Н. В. Глотов, Л. А. Животовский,
Н. В. Хованов, Н. Н. Хромов-Борисов

Биометрия

Учебное пособие

Под редакцией д-ра биол. наук проф.
М. М. Тихомировой

УДК
ББК

Биометрия: Учеб. пособие / Н. В. Глотов, Л. А. Животовский, Н. В. Хованов, Н. Н. Хромов-Борисов; Под ред. М. М. Тихомировой. — 381 с.

В учебном пособии рассматривается аппарат математической статистики, знание которого необходимо для решения разнообразных биологических задач. Приводятся сведения по теории вероятностей, излагаются теории статистического оценивания, проверки статистических гипотез, основы дисперсионного, регрессионного и корреляционного анализа. Описываются конкретные методики для разных областей биологии, каждая глава сопровождается задачами для самостоятельного решения. В отличие от существующих в настоящем пособии проводится четкое разграничение между моделью и ее биологическим приложением.

Пособие предназначено для студентов-биологов, а также для специалистов, использующих методы математической статистики.

ISBN **ББК**

Предисловие

Традиции преподавания биометрии в нашей стране берут начало от Ю. А. Филипченко, С. С. Четверикова, А. А. Сапегина. В 50-е гг существенный вклад в совершенствование преподавания биометрии, в пропаганду биометрических методов внесли П. Ф. Рокицкий, П. В. Терентьев, Н. А. Плохинский. Главная особенность отечественной биометрической школы заключается в стремлении к предельной биологизации статистических методов, в наглядном показе биологического смысла любой статистической процедуры, в попытке дать биологическую интерпретацию значениям любых статистических параметров и изменениям этих значений. Отнюдь не случайно деятелям русской биометрической школы удалось получить ряд изящных имеющих глубокий смысл результатов в генетике, ботанике и зоологии, в растениеводстве и животноводстве, что требовало обширных биологических познаний.

Но где сила, там и слабость. Отечественные биометрики никогда не строили математико-статистическую модель, имеющую, если можно так выразиться, свою «жизнь». Строгая формулировка модели всегда связана с введением упрощающих условий, с возникновением специальных, чисто математических задач. Применение же соответствующего математико-статистического метода в биологии сводится к установлению допустимости применения математической модели к соответствующей биологической ситуации. Такой подход отнюдь не снижает требований к уровню математической подготовки биолога и составляет основу, на которой формируется то, что обозначают словами «статистическое мышление».

Этим путем развивается англо-американская биометрическая школа, по сути дела берущая начало от основополагающей книги Р. А. Фишера «статистические методы для исследователей». Стиль изложения предмета представителями этой школы нашел, пожалуй, наиболее полное отражение в переведенной у нас книге Н. Бейли [1962]. Автор излагает суть статистической модели, ее ограничения и, полностью опуская все процедуры математических выводов и выкладок, сразу дает конечные формулы для вычислений.

Такая чисто словесная формулировка статистической задачи (скорее, рассказ о задаче) и отсутствие указаний на пути статистического выво-

да требуют от читателя очень большого внимания, напряженной мыслительной работы, и даже если эта работа оказывается успешной, подчас остается неясной логика (последовательность) статистического вывода. По-видимому, все-таки невозможно, говоря о математических моделях, полностью избежать использования математических формализмов. Нам представляется, что книги В. Ю. Урбаха [1964; 1975] являются существенным этапом на пути решения этой проблемы.

Точка зрения авторов данного учебного пособия на основные принципы преподавания биометрии сводится к следующему. Курс биометрии — это лишь один из этапов университетского математического образования студентов-биологов. Причем не первый этап. Начало должно быть положено курсом общей математики, что сейчас и делается. Заметим лишь, что принципы и объем курса общей математики по традиции во многом связаны пока со способом и объемом преподавания математики в технических вузах.

Второй этап — курс основ (наверно, правильнее — элементов) теории вероятностей и математической статистики. Логически, исходя из несомненно существующей специфики биологических приложений вероятностно-статистических методов, изложение теории вероятностей относится к этому этапу, а не к курсу общей математики. Второй этап и есть собственно курс биометрии в нашем понимании. Главное здесь — дать понятие о вероятностной и статистической модели и провести четкое разграничение между моделью и ее биологическим приложением. Статистическая модель — это феномен математический, ее структуру, функционирование и вытекающие из нее следствия нельзя описать чисто словесно, полностью отказавшись от формализмов математического языка. Полное упрощение неизбежно повлечет за собой искажение смысла. Поэтому возникает вопрос об уровне строгости изложения. Принять современный уровень строгости теории вероятностей и математической статистики — дело немыслимое. Это означало бы превращение биолога в математика. Поэтому мы избираем уровень строгости изложения, близкий уровню строгости, принятому в технических вузах.

Выбор статистических методов, даваемых в курсе биометрии, сегодня вряд ли вызывает споры. В этом отношении и наше пособие не отличается новаторством. Однако нам представляется важным изложение разных методов с единой точки зрения, поэтому мы строим курс на основе нормального распределения.

Разумеется, методы, излагаемые в курсе биометрии, составляют лишь небольшую часть возможных приложений математико-статистического аппарата. Но ведь биометрия — лишь второй этап математического обра-

зования биологов. Необходим и имеет место третий этап — математические (в том числе и статистические) методы специальных биологических дисциплин (микробиологии, зоологии, ботаники, генетики, биохимии, физиологии и т. д.), — на котором и решаются более сложные, более специальные задачи, опускаемые в курсе биометрии.

Построение и анализ в курсе биометрии математико-статистических моделей не должны лишать нас ярких достоинств русской биометрической школы — биологического взгляда на количественные методы. Нам представляется, что таким духом должен быть проникнут весь курс. В значительной мере это достигается собственно приложениями — практическими занятиями, решением задач. Прежде всего биологические примеры, рассматриваемые в курсе, должны быть содержательными: не просто выдуманными примерами, но реально возникающими в разных областях науки задачами. И решение этих задач должно заключаться не в рецептурной подстановке в вычислительные формулы экспериментальных данных, а в обоснованном выборе статистической модели.

Такого рода подход, по-видимому, никогда раньше не проводился сколько-нибудь последовательно в биометрии. И если это привлечет внимание читателя и позволит студентам овладеть не только рецептурой и техникой статистических выкладок, но и основами вероятностно-статистического мышления, то авторы будут считать свою задачу выполненной.

Над всеми разделами пособия авторы работали совместно. Даже если первоначальный вариант той или иной главы был написан одним из нас, в дальнейшем он перерабатывался, дополнялся другими.

Авторы выражают сердечную признательность научному редактору профессору М. М. Тихомировой, сделавшей возможным написание этой книги. Маргарита Михайловна «приняла» курс «Биометрия» в Ленинградском университете от Павла Викторовича Терентьева, определила пути развития и совершенствования этого курса и передала его авторам — Н. В. Глотову и Н. Н. Хромову-Борисову.

Мы благодарны профессору С. Г. Инге-Вечтомову, энергично стимулировавшему нашу работу на всех этапах. Мы признательны О. И. Белозеровой, М. В. Голубковой, Н. В. Носкову за неоценимую помощь в подготовке пособия к печати, М. И. Рахману за вычисление некоторых значений в табл. VI и VII Приложения 1, В. М. Кузнецовой за техническую подготовку рукописи.

Выражаем признательность всем коллегам, предоставившим результаты своих исследований для составления многих задач и примеров.

Основные обозначения

(В скобках указаны параграф и глава, в которых вводится данное обозначение)

Знак $\sim$ («тильда») ставится над буквой и обозначает случайную величину, например, $\widetilde{x}$ (§ 1 гл. I).

Буква без тильды есть значение соответствующей случайной величины, например, x есть значение случайной величины $\widetilde{x}$.

Индекс «эксп» означает, что значение статистики вычислено по экспериментальным (выборочным) данным, например, $g_{\text{эксп}}$ (§ 2 гл. IV).

Обозначение $\widetilde{x} \sim X(\Theta_1, \Theta_2)$ означает, что случайная величина $\widetilde{x}$ имеет X-распределение с параметрами Θ_1 и Θ_2 (§ 3 гл. III).

Знак $\varnothing$ — пустое множество, невозможное событие (§ 2, § 3 гл. I).

Буквами греческого алфавита обозначаются:

α — уровень значимости критерия (§ 2 гл. IV);

β — коэффициент линейной регрессии (§ 1 гл. I);

$\widetilde{\beta}_n$ — статистика критерия Колмогорова для проверки нормальности распределения (§ 4 гл. VII);

γ — комплекс условий случайного испытания $\mathscr{E}$;

δ — параметр рассеяния неизвестного распределения (§ 7 гл. VI);

ζ — медиана распределения случайной величины (§ 1 гл. III);

$\varkappa$ или $\varkappa_0$ — свободный член в уравнении линейной регрессии (§ 1 гл. I);

λ — параметр распределения Пуассона (§ 4 гл. II);

μ — среднее значение нормального распределения (§ 3 гл. III);

ν — число степеней свободы (§ 7 гл. III);

ϱ или $\varrho(\widetilde{x}_1, \widetilde{x}_2)$ — коэффициент корреляции двух случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$ (§ 2 гл. III);

ϱ_S — коэффициент ранговой корреляции Спирмена (§ 9 гл. IX);

ϱ_w — коэффициент внутриклассовой корреляции (§ 4 гл. VIII);

σ — среднее квадратичное отклонение нормального распределения (§ 3 гл. III);

σ^2 — дисперсия нормального распределения (§ 3 гл. III);

τ — параметр положения неизвестного распределения (§ 7 гл. VI);

$\Phi(u)$ — символ функции нормированного нормального распределения (§ 4 гл. III);

$\widetilde{\chi}^2$ — случайная величина, имеющая χ^2-распределение (§ 7 гл. III); статистика критерия χ^2 (гл. VII);

$\chi^2(\nu)$ — символ χ^2-распределения с параметром ν (§ 7 гл. III);

Ω — множество элементарных событий; достоверное событие (§ 1, § 3 гл. I);

ω_i — элементарный исход случайного испытания $\mathscr{E}$ (§ 1 гл. I);

$\{\omega_i\}$ — элементарное случайное событие (§ 3 гл. I).

Буквами латинского алфавита обозначаются:

$\mathscr{A}$ — алгебра случайных событий (§ 2, § 3 гл. I);

$A, B, \ldots$ — случайные события (§ 3 гл. I);

a или a_0 — оценка свободного члена $\varkappa$ или $\varkappa_0$ в уравнении линейной регрессии (§ 2 гл. IX);

b — оценка коэффициента линейной регрессии β (§ 2 гл. IX);

$\mathrm{Cov}(\widetilde{x}_1, \widetilde{x}_2)$ — ковариация случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$ (§ 2 гл. III);

$D\widetilde{x}$ — дисперсия случайной величины $\widetilde{x}$ (§ 1 гл. II);

$\sqrt{D\widetilde{x}}$ — среднее квадратичное отклонение случайной величины $\widetilde{x}$ (§ 1 гл. II);

d_i — разность значений $x_{1i} - x_{2i}$ в случае парных наблюдений (§ 3 гл. IV);

$\widetilde{D}_n$ — статистика критерия Колмогорова (§ 4 гл. VII);

$\mathscr{E}$ — случайное испытание (§ 1 гл. I);

$\widetilde{e}_i$ или $\widetilde{e}_{ij}$ — случайные ошибки в моделях дисперсионного анализа (§ 1, § 4 гл. VII);

$E\widetilde{x}$ — математическое ожидание или среднее значение случайной величины $\widetilde{x}$ (§ 1 гл. II);

$\widetilde{F}$ — случайная величина, имеющая F-распределение (§ 9 гл. III); статистика F-критерия (§ 1 гл. VI, гл. VIII);

$F(\nu_1, \nu_2)$ — символ F-распределения с параметрами ν_1 и ν_2 (§ 9 гл. III);

$F(x)$ — функция распределения случайной величины $\widetilde{x}$ (§ 10 гл. I);

$f(x)$ — плотность распределения случайной величины $\widetilde{x}$ (§ 1 гл. III);

$G(y)$ — производящая функция целочисленной случайной величины (§ 3 гл. II);

$\widetilde{H}$ — статистика критерия Крускала–Уоллиса (§ 7 гл. VIII);

H_0 — символ нулевой (проверяемой) гипотезы (§ 2 гл. IV);

H_1 — символ альтернативной (конкурирующей) гипотезы (§ 2 гл. IV);

h — частота, оценка параметра p биномиального распределения (§ 1 гл. V);

$h(A)$ — частота случайного события A (§ 4 гл. I);

$h(A; n)$ — частота случайного события A в последовательности n случайных испытаний (§ 4 гл. I);

k — значение целочисленной случайной величины, имеющей пуассоновское или биномиальное распределение (§ 4, § 5 гл. II);

L — символ функции правдоподобия (§ 1 гл. V);

m — оценка среднего значения μ нормального распределения или оценка параметра λ распределения Пуассона (§ 1 гл. V);

m_d — оценка среднего значения для разностей $\widetilde{x_{1i}} - \widetilde{x}_{2i}$ в случае парных наблюдений (§ 3 гл. IV);

MS_a — символ среднего квадрата между блоками (§ 5 гл. VIII);

MS_b — символ межгруппового среднего квадрата (§ 1 гл. VIII);

MS_w — символ внутригруппового среднего квадрата (§ 1 гл. VIII);

$N(\mu; \sigma^2)$ — символ нормального распределения с параметрами μ и σ^2 (§ 3 гл. III);

$N(0; 1)$ — символ нормированного нормального распределения (§ 3 гл. III);

$P(A)$ — вероятность случайного события A (§ 5 гл. I);

$P(A/B)$ — условная вероятность события A при реализации события B (§ 6 гл. I);

p — параметр биномиального распределения (§ 5 гл. II);

$\{p_i\}$ — распределение вероятностей целочисленной случайной величины (§ 1 гл. II);

$R(x_i)$ или R_i — ранг выборочного значения x_i (§ 9 гл. IX);

$\widetilde{r}$ — статистика коэффициента корреляции Пирсона (§ 6 гл. IX);

$\widetilde{r}_S$ — статистика рангового коэффициента корреляции Спирмена (§ 9 гл. IX);

r_w — оценка коэффициента внутриклассовой корреляции (§ 4 гл. VIII);

SS_a — символ суммы квадратов между блоками (§ 5 гл. VIII);

SS_b — символ межгрупповой суммы квадратов (§ 1 гл. VIII);

SS_t — символ общей (полной) суммы квадратов (§ 1 гл. VIII);

SS_w — символ внутригрупповой суммы квадратов (§ 1 гл. VIII);

s — оценка среднего квадратичного отклонения σ нормального распределения (§ 1 гл. V);

s^2 — оценка дисперсии σ^2 нормального распределения (§ 1 гл. V);

s_b — стандартная ошибка выборочного коэффициента линейной регрессии b (§ 2 гл. IX);

s_m — стандартная ошибка выборочного среднего m (§ 3 гл. V);

$\widetilde{t}$ — случайная величина, имеющая t-распределение (§ 8 гл. III); статистика t-критерия (§ 2,§ 3 гл. VI);

$t(\nu)$ — символ t-распределения с параметром ν (§ 8 гл. III);

U — статистика критерия Вилкоксона–Манна–Уитни (§ 7 гл. VI);

$\widetilde{u}$ — нормированная нормальная случайная величина (§ 3 гл. III); статистика u-критерия (§ 4 гл. VI);

$\widetilde{W}$ — статистика парного критерия Вилкоксона (§ 8 гл. VI);

$\{x_i, p_i\}$ — распределение вероятностей дискретной случайной величины $\widetilde{x}$ (§ 1 гл. II);

$x_{(i)}$ — выборочное значение, имеющее ранг R_i (§ 1 гл. IV);

Z — оценка медианы ζ распределения (§ 6 гл. V);

z — фишеровское преобразование коэффициента корреляции Пуассона r (§ 8 гл. IX);

z^* — преобразование Г. Хотеллинга для коэффициента корреляции Пирсона (§ 8 гл. IX).

Введение
Математические идеи в биологии и предмет биометрии

Эволюционная теория Ч. Дарвина явилась, по существу, первой естественно-научной теорией, которая привнесла в исследования вероятностный дух. Анализ взаимоотношений между такими исходными понятиями эволюционной теории, как изменчивость, наследственность и отбор, оказался бы несостоятельным без того, что сейчас называется вероятностным стилем мышления. Развитие математических идей в биологии неразрывно связано также с именем Г. Менделя и возникновением генетики, законы которой по самой своей природе являются вероятностными. И сегодня исследование проблем организации, функционирования, взаимодействия и эволюции живых систем уже немыслимо без привлечения идей и методов теории вероятностей, математической статистики и других разделов математики.

Еще ученые и мыслители прошлых столетий (И. Ньютон, П. Лаплас, И. Кант, Д. И. Менделеев и многие другие) пришли к выводу, что точность и уровень той или иной области человеческих знаний определяются степенью использования соответствующим разделом науки математических методов. Последнее возможно, как правило, при условии проведения количественных экспериментов и наблюдений, когда достаточно строго определены основные исходные понятия, когда известно, какие величины и как следует измерять. Чрезвычайное многообразие проявлений жизни на Земле, изобилие биологических форм и функциональных связей приводят, однако, к тому, что биология сегодняшнего дня имеет перед собой (а биология дня завтрашнего, по-видимому, будет иметь) целый ряд качественных задач, где применение математических, в частности, математико-статистических, методов вовсе не требуется. Хороший пример тому — сравнительно недавнее исследование ленинградского зоолога проф. А. В. Иванова, сумевшего выделить новый тип животных Pogonophora. Подчеркнем, что речь идет об открытии типа — высшей систематической единицы в царстве животных! Недаром эта работа была удостоена Ленинской премии. Несомненно,

качественные задачи существуют и будут существовать не только в классических областях биологии — в систематике, анатомии, гистологии, — но и в таких относительно новых, бурно развивающихся отделах биологии, как генетика, биохимия и биофизика, физиология.

В то же время ни один биолог, в том числе начинающий, знакомый с нашей наукой лишь по популярным статьям и книгам, не станет отрицать, что ее стволовым направлением, ее будущим является развитие мышления, количественного по сути, связанного с внедрением методологии и методов наиболее развитых отраслей естествознания — математики, физики, химии.

В самых различных областях теоретической и прикладной биологии — биогеоценологии и почвоведении, систематике и экологии, генетике и биохимии, биофизике, физиологии, в частных отделах зоологии, ботаники, микробиологии — в настоящее время эффективно используется разнообразный математический аппарат. Это теория матриц, дифференциальные и интегральные уравнения, теория вероятностей, математическая статистика и т. д. По-видимому, можно утверждать, что практически любой раздел математики (даже столь абстрактный, как топология) используется сегодня для решения тех или иных биологических задач.

Биометрия — область научного знания, охватывающая планирование и анализ результатов количественных биологических экспериментов и наблюдений методами математической статистики.

Современный количественный эксперимент включает в себя самостоятельное математико-статистическое исследование, которое начинается со статистического планирования эксперимента, т. е. с организации его постановки, и завершается статистической обработкой полученных результатов. Поэтому биометрия находит себе все более широкое общебиологическое применение, ибо задачи, которые она решает — планирование экспериментов и анализ их результатов, — составляют основу экспериментальной работы в любой частной области биологии. Можно сказать, что биометрия есть математическая культура биологического эксперимента.

Базой для построения статистических моделей в биологии, как и в любой другой области естествознания, техники, экономики, социологии, является теория вероятностей. Поэтому мы начинаем курс с изложения основ теории вероятностей.

Глава I

Основные представления теории вероятностей

> «Замечательно, что наука, которая начала с рассмотрения азартных игр, обещает стать наиболее важным объектом человеческого знания... Ведь большей частью важнейшие жизненные вопросы являются на самом деле лишь задачами из теории вероятностей».
>
> П. Лаплас

Понятия случайности и вероятности обсуждаются в течение последних двух тысяч лет, что привело к накоплению многочисленных и дополняющих друг друга точек зрения. К настоящему времени теория вероятностей сформировалась в виде обширной и сложной математической дисциплины, опирающейся на методы, заимствованные из различных отделов современной математики. Для определения вероятностных понятий используется аппарат теории множеств и теории меры, функционального анализа, линейной алгебры, топологии и т. п. Это превращает теорию вероятностей в высшей степени абстрактную науку. Глубина развития математического аппарата обеспечивает успешное применение теории вероятностей во многих областях науки и техники. На вероятностных методах основаны контроль промышленной продукции, расчеты артиллерийской и ракетной стрельбы и траекторий космических аппаратов, модели экономических, физических, химических и биологических процессов. Теория вероятностей успешно применяется при планировании и обработке результатов научных экспериментов, при конкретных социологических и психологических исследованиях, при проектировании сложных технических систем. Трудно назвать хотя бы одну область науки и техники, в которую в той или иной мере не проникали бы вероятностные методы. Это дает повод некоторым ученым говорить о стохастической революции в современном научном мышлении.

Биометрические методы представляют собой одну из важных областей применения теории вероятностей, имеющей свою специфику. Поэтому далее мы будем излагать тот фрагмент общей теории, следствия из которого непосредственно используются в биометрии и понимание которого необходимо для интерпретации биометрических схем. Ряд других понятий, вводимых в первых главах, необходим для целостного представления и изложения основных идей теории вероятностей и служит, образно выражаясь, вратами в стохастический мир.

§ 1. Случайное испытание

Начнем с представления об *испытании*, которое будем обозначать символом $\mathscr{E}$. Понятие испытания $\mathscr{E}$ предполагает определение некоторого *комплекса условий* испытания — обозначим его γ — и регистрацию связанного с этим комплексом условий наблюдаемого *исхода* испытаний, который мы будем обозначать ω. Комплексы условий и наблюдаемые исходы испытаний могут иметь различную природу. Это физические и биологические явления и процессы, «мысленные» эксперименты и т. д. Поясним сказанное несколькими примерами.

Пример I-1. Испытание $\mathscr{E}$ состоит в бросании монеты и наблюдении стороны, оказавшейся сверху. Здесь комплекс условий γ описывается хотя и довольно неопределенными, на практически одинаково понимаемыми требованиями «правильности» монеты (т. е. предполагается, что монета симметричная, изготовлена из однородного материала и т. д.), необходимости «кувыркания» монеты при падении, отсутствия препятствий на плоской поверхности в окрестности падения монеты и т. п. Исходом ω этого испытания $\mathscr{E}$ служит наименование верхней стороны упавшей монеты, т. е. имеются два варианта исхода ω: «герб» ($\omega = \omega_1$) и «решка» ($\omega = \omega_2$).

Различные варианты исхода ω мы будем называть *элементарными исходами* испытания $\mathscr{E}$ и обозначать той же буквой ω, но с различными индексами $\omega_t, \omega_i, \ldots$

Пример I-2. Испытание $\mathscr{E}$ состоит в бросании кубика (игральной кости), на каждой грани которого указано число очков (от 1 до 6), и наблюдении стороны, оказавшейся сверху. Комплекс условий γ здесь имеет тот же характер, что и в примере I-1, а исходом ω испытания служит число очков, указанное на верхней грани упавшей кости. Таким образом, имеются шесть элементарных исходов испытания $\mathscr{E}$: выпало одно очко, выпало два очка и т. д., выпало шесть очков ($\omega_1 = 1; \omega_2 = 2; \ldots; \omega_6 = 6$).

Приведенные примеры показывают, что понятия испытания $\mathscr{E}$, комплекса условий γ и исхода ω являются взаимосвязанными понятиями. Исход испытания, а следовательно, и само испытание можно определить различными способами в зависимости от цели или возможностей исследователя. Допустим, что по условиям игры исследователь различает грани только с четными (2, 4, 6) и нечетными (1, 3, 5) числами. Тогда исход ω' включает лишь два элементарных исхода: ω'_1 — «чет» и ω'_2 — «нечет», и при том же комплексе условий γ мы должны говорить об испытании $\mathscr{E}'$.

ПРИМЕР I-3. Испытание $\mathscr{E}$ состоит в скрещивании гетерозиготных особей Aa и наблюдении генотипа потомка. Комплекс условий γ включает образование яйцеклетки, несущей аллель A или a, в процессе онтогенеза; образование спермия, несущего аллель A или a, в процессе сперматогенеза; слияние гамет с образованием зиготы; процесс развития потомка до той стадии онтогенеза, на которой будет учитываться его генотип, и т. д. Исход ω есть генотип потомка. Здесь возможны три элементарных исхода: ω_1 — доминантная гомозигота AA, ω_2 — гетерозигота Aa и ω_3 — рецессивная гомозигота aa. Если исследователь учитывает лишь сумму гомозигот AA и aa, мы должны говорить об испытании $\mathscr{E}'$, комплексе условий γ и исходе ω' с двумя элементарными исходами: ω'_1 — гомозигота AA или aa и ω'_2 — гетерозигота Aa. Если исследователь не в состоянии различить потомков AA и Aa, мы должны рассматривать испытание $\mathscr{E}''$, комплекс условий γ и исход ω'' с двумя элементарными исходами: ω''_1 — потомок имеет доминантный признак (генотип AA или Aa) и ω''_2 — потомок имеет рецессивный признак (генотип aa).

ПРИМЕР I-4. Испытание $\mathscr{E}$ состоит в измерении роста призывника в военкомате. Здесь комплекс условий γ описывается инструкцией по измерению антропометрических показателей, а исход ω есть записываемое в журнал число, характеризующее рост призывника в сантиметрах. Элементарными исходами ω_t служат целые числа из некоторого диапазона, определяемого наименьшим и наибольшим измеренным ростом.

