0,8107	0,8291	0,8480
7	0,8673	0,8872	0,9076	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
8	1,0986	1,1270	1,1518	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
9	1,4722	1,5275	1,5890	1,6584	1,7380	1,8318	1,9459	2,0923	2,2976	2,6467

Таблица XXVI

Значения r для z от 0,00 до 2,99 [гл. 8, § 5]

	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	6527	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	9327	9830	9834
4	9837	9340	9843	9846	9349	9352	9355	9858	9861	9864
5	9866	9869	9871	9874	9876	9879	9381	9884	9886	9888
6	9890	9392	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	9335	9936	9937	9938
9	9940	9941	9942	9943	9944	9945	9946	9947	9948	9949

Таблица XXVII

Критические значения выборочного коэффициента корреляции r_α
[гл. 8, § 5]

	$\alpha = 5\%$	$\alpha = 1\%$		$\alpha = 5\%$	$\alpha = 1\%$
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

r незначим при $r \leqslant r_{05}$ и значим при $r > r_{01}$.

Таблица XXVIII

Критические значения выборочного показателя корреляции рангов r_α^S
[гл. 10, § 1]

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 $r^S \leqslant r^S_{05}$ и значим при $r^S > r^S_{01}$.

Таблица XXIX

Квадраты трехзначных чисел

	0	1		3	4	5	6	7	8	9
0	0	1	4	9	16	25	36	49	64	81
10	100	121	144	169	196	225	256	289	324	361
20	400	441	484	529	576	625	676	729	784	841
30	900	961	1 024	1 089	1 156	1 225	1 296	1 369	1 444	1 521
40	1 600	1 681	1 764	1 849	1 936	2 025	2 116	2 209	2 304	2 401
50	2 500	2 601	2 704	2 809	2 916	3 025	3 136	3 249	3 364	3 481
60	3 600	3 721	3 844	3 969	4 096	4 225	4 356	4 489	4 624	4 761
70	4 900	5 041	5 184	5 329	5 476	5 625	5 776	5 929	6 084	6 241
80	6 400	6 561	6 724	6 889	7 056	7 225	7 396	7 569	7 744	7 921
90	8 100	8 281	8 464	8 649	8 836	9 025	9 216	9 409	9 604	9 801
100	10 000	10 201	10 404	10 609	10 816	11 025	11 236	11 449	11 664	11 881
110	12 100	12 321	12 544	12 769	12 996	13 225	13 456	13 689	13 924	14 161
120	14 400	14 641	14 884	15 129	15 376	15 625	15 876	16 129	16 384	16 641
130	16 900	17 161	17 424	17 689	17 956	18 225	18 496	18 769	19 044	19 321
140	19 600	19 881	20 164	20 449	20 736	21 025	21 316	21 609	21 904	22 201
150	22 500	22 801	23 104	23 409	23 716	24 025	24 336	24 649	24 964	25 281
160	25 600	25 921	26 244	26 569	26 896	27 225	27 556	27 889	28 224	28 561
170	28 900	29 241	29 584	29 929	30 276	30 625	30 976	31 329	31 684	32 041
180	32 400	32 761	33 124	33 489	33 856	34 225	34 596	34 965	35 344	35 721
190	36 100	36 481	36 864	37 249	37 636	38 025	38 416	38 809	39 204	39 601
200	40 000	40 401	40 804	41 209	41 616	42 025	42 436	42 849	43 264	43 681
210	44 100	44 521	44 944	45 369	45 796	46 225	46 656	47 089	47 524	47 961
220	48 400	48 841	49 284	49 729	50 176	50 625	51 076	51 529	51 984	52 441
230	52 900	53 361	53 824	54 289	54 756	55 225	55 696	56 169	56 644	57 121
240	57 600	58 081	58 564	59 049	59 536	60 025	60 516	61 009	61 504	62 001
250	62 500	63 001	63 504	64 009	64 516	65 025	65 536	66 049	66 564	67 081
260	67 600	68 121	68 644	69 169	69 696	70 225	70 756	71 289	71 824	72 361
270	72 900	73 441	73 984	74 529	75 076	75 625	76 176	76 729	77 284	77 841
280	78 400	78 961	79 524	80 089	80 656	81 225	81 796	82 369	82 944	83 521
290	84 100	84 681	85 264	85 849	86 436	87 025	87 616	88 209	88 804	89 401
300	90 000	90 601	91 204	91 809	92 416	93 025	93 636	94 249	94 864	95 481
310	96 100	96 721	97 344	97 969	98 596	99 225	99 856	100 489	101 124	101 761
320	102 400	103 041	103 684	104 329	104 976	105 625	106 276	106 929	107 584	108 241
330	108 900	109 561	110 224	110 889	111 556	112 225	112 896	113 569	114 244	114 921
340	115 600	116 281	116 964	117 649	118 336	119 025	119 716	120 409	121 104	121 801
350	122 500	123 201	123 904	124 609	125 316	126 025	126 736	127 449	128 164	128 881