Любое измерение количественного, или мерного, признака, производимое, скажем, с помощью линейки, торзионных весов, спектрофотометра и т. п., может быть представлено как испытание $\mathscr{E}$. Комплекс условий γ будет определяться всеми теми предосторожностями, которые предусматривает научная методология измерения данного показателя. Наблюдаемый исход ω есть регистрируемое числовое значение. Вследствие конечной точности любого измерительного прибора элементарными исходами такого испытания являются рациональные числа. Однако теория дальнейшей мате-

матической обработки результатов измерений становится, как правило, гораздо проще, если допустить возможность бесконечно точного измерения, т. е. допустить в качестве элементарных исходов ω_t любые действительные числа.

Рассматривая испытание $\mathscr{E}$, в теории вероятностей предполагают, что это испытание может быть реализовано бесконечное число раз. Мы абстрагируемся, отвлекаемся от ограниченности наших сил и средств и предполагаем возможность бесконечного повторения одного и того же комплекса условий γ.

При этом возможны два принципиально разных отношения между комплексом условий γ и исходом ω. С одной стороны, при повторении комплекса условий γ мы можем всегда получить один и тот же единственный элементарный исход ω_0; такое испытание $\mathscr{E}$ будем называть *детерминированным*, т. е. однозначно определенным. С другой стороны, при повторении комплекса условий γ мы можем в разных реализациях получать различные элементарные исходы ω_0: иногда — элементарный исход ω_1, иногда — ω_2 и т. д.; такое испытание $\mathscr{E}$ будем называть *недетерминированным*.

Недетерминированное испытание часто называют *случайным испытанием* (все приведенные выше примеры — это примеры случайных испытаний). Предварительно мы также воспользуемся этой терминологией, отложив на дальнейшее более строгое определение понятия случайности. Здесь лишь отметим два смысловых оттенка, связанных со словом «случай». Во-первых, случайность может означать непредсказуемость того или иного явления. Этот смысловой оттенок используется, когда мы говорим, что, зная комплекс условий γ случайного испытания $\mathscr{E}$, нельзя однозначно предсказать элементарный исход ω_t в какой-то конкретной реализации. Во-вторых, — и такой смысловой оттенок нам представляется главным в рамках рассматриваемых далее математических схем — случайность испытания $\mathscr{E}$ указывает просто на то, что при одном и том же комплексе условий в разных реализациях возможны разные случаи, приводящие к различным элементарные исходам ω_t. Другими словами, в комплекс условий γ не входит ряд характеристик, воздействующих на исход конкретного испытания.

Нередко можно указать множество $\Omega = \{\omega_t\}$ всех возможных элементарных исходов случайного испытания. Такая ситуация имеет место в примерах I-1, I-2 и I-4. В примере I-4 диапазон допустимых значений признака может быть выбран достаточно широким, чтобы наверняка включить в себя значение роста любого призывника, тогда указанный числовой диапазон будет служить в качестве множества элементарных исходов Ω. Приведем еще несколько примеров.

ПРИМЕР I-5. Наблюдение пола новорожденного есть случайное испытание с двумя возможными элементарными исходами: ω_1 — родился мальчик, ω_2 — родилась девочка.

ПРИМЕР I-6. Измерение диаметра колонии дрожжей в чашке Петри может дать любое числовое значение, заключенное между нулем и величиной диаметра самой чашки. В этом случае множество элементарных исходов есть отрезок числовой оси $\Omega = [0, d]$, где d — диаметр чашки Петри.

ПРИМЕР I-7. Определяется процент насекомых, погибших от некоторой дозы ядохимиката. Хотя от опыта к опыту показатель может варьировать, очевидно, что множество возможных элементарных исходов такого случайного испытания есть множество действительных чисел, заключенных между 0 и 100.

В дальнейшем мы будем предполагать, что всегда выполняется описанное требование полной определенности множества элементарных исходов $\Omega = \{\omega_t\}$.

Многие биологические явления, описываемые при помощи понятия случайного испытания, допускают представление в виде простых и наглядных моделей, которые в дальнейшем мы будем называть модель «урна» и модель «мишень».

Модель «урна». В урне находятся неразличимые на ощупь шары. Каждый шар имеет свой цвет (или какой–то другой отличительный признак). Имеется n_1 шаров первого цвета, n_2 — второго, ..., n_r — r-го цвета; всего в урне $n = n_1 + \ldots + n_r$ шаров. Случайное испытание состоит в выборе наугад шара из урны с регистрацией в качестве исхода цвета шара. Если выбирается каждый раз по одному шару, то исход такого случайного испытания можно представить множеством элементарных исходов $\Omega = \{\omega_1, \omega_2, \ldots, \omega_r\}$, каждый элемент ω_i которого соответствует вытаскиванию из урны шара i-го цвета. Модель «урна» допускает и другие правила извлечения шаров, что делает ее весьма гибким инструментом интерпретации теоретико-вероятностных понятий (см. задачу I-7).

Модель «мишень». Имеется плоская мишень — квадрат единичной площади, левый нижний угол которого расположен в начале координат, а стороны совпадают с лучами, ограничивающими первый квадрант (рис. 1). В мишень бросается «точка», т. е. объект, размерами которого можно пренебречь.

Исходом испытания является попадание точки в одну из точек квадрата. Таким образом, элементарный исход такого испытания представляет собой

точку с координатами $\omega_t = (x_t, y_t)$, где $0 \leqslant x_t \leqslant 1$ и $0 \leqslant y_t \leqslant 1$, а само множество элементарных исходов Ω совпадает с этим квадратом. Модель «мишень» может иметь несколько модификаций, например, множество Ω может быть не квадратом, а произвольной частью плоскости, частью прямой линии, частью трехмерного пространства и т. д.

До сих пор случайные испытания мы описывали языком, «естественным» для исследователя, работающего в той или иной конкретной области. Этот язык обладает несомненными достоинствами наглядности и общедоступности. Однако названные достоинства затрудняют его использование при построении строгих математических схем. Поэтому, по крайней мере, на время мы должны пожертвовать наглядностью и обратиться к более формальному языку математики. Вспомним вначале элементы теории множеств, которые мы уже «незаметно» привлекали в ходе предшествующего изложения.

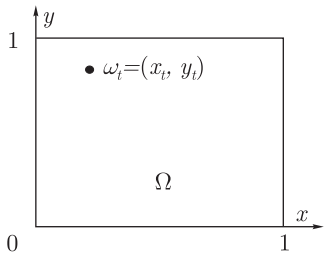


Рис. 1. Модель «мишень». Элементарными исходами случайного испытания являются все точки единичного квадрата Ω. Каждый элементарный исход представляет собой точку с координатами (x, y): $\omega_t = (x_t, y_t)$, $0 \leqslant x_t \leqslant 1$, $0 \leqslant y_t \leqslant 1$

§ 2. Элементы теории множеств

Исходное «универсальное» множество, с которым «работают» в теории вероятностей, есть совокупность Ω элементарных исходов ω_t, что обозначается при помощи записи $\Omega = \{\omega_t\}$. Тот факт, что элементарный исход ω_t является *элементом* множества Ω, записывается как $\omega_t \in \Omega$.

Используя только часть элементов множества Ω, можно образовывать другие, отличные от Ω, множества. Множество A будет считаться определенным, если для любого элемента $\omega \in \Omega$ можно сказать, принадлежит элемент ω множеству A ($\omega \in A$) или элемент ω не принадлежит множеству A. Случай непринадлежности элемента ω множеству A будем обозначать $\overline{\in}$.

Пример I-8. При бросании игральной кости универсальным множеством $\Omega = \{\omega_1, \ldots, \omega_6\}$ является множество, состоящее из чисел 1, 2, ..., 6. Подмножеством этого множества будет, например, множество $A = \{2, 4, 6\}$, соответствующее тому, что в результате случайного испытания выпало четное число очков. Здесь множество A определено прямым перечислением

входящих в него элементов: $2 \in A$, $4 \in A$, $6 \in A$. Остальные элементы не принадлежат A: $1 \overline{\in} A$, $3 \overline{\in} A$, $5 \overline{\in} A$.

Если множества Ω и A бесконечны, множество A нельзя задать прямым перечислением элементов. Тогда A считается заданным, если указано свойство элементов $\omega \in \Omega$, делающее их элементами множества A.

ПРИМЕР I-9. Отрезок $A = [0, 1]$, состоящий из бесконечного числа точек, определен как множество указанием, что он содержит все действительные числа, обладающие свойством $0 \leqslant \omega_t \leqslant 1$. Здесь роль Ω выполняет множество всех действительных чисел. И для любого $\omega_t \in \Omega$ мы можем указать, принадлежит ли этот элемент множеству $A = [0, 1]$, проверив для этого элемента выполнение указанного неравенства.

Если каждый элемент множества A является элементом множества B, то это соотношение условно записывается $A \subseteq B$ и читается «множество A содержится в множестве B» или «множество A есть *подмножество* множества B». Очевидно выполнение следующих свойств отношения $\subseteq$, которые делают его очень похожим на отношение обычного неравенства для чисел:

1) для любого множества A имеет место $A \subseteq A$;

2) если $A \subseteq B$ и $B \subseteq A$, то $A = B$;

3) если $A \subseteq B$ и $B \subseteq C$, то $A = C$.

Запись $A = C$ означает, что все элементы множеств A и B совпадают. В ситуации, когда $A \subseteq B$ и $A \neq B$, используется обозначение $A \subset B$.

ПРИМЕР I-10. При подбрасывании игральной кости множество четных чисел $A = \{2, 4, 6\}$ есть подмножество множеств $B_1 = \{1, 2, 3, 4, 5, 6\}$ и $B_2 = \{2, 3, 4, 6\}$. При этом можно видеть, что $A \subseteq B_1$, $B_2 \subseteq B_1$ и $A \subseteq B_2$.

Для любого множества A имеет место соотношение $A \subseteq \Omega$, т. е. все множества суть подмножества универсального множества Ω. Введем особое множество, обозначаемое далее знаком $\varnothing$, которое обладает противоположным, сравнительно с множеством Ω, свойством: для всех множеств A имеет место соотношение $\varnothing \subseteq A$. Поскольку множество $\varnothing$, обладающее таким свойством, не может содержать ни одного элемента, постольку множество $\varnothing$ получает наименование пустого множества. Заметим, что любое множество A «заключено» между пустым множеством $\varnothing$ и универсальным множеством Ω в смысле выполнения соотношения $\varnothing \subseteq A \subseteq \Omega$.

Пример I-11. Элементарные исходы при бросании монеты суть ω_1 — «герб» и ω_2 — «решка», а универсальное множество $\Omega = \{\omega_1, \omega_2\}$. Перечислим все подмножества множества Ω: $\varnothing, \{\omega_1\}, \{\omega_2\}$ и $\Omega = \{\omega_1, \omega_2\}$. Здесь $\{\omega_i\}$ — множество, состоящее из единственного элемента $\omega_i \subseteq \Omega$.

Необходимо различать элемент ω_i и *одноэлементное множество* $\{\omega_i\}$, так как они обладают совершенно разными теоретико-множественными свойствами и по-разному относятся к универсальному множеству Ω: ω_i принадлежит Ω как элемент, что записывается как $\omega_i \in \Omega$; $\{\omega_i\}$ входит в Ω как подмножество, что записывается $\{\omega_i\} \subseteq \Omega$.

По аналогии с функцией многих переменных $y = f(x_1, \ldots, x_n)$ в теории множеств вводятся операции $B = f(A_1, \ldots, A_n)$, ставящие в соответствие множествам $A_i, \ldots, A_n$ множество B.

Важной операцией является взятие дополнения $\overline{A}$ множества A до всего множества Ω. *Дополнение* $\overline{A}$ множества A есть множество, состоящее из всех элементов Ω, не входящих в множество A.

Пример I-12. При бросании игральной кости дополнением множества $A = \{2, 4, 6\}$ будет множество $B = \overline{A} = \{1, 3, 5\}$, что соответствует выпадению нечетного числа очков.

Вводятся операция объединения множеств A_1 и A_2, обозначаемая $B = A_1 \cup A_2$, и операция пересечения множеств A_1 и A_2, обозначаемая $B = A_1 \cap A_2$. *Объединение* $B = A_1 \cup A_2$ множеств A_1 и A_2 есть множество, состоящее из всех тех элементов множества Ω, которые входят, по крайней мере в одно из множеств A_1 или A_2. *Пересечение* $B = A_1 \cap A_2$ множеств A_1 и A_2 есть множество, состоящее из всех тех элементов множества Ω, которые одновременно входят в множество A_1 и в множество A_2.

Пример I-13. При бросании игральной кости объединением множеств $A_1 = \{2, 4\}$ и $A_2 = \{4, 6\}$ будет уже знакомое нам множество $B = A_1 \cup A_2 = \{2, 4, 6\}$, соответствующее выпадению четного числа очков, а пересечением этих множеств будет одноэлементное множество $B = A_1 \cap A_2 = \{4\}$, состоящее из одного элементарного исхода $4 \in \Omega$.

Операции объединения и пересечения множеств естественно обобщаются на случай любого числа множеств: $B = A_1 \cup A_2 \cup \ldots \cup A_n \cup \ldots$ и $B = A_1 \cap A_2 \cap \ldots \cap A_n \cap \ldots$. Если число множеств конечно и равно n, то их объединение (пересечение) обозначается $\cup_{i=1}^{n} A_i$ ($\cap_{i=1}^{n} A_i$); аналогично записывается объединение (пересечение) бесконечного числа множеств: $\cup_{i=1}^{\infty} A_i$ ($\cap_{i=1}^{\infty} A_i$); когда объединяться (пересекаться) может как конечное число множеств, так и бесконечное, тогда мы будем использовать обозначение $\cup_i A_i$ ($\cap_i A_i$).

В теории множеств обычно рассматриваются не изолированные множества, а их совокупности, в которых отдельные множества связаны между собой определенными операциями и соотношениями. Среди таких совокупностей особый интерес для теории вероятностей представляет так называемая алгебра множеств. Совокупность множеств $\mathscr{A}$ называют *алгеброй*, если для нее выполнены следующие условия:

1) $\Omega \in \mathscr{A}$ (принадлежность множества алгебре обозначается тем же символом, что и принадлежность элемента множеству);

2) $\varnothing \in \mathscr{A}$;

3) если $A, B \in \mathscr{A}$, то и $A \cup B \in \mathscr{A}$;

4) если $A, B \in \mathscr{A}$, то и $A \cap B \in \mathscr{A}$;

5) если $A \in \mathscr{A}$, то и $\overline{A} \in \mathscr{A}$.

Иными словами, алгебра $\mathscr{A}$ «замкнута» относительно операций объединения, пересечения и взятия дополнений множеств. Математик сказал бы, что это не просто алгебра, а сигма-алгебра, так как замкнутость $\mathscr{A}$ имеет место для бесконечных пересечений и объединений. Но мы позволим себе небольшую неточность, чтобы не делать терминологию слишком громоздкой.

§ 3. Случайное событие

Одним из важнейших понятий теории вероятностей является понятие случайного события. Теперь, используя результаты двух предыдущих параграфов, мы можем его определить.

Выше мы ввели $\Omega = \{\omega_t\}$ — универсальное множество элементарных исходов случайного испытания $\mathscr{E}$, рассмотрели подмножество $A \subseteq \Omega$ и задали алгебру $\mathscr{A}$. *Случайное событие* A есть такое подмножество A множества элементарных исходов Ω, которое одновременно является элементом алгебры $\mathscr{A}$. В связи с этим определением алгебру $\mathscr{A}$ в теории вероятностей называют *алгеброй случайных событий*.

ПРИМЕР I-14. Рассмотрим множество элементарных исходов случайного испытания, состоящего в бросании монеты $\Omega = \{\omega_1, \omega_2\}$. Множество всех подмножеств множества Ω естественно замкнуто относительно теоретико-множественных операций:

1) $\overline{\varnothing} = \Omega$, $\{\overline{\omega_1}\} = \{\omega_2\}$, $\{\overline{\omega_2}\} = \{\omega_1\}$, $\overline{\Omega} = \varnothing$;

2) $\varnothing \cup \{\omega_1\} = \{\omega_1\}$, $\varnothing \cup \{\omega_2\} = \{\omega_2\}$, $\varnothing \cup \Omega = \Omega$, $\{\omega_1\} \cup \{\omega_2\} = \Omega$, $\{\omega_1\} \cup \Omega = \Omega$, $\{\omega_2\} \cup \Omega = \Omega$;

3) $\varnothing \cap \{\omega_1\} = \varnothing$, $\varnothing \cap \{\omega_2\} = \varnothing$, $\varnothing \cap \Omega = \varnothing$, $\{\omega_1\} \cap \{\omega_2\} = \varnothing$, $\{\omega_1\} \cap \Omega = \{\omega_1\}$, $\{\omega_2\} \cap \Omega = \{\omega_2\}$.

Следовательно, множество всех подмножеств множества элементарных исходов Ω является алгеброй случайных событий $\mathscr{A}$. Поэтому подмножества $\varnothing, \{\omega_1\}, \{\omega_2\}, \Omega = \{\omega_1, \omega_2\}$ суть случайные события. Однако множество всех подмножеств Ω отнюдь не единственная система, образующая алгебру $\mathscr{A}$.

Пример I-15. Рассмотрим множество элементарных исходов случайного испытания, состоящего в бросании игральной кости $\Omega = \{\omega_1, \ldots, \omega_6\}$. Предложим в качестве случайных событий следующую совокупность подмножеств множества Ω: $\varnothing, A = \{1, 3, 5\}, B = \{2, 4, 6\}, \Omega$. С помощью прямого перечисления, как это было сделано в предыдущем примере, нетрудно убедиться в замкнутости данной совокупности относительно операций объединения, пересечения и дополнения множеств, т. е. она образует алгебру $\mathscr{A}$, и, следовательно, $\varnothing, A, B, \Omega$ можно рассматривать в качестве случайных событий. Такая совокупность случайных событий является естественной математической моделью случайного испытания, когда нас интересует лишь выпадение четного или нечетного числа очков.

Пример I-16. При определении содержания ДДТ в траве для установления пригодности поля в качестве пастбища важно не само измеренное значение, а ответ на вопрос: превышает ли оно допустимый санитарный уровень C? Иными словами, интерес представляет не какой-то элементарный исход $\{\omega_t\}$, но случайное событие $\{\omega_t < C\}$.

Таким образом, необязательное совпадение совокупности (алгебры) случайных событий $\mathscr{A}$ с множеством всех подмножеств Ω дает в руки исследователя гибкий аппарат описания исходов случайного испытания, позволяющий приспосабливать этот аппарат к решению конкретных задач.

Введем следующий принцип осуществления событий: случайное событие $A \subseteq \Omega$ происходит тогда и только тогда, когда при случайном испытании имеет место элементарный исход ω_t, входящий в множество A, т. е. $\omega_t \in A$ (рис. 2*а*).

Про элементарные исходы, входящие в множество A, говорят, что они благоприятствуют событию A.

Заметим, что каждому элементарному исходу $\omega_i \subseteq \Omega$ можно сопоставить событие $A_i \subseteq \Omega$, представляющее собой одноэлементное множество A_i: $\{\omega_i\} \subseteq \Omega$. Такие одноэлементные случайные события называют

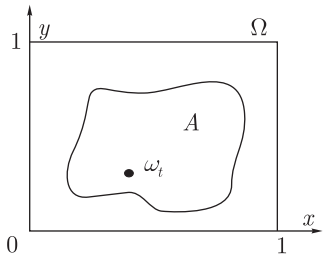


Рис. 2*а*

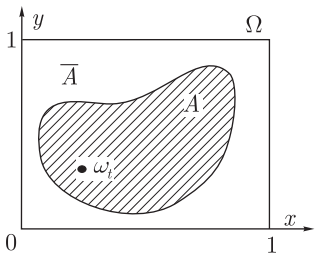


Рис. 2*б*

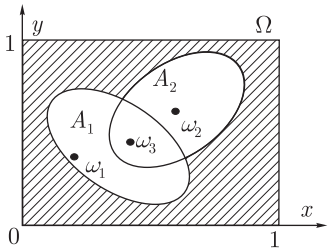


Рис. 2*в*

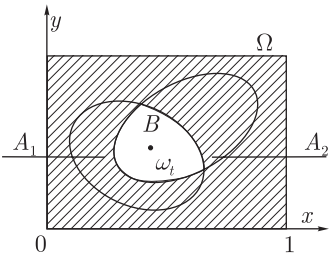


Рис. 2*г*

элементарными случайными событиями. Таким образом, мы получаем возможность интерпретировать элементарные исходы $\omega_i \in \Omega$ как элементарные события $\{\omega_i\} \subseteq \Omega$.

Опишем теперь в теоретико-вероятностных терминах операции над событиями.

Случайное событие $B = A$ происходит тогда и только тогда, когда не происходит событие A (рис. 2*б*).

Случайное событие $B = A_1 \cup A_2$ происходит тогда и только тогда, когда происходит одно из событий A_1 и A_2 или оба они происходят вместе (рис. 2*в*). Аналогично, случайное событие $B = \cup_i A_i$ происходит тогда и только тогда, когда происходит, по крайней мере, одно из событий $A_1, \ldots, A_k, \ldots$.

Случайное событие $B = A_1 \cap A_2$ происходит тогда и только тогда, когда происходят одновременно события A_1 и A_2 (рис. 2*г*). Аналогично, событие $B = \cap_i A_i$ происходит тогда и только тогда, когда все события $A_1, \ldots, A_k, \ldots$ происходят совместно.

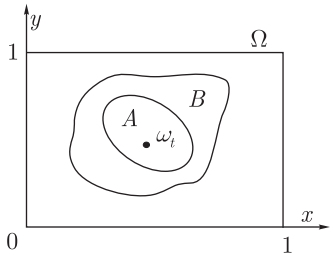


Рис. 2*д*. Случайные события и операции над ними:

а — событие, состоящее в попадании в множество A, осуществляется тогда и только тогда, когда элементарный исход ω_t случайного испытания есть элемент множества A: $\omega_t \in A$;

б — событие, состоящее в попадании в множество $\overline{A}$ (в незаштрихованную часть множества Ω), происходит тогда и только тогда, когда не происходит событие A, состоящее в попадании в множество A;

в — событие, состоящее в попадании в множество $B = A_1 \cup A_2$ (в незаштрихованную часть множества Ω), происходит тогда и только тогда, когда или происходит событие, состоящее в попадании в множество A_1 (исход ω_1), или событие, состоящее в попадании в множество A_2 (исход ω_2), или событие, состоящее в попадании в множество A_1 и в множество A_2 (исход ω_3);

г — событие, состоящее в попадании в множество $B = A_1 \cap A_2$ (в незаштрихованную часть множества Ω), происходит тогда и только тогда, когда одновременно происходят событие, состоящее в попадании в множество A_1, и событие, состоящее в попадании в множество A_2 (исход ω_t);

д — из осуществления события, состоящего в попадании в множество A (исход ω_t), следует осуществление события, состоящего в попадании в множество B, содержащее множество A: $A \subseteq B$

Если $A_1 \cap A_2 = \varnothing$, то эти события называются *несовместными*. Например, событие A и его дополнение $\overline{A}$ несовместны.

Если произошло событие A и имеет место соотношение $A \subseteq B$, то произошло и событие B (рис. 2*д*). Поэтому, когда имеет место соотношение $A \subseteq B$, мы будем говорить, что событие A влечет событие B. Понятно, что событие Ω следует из любого элементарного события $\{\omega_i\}$, которое происходит при появлении элементарного исхода ω_i. Иными словами, со-

бытие Ω осуществляется при любом исходе случайного испытания и может быть названо *достоверным событием*. Событие $\varnothing$ не следует ни из одного элементарного события, т. е. не осуществляется ни при каком исходе случайного испытания. Это позволяет назвать событие $\varnothing$ *невозможным событием*.

§ 4. Частота случайного события

Случайное испытание $\mathscr{E}$ повторим n раз $(\mathscr{E}_1, \mathscr{E}_2, \ldots, \mathscr{E}_n)$, регистрируя каждый раз наступление некоторого случайного события A из алгебры событий $\mathscr{A}$. Величина $h(A; n; \mathscr{E}_1, \ldots, \mathscr{E}_n)$, равная частному от деления числа испытаний n_A, в которых произошло событие A, на общее число испытаний n, называется *относительной частотой*, или просто *частотой*, события A в данной серии из n реализаций случайного испытания $\mathscr{E}$. Для простоты обозначение $h(A; n; \mathscr{E}_1, \ldots, \mathscr{E}_n)$ мы будем заменять на $h(A; n)$.