Таблица XXIX (продолжение)

		1		3	4					
360	129 600	130 321	131 044	131 769	132 496	133 225	133 956	134 689	135 424	136 161
370	136 900	137 641	138 384	139 129	139 876	140 625	141 376	142 129	142 884	143 641
380	144 400	145 161	145 924	146 689	147 456	148 225	148 996	149 769	150 544	151 321
390	152 100	152 881	153 664	154 449	155 236	156 025	156 816	157 609	158 404	159 201
400	160 000	160 801	161 604	162 409	163 216	164 025	164 836	165 649	166 464	167 281
410	168 100	168 921	169 744	170 569	171 396	172 225	173 056	173 889	174 724	175 561
420	176 400	177 241	178 084	178 929	179 776	180 625	181 476	182 329	183 184	184 041
430	184 900	185 761	186 624	187 489	188 356	189 225	190 096	190 969	191 844	192 721
440	193 600	194 481	195 364	196 249	197 136	198 025	198 916	199 809	200 704	201 601
450	202 500	203 401	204 304	205 209	206 116	207 025	207 936	208 849	209 764	210 681
460	211 600	212 521	213 444	214 369	215 296	216 225	217 156	218 089	219 024	219 961
470	220 900	221 841	222 784	223 729	224 676	225 625	226 576	227 529	228 484	229 441
480	230 400	231 361	232 324	233 289	234 256	235 225	236 196	237 169	238 144	239 121
490	240 100	241 081	242 064	243 049	244 036	245 025	246 016	247 009	248 004	249 001
500	250 000	251 001	252 004	253 009	254 016	255 025	256 036	257 049	258 064	259 081
510	260 100	261 121	262 144	263 169	264 196	265 225	266 256	267 289	268 324	269 361
520	270 400	271 441	272 484	273 529	274 576	275 625	276 676	277 729	278 784	279 841
530	280 900	281 961	283 024	284 089	285 156	286 225	287 296	288 369	289 444	290 521
540	291 600	292 681	293 764	294 849	295 936	297 025	298 116	299 209	300 304	301 401
550	302 500	303 601	304 704	305 809	306 916	308 025	309 136	310 249	311 364	312 481
560	313 600	314 721	315 844	316 969	318 096	319 225	320 356	321 489	322 624	323 761
570	324 900	326 041	327 184	328 329	329 476	330 625	331 776	332 929	334 084	335 241
580	336 400	337 561	338 724	339 889	341 056	342 225	343 396	344 569	345 744	346 921
590	348 100	349 281	350 464	351 649	352 836	354 025	355 216	356 409	357 604	358 801
600	360 000	361 201	362 404	363 609	364 816	366 025	367 236	368 449	369 664	370 881
610	372 100	373 321	374 544	375 769	376 996	378 225	379 456	380 689	381 924	383 161
620	384 400	385 641	386 884	388 129	389 376	390 625	391 876	393 129	394 384	395 641
630	396 900	398 161	399 424	400 689	401 956	403 225	404 496	405 769	407 044	408 321
640	409 600	410 881	412 164	413 449	414 736	416 025	417 316	418 609	419 904	421 201
650	422 500	423 801	425 104	426 409	427 716	429 025	430 336	431 649	432 964	434 281
660	435 600	436 921	438 244	439 569	440 896	442 225	443 556	444 889	446 224	447 561
670	448 900	450 241	451 584	452 929	454 276	455 625	456 976	458 329	459 684	461 041
680	462 400	463 761	465 124	466 489	467 856	469 225	470 596	471 969	473 344	474 721
690	476 100	477 481	478 864	480 249	481 636	483 025	484 416	485 809	487 204	488 601
700	490 000	491 401	492 804	494 209	495 616	497 025	498 436	499 849	501 264	502 681

Таблица XXIX (окончание)