Очевидны следующие свойства частоты, верные для любой серии испытаний (см. также задачу I-4):

1) для любого события $A \in \mathscr{A}$ имеет место неравенство $0 \leqslant h(A; n) \leqslant 1$;

2) если $A = \Omega$, то $h(A; n) = 1$; однако обратное неверно: если частота $h(A; n) = 1$, это не означает, что A есть событие достоверное;

3) если $A = \varnothing$, то $h(A; n) = 0$; обратное неверно: если частота $h(A; n) = 0$, то это не означает, что A — невозможное событие;

4) если любые два события A_i, A_j из группы событий $A_1, \ldots, A_k$ попарно несовместны $(A_i \cap A_j = \varnothing, i \neq j)$, то

$$h(\cup_{i=1}^{k} A_i; n) = \sum_{i=1}^{k} h(A; n).$$

Повторяя серии из n испытаний, мы будем получать в разных сериях, вообще говоря, разные частоты события A. В самом деле, кого удивит, что в одной серии из трех бросаний монеты «герб» выпадает два раза, а в другой — один раз? Частота события A, состоящего в выпадении «герба», в первой серии будет равна $h_1(A; 3) = 2/3$, а во второй — $h_2(A; 3) = 1/3$. Однако для достаточно обширного класса случайных испытаний частоты событий при увеличении длины серии имеют тенденцию стабилизироваться около

определенного числового значения $h(A)$ и тем самым становиться независимыми от конкретной серии испытаний. Таким образом, частота события A становится характеристикой самого случайного события $A \in \mathscr{A}$ безотносительно к конкретной реализации случайного испытания $\mathscr{E}$. Случайное испытание, состоящее в бросании монеты, — это типичный пример стабилизации частот: для «правильной» монеты неоднократно устанавливалось, что в длинных сериях частота появления «герба» мало отличается от 1/2.

Так, например, Ж. Бюффон в 1777 г. получил $h(A; 4\,040)=0,5069$; А. де Морган (начало XIX в.) — $h(A; 4\,092)=0,5005$; У. Джевонс (1887 г.) — $h_1(A; 10\,240)=0,5036$ и $h_2(A; 10\,240)=0,5100$; В. И. Романовский (1912 г.) — $h(A; 80\,640) = 0,4923$; К. Пирсон (1926 г.) — $h_1(A; 12\,000) = 0,5016$ и $h_2(A; 24\,000)=0,5005$; Дж. Керрих (1946 г.) — $h(A; 10\,000)=0,5067$ (см. также рис. 3).

Подобными «генераторами» стабильности частот являются бросание игральной кости, вращение рулетки и т. п.

Многие биологические процессы сопровождаются стабилизацией частот некоторых событий: довольно стабильны частота появления новорожденных определенного пола, частота различных заболеваний, частота вступления в брак в разных возрастных категориях, частота появления определенных генотипов в определенных скрещиваниях и т. п. Физические процессы также служат обильным источником примеров генераторов стабильных частот: попадания космических частиц в счетчик Гейгера–Мюллера; появления флуктуации заданной величины во всех процессах, изучаемых статистической физикой; попадания капель дождя на некоторую поверхность и т. д.

Выбранные нами модели случайных испытаний — модель «урна» и модель «мишень» — одновременно могут служить и генераторами стабильных частот. Многочисленные эксперименты, состоящие в извлечении (наугад, с последующим возвращением и тщательным перемешиванием) шара из урны, показывают, что частота появления шаров данного цвета устойчива и колеблется около числа, равного доле шаров этого цвета в урне. Например, А. Кетле (1846 г.) вынимал 4 096 раз с возвращением шар из урны, содержащей 20 белых и 20 черных шаров, и получил $h(A; 4\,096) = 0,5044$. Если же наугад бросать точку в мишень, то частота попадания точки в какую-либо часть мишени стабилизируется около величины, пропорциональной площади этой части. Проведите эксперимент: выставьте несколько раз на моросящий дождик плоскую фанеру с очерченной на ней произвольной фигурой и подсчитайте частоту попаданий дождинок в эту фигуру. Дождик должен быть слабым, а экспозиция — недолгой, чтобы следы от капелек не сливались. Практическая работа статистика (в частности, биометрика) тре-

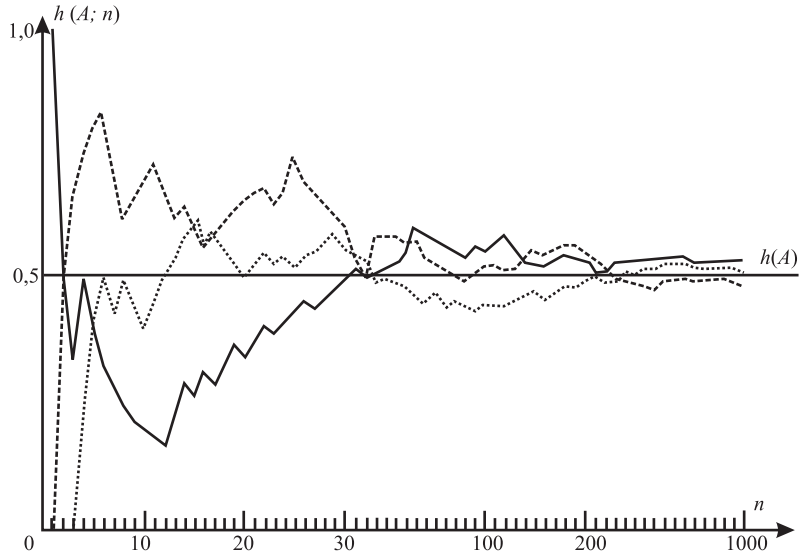


Рис. 3. Результаты трех серий экспериментов с монетой (до 1000 бросаний в каждой серии); $h(A;n)$ — частота выпадения «герба»

бует простого и доступного генератора стабильных частот; эту роль играет таблица случайных чисел (табл. I Приложения 1, см. гл. IV, § 3).

Об эмпирической стабилизации частоты события A мы будем говорить только в случае выполнения следующих требований:

1) проведено достаточно много достаточно длинных серий реализаций испытания $\mathscr{E}$;

2) среди всех серий реализаций случайного испытания $\mathscr{E}$ доля тех серий, в которых частота $h(A;n)$ события A отличается от числа $h(A)$ больше, чем на некоторое достаточно малое число $\epsilon > 0$, пренебрежимо мала.

Далее, говоря о случайных испытаниях, мы всегда будем предполагать стабильность частот всех событий, входящих в алгебру случайных событий $\mathscr{A}$.

В приведенном описании стабилизации частоты события A специально подчеркнута неопределенность: «достаточно много» серий, «достаточная

длина» серии, «достаточно малое число $\epsilon > 0$», «пренебрежимо малая» доля. В каждой конкретной сфере исследования указанная неопределенность получает свою количественную оценку исходя из специфики требований к точности исследования и из возможностей исследователя.

В заключение параграфа резюмируем все ограничения, которые были наложены на употребление слов «случайное испытание $\mathscr{E}$»:

1) в случайном испытании определено множество элементарных исходов Ω;

2) зафиксирована алгебра случайных событий $\mathscr{A}$, состоящая из подмножеств множества Ω;

3) для каждого случайного события A из алгебры случайных событий $\mathscr{A}$ определена стабильная частота $h(A)$, зависящая только от самого испытания $\mathscr{E}$ и от события A, но не от конкретной серии реализаций случайного испытания $\mathscr{E}$; тем самым из очень широкого класса недетерминированных испытаний выделена сравнительно узкая область явлений, которые мы и называем случайным испытанием.

§ 5. Вероятность случайного события

Развивая представления о случайных испытаниях, мы начали с обсуждения некоторых экспериментальных схем и ввели понятие множества элементарных исходов Ω (§ 1). Необходимость обобщения этих представлений потребовала прибегнуть к аппарату теории множеств (§ 2) и ввести понятие алгебры $\mathscr{A}$ и случайного события A (§ 3). В рамках конкретных экспериментальных ситуаций мы рассмотрели затем важнейшую характеристику случайного испытания — частоту случайного события $h(A)$ (§ 4). Теперь необходим следующий этап построения математического аппарата: на смену понятию частоты мы должны ввести строго определенное, аксиоматическое понятие вероятности случайного события $P(A)$. При этом понятие и свойства вероятности естественно формулировать таким образом, чтобы они практически совпадали с понятием и свойством частоты, т. е. были бы приложимы к анализу широкого класса экспериментов. Введем аксиоматику теории вероятностей, предложенную замечательным отечественным математиком А. Н. Колмогоровым в 1933 г.

Вероятностью случайного события A называется функция $P(A)$, заданная на алгебре случайных событий $\mathscr{A}$ и обладающая следующими свойствами

АКСИОМА 1. Вероятность любого случайного события есть неотрицательное число: $P(A) \geqslant 0$.

АКСИОМА 2. Вероятность достоверного события равна единице: $P(\Omega) = 1$.

АКСИОМА 3. Если события $A_1, \ldots, A_k, \ldots$ попарно несовместны ($A_i \cap A_j = \varnothing$, $i \neq j$), то вероятность объединения этих событий равна сумме их вероятностей:

$$P(\cup_i A_i) = \sum_i P(A_i).$$

Тройка математических объектов $(\Omega, \mathscr{A}, P)$, сопоставляемых случайному испытанию $\mathscr{E}$, называется *вероятностным пространством*. Теперь, с появлением понятия вероятностного пространства, можно приступить к построению теории вероятностей: вероятностное пространство $(\Omega, \mathscr{A}, P)$ есть математический образ реального случайного испытания $\mathscr{E}$ и далее будет отождествляться с этим испытанием.

Еще раз подчеркнем, что математическая схема $(\Omega, \mathscr{A}, P)$ извлекает из реального случайного эксперимента $\mathscr{E}$ лишь наличие множества возможных исходов и стабильных частот, оставляя за своими рамками «непредсказуемость» исхода конкретной реализации случайного испытания.

Рассмотрим простейшие следствия из принятых аксиом.

1. Вероятность $P(\overline{A})$ — дополнительного события $\overline{A}$ к событию A — определяется формулой
$$P(\overline{A}) = 1 - P(A).$$

   Доказательство:

   Рассмотрим некоторое событие $A \in \mathscr{A}$. Из замкнутости $\mathscr{A}$ относительно операции взятия дополнения следует, что $\overline{A} \in \mathscr{A}$. Поскольку A, $\overline{A} \in \mathscr{A}$, то по замкнутости $\mathscr{A}$ относительно объединения имеем $A \cup \overline{A} \in \mathscr{A}$. Следовательно, определены вероятности событий $P(A)$, $P(\overline{A})$, $P(A \cup \overline{A})$. Так как $A \cup \overline{A} = \varnothing$ (см. § 3), то по аксиоме 3 $P(A \cup \overline{A}) = P(A) + P(\overline{A})$. Однако $A \cup \overline{A} = \Omega$ и $P(\Omega) = 1$. Следовательно, $1 = P(A) + P(\overline{A})$, откуда имеем искомый результат. ■

2. Вероятность невозможного события равна нулю: $P(\varnothing) = 0$.

3. Если $A \subseteq B$, то $P(A) \leqslant P(B)$.

4. Вероятность любого события заключена между нулем и единицей: $0 \leqslant \leqslant P(A) \leqslant 1$.

5. Для любых событий A и B имеет место соотношение

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Проведем подробное доказательство этой важной формулы. Очевидны два следующих соотношения: $A \cup B = B \cup (A \cap \overline{B})$, $A = (A \cap \overline{B}) \cup (A \cap B)$, в которых правые части имеют вид объединения несовместных событий. Несовместность событий в правых частях позволяет применить к обоим соотношениям аксиому 3, что дает два уравнения: $P(A \cup B) = P(B) + + P(A \cap \overline{B})$ и $P(A) = P(A \cap \overline{B}) + P(A \cap B)$. Вычитая из первого уравнения второе и приводя подобные члены, получим искомую формулу, которая носит название *формулы сложения вероятностей*.

§ 6. Условная вероятность и независимость событий

Пусть $(\Omega, \mathscr{A}, P)$ — вероятностное пространство, описывающее случайное испытание $\mathscr{E}$. Выберем из алгебры событий $\mathscr{A}$ некоторое событие B, имеющее ненулевую вероятность: $P(B) > 0$. Рассмотрим функцию $P(A/B)$, определенную для любого события $A \in \mathscr{A}$ соотношением

$$P(A/B) = \frac{P(A \cap B)}{P(B)}.$$

Можно показать, что $P(A/B)$ удовлетворяет аксиомам теории вероятностей, т. е. является вероятностью (задача I-5).

Выясним содержательный смысл новой вероятности $P(A/B)$, полученной пока формальным путем. Обратимся с этой целью к модели «мишень» (см. рис. 4) и предположим, что произошло событие B, т. е. наугад брошенная точка попала в множество B. Как влияет реализация события B на вероятность появления события A? Когда событие B имеет с событием A непустое пересечение $(A \cap B \neq \varnothing)$ и одновременно не полностью содержится в A $(B \cap \overline{A} \neq \varnothing)$, тогда появление события B не определяет однозначно реализацию события A: если элементарным исходом испытания была точка $\omega_1 \in A \cap B$, то событие A произошло, а если таким исходом была точка $\omega_2 \in \overline{A} \cap B$, то событие A не произошло. Информация о появлении события B сужает множество возможных элементарных исходов случайного испытания: вместо множества Ω теперь достоверным событием выступает множество $\Omega' = B$.

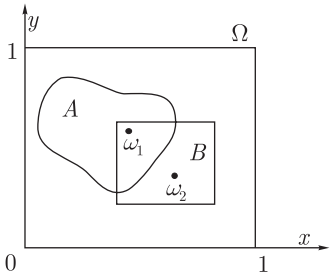


Рис. 4. Вероятность попадания наугад брошенной точки в множество A, когда известно, что эта точка попала в множество B, равна отношению площади $A' = A \cap B$ к площади $\Omega' = B$. Эта вероятность и есть условная вероятность $P(A/B)$ события A при реализации события B

Аналогично отсекаются элементарные исходы, принадлежащие множеству $A \cap \overline{B}$, и теперь событию A благоприятствуют не все элементарные исходы, составляющие A, а лишь те, которые входят в пересечение $A' = A \cap B$. Поскольку в модели «мишень» устойчивость частоты событий, а следовательно, и их вероятности определяются долей площади $S(A)$ соответствующего множества A от площади $S(\Omega)$ всего множества элементарных исходов, то появление события B изменяет вероятность $P(A) = S(A)/S(\Omega')$ на новую величину $P'(A) = S(A')/S(\Omega') = S(A \cap B)/S(B)$. Поделив числитель и знаменатель последней дроби на $S(\Omega)$, получаем для вероятности появления события A при условии реализации события B выражение $P'(A) = P'(A) = P(A \cap B)/P(B)$.

Приведенные выше соображения позволяют, таким образом, интерпретировать $P(A/B)$ как *условную вероятность* события A при реализации B. Из определения условной вероятности сразу же следует *формула умножения вероятностей*:

$$P(A \cap B) = P(A/B) \cdot P(B).$$

Нельзя забывать, что формула умножения вероятностей определена лишь тогда, когда $P(B) > 0$.

Из полученной формулы следует также, что

$$P(A \cap B) = P(B/A) \cdot P(A),$$

при условии, что $P(A) > 0$.

Пусть $P(A) > 0$, $P(B) > 0$. Будем говорить, что вероятность события B не зависит от реализации события A, если условная вероятность события A при реализации события B совпадает с «безусловной» вероятностью события A: $P(A/B) = P(A)$. Нетрудно показать, что в таком случае верно и соотношение $P(B/A) = P(B)$, т. е. и вероятность события B не

зависит от реализации события A. Поэтому мы будем называть A и B *независимыми*, если выполнены эти соотношения, эквивалентные соотношению

$$P(A \cap B) = P(A) \cdot P(B).$$

Заметим, что независимость и несовместность событий A, B, имеющих положительную вероятность, являются несовместимыми понятиями. Допустим противное, полагая события A, B одновременно несовместными $A \cap B = \varnothing$ и независимыми. Однако тогда получаем соотношение $P(\varnothing) = P(A) \cdot P(B)$, которое ложно, так как левая его часть равна нулю, а правая — строго больше нуля. Полученное противоречие и доказывает исходное утверждение.

Понятие независимости можно сформулировать и для группы событий $A_1, \ldots, A_n$, но здесь условие независимости может быть модифицировано различными способами, что дает разные виды независимости, три из которых мы рассмотрим.

1. События $A_1, \ldots, A_n$ называются *попарно независимыми*, если любые два события A_i, A_j, $i \neq j$, независимы, т. е. если для любых A_i, A_j, $i \neq \neq j$, выполнено соотношение

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j).$$

2. События $A_1, \ldots, A_n$ называются *совместно независимыми*, если выполнено соотношение

$$P(\cap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i).$$

3. События $A_1, \ldots, A_n$ называются *независимыми в совокупности*, если для любого набора различных индексов $i_1, \ldots, i_k$, $1 \leqslant i_s \leqslant n$, $2 \leqslant k \leqslant \leqslant n$, имеет место соотношение

$$P(\cap_{s=1}^{n} A_{i_s}) = \prod_{s=1}^{k} P(A_{i_s}).$$

Другими словами, независимость в совокупности означает независимость двоек событий (попарную независимость), троек, четверок и т. д. до n событий.

Совместная и попарная независимости не следуют друг из друга и не влекут независимости в совокупности. Это утверждение доказывается следующими двумя примерами.

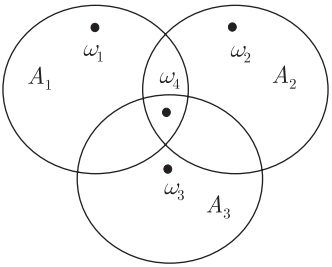


Рис. 5. События A_1, A_2, A_3 попарно независимы, но не являются ни совместно независимыми, ни независимыми в совокупности (пример I-17)

ПРИМЕР I-17. Имеется вероятностное пространство $(\Omega, \mathscr{A}/P)$, где множество элементарных исходов состоит из четырех элементов: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ (рис. 5). Каждому элементарному событию $\{\omega_i\}$ приписывается одинаковая вероятность $P(\{\omega_i\}) = 1/4$. Рассмотрим события $A_1 = \{\omega_1, \omega_4\}$, $A_2 = \{\omega_2, \omega_4\}$, $A_3 = \{\omega_3, \omega_4\}$, взятые из алгебры событий $\mathscr{A}$, включающей в себя все подмножества Ω (см. рис. 5).

Используя аксиому 3, получаем вероятности $P(A_1)=P(A_2)=P(A_3) = 1/2$. Очевидны следующие соотношения: $P(A_1 \cap A_2) = 1/4 = P(A_1) \times P(A_2)$, $P(A_1 \cap A_3) = 1/4 = P(A_1) \cdot P(A_3)$, $P(A_2 \cap A_3) = 1/4 = P(A_2) \cdot P(A_3)$, показывающие наличие попарной независимости событий A_1, A_2, A_3. Одновременно имеет место соотношение $P(A_1 \cap A_2 \cap A_3) = P(\{\omega_4\}) = 1/4 \neq P(A_1) \cdot P(A_2) \times \times P(A_3)$, указывающее на отсутствие совместной независимости и независимости в совокупности. Итак, попарная независимость не влечет ни независимости в совокупности, ни совместной независимости.

ПРИМЕР I-18. Аналогично предыдущему примеру рассмотрим вероятностное пространство $(\Omega, \mathscr{A}, P)$: $\Omega = \{\omega_1, \ldots, \omega_8\}$, $P(\{\omega_i\}) = 1/8$, $A_1 = \{\omega_1, \omega_2, \omega_3, \omega_8\}$, $A_2 = \{\omega_2, \omega_4, \omega_5, \omega_8\}$, $A_3 = \{\omega_5, \omega_6, \omega_7, \omega_8\}$ (рис. 6).

Имеем $P(A_1) = P(A_2) = P(A_3) = 1/2$; $P(A_1 \cap A_2 \cap A_3) = P(\{\omega_8\}) = 1/8 = P(A_1) \cdot P(A_2) \cdot P(A_3)$, следовательно, события A_1, A_2, A_3 совместно независимы. Но $P(A_1 \cap A_3) = P(\{\omega_8\}) = 1/8 \neq P(A_1) \cdot P(A_3)$, что указывает на отсутствие попарной независимости и независимости в совокупности. Итак, совместная независимость не влечет ни независимости в совокупности, ни попарной независимости.

Поясним выявленные связи между видами независимости на примере трех событий A, B, C. Их совместная независимость эквивалентна выполнению соотношения НС, попарная независимость — соотношений НП, а независимость в совокупности (НВС) — выполнению всех этих соотно-

шений одновременно:

$$\text{НВС}\left\{\begin{array}{l} P(A\cap B\cap C)=P(A)\cdot P(B)\cdot P(C) \\ \left.\begin{array}{l} P(A\cap B)=P(A)\cdot P(B) \\ P(A\cap C)=P(A)\cdot P(C) \\ P(B\cap C)=P(B)\cdot P(C) \end{array}\right\} \end{array}\right. \begin{array}{l} - \text{ НС} \\ \\ - \text{ НП.} \end{array}$$

Для решения задач полезно знать о легко показываемой эквивалентности следующих четырех утверждений:

1) события A, B независимы;

2) события $\overline{A}$, B независимы;

3) события A, $\overline{B}$ независимы;

4) события $\overline{A}$, $\overline{B}$ независимы.

§ 7. Классическое определение вероятности

Математическая модель $(\Omega, \mathscr{A}, P)$ случайного испытания Ω носит довольно общий и абстрактный характер: множество элементарных исходов может быть произвольным множеством с элементами любой природы; нет никаких содержательных ограничений на задание алгебры $\mathscr{A}$ и не конкретизирован способ задания вероятности P. Однако зачастую мы обладаем дополнительной информацией, позволяющей конкретизировать абстрактные объекты Ω, $\mathscr{A}$, P. С этой точки зрения рассмотрим два подхода к определению понятия вероятности, исторически предшествовавших аксиоматическому определению.

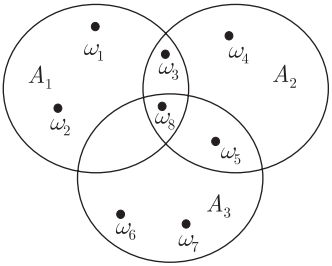


Рис. 6. События A_1, A_2, A_3 совместно независимы, но не являются независимыми ни в совокупности, ни попарно (пример I-18)

Возникшее в XVII в. в работах П. Ферма, Б. Паскаля, Х. Гюйгенса и Я. Бернулли классическое определение вероятности получается как частный случай аксиоматического, когда введены два ограничения:

1) множество элементарных исходов Ω конечно: $\Omega = \{\omega_1, \ldots, \omega_n\}$;

2) все элементарные события $\{\omega_1\}, \ldots, \{\omega_n\}$ равновероятны: $P(\{\omega_i\}) = P(\{\omega_j\})$.

Алгебра $\mathscr{A}$ совпадает с множеством всех подмножеств множества Ω. Тогда вероятность любого случайного события A, состоящего из элементарных исходов $\omega_{i_1}, \ldots, \omega_{i_m}$, вычисляется по формуле:

$$P(A) = \sum_{s=1}^{m} P(\{\omega_{i_s}\}) = \frac{n_A}{n}.$$

Классическое определение вероятности есть не что иное, как формула, записанная в словесной форме: вероятность любого события $A \in \mathscr{A}$ есть отношение числа n_A элементарных исходов, благоприятствующих данному событию A, к общему числу n равновероятных элементарных событий.

В рамках классической схемы вычисление вероятностей сводится к подсчету комбинаций элементарных исходов, благоприятствующих событию, и подсчету общего числа равновероятных элементарных событий. Очевидно, что классическому определению вероятности соответствует вероятностное пространство, описываемое моделью «урна».

ПРИМЕР I-19. При бросании «правильной» монеты мы принимаем равновероятным выпадение «герба» и «решки», т. е. $P(\text{«герб»}) = P(\text{«решка»}) = \frac{1}{2}$.

ПРИМЕР I-20. При бросании «правильной» игральной кости мы принимаем равновероятными все элементарные исходы, т. е. $P(1) = P(2) = \ldots = P(6) = \frac{1}{6}$. Тогда вероятность, скажем, события A, заключающегося в выпадении цифры $\leqslant 5$, равна $P(A) = \frac{5}{6}$, так как из общего числа $n = 6$ равновероятных исходов $n_A = 5$ благоприятствуют событию A.

ПРИМЕР I-21. При скрещивании гетерозигот Aa мы предполагаем равновероятность образования четырех типов зигот (AA, Aa, aA, aa), считая разными элементарными исходами появление потомка Aa от слияния:

1) яйцеклетки A и спермия a;

2) яйцеклетки a и спермия A.

Тогда вероятность события B, заключающегося в появлении потомков с доминантным признаком, равна $P(B) = \frac{3}{4}$.

§ 8. Геометрическое определение вероятности

Рассмотрим теперь вероятностное пространство, соответствующее модели «мишень». Введем следующие ограничения на вероятностное пространство $(\Omega, \mathscr{A}, P)$:

1) множество элементарных исходов Ω есть ограниченная фигура евклидовой плоскости, имеющая конечную площадь $S(\Omega) < +\infty$;

2) алгебра событий $\mathscr{A}$ состоит из тех подмножеств A множества Ω, для которых можно определить площадь $S(A)$;

3) вероятность $P(A)$ события $A \in \mathscr{A}$ пропорциональна площади множества A: $P(A) = k \cdot S(A)$, где k — некоторый положительный коэффициент пропорциональности.

При таких ограничениях нетрудно получить явное задание вероятности $P(A)$ события $A \in \mathscr{A}$: поскольку $P(\Omega) = 1$ и $P(\Omega) = k \cdot S(\Omega)$, то $k = \dfrac{1}{S(\Omega)}$. В результате получаем формулу, называемую *геометрическим определением вероятности*:

$$P(A) = \frac{S(A)}{S(\Omega)}.$$

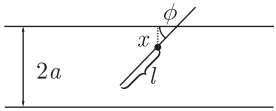


Рис. 7. Случайное испытание «игла Бюффона» состоит в бросании наугад иглы длиной $2l$ на плоскость, расчерченную параллельными прямыми, расстояние между которыми равно $2a$, где $l < a$

Очевидны видоизменения, которые надо ввести в геометрическое определение вероятности, чтобы оно стало применимым к случаю одномерной мишени (отрезок прямой) и к случаям мишеней, являющихся подмножествами пространств, имеющих более двух измерений.

Подчеркнем особо, что пространство, подмножеством которого является множество элементарных исходов Ω, не обязательно совпадает с реальным физическим пространством, в котором происходит случайное испытание; напротив, пространство может быть образовано любыми параметрами наблюдаемого объекта. Продемонстрируем это на классическом примере случайного испытания, носящего название «игла Бюффона».

ПРИМЕР I-22. Случайное испытание состоит в бросании наугад иглы длиной $2l$ на плоскость, расчерченную прямыми, расстояния между которыми равны $2a$ ($l < a$). Какова вероятность того, что упавшая игла пересечет одну из прямых?

Чтобы использовать геометрическое определение вероятности, параметризуем наш случайный эксперимент, выбрав в качестве параметров расстояние x от центра иглы до ближайшей прямой и угол ϕ между иглой и этой прямой (см. рис. 7). Иными словами, мы сопоставили случайному испытанию двумерное пространство параметров (ϕ, x), где $0 \leqslant \phi \leqslant \pi$, $0 \leqslant x \leqslant a$. Любой исход случайного бросания иглы на плоскость представляет собой точку внутри соответствующего прямоугольника с основанием n и высотой a (рис. 8).

Итак, исходному случайному испытанию соответствует модель «мишень», в которой множество элементарных исходов есть прямоугольник $\Omega = \{(\phi, x)\}$: $0 \leqslant \phi \leqslant \pi$, $0 \leqslant x \leqslant a$, а вероятность попадания в подмножество A множества Ω пропорциональна площади этого подмножества. Заметим, что последние слова о вероятности попадания в подмножество множества элементарных исходов и есть формализация употреблявшегося ранее выражения «игла бросается наугад». В множестве Ω можно указать множество исходов A (незаштрихованная часть множества Ω на рис. 8), благоприятствующих событию, состоящему в пересечении иглой линии: $\Omega = \{(\phi, x)\colon x \leqslant l \sin\phi\}$. Чтобы определить вероятность события A, остается найти площадь всей мишени Ω: $S(\Omega) = \pi a$, площадь множества A: $S(A) = \int_0^\pi l \sin\phi d\phi = 2l$ и поделить вторую площадь на первую: $P(A) = \frac{2l}{\pi a}$.

Отметим, что различные варианты и обобщения задачи Бюффона находят сейчас широкое применение в разных областях биологии, в частности, при исследовании трехмерной структуры объектов, когда известны только их сечения или проекции на плоскость[1].

§ 9. Последовательность случайных испытаний

Рассмотрим еще один важный частный случай вероятностного пространства $(\Omega, \mathscr{A}, P)$, который позволит нам ввести понятие независимых испытаний.

Обратимся к модели «мишень». Как мы уже знаем, множество элементарных исходов Ω в этом случае есть совокупность точек плоскости,

[1] См.: Число и мысль. — М.: Знание, 1977. — С. 42–48.

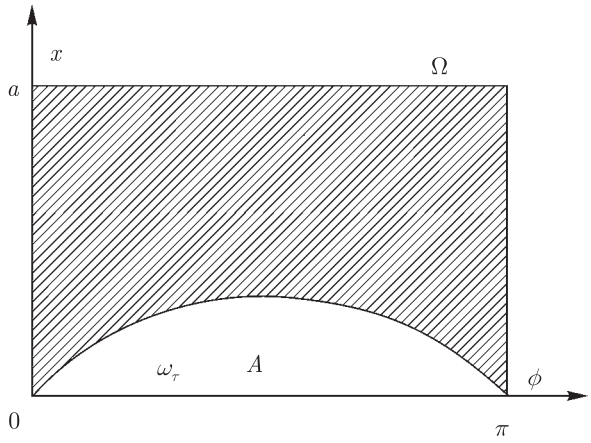


Рис. 8. Множество элементарных исходов случайного испытания, состоящего в бросании наугад иглы на расчерченную параллельными прямыми плоскость $\Omega = \{(\phi, x)\}$: $0 \leqslant \phi \leqslant \pi, \quad 0 \leqslant x \leqslant a$

попадающих в единичный квадрат. Элементарный исход ω можно представить в виде двумерного вектора $\omega = (\omega^{(1)}, \omega^{(2)})$, первая координата которого есть число $\omega^{(1)}$, взятое из множества $\Omega^{(1)}$, представляющего собой отрезок $[0, 1]$ горизонтальной оси, вторая координата — число $\omega^{(2)}$, взятое из множества $\Omega^{(2)}$, представляющего собой отрезок $[0, 1]$ вертикальной оси (См. рис. 9).

Выделим из алгебры $\mathscr{A}$, соответствующей универсальному множеству Ω, алгебру $\widehat{\mathscr{A}^{(1)}}$, включающую множества $\widehat{A}^{(1)}$ двумерных векторов $\omega = (\omega^{(1)}, \omega^{(2)})$, у которых первая координата принадлежит некоторому подмножеству $A^{(1)} \subseteq \Omega^{(1)}$, а вторая координата может принимать любые значения из множества $\Omega^{(2)}$. На рис. 9 одно из множеств $\widehat{A}^{(1)}$ показано вертикально заштрихованным прямоугольником. Аналогично из $\mathscr{A}$ можно выделить алгебру $\widehat{\mathscr{A}^{(2)}}$, включающую множества $\widehat{A}^{(2)}$ двумерных векторов, первая координата которых может принимать любые значения из множества $\Omega^{(1)}$, а вторая принадлежит некоторому подмножеству $A^{(2)} \subseteq \Omega^{(2)}$. На рис. 9 одно из множеств $\widehat{A}^{(2)}$ показано горизонтально заштрихованным прямоугольником. Для каждого $\widehat{A}^{(i)}$, принадлежащего $\mathscr{A}$, можно указать вероятность $P(\widehat{A}^{(i)})$; поскольку квадрат на рис. 9 имеет площадь, равную единице, эта вероятность равна просто площади соответствующего

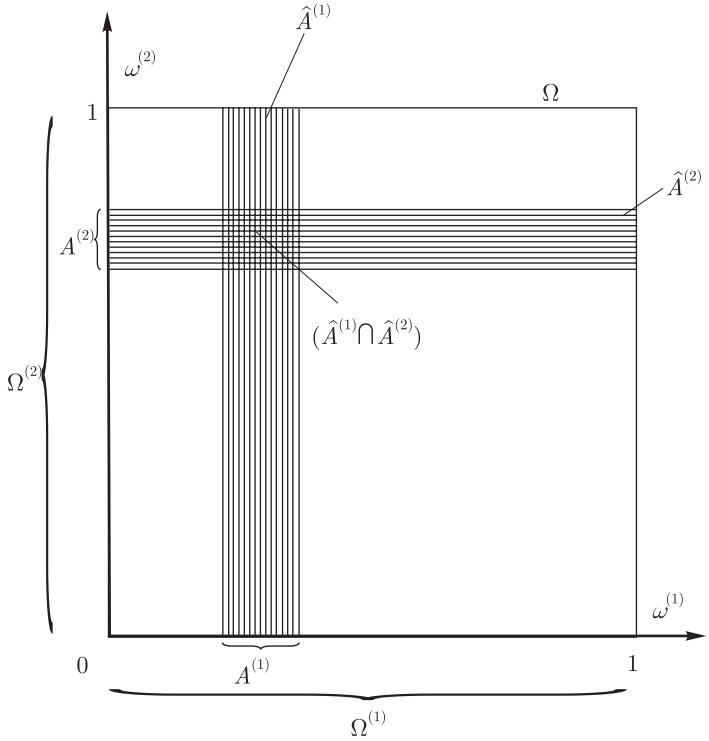


Рис. 9. Последовательность случайных испытаний. Множество элементарных исходов Ω случайного испытания $\mathscr{E}$, состоящего в бросании наугад точки в единичный квадрат, есть сам этот квадрат: $\Omega = (\omega^{(1)}, \omega^{(2)})$, $0 \leqslant \omega^{(1)} \leqslant 1$, $0 \leqslant \omega^{(2)} \leqslant 1$. Множество $\widehat{A}^{(1)}$ включает элементарные исходы, для которых $\omega^{(1)} \in A^{(1)}$ из $\Omega^{(1)}$ и $\omega^{(2)} \in \Omega^{(2)}$. Множество $\widehat{A}^{(2)}$ включает элементарные исходы, для которых $\omega^{(1)} \in \Omega^{(1)}$ и $\omega^{(2)} \in A^{(2)}$ из $\Omega^{(2)}$. Испытание можно представить как последовательность двух независимых испытаний $\mathscr{E}_1$ и $\mathscr{E}_2$, заключающихся в бросании точки на горизонтальную $\Omega^{(1)} = [0, 1]$ и вертикальную $\Omega^{(2)} = [0, 1]$ оси

прямоугольника: $P(\widehat{A}^{(i)}) = S(\widehat{A}^{(i)})$. Но событие $\widehat{A}^{(i)} \in \widehat{\mathscr{A}^{(2)}} \subseteq \widehat{\mathscr{A}}$ находится во взаимно-однозначном соответствии с событием $A^{(i)} \subseteq \Omega^{(i)}$. Поэтому события вида $A^{(i)} \subseteq \Omega^{(i)}$ также образуют некоторую алгебру $\mathscr{A}^{(i)}$, на которой можно задать вероятность $P^{(i)}(A^{(i)}) = P(\widehat{A}^{(i)})$. В случае, изобра-

женном на рис. 9, эта вероятность равна просто длине соответствующего отрезка: $P^{(i)}(A^{(i)}) = L(A^{(i)})$.

Таким образом, от случайного испытания $\mathscr{E}$, описываемого вероятностным пространством $(\Omega, \mathscr{A}, P)$, мы перешли к двум случайным испытаниям $\mathscr{E}_1$ и $\mathscr{E}_2$, описываемым вероятностными пространствами $(\Omega^{(1)}, \mathscr{A}^{(1)}, P^{(1)})$ и $(\Omega^{(2)}, \mathscr{A}^{(2)}, P^{(2)})$. Другими словами, двумерное случайное испытание $\mathscr{E}$ порождает два одномерных случайных испытания $\mathscr{E}_1$ и $\mathscr{E}_2$. Испытания $\mathscr{E}_1$ и $\mathscr{E}_2$ будем называть *маргинальными*[2] (частными) испытаниями. Соответственно, вероятностное пространство $(\Omega, \mathscr{A}, P)$ будем называть *совместным вероятностным пространством*, а пространства $(\Omega^{(i)}, \mathscr{A}^{(i)}, P^{(i)})$ — *маргинальными вероятностными пространствами*.

Среди событий из алгебры $\mathscr{A}$ наибольший интерес представляют события, состоящие в том, что в первом маргинальном испытании произошло событие $A^{(1)} \in \mathscr{A}^{(1)}$, а во втором — событие $A^{(2)} \in \mathscr{A}^{(2)}$. Эти события можно представить в виде пересечения $\widehat{A}^{(1)} \cap \widehat{A}^{(2)}$. Используя формулу умножения вероятностей, имеем:

$$P(\widehat{A}^{(1)} \cap \widehat{A}^{(2)}) = P(\widehat{A}^{(1)}/\widehat{A}^{(2)}) \cdot P^{(2)}(A^{(2)}) = P(\widehat{A}^{(2)}/\widehat{A}^{(1)}) \cdot P^{(1)}(A^{(1)}).$$

Из этой формулы видно, что в общем случае для вычисления вероятности событий $\widehat{A}^{(1)} \cap \widehat{A}^{(2)}$ нужно знать маргинальные вероятности и условные совместные вероятности. Однако если события $\widehat{A}^{(1)}$ и $\widehat{A}^{(2)}$ независимы, то достаточно знания только маргинальных вероятностей:

$$P(\widehat{A}^{(1)} \cap \widehat{A}^{(2)}) = P^{(1)}(A^{(1)}) \cdot P^{(2)}(A^{(2)}).$$

Полное сведение совместной вероятности к маргинальным вероятностям очень важно в приложениях теории вероятностей; оно положено в основу понятия независимости испытаний. Испытания $\mathscr{E}_1$ и $\mathscr{E}_2$, являющиеся маргинальными испытаниями для испытания $\mathscr{E} = (\mathscr{E}_1, \mathscr{E}_2)$, называются *независимыми*, если для любых событий $A^{(1)} \in \mathscr{A}^{(1)}$ и $A^{(2)} \in \mathscr{A}^{(2)}$ имеет место указанное равенство. Следует обратить внимание, что приведенные рассуждения о связи маргинальных и совместных вероятностей дают метод установления связи между разными вероятностными пространствами. Это особенно важно при изучении последовательностей случайных испытаний. Введение маргинальных вероятностей в модель «мишень» можно представить в виде последовательного бросания точки вначале на $\Omega^{(1)}$, а затем

[2]От англ. marginal — крайний, предельный, граничный.

на $\Omega^{(2)}$. Точно так же можно интерпретировать бросание двух монет (см. задачу I-11), игральных костей, появление двух потомков в скрещивании и т. д.

Заметим, что все приведенные выше рассуждения могут быть обобщены на последовательность любого числа испытаний $\mathscr{E} = (\mathscr{E}_1, \mathscr{E}_2, \ldots, \mathscr{E}_n)$. При этом независимость испытаний $\mathscr{E}_1, \mathscr{E}_2, \ldots, \mathscr{E}_n$ в дальнейшем мы будем понимать как независимость событий $\widehat{A}^{(1)}, \widehat{A}^{(2)}, \ldots, \widehat{A}^{(n)}$ в совокупности.

§ 10. Случайная величина

Рассмотрим числовое множество элементарных исходов Ω. Будем рассматривать случайное испытание, результатом которого может быть любое число из множества действительных чисел $\Omega = R^1 = \{-\infty, \infty\}$. Если реальные случайные испытания не охватывают всего множества R^1, соответствующим множествам будем приписывать нулевую вероятность.

Обозначим $< a, b >$ диапазон чисел от a до b. Запись $< a, b >$ означает любой вариант расстановки круглых и квадратных скобок: $(a, b), [a, b), (a, b]$ и $[a, b]$. При изучении случайных испытаний, имеющих в качестве исходов числа, интерес представляют события, состоящие в том, что число, являющееся исходом, попадает в тот или иной диапазон $< a, b >$. Поэтому обязательным требованием к алгебре случайных событий испытания с числовыми исходами будет требование содержать любой диапазон $< a, b >$. Наименьшая из алгебр, содержащих все диапазоны указанного вида, будет обозначаться $\mathscr{B}_1$.

Для задания вероятности P случайных событий из алгебры $\mathscr{B}_1$ достаточно рассмотреть лишь события вида $[a, b)$. В свою очередь, задание вероятности $P\{[a, b)\}$ можно упростить: представляя множество $(-\infty, b)$ в виде двух непересекающихся множеств $(-\infty, a)$ и $[a, b)$, получаем соотношение

$$P\{[a, b)\} = P\{(-\infty, b)\} - P\{(-\infty, a)\}.$$

Величину

$$F(x) = P\{(-\infty, x)\},$$

определенную для любого действительного числа x, называют *функцией распределения вероятностей*. Тогда

$$P\{[a, b)\} = F(b) - F(a).$$

Таким образом, знание функции распределения $F(x)$ позволяет задать вероятность P на алгебре $\mathscr{B}_1$ подмножеств множества R^1. Поскольку различные

случайные испытания с числовыми исходами имеют одинаковые $\Omega = R^1$ и $\mathscr{B}_1$, различаясь только заданием вероятности, постольку функция распределения $F(x)$ несет полную информацию о случайном испытании с числовым исходом.

Задание вероятностного пространства $[R^1, \mathscr{B}_1, F(x)]$, описывающего случайное испытание $\mathscr{E}$ с числовыми исходами, будем называть заданием *случайной величины*[3] $\widetilde{x}$, пробегающей множество значений $\Omega = R^1$. При этом вероятность события A будет интерпретироваться как вероятность попадания случайной величины $\widetilde{x}$ в множество A. Вероятность попадания $\widetilde{x}$ в отрезок $[a, b)$, т. е. вероятность события $A = \{a \leqslant \widetilde{x} < b\}$, определяется функцией распределения случайной величины $\widetilde{x}$:

$$P(a \leqslant \widetilde{x} < b) = F(b) - F(a).$$

Пример I-23. Случайное испытание состоит в бросании на ровную поверхность симметричного кружка, на одну сторону которого нанесена цифра 0, а на другую — 1. Чтобы применить введенную выше схему, будем считать, что возможными исходами испытания являются все действительные числа $\Omega = R^1$. Соответствие же этой схемы реальной экспериментальной ситуации обеспечим следующим способом задания вероятности. Вероятность события A из алгебры $\mathscr{B}_1$ определим так, что:

1) $P(A) = 0$, если множество A не содержит чисел 0 и 1;

2) $P(A) = 0{,}5$, если один из элементов (0 или 1) содержится в A;

3) $P(A) = 1$, если оба элемента (0 и 1) содержатся в A.

Отсюда следует выражение для функции распределения случайной величины $\widetilde{x}$:

$$F(x) = 0, \text{ когда } -\infty < \widetilde{x} \leqslant 0;$$
$$F(x) = 0{,}5, \text{ когда } 0 < \widetilde{x} \leqslant 1;$$
$$F(x) = 1, \text{ когда } 1 < \widetilde{x} < \infty.$$

Заметим, что для использования понятия случайной величины мы должны рассматривать числовое множество. Поэтому при использовании случайных величин для описания, например, случайного бросания монеты следует сопоставить исходам «герб», «решка» какие-либо числа, скажем, «решка» — 0, «герб» — 1.

[3] Над буквой, обозначающей случайную величину, мы всегда будем ставить знак $\sim$ («тильда»).

ПРИМЕР I-24. Случайное испытание состоит в выборе наугад семьи и установлении в ней числа детей $k = 0, 1, 2, \ldots$. Сколь велико может быть число детей в семье? Ясно, что у данной супружеской пары не может быть 1000 детей. Но, исключив границу 1000 как слишком завышенную, мы тут же снова ставим вопрос о максимально возможном числе детей. И на него нет ответа, поскольку нет оснований полагать, что могут быть семьи с k детьми и не может быть семей с $(k+1)$ детьми. Поэтому лучше не ограничивать множество исходов, считая, что оно охватывает все множество R^1. Ясно, что для $x < 0$ нужно задать нулевые вероятности; это же касается и положительных дробных чисел. Но как определить вероятности для целых чисел от 1 до ∞? Это потребует привлечения данных демографии и построения некоторой статистической теории, изложенной в § 4 гл. II.

Если случайное испытание Ω имеет элементарные исходы, характеризуемые двумя числами (x_1, x_2), то по аналогии с одномерным испытанием будем считать множеством элементарных исходов всю плоскость R^2, т. е. множество всех пар действительных чисел $-\infty < x_1 < \infty$, $-\infty < x_2 < \infty$. В качестве алгебры случайных событий выберем минимальную алгебру $\mathscr{B}_2$, содержащую все прямоугольники вида $A = \{(x_1, x_2)\colon a_1 \leqslant x_1 < b_1, a_2 \leqslant x_2 < b_2\}$. По аналогии с одномерной случайной величиной введем функцию распределения $F(x_1, x_2)$ двумерной случайной величины, определенную для пар действительных чисел:

$$F(x_1, x_2) = P\{(t_1, t_2)\colon \quad -\infty < \widetilde{t}_1 < x_1, \quad -\infty < \widetilde{t}_2 < x_2\}.$$

Таким образом, знание двумерной функции распределения $F(x_1, x_2)$ позволяет задать вероятность P на алгебре $\mathscr{B}_2$ подмножеств множества R^2. Поскольку различные случайные испытания с двумерными числовыми исходами (x_1, x_2) различаются только заданием вероятности (множество R^2 и алгебра $\mathscr{B}_2$ всегда одни и те же), то функция распределения $F(x_1, x_2)$ несет полную информацию о таком случайном испытании.

Подобно тому, как это было сделано в предыдущем параграфе, случайное испытание $\mathscr{E}$ с двумерными числовыми исходами можно представить в виде последовательности двух одномерных случайных испытаний $\mathscr{E}_1$ и $\mathscr{E}_2$. Вероятностное пространство $(R^2, \mathscr{B}_2, P)$ позволяет определить маргинальные вероятностные пространства $(R^1, \mathscr{B}_1, P^{(1)})$ и $(R^1, \mathscr{B}_1, P^{(2)})$, описывающие случайные испытания $\mathscr{E}_1$ и $\mathscr{E}_2$. Вероятностное пространство $(R^2, \mathscr{B}_2, P)$ соответствует одномерной маргинальной случайной величине $\widetilde{x}_i$, имеющей функцию распределения $F_i(x)$. Последняя позволяет задавать вероятность $P^{(i)}$ на алгебре $\mathscr{B}_1$, $i = 1, 2$. Мы будем обозначать случайную величину $\widetilde{x}$ в виде вектора $(\widetilde{x}_1, \widetilde{x}_2)$, компонентами которого яв-

ляются маргинальные случайные величины. Тогда попадание двумерного исхода в прямоугольник $A = \{(x_1, x_2)\colon a_1 \leqslant x_1 < b_1, \quad a_2 \leqslant x_2 < b_2\}$ можно интерпретировать как событие, состоящее в одновременном выполнении неравенств $a_1 \leqslant \widetilde{x}_1 < b_1, \quad a_2 \leqslant \widetilde{x}_2 < b_2$, а функцию распределения можно условно записать в таком виде:

$$F(x_1, x_2) = P\{-\infty < \widetilde{x}_1 < x_1, \quad -\infty < \widetilde{x}_2 < x_2\} = \\ = P\{\widetilde{x}_1 < x_1, \widetilde{x}_2 < x_2\}.$$

В простейшем случае независимости случайных испытаний функция распределения $F(x_1, x_2)$ случайной величины $\widetilde{x}$ однозначно определяется функциями распределения $F_1(x)$, $F_2(x)$ маргинальных случайных величин $\widetilde{x}_1$, $\widetilde{x}_2$:

$$F(x_1, x_2) = F_1(x_1) \cdot F_2(x_2).$$

При этом $\widetilde{x}_1$ и $\widetilde{x}_2$ называются *независимыми случайными величинами*.

По аналогии можно записать многомерную (n-мерную) функцию распределения

$$F(x_1, \ldots, x_n) = P\{(t_1, \ldots, t_n)\colon -\infty < \widetilde{t}_1 < x_1, \ldots, -\infty < \widetilde{t}_n < x_n\} = \\ = P\{\widetilde{x}_1 < x_1, \ldots, \widetilde{x}_n < x_n\},$$

задающую вероятность P на алгебре случайных событий $\mathscr{B}_n$, представляющих собой подмножества $\Omega = R^n$. По аналогии же в случае независимости маргинальных случайных величин $\widetilde{x}_1, \ldots, \widetilde{x}_n$

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i).$$

Чтобы случайные величины были независимыми, это соотношение должно выполняться для всех их пар, троек, четверок и т. д. (ср. с понятием независимости n случайных событий — § 6).

В следующих двух главах подробно изучаются случайные величины, важные для биометрии, и действия над ними.

Задачи

I-1. Постройте модель «урна», описывающую случайные испытания в примерах I-3 и I-5.

I-2. Постройте модель «мишень», описывающую случайные испытания в примерах I-4, I-6 и I-7.

I-3. Пусть элементарные события представляют собой точки действительной оси. Для любых $a < b$ определим события $A = (a, b]$ как множество точек x, таких, что $a < x \leqslant b$. Пусть $A_1 = (a_1, b_1]$ и $A_2 = (a_2, b_2]$ — два события. Тогда:

а) что графически означает $A_1 \cap A_2 = \varnothing$, чему равно при этом $A_1 \cup A_2$?

б) при каких условиях, накладываемых на числа a_1, b_1, a_2, b_2, будут иметь место события: $A_1 \cap A_2 = A_1$, $A_1 \cup A_2 = A_2$, $A_1 \cap A_2 = \varnothing$?

I-4. Покажите справедливость следующих свойств частоты случайного события:

а) если $A = \varnothing$, то $h(A; n) = 0$;

б) если $A \subseteq B$, то $h(A; n) \leqslant h(B; n)$;

в) $h(\overline{A}; n) = 1 - h(A; n)$;

г) если $A \cap B = \varnothing$, то $h(A \cup A; n) = h(A; n) + h(B; n)$;

д) $h(A \cup B; n) = h(A; n) + h(B; n) - h(A \cap B; n)$.

I-5. Покажите, что для условной вероятности $P(A/B)$ выполняются аксиомы теории вероятностей.

I-6. (Задача С. Н. Бернштейна.) На стол бросают правильный тетраэдр, одна грань которого белая, другая — черная, третья — красная, а четвертая раскрашена всеми тремя красками. Вероятность выпадения граней одинакова. Событием Б (Ч или К) назовем выпадение грани, на которую нанесена белая (черная или красная) краска. Покажите, что эти события попарно независимы, т. е. $P(\text{БЧ}) = P(\text{Б}) \cdot P(\text{Ч})$ и т. д. Покажите, что эти события в совокупности зависимы, т. е. $P(\text{БЧК}) \neq P(\text{Б}) \cdot P(\text{Ч}) \cdot P(\text{К})$. Сравните эту задачу с примером I-17.