	0	1		3	4	5	6	7	8	9
710	504 100	505 521	506 944	508 369	509 796	511 225	512 656	514 089	515 524	516 961
720	518 400	519 841	521 284	522 729	524 176	525 625	527 076	528 529	529 984	531 441
730	532 900	534 361	535 824	537 289	538 756	540 225	541 696	543 169	544 644	546 121
740	547 600	549 081	550 564	552 049	553 536	555 025	556 516	558 009	559 504	561 001
750	562 500	564 001	565 504	567 009	568 516	570 025	571 536	573 049	574 564	576 081
760	577 600	579 121	580 644	582 169	583 696	585 225	586 756	588 289	589 824	591 361
770	592 900	594 441	595 984	597 529	599 076	600 625	602 176	603 729	605 284	606 841
780	608 400	609 961	611 524	613 089	614 656	616 225	617 796	619 369	620 944	622 521
790	624 100	625 681	627 264	628 849	630 436	632 025	633 616	635 209	636 804	638 401
800	640 000	641 601	643 204	644 809	646 416	648 025	649 636	651 249	652 864	654 481
810	565 100	657 721	659 344	660 929	662 596	664 225	665 856	667 489	669 124	670 761
820	672 400	674 041	675 684	677 329	678 976	680 625	682 276	683 929	685 584	687 241
830	688 900	690 561	692 224	693 889	696 556	697 225	698 896	700 569	702 244	703 921
840	705 600	707 281	708 964	710 649	712 336	714 025	715 716	717 409	719 104	720 801
850	722 500	724 201	725 904	727 609	729 316	731 025	732 736	734 449	736 164	737 881
860	739 600	741 321	743 044	744 769	746 496	748 225	749 956	751 689	753 424	755 161
870	756 900	758 641	760 384	762 129	763 876	765 625	767 376	769 129	770 884	772 641
880	774 400	776 161	777 924	779 689	781 456	783 225	784 996	786 769	788 544	790 321
890	792 100	793 881	795 664	797 449	799 236	801 025	802 816	804 609	806 404	808 201
900	810 000	811 801	813 604	815 409	817 216	819 025	820 836	822 649	824 464	826 281
910	828 100	829 921	831 744	833 569	835 396	837 225	839 056	840 889	842 724	844 561
920	846 400	848 241	850 084	851 929	853 776	855 625	857 476	859 329	861 184	863 041
930	864 900	866 761	868 624	870 489	872 356	874 225	876 096	877 969	879 844	881 721
940	883 600	885 481	887 364	889 249	891 136	893 025	894 916	896 809	898 704	900 601
950	902 500	904 401	906 304	908 209	910 116	912 025	913 936	915 849	917 764	919 681
960	921 600	923 521	925 444	927 369	929 296	931 225	933 156	935 089	937 024	938 961
970	940 900	942 841	944 784	946 729	948 676	950 625	952 576	954 529	956 484	958 441
980	960 400	962 361	964 324	966 289	968 256	970 225	972 196	974 169	976 144	978 121
990	980 100	982 081	984 064	986 049	988 036	990 025	992 016	994 009	996 004	998 001

Пример. 1) $437^2 = 190\,969$. Квадрат четырехзначного числа может быть подсчитан по формуле $(a \pm b)^2 = a^2 \pm 2ab + b^2$; 2) $437{,}2^2 = (437+0{,}2)^2 = 190\,969 + 174{,}8 + 0{,}04 = 191143{,}84$.

ВАЖНЕЙШИЕ ФОРМУЛЫ

(Числа в скобках указывают страницы текста)

Среднее значение (28):

$$\hat{x} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i. \qquad N = \sum_{i=1}^{k} n_i.$$

Среднее значение от функций (30, 33):

$$\langle ax \rangle = a\hat{x};$$
$$\langle x + a \rangle = \hat{x} + a;$$
$$\langle y - x \rangle = \hat{y} - \hat{x};$$
$$\langle y + x \rangle = \hat{y} + \hat{x}.$$

Дисперсия (40):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \hat{x})^2 = \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - \hat{x}^2.$$

Дисперсия функции (43, 44):

$$\sigma^2 \{ax\} = a^2\sigma^2 \{x\};$$
$$\sigma^2 \{x + a\} = \sigma^2 \{x\};$$
$$\sigma^2 \{x + y\} = \sigma^2 \{x - y\} = \sigma^2 \{x\} + \sigma^2 \{y\}$$

(при отсутствии корреляции между x и y).

Выборочная оценка дисперсии (105):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2.$$

Среднее значение и дисперсия объединенной выборки (116, 118):

$$\bar{x} = \sum_{j=1}^{w} \frac{\bar{x}^{(j)}}{s^2_{\bar{x}(j)}} \Bigg/ \sum_{j=1}^{w} \frac{1}{s^2_{\bar{x}(j)}};$$

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

Коэффициент вариации (45):

$$v = \frac{\sigma}{\hat{x}} 100\%.$$

Связь между центральными и начальными моментами (49, 50):

$$\mu_2 = m_2 - m_1^2;$$
$$\mu_3 = m_3 - 3m_2m_1 + 2m_1^3 = m_3 - 3\mu_2 m_1 - m_1^3;$$
$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4 =$$
$$= m_4 - 4\mu_3 m_1 - 6\mu_2 m_1^2 - m_1^4.$$

Коэффициент асимметрии (47):

$$\hat{A} = \rho_3 = \frac{\mu_3}{\sigma^3}.$$

Коэффициент эксцесса (71):

$$\hat{E} = \rho_4 - 3 = \frac{\mu_4}{\sigma^4} - 3.$$

Оценка стандартного отклонения (стандартной ошибки) среднего значения (107):

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

Стандартная ошибка среднего значения при альтернативном распределении (129):

$$\sigma_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Стандартная ошибка суммы и разности средних значений (120):

$$\sigma_{\bar{x} \pm \bar{y}} = \sqrt{\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2}$$

Стандартная ошибка произведения и частного средних значений (122):

$$\sigma_{\bar{z}} = |\bar{z}| \sqrt{\left(\frac{\sigma_{\bar{x}}}{\bar{x}}\right)^2 + \left(\frac{\sigma_{\bar{y}}}{\bar{y}}\right)^2}$$

Стандартные ошибки стандартного отклонения, коэффициентов асимметрии и эксцесса (108):

$$\sigma_s \approx \frac{s}{\sqrt{2n}} \qquad \sigma_A \approx \sqrt{\frac{6}{n+3}}; \qquad \sigma_E \approx \sqrt{\frac{24}{n+5}}$$

Скользящие средние: простое (339)

$$\tilde{y}_i = (y_{i-1} + y_i + y_{i+1}) : 3; \quad \tilde{y}_1 = (7y_1 + 4y_2 - 2y_3) : 9;$$

взвешенное (344)

$$\tilde{y}_i = (y_{i-2} + 2y_{i-1} + 4y_i + 2y_{i+1} + y_{j+2}) : 10;$$
$$\tilde{y}_1 = (7y_1 + 5y_2 - y_3 - y_4) : 10;$$
$$\tilde{y}_2 = (3y_1 + 5y_2 + y_3 + y_4) : 10.$$

Критерий принадлежности крайних вариант к совокупности (139):

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

Критерий Стьюдента (149):

$$t \quad \frac{\bar{x} - \bar{y}}{s_{\bar{x}} \; \bar{y}};$$

$$s_{\bar{x} \cdot \bar{y}} \quad \sqrt{\frac{\Sigma (x_i - \bar{x})^2 \quad \Sigma (y_i - \bar{y})^2}{n_x \; n_y - 2} \quad \frac{n_x + n_y}{n_x n_y}}.$$

Критерий Фишера для сравнения дисперсий (164):

$$F \quad \frac{s_1^2}{s_1^2}$$

Критерий χ^2:
сравнение эмпирического и теоретического распределений (222)

$$\chi^2 = \sum_i \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = \sum_i \frac{n_i^2}{\hat{n}_i} - n;$$

сравнение двух эмпирических распределений с одинаковыми объемами (231)

$$\chi^2 = \sum_i \frac{(n_i' - n_i'')^2}{n_i' + n_i''};$$

то же с разными объемами (234)

$$\chi^2 = \frac{1}{n' n''} \sum_i \frac{(n_i' n'' - n_i'' n')^2}{n_i' + n_i''};$$

то же для альтернативного распределения (237)

$$\chi^2 = \frac{(|\,ad - bc\,| - n/2)^2\, n}{(a+b)(c+d)(a+c)(b+d)}$$

Критерий Фишера для таблиц 2×2 (239):

$$P = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!a!b!c!d!}$$

Критерий ван дер Вардена (252):

$$X = \sum_k \Psi\left(\frac{\nu_k}{n+1}\right).$$

Критерий Колмогорова—Смирнова (257):

$$\lambda^2 = D^2 \frac{n_x n_y}{n_x + n_y}; \qquad D = \max |\, z_i\{x\} - z_i\{y\}\,|.$$