I-7. В урне содержится n_1 шаров первого сорта, n_2 — второго, ..., n_r — r-го сорта; $n_1 + \ldots + n_r = n$ — общее число шаров. Рассмотрите три способа выбора шаров:

а) наугад извлекаются сразу m шаров (неупорядоченная выборка объема m);

б) наугад выбирается один шар, фиксируется его сорт, шар возвращается в урну, содержимое урны тщательно перемешивается, снова вынимается один шар и т. д. (последовательный выбор с возвращением m шаров);

в) m шаров вынимаются один за другим и не возвращаются в урну, благодаря чему при каждой выемке изменяются объем и состав урны (выбор без возвращения).

Какова вероятность того, что в выборке объема m имеется m_1 шаров первого сорта, m_2 — второго, ..., m_r — r-го сорта?

I-8. (Парадокс де Мере.) «Толчком к появлению интереса Паскаля к задачам теории вероятностей послужили встречи и беседы с одним из придворных французского королевского двора — шевалье де Мере (1607–1648). Де Мере интересовался философией, литературой и одновременно был азартным игроком. В этом можно видеть истоки тех теоретических вопросов, которые он предложил Паскалю»[4]. Вот один из них. Сколько раз надо подбросить две игральные кости, чтобы вероятность выпадения хотя бы один раз двух шестерок была бы больше $1/2$, т. е. была бы больше вероятности их невыпадения?

Задача эта возникла в связи со следующей игрой. Две кости подбрасывают заранее оговоренное число раз. Можно ставить либо на появление «дубля» хотя бы один раз, либо против такого результата. Де Мере знал, что при бросании одной кости 4 раза вероятность выпадения шестерки превышает $1/2$. Следовательно, рассуждал он, для двух костей достаточно в 6 раз больше бросаний, т. е. 24 бросания.

Так ли это?

I-9. (Парадокс Бертрана.) «Бертран (а затем и Пуанкаре) рассматривает следующий вопрос. Три одинаковых ящика имеют каждый по два отделения. Первый содержит в каждом отделении по золотой медали, второй — по серебряной, а третий — в одном отделении золотую, а в другом — серебряную. Взят один из ящиков. [...] Какова вероятность того, что, взяв ящик и вскрыв одно отделение, во втором отделении этого ящика обнаружим медаль другого металла, чем во вскрытом отделении?»[5]

[4] Гнеденко Б. В. Из истории науки о случайном: Из истории математических идей. — М.: Знание, 1981. — С. 17.

[5] Майстров Л. Е. Развитие понятия вероятности. — М.: Наука, 1980. — С. 215.

I-10. Чему равна вероятность того, что:

а) дни рождения 12 человек придутся на разные месяцы года (при условии, что вероятности попадания дня рождения на каждый месяц остаются равными для всех месяцев);

б) дни рождения 6 человек придутся в точности на два месяца?

I-11. Проинтерпретируйте, следуя § 9, случайное испытание $\mathscr{E}$, состоящее в бросании двух монет, в терминах двух последовательных испытаний $\mathscr{E}_1$ и $\mathscr{E}_2$, состоящих в двух независимых бросаниях одной монеты.

I-12. Покажите, что вычисление вероятности попадания двумерной случайной величины в прямоугольник

$$A = \{(x_1, x_2)\colon a_1 \leqslant x_1 < b_1, \quad a_2 \leqslant x_2 < b_2\}$$

проводится по формуле:

$$P(A) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2).$$

Глава II

Дискретные случайные величины

Поскольку случайная величина $\widetilde{x}$ полностью определяется своей функцией распределения $F(x)$, постольку различные частные виды случайных величин можно получить, налагая определенные ограничения на $F(x)$.

Среди многообразия случайных величин можно выделить два обширных класса: дискретные и непрерывные случайные величины.

Дискретными случайными величинами называются такие, функции распределения которых имеют вид:

$$F(x) = \sum\nolimits_{i:\ x_i<x} p_i,$$

где $p_i < 0$, $p_0 + \ldots + p_n + \ldots = 1$, $x_0 < x_1 < \ldots < x_n < \ldots$, а суммирование ведется по всем индексам i, для которых числовые значения x_i не превосходят величины x.

Выясним смысл величин p_i и x_i. Для этого найдем вероятность того, что случайная величина x принимает значения x_i:

$$P\{\widetilde{x} = x_i\} = P\{x_i \leqslant \widetilde{x} < x_{i+1}\} = F(x_{i+1}) - F(x_i) = p_i.$$

Таким образом, p_i суть вероятности того, что случайная величина $\widetilde{x}$ принимает значения x_i. Нетрудно показать, что для $x \neq x_i$ (для любого значения x_i) имеет место $P\{\widetilde{x} = x\} = 0$. Отсюда следует, что дискретная случайная величина $\widetilde{x}$ принимает с ненулевыми вероятностями p_i лишь значения x_i. Число таких значений может быть или конечно $(x_0, \ldots, x_n)$, или счетно $(x_0, \ldots, x_n, \ldots)$. Совокупность значений $\{x_i, p_i\}$ называется *распределением вероятностей* дискретной случайной величины $\widetilde{x}$. Так как распределение $\{x_i, p_i\}$ полностью определяет функцию распределения $F(x)$, то оно однозначно задает дискретную случайную величину $\widetilde{x}$.

§ 1. Целочисленные случайные величины и их свойства

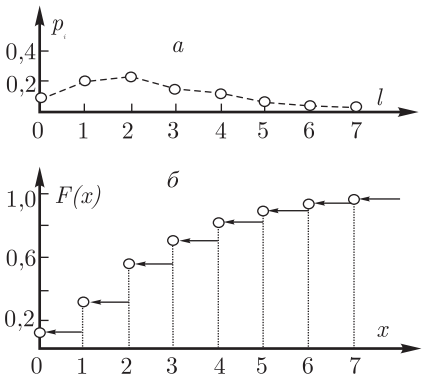


Рис. 10. Целочисленная случайная величина: *а* — распределение вероятностей $\{p_i\}$; *б* — функция распределения $F(x)$

В большинстве биометрических исследований можно ограничиться рассмотрением частного случая дискретной случайной величины, когда значения x_i целочисленны: $x_0 = 0, x_1 = 1, \ldots, x_n = n, \ldots$. У таких *целочисленных случайных величин* значения x_i совпадают с их индексами $x_i = i$, что позволяет понимать под распределением вероятностей целочисленной случайной величины набор $\{p_i\}$ одних только вероятностей $p_i \geqslant 0$.

Функция распределения целочисленной случайной величины имеет вид:

$$F(x) = \sum_{i<x} p_i,$$

где суммирование ведется по всем индексам i, меньшим значения x. Геометрически $F(x)$ представляет собой монотонно неубывающий ступенчатый график с разрывами и, соответственно, скачками величиной $p_0, p_1, \ldots, p_n, \ldots$ в точках $0, 1, \ldots, n, \ldots$ (см. рис. 10). Для $x \leqslant 0$ значение $F(x) = 0$, а при $x \to \infty$ значение $F(x) \to 1$.

Распределение вероятностей и функция распределения могут быть заданы также в виде таблицы.

ПРИМЕР II-1. Производится однократное подбрасывание игральной кости. Тогда, с одной стороны, в соответствии с результатами гл. I имеем множество элементарных исходов $\Omega = \{1, 2, \ldots, 6\}$. Для элементарных событий, соответствующих этим исходам (в предположении симметрии кости), можно задать вероятности $p_i = 1/6$, где $i = 1, 2, \ldots, 6$. С другой стороны, в силу того, что элементы множества Ω — числа натурального ряда, можно говорить о целочисленной случайной величине $\widetilde{x}$:

1	0	1	2	3	4	5	6	7	8	...
p_i	0	1/6	1/6	1/6	1/6	1/6	1/6	0	0	...
$F(x)$	0	0	1/6	2/6	3/6	4/6	5/6	1	1	... .

Вычислим вероятность того, что такая случайная величина примет значение, большее трех:

$$P\{\widetilde{x} > 3\} = P\{\widetilde{x} = 4\} + P\{\widetilde{x} = 5\} + P\{\widetilde{x} = 6\} + P\{\widetilde{x} = 7\} + \ldots = \\ = p_4 + p_5 + p_6 + p_7 + \ldots = 1/6 + 1/6 + 1/6 + 0 + \ldots = 1/2.$$

Эту вероятность можно вычислить также, обратившись к функции распределения:

$$P\{\widetilde{x} > 3\} = F(\infty) - F(4) = 1 - 1/2 = 1/2.$$

Однократное подбрасывание игральной кости — это пример *целочисленной равномерно распределенной случайной величины*, принимающей конечное число значений. Целочисленное равномерное распределение может быть получено и из модели «урна». В урне находятся N шаров, пронумерованных от 1 до N. Если производится выбор шаров по одному с возвращением, то номер i извлеченного шара есть целочисленная равномерно распределенная случайная величина $\widetilde{x}$:

$$p_i = P\{\widetilde{x} = i\} = 1/N, \quad i = 1, \ldots, N.$$

Зачастую помимо полного описания распределения желательно иметь небольшое число характеристик, выявляющих основные его черты. Такие характеристики носят название *параметров распределения*.

Параметром в математическом анализе обычно называют величину, входящую в формулы и математические выражения, значения которой считаются фиксированными в пределах рассматриваемой задачи.

В математической статистике термин «параметр» употребляется в двух значениях:

а) как фиксированная переменная, входящая в формулу того или иного распределения;

б) как величина, характеризующая некоторое свойство распределения, — то же, что показатель, характеристика.

В этом смысле говорят о параметрах (числовых характеристиках) положения и рассеяния.

Параметры положения отражают группировку значений случайной величины вокруг некоторого «центрального» значения, к ним относится, прежде всего, среднее значение.

Величина

$$E\widetilde{x} = \sum_i i p_i$$

называется *средним значением*, или *математическим ожиданием*, целочисленной случайной величины $\widetilde{x}$, имеющей распределение вероятностей $\{p_i\}$.

В примере II-1

$$E\widetilde{x} = \sum_{i=1}^{6} ip_i = 1 \cdot 1/6 + \ldots + 6 \cdot 1/6 = 3{,}5.$$

ПРИМЕР II-2. Обратимся к анализу числа детей в семье. Пусть семьи, имеющие 0 детей, встречаются с вероятностью p_0; имеющие 1 ребенка — с вероятностью p_1; i детей — с вероятностью p_i. Тогда $E\widetilde{x} = \sum_{i=1}^{\infty} ip_i$.

В общем случае дискретной случайной величины $\widetilde{x}$ ее *математическое ожидание* $E\widetilde{x}$ определяется формулой

$$E\widetilde{x} = \sum_i x_i p_i.$$

Параметры рассеяния характеризуют степень варьирования значений случайной величины, степень ее «изменчивости». Важнейший среди них — дисперсия (иногда говорят «варианса» — от англ. variance).

Дисперсия целочисленной случайной величины x определяется формулой

$$D\widetilde{x} = E[\widetilde{x} - W\widetilde{x}]^2 = \sum_i p_i (i - E\widetilde{x})^2.$$

Простые преобразования дают следующее выражение для дисперсии (см. задачу II-3):

$$D\widetilde{x} = E\widetilde{x}^2 - (E\widetilde{x})^2 = \sum_i i^2 p_i - \left(\sum_i ip_i \right)^2.$$

В примере II-1

$$D\widetilde{x} = [1 + 4 + \ldots + 36]1/6 - 3{,}5^2 \approx 2{,}92.$$

В более общем случае дискретной случайной величины $\widetilde{x}$ ее дисперсия определяется формулой

$$D\widetilde{x} = \sum_i (\widetilde{x}_i - E\widetilde{x})^2 p_i = \sum_i \widetilde{x}_i^2 p_i - \left(\sum_i \widetilde{x}_i p_i \right)^2 .$$

Величина $\sqrt{D\widetilde{x}}$ называется *средним квадратичным отклонением*, или *стандартным отклонением*, случайной величины $\widetilde{x}$.

§ 2. Совместное распределение, сумма, независимость дискретных случайных величин

Число детей в семье, рассматривавшееся выше, есть одномерная случайная величина. Однако если в семье учитывается число мальчиков и число девочек, то это будет уже двумерная случайная величина. Множество значений, принимаемых с ненулевой вероятностью двумерной целочисленной случайной величиной $\widetilde{x} = (\widetilde{x}_1, \widetilde{x}_2)$, описывается набором пар чисел (i, j), где $i = 0, 1, 2, \ldots$; $j = 0, 1, 2, \ldots$, графически представленных в виде целых точек положительного квадранта (рис. 11). Множество этих точек и есть пространство элементарных исходов $\Omega = \{\omega_{ij}\} = \{(0,0), (0,1), (1,0), (1,1), \ldots\}$. Каждому элементарному событию $\{\omega_{ij}\} = \{i, j\}$ отвечает вероятность $p_{ij} \geqslant 0$, причем $\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p_{ij} = 1$. Говорят, что $\{p_{ij}\} = \{p_{00}, p_{01}, p_{10}, p_{11}, \ldots\}$ — распределение вероятностей двумерной целочисленной случайной величины $\widetilde{x} = (\widetilde{x}_1, \widetilde{x}_2)$, или совместное распределение случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$.

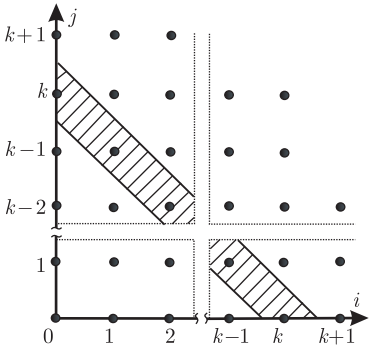


Рис. 11. Сумма целочисленных случайных величин. Целочисленные точки, попавшие в заштрихованную область, образуют событие A_k, состоящее в том, что сумма случайных величин $\widetilde{x}_1 + \widetilde{x}_2$ равна k

Выясним, каким образом, зная совместное распределение целочисленных случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$, можно найти распределение каждой из них в отдельности, т. е. их *маргинальные распределения*. Введем удобное

обозначение:

$$p_{i\cdot} = p_{i0} + p_{i1} + p_{i2} + \ldots = \sum_{j=0}^{\infty} p_{ij};$$

$$p_{\cdot j} = p_{0j} + p_{1j} + p_{2j} + \ldots = \sum_{i=0}^{\infty} p_{ij},$$

т. е. точка на месте определенного индекса означает суммирование по этому индексу; очевидно, $\sum_{i=0}^{\infty} p_i = 0$ и $\sum_{i=0}^{\infty} p_j$. Очевидно, что $\{p_i\}$ — это распределение вероятностей случайной величины $\widetilde{x}_i$, а $\{p_j\}$ — распределение вероятностей случайной величины $\widetilde{x}_2$. Действительно, обозначим A_i событие, состоящее в том, что $\widetilde{x}_1$ принимает значение, равное i. Имеем $A_i = \{(i,0),(i,1),(i,2),\ldots\}$, т. е. $A_i = \cup_{j=0}^{\infty}\{\omega_{ij}\}$. Следовательно,

$$P\{\widetilde{x} = i\} = P(A_i) = \sum_{j=0}^{\infty} P(\{\omega_{ij}\}) = \sum_{j=0}^{\infty} p_{ij} = p_{i\cdot}.$$

Аналогично доказывается, что $p_{\cdot j}$ — это вероятность события $\{\widetilde{x}_2 = j\}$.

В дальнейшем будут возникать вопросы, связанные с *суммой случайных величин*. Рассмотрим наряду с $\widetilde{x}_1$ и $\widetilde{x}_2$ новую случайную величину $\widetilde{x} = \widetilde{x}_1 + \widetilde{x}_2$. Каково распределение вероятностей $\{p_k\}$, $k = 0, 1, \ldots$, случайной величины $\widetilde{x}$?

Теорема II-1. $p_k = \sum_{i=0}^{k} p_{i,k-i}$.

Доказательство:

Обозначим A_k событие, состоящее в том, что $\widetilde{x}$ принимает значение, равное k. Имеем (см. рис. 11) $A_k = \{(o,k),(1,k-1),\ldots,(k-1,1),(k,0)\}$, т. е. $A_k = \cup_{i=0}^{k}\{\omega_{i,k-i}\}$. Следовательно,

$$p_l = P(A_k) = \sum_{i=0}^{k} P(\{\omega_{i,k-i}\}) = \sum_{i=0}^{k} p_{i\cdot,k-i},$$

т. е. суммируются вероятности тех точек, сумма индексов которых равна k (заштрихованная область на рис. 11). ■

Сумма случайных величин является важным частным случаем *функции случайных величин* $\phi(\widetilde{x}_1, \widetilde{x}_2)$. Для функции ϕ от случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$ можно ввести понятие математического ожидания:

$$E[\phi(\widetilde{x}_1, \widetilde{x}_2)] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \phi(i,j) \cdot p_{ij}.$$

Решение ряда задач количественной биологии опирается на понятие независимости случайных величин. Пусть $P\{\widetilde{x}_1 = k/\widetilde{x}_2 = l\}$ — условная вероятность того, что случайная величина $\widetilde{x}_1$ примет значение k, если случайная величина $\widetilde{x}_2$ приняла значение l. Случайные величины $\widetilde{x}_1$ и $\widetilde{x}_2$ независимы, если для любых i и j

$$P\{\widetilde{x}_1 = i/\widetilde{x}_2 = j\} = P\{\widetilde{x}_1 = i\}.$$

Каково совместное распределение $\{p_{ij}\}$ двух независимых случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$? Справедлива следующая теорема.

Теорема II-2. *Целочисленные случайные величины $\widetilde{x}_1$ и $\widetilde{x}_2$ независимы тогда и только тогда, когда для всех $i, j = 0, 1, \ldots$ будет $p_{ij} = p_{i\cdot}p_{\cdot j(k-1)}$.*

Пользуясь результатами двух приведенных выше теорем, ответим, наконец, на вопрос: каково распределение суммы двух независимых случайных величин?

Теорема II-3. *Если $\widetilde{x}_1$ и $\widetilde{x}_2$ — независимые целочисленные случайные величины, то распределение $\{p_{ij}\}$ их суммы $\widetilde{x} = \widetilde{x}_1 + \widetilde{x}_2$ имеет вид $p_k = \sum_{i=0}^{k} p_{i\cdot}p_{\cdot j(k-1)}$.*

Доказательство:

Действительно, как следует из рис. 11,

$$p_k = \sum_{i=0}^{k} p_{i,k-i} = \sum_{i=0}^{k} p_{i\cdot}p_{\cdot(k-i)}.$$

■

Соотношения между понятиями совместного распределения, суммы и независимости целочисленных случайных величин проиллюстрируем на примере.

Пример II-3. Дважды бросают игральную кость; число очков, выпавшее в первый раз, — значение случайной величины $\widetilde{x}_1$, во второй раз — $\widetilde{x}_2$. Тогда совокупность вероятностей событий $\{1,1\}, \{1,2\}, \ldots, \{6,6\}$ — это совместное распределение случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$. Совокупность вероятностей событий «сумма очков, выпавших оба раза» — $\{2\}, \{3\}, \ldots, \{12\}$ — это распределение случайной величины $\widetilde{x}$, являющейся суммой $\widetilde{x}_i$ и $\widetilde{x}_2$. Если результаты первого эксперимента (выпадение определенного числа очков в первый раз) не изменяют распределения вероятностей во втором эксперименте, то $\widetilde{x}_1$ и $\widetilde{x}_2$ независимы.

§ 3. Производящие функции

При изучении целочисленных случайных величин часто пользуются аппаратом производящих функций, позволяющим эффективно решать вопросы, связанные, в частности, с анализом распределений сумм случайных величин. Степенной ряд

$$G(y) = p_0 + p_1 y + p_2 y^2 + \dots,$$

определенный при $0 \leqslant y \leqslant 1$, называется *производящей функцией* целочисленной случайной величины $\widetilde{x}$ с распределением вероятностей $\{p_i\}$.

Выясним свойства производящих функций. Обозначим G' производную производящей функции G. Убедимся вначале, что $G(y)$ и $G'(y)$ положительны при $y > 0$. Действительно, пусть $p_l > 0$ ($l \neq 0$). Тогда $G(y) \geqslant p_l y^l > 0$. Первое утверждение доказано. Далее $G'(y) = \sum_{k=1}^{\infty} k p_k y^{k-1} \geqslant l p_l y^{l-1} > 0$ при $y > 0$. Итак, поскольку $G(0) = p_0$ и $G(1) = \sum_{i=0}^{\infty} p_i = 1$, можно утверждать, что производящая функция при увеличении y монотонно возрастает от p_0 до 1.

Обозначим $G^{(i)}(y)$ i-ю производную функции $G(y)$. Укажем связь между распределением вероятностей целочисленной случайной величины и производными соответствующей производящей функции.

Теорема II-4. $p_i = \frac{1}{i!} G^{(i)}(0)$.

Доказательство:
Выпишем последовательно выражения для производных:

$$\begin{aligned}
&G(y) = p_0 + p_1 y + p_2 y^2 + \dots + p_k y^k + \dots \\
&G'(y) = p_1 + 2p_2 y + \dots + k p_k y^{k-2} + \dots \\
&G''(y) = 2p_2 + \dots + k(k-1) p_k y^{k-2} + \dots \\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\
&G^{(k)}(y) = k! p_k + \dots
\end{aligned}$$

Отсюда следует, что

$$G(0) = p_0, \quad G'(0) = p_1, \quad G''(0) = 2p_2, \dots, G^{(k)}(0) = k! p_k,$$

т. е. для любого k имеем $p_k = \frac{1}{k!} G^{(k)}(0)$, что и утверждалось. ■

Пусть $\widetilde{x}_1$ и $\widetilde{x}_2$ — независимые случайные величины с распределениями вероятностей $\{p'_i\}$ и $\{p''_i\}$ и с производящими функциями $G_1(y)$ и $G_2(y)$ соответственно. Обозначим $G(y)$ производящую функцию суммы $\widetilde{x} = \widetilde{x}_1 + \widetilde{x}_2$.

Теорема II-5. $G(y) = G_1(y)G_2(y)$.

Доказательство:
Рассмотрим правую часть этого равенства. Имеем

$$G_1(y)G_2(y) = (p'_0 + p'_1 y + p'_2 y^2 + \ldots)(p''_0 + p''_1 y + p''_2 y^2 + \ldots) = \\ = p'_0 p''_0 + (p'_0 p''_1 + p'_1 p''_0)y + (p'_0 p''_2 + p'_1 p''_1 + p'_2 p''_0)y^2 + \ldots$$

Следовательно, для любого k коэффициент при y^k, т. е. вероятность p_k, будет $p_k = \sum_{i=0}^{k} p'_i p''_{k-i}$. Сопоставляя этот результат с теоремой II-1, убеждаемся в справедливости доказанного. ■

Теорему II-5 можно обобщить на случай n независимых случайных величин.

Математическое ожидание и дисперсию случайной величины можно выразить через производные производящей функции.

Теорема II-6. $E\widetilde{x} = G'(1)$, $D\widetilde{x} = G''(1) + G'(1) - [G'(1)]^2$.

Доказательство:
Имеем $G'(y) = \sum_{k=1}^{\infty} p_k k y^{k-1}$. Следовательно, $G'(1) = \sum_{k=1}^{\infty} k p_k$, что совпадает с определением математического ожидания. Далее $G''(y) = \\ = \sum_{k=2}^{\infty} p_k k(k-1) y^{k-2}$, откуда

$$G''(1) = \sum_{k=2}^{\infty} p_k k(k-1) = \sum_{k=2}^{\infty} p_k k^2 - \sum_{k=2}^{\infty} k p_k = \\ = \left(\sum_{k=2}^{\infty} p_k k^2 + p_1\right) - \left(\sum_{k=2}^{\infty} k p_k + p_1\right) = \sum_{k=1}^{\infty} p_k k^2 - \sum_{k=1}^{\infty} k p_k.$$

Так как $D\widetilde{x} = \sum_{k=1}^{\infty} k^2 p_k - \left(\sum_{k=1}^{\infty} k p_k\right)^2$, а $\sum_{k=1}^{\infty} k p_k = G'(1)$, то $D\widetilde{x} = \\ = G''(1) + G'(1) - [G'(1)]^2$. Следовательно, теорема доказана. ■

§ 4. Распределение Пуассона

Будем изучать распределение числа колоний микроорганизмов по чашкам Петри и попробуем представить, какой вид может иметь это распределение. Рассмотрим отдельную чашку. Разделим ее мысленно на n участков столь малой площади, что на каждый из них может попасть не больше одной клетки — «основателя» колонии. Вероятность того, что на участке номера i вырастет колония, обозначим λ_i; так как участок выбран малым, то можно считать, что λ_i достаточно мала. Следовательно, участок i описывается случайной величиной $\widetilde{x}_i$, принимающей значения 0 или 1 с вероятностями $q^{(i)} = 1 - \lambda_i$ и $p^{(i)} = \lambda_i$. Допустим далее, что вероятность, вырастет ли колония на данном участке, не зависит от того, есть ли колонии на других участках. Тогда изучаемая случайная величина $\widetilde{x}$ — число колоний на данной чашке — является суммой независимых случайных величин: $\widetilde{x} = \widetilde{x}_1 + \widetilde{x}_2 + \ldots + \widetilde{x}_n$. Предположим, что чашки не отличаются друг от друга: величина $\lambda = \lambda_1 + \ldots + \lambda_n$ одна и та же для всех чашек.

Данная модель описывает случайную величину $\widetilde{x}$, представляющую собой сумму большого числа неких «элементарных единичных актов», вероятность каждого из которых достаточно мала. Эта модель оказывается пригодной для описания распределений самых разнообразных биологических и небиологических явлений: распределения числа сорняков на учетных площадках поля, числа клеток крови по квадратам счетной камеры, числа мутаций, возникающих за определенный промежуток времени, числа несчастных случаев, телефонных вызовов за единицу времени, числа звезд в определенном объеме пространства, числа распадов радиоактивных атомов и т. д., и т. п.

Перейдем к анализу модели — изучению случайной величины $\widetilde{x}$. Каково распределение вероятностей $\{p_k\}$, $k = 0, 1, 2, \ldots$, случайной величины $\widetilde{x}$? Проведем некоторые рассуждения, основываясь на свойствах производящих функций.

Выпишем производящую функцию случайной величины $\widetilde{x}_i$:

$$G_i(y) = (1 - \lambda_i) + \lambda_i y = 1 + \lambda_i(y - 1).$$

Для суммы $\widetilde{x} = \widetilde{x}_1 + \ldots + \widetilde{x}_n$ производящая функция G в силу обобщения теоремы II-5 равна

$$G(y) = G_1(y) \cdot G_2(y) \cdot \ldots \cdot G_n(y).$$

Прологарифмировав это выражение и воспользовавшись приближенным равенством

$$\ln[1 + \lambda_i(y - 1)] \approx \lambda_i(y - 1),$$

справедливым в силу малости величины λ_i, получим

$$\ln G(y) = \sum_{i=1}^{n} \ln G_i(y) = \sum_{i=1}^{n} \ln[1 + \lambda_i(y-1)] \approx$$
$$\approx \sum_{i=1}^{n} \lambda_i(y-1) = \lambda(y-1).$$

Следовательно, приближенное выражение производящей функции имеет вид:

$$G(y) \approx e^{\ln G(y)} = e^{\lambda(y-1)}$$

и обращается в точное равенство при $\lambda_i \to 0, n \to \infty$.

Воспользуемся теперь теоремой II-4. Имеем

$$G'(y) = \lambda e^{\lambda(y-1)}, \quad G''(y) = \lambda^2 e^{\lambda(y-1)}, \ldots, G^{(k)} = \lambda^k e^{\lambda(y-1)}.$$

Следовательно, $G^{(k)}(0) = \lambda^k e^{-\lambda}$, откуда

$$p_k = \frac{1}{k!} G^{(k)}(0) = \frac{1}{k!} \lambda^k e^{-\lambda}.$$

Выведенное нами распределение

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots,$$

называется *распределением Пуассона* с параметром λ. Параметрами распределения определенного вида — в нашем случае распределения Пуассона — называют константы, значения которых однозначно определяют распределение вероятностей. Вид распределения Пуассона для различных значений параметра λ показан на рис. 12.

Следует помнить, что распределение Пуассона получено в предположении, что случайные величины $\widetilde{x}_i$ независимы, λ_i малы, а λ постоянно для единиц наблюдения (разных чашек Петри, учетных площадок, отрезков времени и т. п.). Фактические распределения могут значительно отличаться от распределения Пуассона в силу нарушения этих условий. Например, при плохом перемешивании суспензии клеток $\widetilde{x}_i$ уже не будут независимыми (образование «комков» клеток), а λ будет варьировать от одного опыта к другому. Отклонение эмпирического (экспериментального) распределения

от гипотетического (теоретического) позволяет подчас не только обнаружить методические ошибки в проведении опыта, но и вскрыть существование более сложных механизмов, не описываемых простыми вероятностными моделями.

В заключение этого параграфа отметим, что математическое ожидание и дисперсия пуассоновского распределения определяются параметром λ:

$$E\widetilde{x} = \lambda, \quad D\widetilde{x} = \lambda.$$

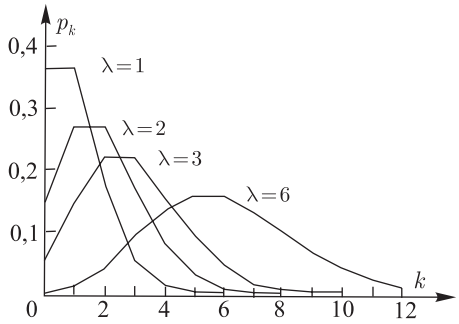


Рис. 12. Распределение Пуассона для некоторых значений параметра λ

Действительно, как мы только что показали, $G'(1) = \lambda$, $G''(1) = \lambda^2$. В силу теоремы II-6 имеем:

$$E\widetilde{x} = G'(1) = \lambda,$$
$$D\widetilde{x} = G''(1) + G'(1) - [G'(1)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

§ 5. Биномиальное распределение

Рассмотрим случайную величину — число девочек в семье из n детей, — принимающую значение 0, если родился мальчик, и значение 1, если родилась девочка. Спрашивается, каково распределение вероятностей для числа девочек в семье из n детей? Эта и многие другие задачи описываются следующей урновой схемой, называемой *схемой Бернулли*.

В урне находятся черные и белые шары в пропорции p: q $(p + q = 1)$. Каждый эксперимент состоит в случайном последовательном выборе

с возвращением n шаров. Каково распределение $\{p_i\}$, $i = 0, 1, 2, \ldots, n$, числа черных шаров?

Проведем рассуждения, подобные рассуждениям предыдущего параграфа. Обозначим $\widetilde{x}_i$ случайную величину, принимающую значение 0, если i-й вынутый шар белый, и значение 1, если он черный. По условию задачи $P\{x_i = 0\} = q$, $P\{x_i = 1\} = p$. Количество черных шаров в отдельном эксперименте описывается суммой $\widetilde{x} = \widetilde{x}_1 + \ldots + \widetilde{x}_n$. Так как каждая из $\widetilde{x}_i$ имеет производящую функцию $G_i(y) = q + py$, то произвадящая функция суммы $G(y)$ в силу обобщения теоремы II-5 будет $G(y) = (q + py)^n$. Далее

$$G'(y) = np(q + py)^{n-1}, G''(y) = n(n-1)p^2(q + py)^{n-2}, \ldots$$
$$\ldots, G^k(y) = n(n-1)\ldots(n-k+1)p^k(q + py)^{n-k}.$$

Поскольку для любого $k = 0, 1, \ldots, n$

$$n(n-1)\ldots(n-k+1) = \frac{n!}{(n-k)!}p^k q^{n-k},$$

$$\text{а } G^{(k)}(0) = \frac{n!}{(n-k)!}p^k q^{n-k},$$

то

$$p_k = \frac{1}{k!}G^{(k)}(0) = C_n^k p^k q^{n-k},$$

где $C_n^k = \dfrac{n!}{k!(n-k)!}$ — число сочетаний из n элементов по k.

Полученное распределение

$$p_k = C_n^k p^k q^{n-k}, \quad k = 0, 1, 2, \ldots, n,$$

называется *биномиальным распределением* с параметрами p и n.

Биномиальным распределением описывается широкий класс биологических явлений. К биномиальному распределению мы обращаемся всякий раз, когда изучаемый объект имеет одно из двух значений качественного альтернативного признака — 0 или 1 («да» или «не да»): семя проросло–не проросло, человек здоров–не здоров (болен), окраска крыльев бабочки светлая–не светлая (темная), клетка повреждена–не повреждена и т. д. Заметим, что биномиальное распределение мы получили в предположении, что $\widetilde{x}_i$ независимы и p остается постоянным. Если, например, речь идет об инфекционном заболевании, то альтернативный характер признака отнюдь не приведет к биномиальному распределению: вероятность заболеть разная для человека находящегося и не находящегося в контакте с больным.

Вид биномиального распределения для $p = 0,1$ при различных значениях n показан на рис. 13.

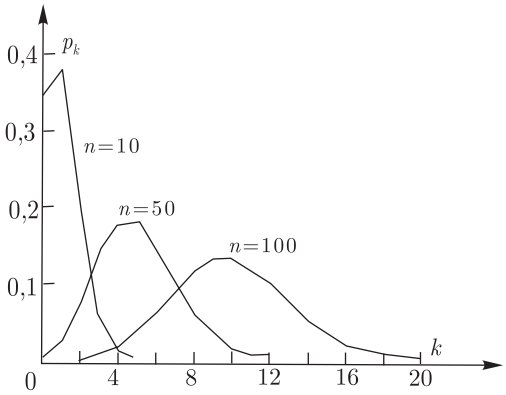


Рис. 13. Биномиальное распределение при $p = 0,1$ для некоторых значений n

Математическое ожидание и дисперсия случайной величины $\widetilde{x}$, распределенной по биномиальному закону, выражаются через параметры p и n:

$$E\widetilde{x} = np, \quad D\widetilde{x} = npq.$$

Действительно, $G'(1) = np$, $G''(1) = n(n-1)p^2$. В силу теоремы II-6 имеем:

$$E\widetilde{x} = np,$$
$$D\widetilde{x} = G''(1) + G'(1) - [G'(1)]^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p) = npq.$$

Укажем на связь между пуассоновским и биномиальным распределениями. Оказывается, что при малых p и достаточно большом n распределение Пуассона является приближением биномиального распределения (говорят: биномиальное распределение аппроксимируется пуассоновским), а именно: если p достаточно мало, то

$$C_n^k p^k q^{n-k} \sim e^{-\lambda} \frac{\lambda^k}{k!},$$

где $\lambda = np$; причем соответствие тем лучше, чем больше n (при одном и том же λ). Это следует из того, что производящая функция биномиального

распределения $G(y)$ равняется

$$G(y) = [q + py]^n = [1 + p(y-1)]^n = \left[1 + \frac{\lambda(y-1)}{n}\right]^n.$$

Как известно из курса математического анализа, функция $(1 + \lambda/n)^n$ стремится к e^λ при $n \to \infty$. Поэтому при $n \to \infty$ имеем $G(y) \to e^{\lambda(y-1)}$, что является производящей функцией распределения Пуассона.

Итак, при малых p биномиальное распределение становится близким к распределению Пуассона. Однако эти два распределения соответствуют разным статистическим схемам. Для пояснения сказанного рассмотрим производящую функцию $G(y)$, общую для пуассоновского и биномиального распределений:

$$G(y) = [1 + \lambda_1(y-1)] \cdot \ldots \cdot [1 + \lambda_n(y-1)].$$

Производящая функция распределения Пуассона

$$e^{\lambda(y-1)}$$

и производящая функция биномиального распределения

$$[q + py]^n = [1 + p(y-1)]^n$$

получаются из $G(y)$ при разных предположениях относительно $\lambda_1, \ldots, \lambda_n$. Распределение Пуассона получается в предположении, что $\lambda_1, \ldots, \lambda_n$ — достаточно малые числа; биномиальное — в предположении, что все λ_i одинаковы: $\lambda_1 = \lambda_2 = \ldots \lambda_n = p$, где p — их общее значение.

Из изложенного следует, что распределение Пуассона и биномиальное распределение хотя и родственны по природе, но описывают разные процессы. Распределение Пуассона описывает сумму гетерогенных событий с малыми вероятностями отдельных событий $(\lambda_1, \lambda_2 \ldots, \lambda_n)$. Распределение биномиальное описывает сумму однородных событий, но без ограничения на величину вероятности отдельного события.

§ 6. Полиномиальное распределение

К биномиальному распределению мы пришли, рассматривая урновую схему с шарами двух типов. Однако нередки задачи, когда число типов шаров в урне больше двух. Например, наличие четырех основных групп

крови в системе $AB0$ человека означает, что изучение их распределения приведет к урновой схеме с четырьмя типами шаров.

Рассмотрим общий случай. Пусть в урне находятся шары l типов в пропорциях $p_1\colon p_2\colon \ldots\colon p_l$, так что $p_1+p_2+\ldots+p_l=1$. Каждый эксперимент состоит в последовательном выборе с возвращением n шаров. Какова вероятность иметь в данной выборке k_1 шаров первого типа, k_2 — второго, k_l — l-го типа $(k_1+k_2+\ldots+k_l=n)$? Обозначим эту вероятность:

$$p(k_1,\ldots,k_l)=P\{\widetilde{x}_1=k_1,\quad \widetilde{x}_2=k_2,\quad \ldots,\quad \widetilde{x}_l=k_l\}.$$

По сути дела, это распределение многомерной случайной величины $(\widetilde{x}_1,\widetilde{x}_2,\ldots,\widetilde{x}_l)$. Проводя рассуждения, подобные тем, что использовались в § 5 при выводе биномиального распределения, можно показать, что

$$P(K_1,\ldots,k_l)=\frac{n!}{k_1!,k_2!,\ldots,k_l!}p_1^{k_1}p_2^{k_2}\ldots p_l^{k_l}.$$

Такое распределение называется *полиномиальным распределением* с параметрами $n,p_1,p_2,\ldots,p_{l-1}$. Полиномиальное распределение является обобщением биномиального, последнее можно рассматривать как полиномиальное при $l=2$.

ПРИМЕР II-4. При скрещивании гетерозигот Aa вероятности появления в потомстве доминантной гомозиготы AA, гетерозиготы Aa и рецессивной гомозиготы aa равны, соответственно, $p_1=1/4$, $p_2=2/4$, $p_3=1/4$. Чему равна вероятность того, что в семье, содержащей двух потомков, не будет ни одной доминантной гомозиготы?

Обозначим p искомую вероятность и пусть $p(k_1,k_2,k_3)$ — вероятность того, что k_1 потомков являются доминантными гомозиготами, k_2 — гетерозиготами и k_3 — рецессивными гомозиготами. Тогда $p=p(0,0,2)+{}+p(0,1,1)+p(0,2,0)$. Вычислим каждую из этих вероятностей:

$$p(0,0,2)=\frac{2!}{0!0!2!}(1/4)^0(1/2)^0(1/4)^2=1/16;$$

$$p(0,1,1)=\frac{2!}{0!1!1!}(1/4)^0(1/2)^1(1/4)^1=1/4;$$

$$p(0,2,0)=\frac{2!}{0!2!0!}(1/4)^0(1/2)^2(1/4)^0=1/4.$$

В итоге $p=1/16+1/4+1/4=9/16$.

Задачи

II-1. Докажите, что:

а) $EC = C$;

б) $E(C\widetilde{x}) = CE\widetilde{x}$;

в) $E(C + \widetilde{x}) = C + E\widetilde{x}$, где C — константа.

II-2. Докажите, что:

а) $DC = 0$;

б) $D(C\widetilde{x}) = C^2 D\widetilde{x}$;

в) $D(C + \widetilde{x}) = D\widetilde{x}$, где C — константа.

II-3. Докажите, что формулу для дисперсии $D\widetilde{x} = E(\widetilde{x} - E\widetilde{x})^2$ можно записать в виде: $D\widetilde{x} = E\widetilde{x}^2 - (E\widetilde{x})^2$.

II-4. Покажите, что математическое ожидание целочисленной равномерно распределенной случайной величины $E\widetilde{x} = \frac{1}{2}(n+1)$. При этом обратите внимание, что

$$S = 1 + 2 + \ldots + (N-1) + N;$$
$$S = N + (N-1) + \ldots + 2 + 1.$$

II-5. Покажите, что дисперсия целочисленной равномерно распределенной случайной величины $D\widetilde{x} = \frac{1}{12}(N^2 - 1)$.

Для вычисления $\sum_{i=1}^{N}$ используйте треугольную таблицу:

II-6. Докажите, что для любых двух случайных величин

$$E(\widetilde{x}_1 + \widetilde{x}_2) = E\widetilde{x}_1 + E\widetilde{x}_2.$$

II-7. Докажите, что если $\widetilde{x}_1$ и $\widetilde{x}_2$ независимы, то

$$D(\widetilde{x}_1 + \widetilde{x}_2) = D\widetilde{x}_1 + D\widetilde{x}_2.$$

II-8. Докажите, что если независимые случайные величины $\widetilde{x}_1$ и $\widetilde{x}_2$ распределены по закону Пуассона с параметрами λ_1 и λ_2 соответственно, то их сумма $\widetilde{x} = \widetilde{x}_1 + \widetilde{x}_2$ также имеет пуассоновское распределение с $\lambda = \lambda_1 + \lambda_2$.

$1 + 2 + 3 + 4 + \ldots + N$	$N(N+1)/2$
$2 + 3 + 4 + \ldots + N$	$(N+2)(N-1)/2$
$3 + 4 + \ldots + N$	$(N+3)(N-2)/2$
$4 + \ldots + N$	$(N+4)(N-3)/2$
$\ldots + N$	$(N+N)(N-N+1)/2$
Сумма по столбцу $1 \quad 2^2 \quad 3^2 \quad 4^2 \quad \ldots \quad N^2$	

Таблица 1

Распределение результатов 26 306 бросаний 12 игральных костей (данные Уэлдона, по В. Феллеру [1967] и М. Кендаллу и А. Стьюарту [1966]) (к задаче II-9)

Число осуществления события $\{5, 6\}$ в одном эксперименте	Число наблюдений	
	эмпирическое	ожидаемое
0	185	203
1	1 149	1 216
2	3 265	3 345
3	5 475	5 576
4	6 114	6 273
5	5 194	5 018
6	3 067	2 927
7	1 331	1 255
8	403	392
9	105	87
10	14	13
11	4	1
12	0	0

II-9. Эксперимент заключается в одновременном подбрасывании 12 игральных костей. Если кости симметричные, то вероятность случайного события $A = \{5, 6\}$ для каждой из них равна 1/3. Определим случайную величину как число костей, на которых выпала цифра 5 или 6. Она имеет биномиальное распределение с параметрами $p = 1/3$, $n = 12$. В табл. 1 приведены результаты 26 306 таких экспериментов. Постройте графики на-

Таблица 2

Распределение числа поврежденных хромосом по клеткам проростков гороха, $n = 1000$ (Н. В. Лучник, 1963 г.). (к задаче II-11)

Число поврежденных хромосом в клетке	Частота клеток
0	0,877
1	0,063
2	0,047
3	0,007
4	0,004
5	0,001
6	0,001

блюдаемой и ожидаемой функций распределения. Сравните две полученные кривые и найдите их максимальное расхождение.

II-10. В одной семье все 9 детей — девочки. Найдите вероятность этого события, предполагая, что появление детей разного пола описывается биномиальным распределением с параметрами $p = 1/2, n = 9$.

II-11. При облучении сухих семян гороха гамма-лучами в клетках проростков регистрировали число поврежденных хромосом (табл. 2). Всего было проанализировано 1 000 клеток. Среднее число поврежденных хромосом на клетку равно $0{,}205$. Сравните графически это эмпирическое распределение с гипотетическим распределением Пуассона, $\lambda = 0{,}205$.

Глава III

Непрерывные случайные величины

В гл. II мы рассмотрели дискретные случайные величины. Другой важный класс составляют случайные величины, множества возможных значений которых представляют отрезки действительной прямой. Примерами непрерывных случайных величин являются масса и размеры особи, диаметр колонии на чашке Петри, уровень активности фермента, количество гемоглобина в миллилитре крови, возраст и т. д.

§ 1. Непрерывные случайные величины и их свойства

Непрерывная случайная величина x определяется тем, что ее функция распределения $F(x)$ имеет следующий вид:

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

где $f(x)$ — неотрицательная функция, называемая *плотностью распределения вероятностей* случайной величины $\widetilde{x}$. Плотность полностью определяет непрерывную случайную величину и позволяет находить вероятность попадания $\widetilde{x}$ в отрезок $[a, b]$ по формуле

$$P\{a \leqslant \widetilde{x} \leqslant b\} = \int_{a}^{b} f(x)dx.$$

Заметим, что вероятность попадания случайной величины x в отрезок $[a, b]$ равна вероятности попадания в любой из отрезков $[a, b), (a, b], (a, b)$, потому что вероятность попадания в любую конкретную точку равна нулю. Последнее утверждение следует из предельного соотношения

$$P\{a \leqslant \widetilde{x} < a + \Delta\} = \int_{a}^{a+\Delta} f(x)dx \xrightarrow[\Delta \to 0]{} P\{\widetilde{x} = a\} = 0.$$

Отметим, что между дискретными и непрерывными случайными величинами существует явная аналогия. Пусть $\widetilde{x}$ — непрерывная случайная величина с плотностью распределения $f(x)$. Разобьем интервал $(a, b]$ (рис. 14) на n отрезков длиной Δ, для простоты примем $a = 0$. Для каждого из интервалов $(0, \Delta], (\Delta, 2\Delta], \ldots$ определим числа $p_0, p_1, \ldots$ как площади построенных на них прямоугольников, высота которых равна значению функции $f(x)$ в серединах отрезков. Например, p_k — это площадь заштрихованного на рис. 14 прямоугольника: $p_k = \Delta \cdot f(x_k)$, где $x_k = k \cdot \Delta + \Delta/2 = (k + 1/2)\Delta$. Таким образом, мы имеем дискретную случайную величину $\widetilde{x}'$ с распределением вероятностей $\{p_k\}, k = 0, 1, 2, \ldots$ Поскольку для непрерывной случайной величины $\widetilde{x}$ имеем $P\{a \leqslant \widetilde{x} \leqslant b\} = \int_a^b f(x)dx$, то в силу определения интеграла

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n-1} f(i\Delta)\Delta = \sum_{i=0}^{n-1} p_i.$$

Следовательно, $P\{a \leqslant \widetilde{x} \leqslant b\} = P\{a \leqslant \widetilde{x}' \leqslant b\}$. Таким образом, дискретная случайная величина $\widetilde{x}'$ служит приближением непрерывной случайной величины $\widetilde{x}$ в такой же степени, в какой частичные суммы (площади прямоугольников) служат приближением интеграла. В пределе, при $\Delta \to 0$, построенная нами дискретная случайная величина $\widetilde{x}'$ стремится к непрерывной случайной величине $\widetilde{x}$.

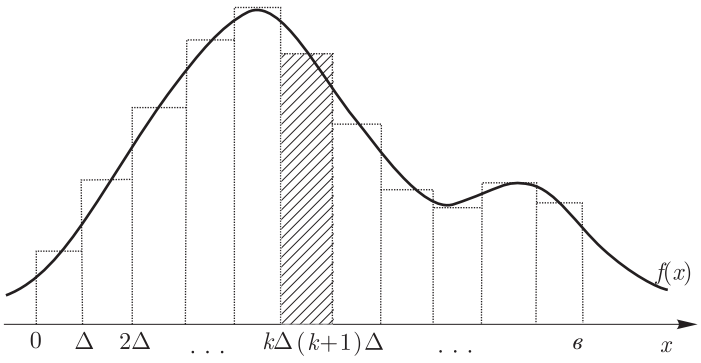


Рис. 14. Приближение непрерывной случайной величины дискретной случайной величиной

Понятия *математического ожидания* $E\widetilde{x}$ и *дисперсии* $D\widetilde{x}$ для непрерывных случайных величин вводятся следующим образом:

$$E\widetilde{x} = \int_{-\infty}^{\infty} xf(x)dx;$$

$$D\widetilde{x} = E(\widetilde{x} - E\widetilde{x})^2 = \int_{-\infty}^{\infty} (x - E\widetilde{x})^2 f(x)dx.$$

Как и для дискретного случая, выражение для дисперсии можно записать по-иному:

$$D\widetilde{x} = E\widetilde{x}^2 - (E\widetilde{x})^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \left[\int_{-\infty}^{\infty} xf(x)dx\right]^2.$$

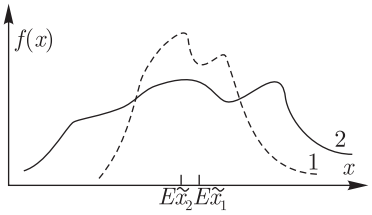


Рис. 15. Дисперсия случайной величины x_1 (кривая 1) меньше дисперсии случайной величины x_2 (кривая 2)

Дисперсия характеризует размах изменчивости изучаемого признака, «ширину» кривой плотности распределения. Например, на рис. 15 распределение 1 имеет меньшую дисперсию, чем распределение 2. Из рисунка можно также сделать качественный вывод о том, что распределение случайной величины «сосредоточено», в основном, в некоторой окрестности математического ожидания. Мы вкладываем в сказанное следующий смысл: вероятность больших отклонений случайной величины от $E\widetilde{x}$ мала, т. е. $P\{|\widetilde{x} - E\widetilde{x}| \geqslant \varepsilon\} \to 0$, если $\varepsilon \to \infty$.

Приведенные качественные соображения можно облечь в количественную форму. Обозначим $\mu = E\widetilde{x}$ и $\sigma = \sqrt{D\widetilde{x}}$. Справедлива следующая теорема, которую дадим без доказательства.

Теорема III-1 (теорема П. Л. Чебышева). *Для любого числа* k

$$P\{|\widetilde{x} - \mu| \geqslant k\tau\} \leqslant \frac{1}{k^2}.$$

Из этой теоремы следует, например, что при $k = 5$ на участке от $\mu - 5\sigma$ до $\mu + 5\sigma$ сосредоточено не менее (а возможно, и более) 96 % значений

случайной величины, т. е.

$$\int_{\mu-5\sigma}^{\mu+5\sigma} f(x)dx \geqslant 0,96.$$

Для практически важных типов распределений 96 % распределения сосредоточено в гораздо меньшей окрестности математического ожидания — в пределах $2\sigma \div 3\sigma$ (см. § 3). Из теоремы III-1 следует также, что при $D\widetilde{x}$, близком к нулю, распределение сосредоточено, в основном, в малой окрестности $E\widetilde{x}$.