Уравнения линейной регрессии (271):

$$\hat{y}_x - \hat{y} = \beta_{y/x}(x - x); \quad \hat{x}_y - \hat{x} = \beta_{x/y}(y - \hat{y}).$$

Коэффициенты регрессии (272):

$$\beta_{y/x} = \frac{\operatorname{cov}\{x, y\}}{\sigma_x^2}; \quad \beta_{x/y} = \frac{\operatorname{cov}\{x, y\}}{\sigma_y^2}.$$

Частные коэффициенты регрессии (321):

$$\beta_{(x/y)z} = \frac{\beta_{x/y} - \beta_{x/z}\beta_{z/y}}{1 - \beta_{y/z}\beta_{z/y}};$$

$$\beta_{x/y} = \frac{\beta_{(x/y)z} + \beta_{(x/z)y}\beta_{(z/y)x}}{1 - \beta_{(x/z)y}\beta_{(z/x)y}}$$

Коэффициент корреляции (283):

$$\rho = \frac{\operatorname{cov}\{x, y\}}{\sigma_x \sigma_y} = \sqrt{\beta_{x/y}\beta_{y/x}}.$$

Коэффициент частной корреляции (318, 320):

$$\rho_{xy(z)} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{(1 - \rho_{xz}^2)(1 - \rho_{yz}^2)}};$$

$$\rho_{xy} = \frac{\rho_{xy(z)} + \rho_{xz(y)}\rho_{yz(x)}}{\sqrt{(1 - \rho_{xz(y)}^2)(1 - \rho_{yz(x)}^2)}}.$$

Ковариация (273):

$$\operatorname{cov}\{x, y\} = \frac{1}{N}\Sigma n_{xy}\,(x - \hat{x})\,(y - \hat{y}).$$

Показатель корреляции рангов (347):

$$\rho^S = 1 - \frac{6\Sigma d^2}{N(N^2-1)}.$$

Коэффициент взаимной сопряженности (357):

$$K = \sqrt{\frac{\chi^2}{N\sqrt{(k_A-1)(k_B-1)}}}.$$

Коэффициент взаимной сопряженности для таблиц 2×2 (360):

$$K = \frac{|\,n_{11}n_{22} - n_{12}n_{21}\,| - N/2}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}}.$$

Нормальное (гауссово) распределение (65):

$$\varphi(x) = \frac{N}{\sqrt{2\pi}\cdot\sigma}\,e^{-\frac{(x-\hat{x})^2}{2\sigma^2}}$$

Распределение Пуассона (85):

$$n_x = N\frac{\hat{x}^x}{x!}e^{-\hat{x}}.$$

Биномиальное распределение (60):

$$n_x = NC_\nu^x\hat{p}^x\,(1-\hat{p})^{\nu-x};$$

$$C_\nu^x = \frac{\nu(\nu-1)(\nu-2)\ldots(\nu-|x-1|)}{x|}.$$

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

Виктор Юльевич Урбах
Биометрические методы

Утверждено к печати
Институтом биологической физики
Академии Наук СССР

Редактор издательства *Л. Н. Большев*
Художник *С. Н. Голубев*
Технический редактор *П. С. Кашин*

Сдано в набор 3/IX 1964 г. Подписано к печати 9/XI 1964 г.
Формат 60×90¹/₁₆. Печ. л. 26. Уч.-изд. л. 22,4
Тираж 8000 экз. Изд. № 3618/₆₄. Тип. зак. № 1147.
Т—15378 Темплан 1965 г., № 705

Цена 1 р. 72 коп.

Издательство «Наука».
Москва, К-62, Подсосенский пер., 21
2-я типография издательства «Наука».
Москва, Г-99, Шубинский пер., 10

ОПЕЧАТКИ И ИСПРАВЛЕНИЯ

Стр.	Строка	Напечатано	Должно быть
248	1 и 2 св.	$\frac{T}{n_x}$	$\frac{T_x}{n_x}$
289	Табл. 113, гр. 1	$N = 12$	$n = 12$
339	1 св.	$b_{\ x}$	$b_{y/x}$
384	Табл. XI, примечание	$\tau_{0,01}$	τ_{01}
384	То же	$\tau_{0,05}$	τ_{05}
407	9 св.	$F = \frac{s_1^2}{s_1^2}$	$F = \frac{s_1^2}{s_2^2}$