В ряде случаев в качестве характеристики «центра» распределения берут *медиану* ζ, которая делит плотность распределения на две равновеликие по площади части. Иными словами, для медианы выполняется соотношение

$$P\{\widetilde{x} < \zeta\} = P\{\widetilde{x} \geqslant \zeta\} = 0{,}5.$$

§ 2. Совместное распределение и независимость непрерывных случайных величин

Пусть $\widetilde{x}_1, \widetilde{x}_2$ — две непрерывные случайные величины. Множество их возможных значений (элементарных событий) — точки плоскости. Любые события представляются геометрически в виде областей на этой плоскости. В дальнейшем для простоты будем одной и той же буквой (например, A) обозначать и область, и соответствующее событие, заключающееся в том, что двумерная случайная величина $\widetilde{x} = (\widetilde{x}_1, \widetilde{x}_2) \in A$.

Совместная функция распределения $F(x_1, x_2)$ двух непрерывных случайных величин $\widetilde{x}_1, \widetilde{x}_2$ задается плотностью двумерного распределения — неотрицательной функцией — только уже двух переменных $f(x_1, x_2)$, такой, что

$$F(x_1, x_2) = \int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f(t_1, t_2)dt_1dt_2.$$

При этом вероятность любого события A, обозначаемая, как и раньше, $P(A)$ или $P\{\widetilde{x} \in A\}$, определяется выражением

$$P\{\widetilde{x} \in A\} = \iint_A f(x_1, x_2)dx_1dx_2,$$

где интегрирование ведется по области A.

Выясним теперь, как связаны маргинальные распределения случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$ с функцией $f(x_1, x_2)$. Пусть $P\{\widetilde{x}_1 < x_1\}$ — это вероятность того, что случайная величина $\widetilde{x}_1$ примет значение, меньшее x_1, вне

зависимости от того, каким будет x_2. Следовательно, $P\{\widetilde{x}_1 < x_1\} = F_1(x_1)$, где $F_1(x_1)$ — маргинальная функция распределения $\widetilde{x}_1$. Обозначим $f_1(x_1)$ соответствующую плотность распределения $\widetilde{x}_1$. Найдем ее:

$$F_1(x_1) = P\{\widetilde{x}_1 > x_1\} = P\{\widetilde{x} \in A\} = \iint_A f(x_1, x_2)dx_1dx_2 =$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{x_1} f(t_1, t_2)dt_1dt_2 = \int_{-\infty}^{x_1}\left[\int_{-\infty}^{\infty} f(t_1, t_2)dt_2\right]dt_1.$$

Так как $f_1(x_1) = F_1'(x_1)$, то по правилу дифференцирования интеграла получим, что плотность $\widetilde{x}_1$ равна

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2.$$

Аналогично плотность распределения $\widetilde{x}_2$ равна

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_2, x_2)dx_1.$$

$$E\widetilde{x}_1 = \int_{-\infty}^{\infty} x_1 f_1(x_1)dx_1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f(x_1, x_2)dx_1dx_2;$$
$$E\widetilde{x}_2 = \int_{-\infty}^{\infty} x_2 f_2(x_1)dx_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2 f(x_1, x_2)dx_1dx_2.$$

Рассмотрим понятие математического ожидания в общем случае. Пусть $\widetilde{x}$ — случайная величина, являющаяся функцией $\widetilde{x}_1$ и $\widetilde{x}_2$: $\widetilde{x} = \phi(\widetilde{x}_1, \widetilde{x}_2)$. Математическим ожиданием случайной величины $\widetilde{x}$ называется число

$$E[\phi(\widetilde{x}_1, \widetilde{x}_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi(\widetilde{x}_1, \widetilde{x}_2) f(x_1, x_2)dx_1dx_2.$$

Из этого равенства немедленно следует свойство суммы случайных величин: $E(\widetilde{x}_1 + \widetilde{x}_2) = E\widetilde{x}_1 + E\widetilde{x}_2$. Действительно,

$$E(\widetilde{x}_1 + \widetilde{x}_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1 + x_2) f(x_1, x_2)dx_1dx_2 =$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f(x_1, x_2)dx_1dx_2 +$$
$$+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2 f(x_1, x_2)dx_1dx_2 = E\widetilde{x}_1 + E\widetilde{x}_2.$$

Как следует из гл. I, непрерывные случайные величины $\widetilde{x}_1$ и $\widetilde{x}_2$ являются независимыми, если для любых интервалов $A_1 = (a_1, b_1]$ и $A_2 = (a_2, b_2]$ условная вероятность $P\{\widetilde{x}_1 \in A_1 / \widetilde{x}_2 \in A_2\}$ не зависит от A_2, т. е. $P\{a_1 < \widetilde{x}_1 \leqslant b_1 / a_2 < \widetilde{x}_2 \leqslant b_2\} = P\{a_1 < \widetilde{x}_1 \leqslant b_1\}$. Пусть $f(x_1, x_2)$ — совместная плотность вероятности $\widetilde{x}_1$ и $\widetilde{x}_2$, а $f_1(x_1)$ и $f_2(x_2)$ — плотности вероятности x_1 и x_2 в отдельности. Приведем без доказательства следующую теорему.

Теорема III-2. *Непрерывные случайные величины $\widetilde{x}_1$ и $\widetilde{x}_2$ независимы тогда и только тогда, когда*

$$f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2).$$

Из этой теоремы следует важное свойство независимых случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$:

$$D(\widetilde{x}_1 + \widetilde{x}_2) = D\widetilde{x}_1 + D\widetilde{x}_2.$$

Действительно,

$$D(\widetilde{x}_1 + \widetilde{x}_2) = E(\widetilde{x}_1 + \widetilde{x}_2)^2 - [E(\widetilde{x}_1 + \widetilde{x}_2)]^2.$$

Имеем

$$E(\widetilde{x}_1 + \widetilde{x}_2)^2 = E(\widetilde{x}_1^2 + 2\widetilde{x}_1\widetilde{x}_2 + \widetilde{x}_2^2) = E\widetilde{x}_1^2 + 2E(\widetilde{x}_1\widetilde{x}_2) + E\widetilde{x}_2^2.$$

Так как

$$[E(\widetilde{x}_1 + \widetilde{x}_2)]^2 = [E\widetilde{x}_1 + E\widetilde{x}_2]^2 = (E\widetilde{x}_1)^2 + 2(E\widetilde{x}_1)(E\widetilde{x}_2) + (E\widetilde{x}_2)^2,$$

то

$$D(\widetilde{x}_1 + \widetilde{x}_2) = [E\widetilde{x}_1^2 - (E\widetilde{x}_1)^2] + [E\widetilde{x}_2^2 - (E\widetilde{x}_2)^2] + \\ + 2[E(\widetilde{x}_1\widetilde{x}_2) - (E\widetilde{x}_1)(E\widetilde{x}_2)].$$

Первые два члена правой части — это по определению $D\widetilde{x}_1$ и $D\widetilde{x}_2$. Третий член равен нулю, так как $\widetilde{x}_1$ и $\widetilde{x}_2$ независимы (см. задачу III-2). Следовательно, $D(\widetilde{x}_1 + \widetilde{x}_2) = D\widetilde{x}_1 + D\widetilde{x}_2$.

Таким образом, независимые случайные величины обладают аддитивным свойством в отношении дисперсии. И в общем случае для n независимых случайных величин $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ будет

$$D\left(\sum_{i=1}^{n} \widetilde{x}_i\right) = \sum_{i=1}^{n} D\widetilde{x}_i.$$

Выше мы рассмотрели свойства математического ожидания и дисперсии для суммы независимых случайных величин. Когда случайные величины зависимы, в качестве меры зависимости используют *ковариацию*:

$$\mathrm{Cov}(\widetilde{x}_1, \widetilde{x}_2) = E[(\widetilde{x}_1 - E\widetilde{x}_1)(\widetilde{x}_2 - E\widetilde{x}_2)].$$

Ее можно записать и иначе:

$$\mathrm{Cov}(\widetilde{x}_1, \widetilde{x}_2) = E(\widetilde{x}_1\widetilde{x}_2) - (E\widetilde{x}_1)(E\widetilde{x}_2).$$

Последнее следует из преобразований

$$\begin{gathered}\mathrm{Cov}(\widetilde{x}_1, \widetilde{x}_2) = E[\widetilde{x}_1\widetilde{x}_2 - \widetilde{x}_1(E\widetilde{x}_2) - \widetilde{x}_2(E\widetilde{x}_1) + (E\widetilde{x}_1)(E\widetilde{x}_2)] = \\ = E(\widetilde{x}_1\widetilde{x}_2) - (E\widetilde{x}_1)(E\widetilde{x}_2) - (E\widetilde{x}_2)(E\widetilde{x}_1) + (E\widetilde{x}_1)(E\widetilde{x}_2) = \\ = E(\widetilde{x}_1\widetilde{x}_2) - (E\widetilde{x}_1)(E\widetilde{x}_2).\end{gathered}$$

Если $f(x_1, x_2)$ — плотность распределения, то

$$\mathrm{Cov}(\widetilde{x}_1, \widetilde{x}_2) = \iint_{-\infty}^{\infty} x_1 x_2 f(x_1, x_2) dx_1 dx_2 - (E\widetilde{x}_1)(E\widetilde{x}_2).$$

Смысл ковариации ясен из сопоставления ее с дисперсиями $\widetilde{x}_1$ и $\widetilde{x}_2$, в которых квадрат $(\widetilde{x} - E\widetilde{x})^2$ запишем как произведение $(\widetilde{x} - E\widetilde{x})(\widetilde{x} - E\widetilde{x})$:

$$\begin{gathered}D\widetilde{x}_1 = E[(\widetilde{x}_1 - E\widetilde{x}_1)(\widetilde{x}_1 - E\widetilde{x}_1)]; \\ D\widetilde{x}_2 = E[(\widetilde{x}_2 - E\widetilde{x}_2)(\widetilde{x}_2 - E\widetilde{x}_2)]; \\ \mathrm{Cov}(\widetilde{x}_1, \widetilde{x}_2) = E[(\widetilde{x}_1 - E\widetilde{x}_1)(\widetilde{x}_2 - E\widetilde{x}_2)].\end{gathered}$$

Аналогичным образом можно определить понятие ковариации и для дискретных случайных величин. Ковариация (буквально «совместная вариация») — это мера совместного варьирования случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$ (английские термины variance и covariance лучше отражают общность смысла этих величин). Можно сказать, что дисперсия случайной величины $\widetilde{x}$ — это ковариация $\widetilde{x} \subset \widetilde{x}$: $D\widetilde{x} = \mathrm{Cov}(\widetilde{x}_1, \widetilde{x}_2)$.

Введем теперь новую случайную величину

$$\widetilde{y} = \frac{\widetilde{x} - E\widetilde{x}}{\sqrt{D\widetilde{x}}},$$

которую будем называть *нормированной случайной величиной*. Смысл такого преобразования (нормирования) заключается в том, что за начало отсчета

значений случайной величины берется среднее значение $E\widetilde{x}$, а стандартное отклонение $\sqrt{D\widetilde{x}}$ используется как единица измерения. В результате получается безразмерная величина, математическое ожидание которой равно нулю:

$$Ey = 0,$$

а дисперсия равна единице:

$$D\widetilde{y} = 1$$

(см. задачу III-3).

Ковариация двух нормированных случайных величин

$$\widetilde{y}_1 = \frac{\widetilde{x}_1 - E\widetilde{x}_1}{\sqrt{D\widetilde{x}_1}} \quad \text{и} \quad \widetilde{y}_2 = \frac{\widetilde{x}_2 - E\widetilde{x}_2}{\sqrt{D\widetilde{x}_2}}$$

называется *коэффициентом корреляции* случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$ и обозначается как $\varrho(\widetilde{x}_1, \widetilde{x}_2)$ или просто ϱ:

$$\varrho = \operatorname{Cov}(\widetilde{y}_1, \widetilde{y}_2) = E(\widetilde{y}_1\widetilde{y}_2) - (E\widetilde{y}_1)(E\widetilde{y}_2) = E(\widetilde{y}_1\widetilde{y}_2) = \\ = \frac{\operatorname{Cov}(\widetilde{x}_1, \widetilde{x}_2)}{\sqrt{(D\widetilde{x}_1)(D\widetilde{x}_2)}}.$$

Как и исходные нормированные величины $\widetilde{y}_1$ и $\widetilde{y}_2$, коэффициент корреляции есть величина безразмерная, и потому он является удобной мерой зависимости (показателем силы связи) двух случайных величин. Его значения по абсолютной величине не превышают единицы:

$$-1 \leqslant \varrho \leqslant 1,$$

и достигают $+1$ или -1 лишь при наличии линейной зависимости между величинами $\widetilde{x}_1$ и $\widetilde{x}_2$. Если случайные величины независимы, то $\varrho = 0$; это следует из того, что при независимости $\operatorname{Cov}(\widetilde{x}_1, \widetilde{x}_2) = 0$. Надо подчеркнуть, что обратное утверждение в общем случае неверно: равенство нулю коэффициента корреляции еще не означает независимости случайных величин. Однако для биометрических задач очень важно, что обратное утверждение справедливо в случае двумерного нормального распределения (см. § 6).

Дисперсия суммы зависимых случайных величин выражается через коэффициент корреляции:

$$D(\widetilde{x}_1 + \widetilde{x}_2) = D\widetilde{x}_1 + D\widetilde{x}_2 + 2\varrho\sqrt{(D\widetilde{x}_1)(D\widetilde{x}_2)}.$$

Докажем это:

$$E(\widetilde{x}_1+\widetilde{x}_2)^2=E(\widetilde{x}_1^2+2\widetilde{x}_1\widetilde{x}_2+\widetilde{x}_2^2)=E\widetilde{x}_1^2+E\widetilde{x}_2^2+2E(\widetilde{x}_1\widetilde{x}_2).$$

Далее

$$[E(\widetilde{x}_1+\widetilde{x}_2)]^2=(E\widetilde{x}_1)^2+(E\widetilde{x}_2)^2+2(E\widetilde{x}_1)(E\widetilde{x}_2),$$

откуда

$$\begin{gathered}D(\widetilde{x}_1+\widetilde{x}_2)=E(\widetilde{x}_1+\widetilde{x}_2)^2-[E(\widetilde{x}_1+\widetilde{x}_2)]^2=[E\widetilde{x}_1^2-(E\widetilde{x}_1)^2]+\\+[E\widetilde{x}_2^2-(E\widetilde{x}_2)^2]+2[E(\widetilde{x}_1\widetilde{x}_2)-(E\widetilde{x}_1)(E\widetilde{x}_2)]=\\=D\widetilde{x}_1+D\widetilde{x}_2+2\operatorname{Cov}(\widetilde{x}_1,\widetilde{x}_2).\end{gathered}$$

А так как $\operatorname{Cov}(\widetilde{x}_1,\widetilde{x}_2)=\varrho\sqrt{(D\widetilde{x}_1)(D\widetilde{x}_2)}$, то требуемое равенство доказано.

§ 3. Нормальное распределение

Рассмотрим следующую статистическую модель. Пусть $\widetilde{x}_1,\widetilde{x}_2,\ldots,\widetilde{x}_n$ — очень большое число независимых дискретных случайных величин, каждая из которых может принимать равновероятно одно из двух значений: $-\varepsilon$ и ε. Выясним, как распределена случайная величина $\widetilde{x}$, являющаяся суммой случайных величин $\widetilde{x}_i$: $\widetilde{x}_1+\widetilde{x}_2+\ldots+\widetilde{x}_n$.

Если $(n/2+k)$ случайных величин $\widetilde{x}_i$ приняли значение ε, а остальные $(n/2-k)$ приняли значение $-\varepsilon$, то случайная величина $\widetilde{x}$ примет значение $x=\mu+2\varepsilon k$, где μ — математическое ожидание $\widetilde{x}$. Иначе говоря, для того чтобы $\widetilde{x}$ приняла значение x, нужно, чтобы $k=\dfrac{x-\mu}{2\varepsilon}$. В силу биномиального закона вероятность этого будет

$$p_k=\frac{n!}{(n/2+k)!(n/2-k)!}\left(\frac{1}{2}\right)^n.$$

Если $(n/2+k+1)$ случайных величин приняли значение ε, а остальные $(n/2-k-1)$ приняли значение $-\varepsilon$, то случайная величина $\widetilde{x}$ примет следующее по величине значение $x'=\mu+2\varepsilon k+2\varepsilon$, большее x на 2ε. Вероятность этого равна

$$p_{k+1}=\frac{n!}{(n/2+k+1)!(n/2-k-1)!}\left(\frac{1}{2}\right)^n.$$

Отношение вероятностей p_{k+1} и p_k равно

$$\frac{p_{k+1}}{p_k} = \frac{(n/2+k)!(n/2-k)!}{(n/2+k+1)!(n/2-k-1)!} =$$

$$= \frac{(n/2+k)!(n/2-k)!(n/2-k-1)!}{(n/2+k+1)!(n/2+k)!(n/2-k-1)!} = \frac{n/2-k}{n/2+k+1},$$

а относительный прирост вероятности равен

$$\frac{\Delta p_k}{p_k} = \frac{p_{k+1}-p_k}{p_k} = \frac{p_{k+1}}{p_k} - 1 = \frac{n/2-k}{n/2+k+1} - 1 = -\frac{2k+1}{n/2+k+1}.$$

Если n велико, то очевидно, что случаи, когда $k = 0$ или близко нулю, а также случаи, когда $k = n$ или близко n, достаточно редки; их вероятности близки нулю. Поэтому для простоты рассмотрим случаи, когда k значительно больше нуля, но значительно меньше n, т. е. $0 \ll k \ll n$. Тогда

$$\frac{\Delta p_k}{p_k} \approx -\frac{4k}{n}.$$

Величина

$$\Delta p_k = p_{k+1} - p_k = \frac{p_{k+1}-p_k}{(k+1)-k},$$

следовательно, приближенно равна производной $\frac{dp_k}{dk}$, если p_k рассматривать как функцию непрерывного аргумента k. Отсюда имеем:

$$\frac{dp_k}{p_k dk} \approx -\frac{4k}{n}.$$

Решим это уравнение, полагая его точным:

$$\frac{dp_k}{p_k} = -4\frac{k}{n}dk,$$

откуда $\ln p_k = -2k^2/n + \ln C$, где C — произвольная константа. Следовательно,

$$p_k = Ce^{-2k^2/n},$$

или, заменив k на $\dfrac{x-\mu}{2\varepsilon}$ и обозначив значение p_k через $f(x)$:

$$f(x) = Ce^{-(x-\mu)^2/2\sigma^2},$$

где $\sigma^2 = \varepsilon^2 n$.

Константа C определяется из условия нормировки: $\int_{-\infty}^{\infty} f(x)dx = 1$. Из курса математического анализа известно, что $\int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2}dx = \sqrt{2\pi}$, откуда $C = \dfrac{1}{\sigma\sqrt{2\pi}}$. Таким образом, плотность распределения случайной величины $\widetilde{x}$ имеет вид:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

где μ и σ — параметры распределения. Это распределение называется *нормальным*, или *гауссовым* (рис. 16, 17). Выше мы ввели μ как математическое ожидание $\widetilde{x}$; далее будет показано, что σ^2 есть дисперсия $\widetilde{x}$, а σ — его среднее квадратичное отклонение.

Описанная нами модель является частным случаем более общей ситуации. Может быть сформулировано и доказано утверждение (*центральная предельная теорема*), которое в нестрогой форме звучит следующим образом.

Если случайная величина $\widetilde{x}$ представляет собой сумму большого числа взаимно независимых случайных величин, влияние каждой из которых на всю сумму ничтожно мало, то $\widetilde{x}$ имеет распределение, близкое к нормальному.

Нормальное распределение занимает центральное место в теории математической статистики и в ее биологических приложениях.

Почему статистические модели, основанные на нормальном распределении, находят широкое применение в биометрии? В значительной мере, по-видимому, потому, что в биологии очень часты ситуации, близкие к условиям центральной предельной теоремы. Значения различных признаков, учитываемых биологом, есть реализация длительного и сложного пути их становления на клеточном, организменном, популяционном или биогеоценотическом уровнях организации жизни. Процесс формирования признака включает множество относительно простых действий, каждое из которых подвержено массе случайных, разнонаправленных флуктуаций.

В самом деле, для многих количественных, или мерных, признаков (например, рост, масса, количество определенного продукта и т. п.), могущих

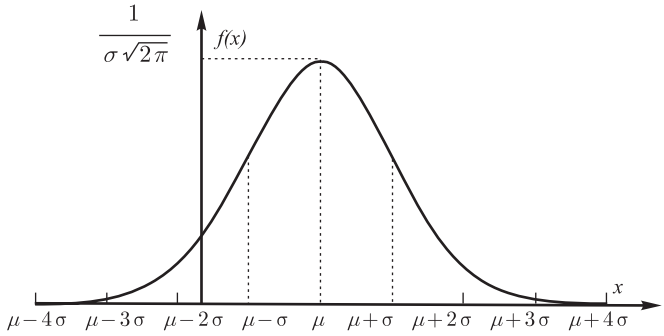


Рис. 16. Плотность нормального распределения. Можно видеть, что $f(x)$ положительна, одновершинна, с максимумом при $x = \mu$ и асимптотически стремится к $-\infty$ и $+\infty$. Кривая имеет две точки перегиба: при $x = \mu - \sigma$ и $x = \mu + \sigma$; она симметрична относительно перпендикуляра, опущенного из вершины на ось абсцисс

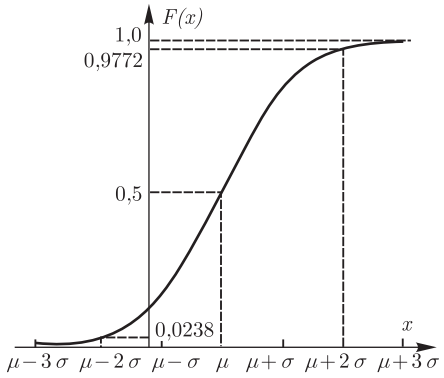


Рис. 17. Функция нормального распределения

принимать непрерывный ряд значений, их распределения близки к нормальному. Это подтверждает анализ фактических данных в больших (несколько тысяч наблюдений) выборках. Например, на рис. 18 показано выборочное распределение по росту 8 585 взрослых мужчин. Налицо хорошее согласие

эмпирического распределения с теоретическим нормальным распределением.

Когда распределение количественного признака отличается от нормального, внимательный анализ выборки позволяет выявить и устранить причину расхождений, нередко заключающуюся в неоднородности исходного материала. На рис. 19 приведены результаты известных опытов В. Иоганнсена, из которых в генетике возникло представление о «чистых линиях». Распределение массы 598 бобов фасоли явно двухвершинно. Однако автору удалось показать, что это является следствием неоднородности — генетической гетерогенности — популяции, состоящей из нескольких «чистых линий». Дальнейший анализ показал, что внутри «чистых линий» распределение признака нормально.

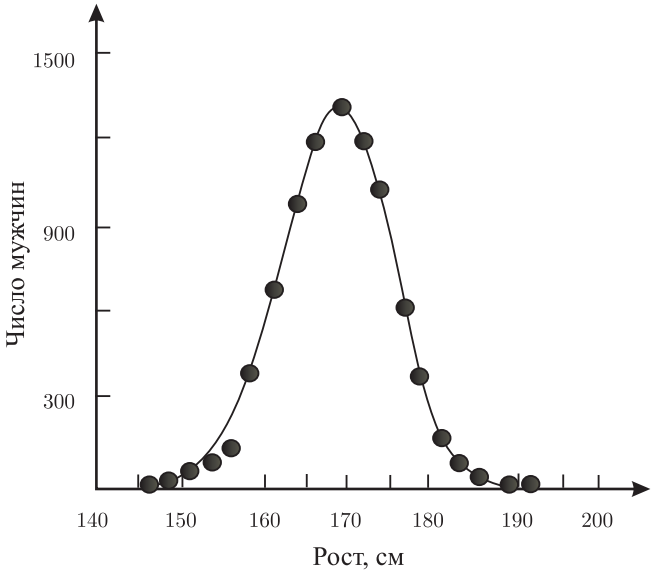


Рис. 18. Распределение 8 585 взрослых мужчин по росту (Дж. О. Юл, М. Дж. Кендал, 1960 г.). Кружками показаны эмпирические данные, непрерывная линия — плотность соответствующего нормального распределения

В других случаях, когда материал несомненно однороден, но распределение признака отличается от нормального, с помощью надлежащего преоб-

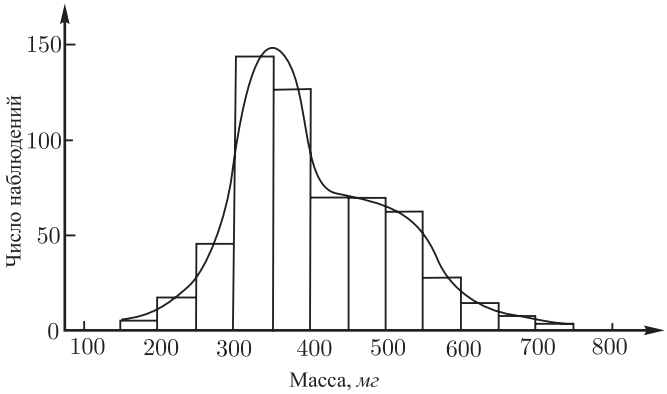


Рис. 19. Распределение массы бобов фасоли в опытах В. Иоганнсена [ван дер Варден, 1960]. Двухвершинность является следствием того, что исходный материал представляет собой смесь нескольких «чистых линий», для каждой из которых характерно нормальное распределение

разования шкалы измерений можно прийти к нормальному распределению. Иногда вид такого преобразования можно получить, исходя из некоторых эвристических рассуждений. Пусть, например, известно, что рост и ряд других промеров распределены по нормальному закону, а масса распределена не нормально. Поскольку масса (x) — это, грубо говоря, произведение трех линейных размеров, то можно ввести новый признак $\sqrt[3]{x}$, который должен быть ближе к нормально распределенному, чем исходный. Практически так и получается. Часто используют также логарифмическое преобразование: вместо значений x анализируют $\log x$.

§ 4. Свойства нормального распределения

1. Поскольку $f(x)$ — плотность распределения, то для любого μ и любого $\sigma > 0$

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx = 1.$$

По сути дела, это было введено в статистическую модель, когда мы провели нормировку, выбирая значение константы C. Таким образом,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2/2\sigma^2} dt$$

— функция нормального распределения.

2. Если $\widetilde{x} \sim N(\mu;\sigma)$ (читается: случайная величина $\widetilde{x}$ распределена нормально с параметрами μ и σ), то

$$E\widetilde{x} = \mu; \quad D\widetilde{x} = \sigma^2.$$

Докажем это. По определению

$$E\widetilde{x} = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} x dx.$$

Сделаем замену переменных $y = \dfrac{x-\mu}{\sigma}$. Тогда

$$E\widetilde{x} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (y\sigma + \mu) e^{-y^2/2} dy =$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-y^2/2} dy + \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy =$$

$$= \frac{\sigma}{2\pi} \int_{-\infty}^{\infty} e^{-y^2/2} dy^2/2 + \mu = -\frac{\sigma}{\sqrt{2\pi}} e^{-y^2/2} \Big|_{-\infty}^{\infty} + \mu = \mu.$$

Далее

$$D\widetilde{x} = E(\widetilde{x} - \mu)^2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2} dx =$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-y^2/2} dy^2/2 =$$

$$= -\frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y de^{-y^2/2}.$$

Интегрируя по частям, получим:

$$D\widetilde{x} = -\frac{\sigma^2}{\sqrt{2\pi}} y e^{-y^2/2}\Big|_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} + \int_{-\infty}^{\infty} e^{-y^2/2} dy = \sigma^2.$$

3. Очевидно, что по аналогии с другими непрерывными распределениями

$$\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx = \Phi(b) - \Phi(a).$$

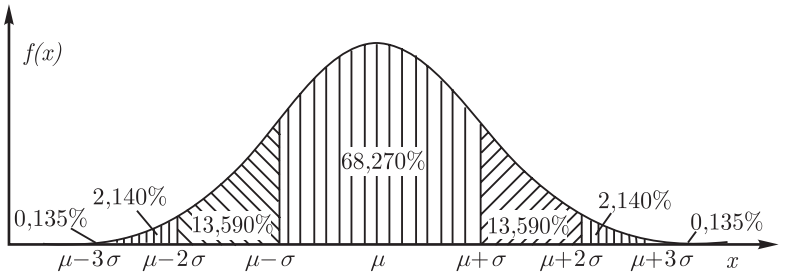


Рис. 20. Плотность нормального распределения со средним значением μ и дисперсией σ^2. Цифры означают процент площади, занимаемой данной частью распределения

4. При рассмотрении теоремы III-1 отмечалось, что для некоторых типов распределений характерна более тесная группировка значений случайной величины вокруг математического ожидания, чем в общем случае. На рис. 20 можно видеть, что

— $68,27\%$ нормального распределения сосредоточено в окрестности $\mu \pm \sigma$;
— $95,45\%$ — в окрестности $\mu \pm 2\sigma$;
— $99,73\%$ — в окрестности $\mu \pm 3\sigma$.

5. Вследствие симметрии нормального распределения

$$\Phi(x) = 1 - \Phi(-x).$$

6. Функция $\Phi(x)$ не выражается через элементарные функции, поэтому ее вычисление будет вызывать каждый раз технические трудности. Однако выход из положения очень прост: нормальное распределение с любыми μ и σ можно свести с помощью уже использовавшейся замены переменных $\widetilde{u} = \frac{\widetilde{x} - \mu}{\sigma}$ к $\Phi(u)$:

$$\Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2} dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-u^2/2\sigma^2} du =$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-t^2/2} dt = \Phi(u).$$

Очевидно, что новая случайная величина $\widetilde{u} \sim N(0;1)$, ее обозначают как *нормированное нормальное распределение*. Это четная функция, т. е. $f(u) = f(-u)$, симметричная относительно оси ординат. Поэтому

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-u^2/2} du = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-u^2/2} du = \frac{1}{2}.$$

Таким образом, достаточно вычислить таблицы для функции нормированного нормального распределения $\Phi(u)$, чтобы вычислять $\Phi(x)$ для нормального распределения с произвольными μ и σ (см. рис. 21, 22 и табл. II Приложения 1).

Приведем, наконец, без доказательства следующую важную теорему.

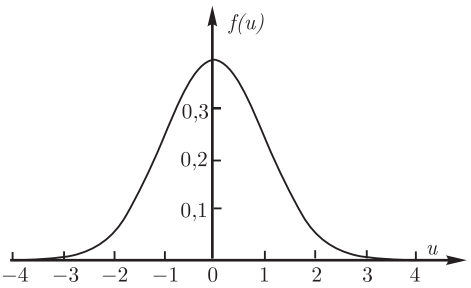


Рис. 21. Плотность нормированного нормального распределения

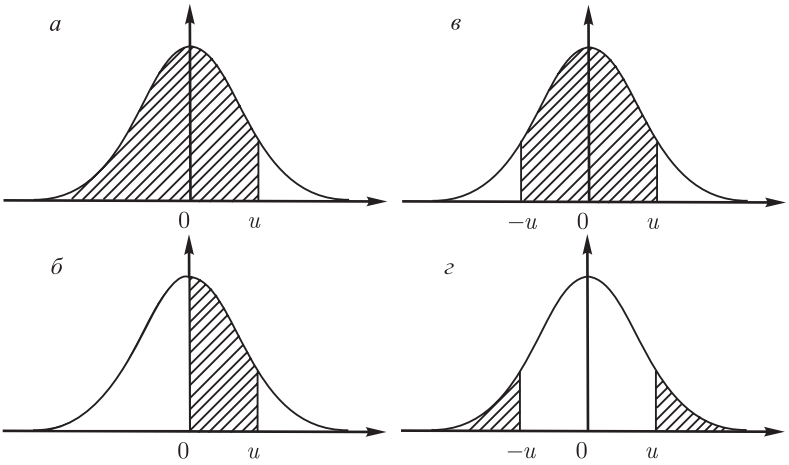


Рис. 22. В разных статистических таблицах и учебниках биометрии интеграл нормального распределения может даваться при разных пределах интегрирования. *Функция нормального распределения, пределы интегрирования*: *а* — от $-\infty$ до u (см. табл. II Приложения 1); *б* — от 0 до u; *в* — от $-u$ до u; *г* — удвоенный интеграл от $-\infty$ до $-u$ (или от u до ∞)

Теорема III-3. *Пусть* $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_k$ *— независимые нормально распределенные случайные величины с параметрами* $(\mu_1, \sigma_1), (\mu_2, \sigma_2), \ldots, (\mu_k, \sigma_k)$ *соответственно. Случайная величина*

$$\widetilde{x} = a_0 + a_1\widetilde{x}_1 + \ldots + a_k\widetilde{x}_k$$

будет тогда распределена нормально с параметрами (μ, σ)*, где*

$$\mu = a_0 + a_1\mu_1 + \ldots + a_k\mu_k;$$
$$\sigma^2 = a_1\sigma_1^2 + \ldots + a_k\sigma_k^2.$$

Рассмотренные свойства нормального распределения частично уже разъясняют, почему это распределение занимает центральное положение в теории математической статистики. Дополнительные аргументы будут приведены в § 5 и § 8 этой главы. Однако полностью ситуация станет ясной лишь после проработки всего курса.

§ 5. Аппроксимация биномиального и пуассоновского распределений нормальным распределением

Упомянем, наконец, еще одно обстоятельство, связанное с широким использованием в биометрии нормального распределения. Выше уже отмечалось, что качественные альтернативные признаки нередко имеют распределение, близкое к биномиальному (§ 5 гл. II), а редкие события часто описываются распределением Пуассона (§ 4 гл. II).

ПРИМЕР III-1. В большой популяции, состоящей из десятков тысяч особей, имеются особи с темной и светлой окраской в равном соотношении. Из популяции извлекается случайная выборка в 100 особей. Какова вероятность того, что число темноокрашенных особей будет не меньше 45 и не больше 55? Налицо урновая схема с $p = q = 1/2$ и $n = 100$. Число темноокрашенных особей в выборке — это реализация случайной величины $\widetilde{x}$, распределенной по биномиальному закону. Единственное отступление от этой схемы — невозвращение отловленных особей в популяцию, однако отловленная часть (100 особей) столь мала по сравнению с численностью всей популяции, что можно пренебречь этим отклонением от схемы выбора с возвращением. Искомая вероятность

$$P\{45 \leqslant \widetilde{x} \leqslant 55\} = \sum_{k=45}^{5} 5P\{\widetilde{x} = k\} = \sum_{k=45}^{5} 5\frac{100!}{(100-k)!k!}\left(\frac{1}{2}\right)^{100}.$$

Очевидно, что найти все члены этой суммы технически сложно. С подобными вычислительными затруднениями сталкиваются и при анализе распределения Пуассона. Хотелось бы как-то упростить вычисления.

Справедливы следующие утверждения относительно пуассоновского и биномиального распределений:

1. Пусть случайная величина $\widetilde{x}$ распределена по закону Пуассона с параметром λ. Тогда при $\lambda \to \infty$

$$P\{a \leqslant \widetilde{x} \leqslant b\} \approx \Phi(u_2) - \Phi(u_1),$$

где $u_2 = \dfrac{b - \lambda + 0,5}{\sqrt{\lambda}}$; $u_1 = \dfrac{a - \lambda - 0,5}{\sqrt{\lambda}}$. Обычно эту аппроксимацию используют при $\lambda > 9$.

2. Пусть случайная величина $\widetilde{x}$ распределена по биномиальному закону с параметрами p и n. Тогда при $n \to \infty$

$$P\{a \leqslant \widetilde{x} \leqslant b\} \approx \Phi(u_2) = \Phi(u_1),$$

где $u_2 = \dfrac{b - np + 0,5}{\sqrt{npq}}$; $u_1 = \dfrac{a - np - 0,5}{\sqrt{npq}}$. Эта аппроксимация применяется обычно при $npq > 9$.

Константы $+0,5$ и $-0,5$ в приведенных выше формулах улучшают аппроксимацию дискретного распределения непрерывным и называются *поправками на дискретность*; естественно, ими можно пренебречь при больших n.

Следует, однако, подчеркнуть, что эти аппроксимации справедливы асимптотически. При небольших значениях параметров λ и n указанные приближения могут быть очень неточными.

Используя полученные результаты, обратимся к решению примера III − 1.

Пример III-1 (окончание). Имеем $npq = 100 \cdot \frac{1}{2} \cdot \frac{1}{2} = 25 > 9$, следовательно, утверждение 2 применимо. В силу этого имеем $np = 50$, $\sqrt{npq} = 5$, откуда $P\{45 \leqslant \widetilde{x} \leqslant 55\} \approx \Phi(u_2) - \Phi(u_1)$, где $u_2 = 1,1$ и $u_1 = -1,1$. Из табл. II Приложения 1 находим $\Phi(1,1) = 0,864$; $\Phi(-1,1) = 0,136$. Следовательно, вероятность того, что число темноокрашенных особей в выборке объема 100 находится в пределах от 45 до 55, равна $P = 0,864 - 0,136 = 0,728$.

§ 6. Двумерное нормальное распределение

Аналогично одномерному случаю можно ввести плотность двумерного нормального распределения. Несмотря на некоторую сложность (вернее, громоздкость) последующих выкладок, обсуждение возникающих здесь вопросов крайне важно в связи с очень частыми в биометрии задачами изучения корреляции и репрессии.

Получим вначале выражение плотности двумерного нормального распределения для независимых случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$, плотности которых суть

$$f_1(x_1) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-(x_1-\mu_1)^2/2\sigma_1^2} \quad \text{и} \quad f_2(2_1) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-(x_2-\mu_2)^2/2\sigma_2^2}.$$

По теореме III-2 плотность их совместного распределения будет

$$f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2) =$$
$$= \frac{1}{\sigma_1 \sigma_2 2\pi} \exp\left\{ -\frac{1}{2} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\}.$$

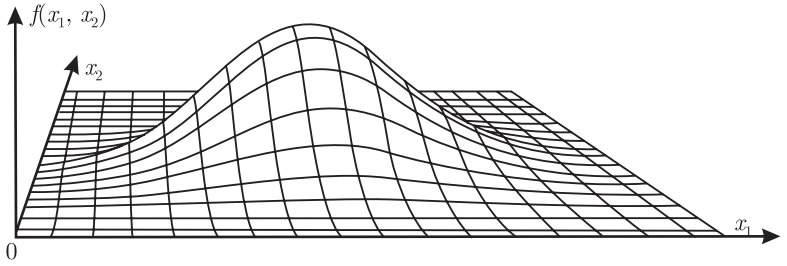


Рис. 23. Поверхность, описываемая плотностью двумерного нормального распределения независимых случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$ [Sokal, Rohlf, 1969]

Вид функции $f(x_1, x_2)$ показан на рис. 23. Поверхность, описываемая функцией $f(x_1, x_2)$, имеет колоколообразную форму; вершина — максимальное значение — отвечает значениям $x_1 = \mu_1$, $x_2 = \mu_2$. Линии $f(x_1, x_2) = C$, образованные пересечением поверхности $f(x_1, x_2)$ с плоскостями, параллельными $x_1 0 x_2$, представляют собой при разных C вложенные друг в друга эллипсы (так называемые контурные эллипсы) с центром (μ_1, μ_2) и осями, параллельными координатным осям (рис. 24). Действительно, уравнение $f(x_1, x_2) = C$ эквивалентно уравнению

$$\left(\frac{x_1 - \mu_1}{\sigma_1}^2 \right) + \left(\frac{x_2 - \mu_2}{\sigma_2}^2 \right) = -2 \ln(2\pi\sigma_1\sigma_2 C).$$

Из аналитической геометрии известно, что общее уравнение эллипса, не обязательно параллельного координатным осям, с центром в точке (μ_1, μ_2) описывается с учетом дополнительного члена

$$\frac{x_1 - \mu_1}{\sigma_1} \cdot \frac{x_2 - \mu_2}{\sigma_2}$$

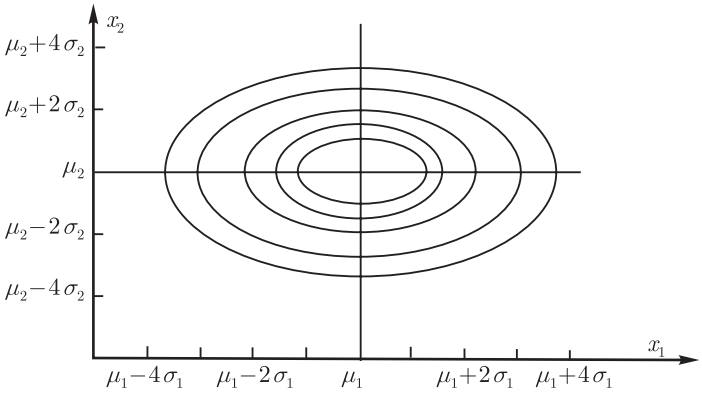


Рис. 24. Контурные эллипсы для двумерного нормального распределения, независимых случайных величин [Хальд, 1956]

с произвольным коэффициентом. Поэтому в общем случае плотность нормального распределения случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$ можно задать в следующем виде:

$$f(x_1, x_2) = \frac{A}{2\pi\sigma_1\sigma_2} e^{-BQ/2},$$

где

$$Q = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + C\frac{x_1 - \mu_1}{\sigma_1} \cdot \frac{x_2 - \mu_2}{\sigma_2} + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2$$

— так называемая квадратичная форма переменных x_1 и x_2, константа $A \geqslant 0$.

Пока еще неопределенные константы A, B и C следует выбрать так, чтобы $f(x_1, x_2)$ являлась плотностью, т. е. потребуем выполнения условия:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)dx_1dx_2 = 1. \tag{1}$$

Потребуем также, чтобы параметры σ_1^2 и σ_2^2, как и раньше, являлись дисперсиями соответствующих случайных величин:

$$D\widetilde{x}_1 = \sigma_1^2; \quad D\widetilde{x}_2 = \sigma_2^2. \tag{2}$$

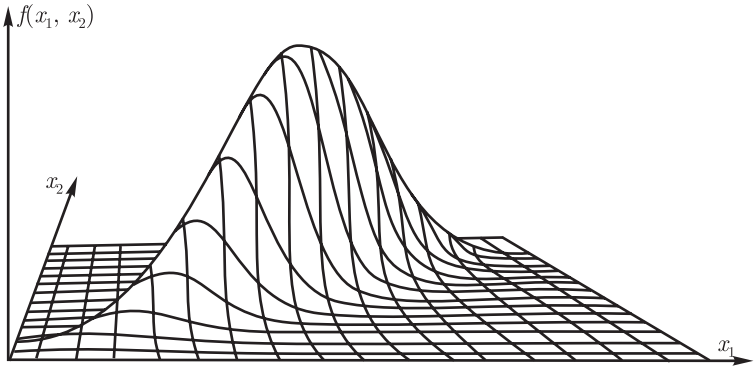


Рис. 25. Поверхность, описываемая плотностью двумерного нормального распределения коррелированных случайных величин, $\rho = 0{,}9$, — ср. рис. 23 [Sokal, Rohlf, 1969]

И, наконец, константы A, B и C должны быть такими, чтобы

$$\operatorname{Cov}(\widetilde{x}_1, \widetilde{x}_2) = \rho\sigma_1\sigma_2. \tag{3}$$

Из этих трех условий можно получить выражения для искомых констант, откуда следует формула для плотности двумерного нормального распределения:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$
$$\times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \right.\right.$$
$$\left.\left. + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right] \right\},$$

где $\exp z = e^z$.

В § 2 говорилось, что для случайных величин, распределенных нормально, равенство коэффициента корреляции нулю означает независимость случайных величин. Действительно, если в приведенном уравнении положить $\rho = 0$, то плотность совместного распределения $\widetilde{x}_1$ и $\widetilde{x}_2$ равна произведению плотностей маргинальных случайных величин.

На рис. 25 показана поверхность $f(x_1, x_2)$ при $\rho = 0{,}9$ и тех же самых σ_1^2 и σ_2^2, что и на рис. 23, где $\rho = 0$. Оси контурных эллипсов при $\rho \neq \infty$ не параллельны осям координат. При фиксированных σ_1 и σ_1 эксцентриситет эллипса — его форма — определяется значениями ρ (рис. 26).

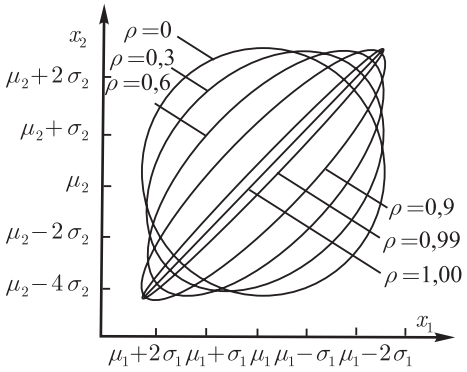


Рис. 26. Контурные эллипсы для двумерного нормального распределения, $\sigma_1 = \sigma_2$ [Хальд, 1956]. Можно видеть, что с увеличением ρ возрастает эксцентриситет эллипса; при $\rho = 1$ эллипс вырождается в прямую

§ 7. Распределение χ^2

Многие задачи биометрии сводятся к задачам следующего типа. Имеются независимые нормально распределенные случайные величины $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_\nu$ с математическими ожиданиями μ_i и дисперсиями σ_i^2 ($i = 1, 2, \ldots, \nu$). Каково распределение новой случайной величины

$$\sum_{i=1}^{\nu} \left(\frac{\widetilde{x}_i - \mu_i}{\sigma_i} \right)^2 ?$$

Поскольку каждая из величин $\widetilde{u}_i = \dfrac{\widetilde{x}_i - \mu_i}{\sigma_i}$ распределена нормально с параметрами 0 и 1 (см. § 4 этой главы), то задача ставится таким образом: найти распределение случайной величины

$$\widetilde{u}_1^2 + \widetilde{u}_2^2 + \ldots + \widetilde{u}_\nu^2,$$

где независимые случайные величины $\widetilde{u}_i \sim N(0;1)$. Эта задача была решена в 1900 г. английским статистиком Карлом Пирсоном, нашедшим плотность и функцию распределения, которое он назвал χ^2-распределением (хи-квадрат). Единственным параметром распределения хи-квадрат является число степеней свободы ν.

Мы не будем рассматривать в дальнейшем вопросов, в которых использовались бы вид плотности и функции распределения χ^2. Кроме того, для разных значений ν имеются таблицы (см. табл. V Приложения 1). Поэтому дадим лишь качественную характеристику плотности. Поскольку величина χ^2 не отрицательна, то это распределение сосредоточено на полуоси $[0,\infty)$. При $\nu = 1, 2$ плотности χ^2-распределения монотонно убывают, при $\nu > 2$ это унимодальные асимметричные кривые (рис. 27, 28). Асимметрия уменьшается с увеличением ν, и при достаточно больших ν (порядка 100) распределение χ^2 аппроксимируется нормальным распределением (вновь нормальным!) с математическим ожиданием ν и дисперсией 2ν. Аппроксимация еще лучше для преобразованной случайной величины

$$\sqrt{2\widetilde{\chi^2}} \sim N(\sqrt{2\nu - 1}, 1).$$

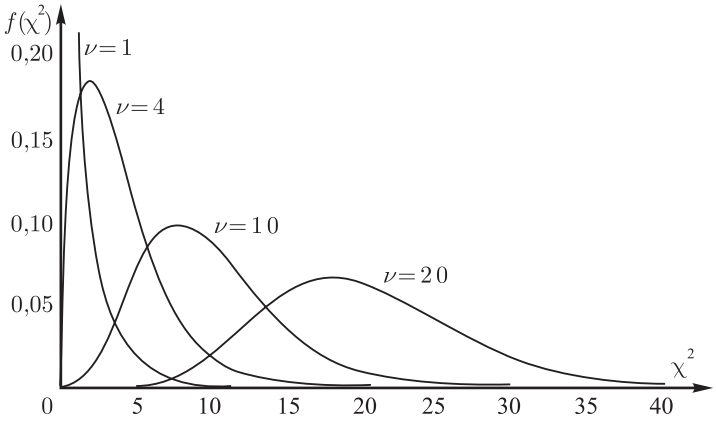


Рис. 27. Плотность χ^2-распределения при разном числе степеней свободы ν

Поэтому случайная величина

$$\widetilde{u} = (\sqrt{2\widetilde{\chi^2}} - \sqrt{2\nu - 1}) \sim N(0;1).$$

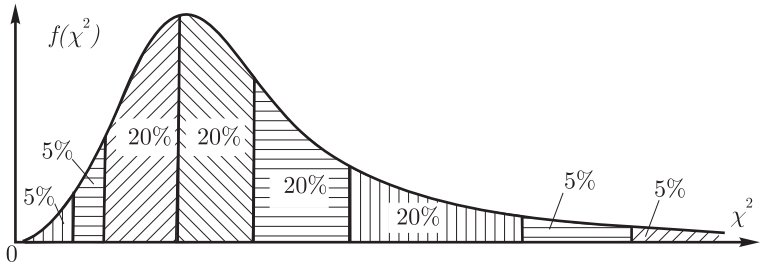


Рис. 28. Плотность χ^2-распределения при $\nu = 5$. Цифры указывают процент площади, занимаемой данной частью распределения

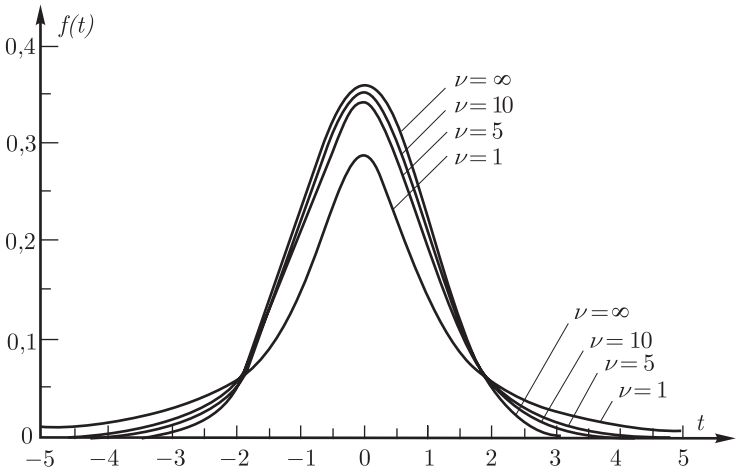


Рис. 29. Плотность t-распределения при разном числе степеней свободы ν. Когда $\nu \to \infty$, t-распределение аппроксимируется нормированным нормальным

Из определения χ^2 следует важная теорема.

Теорема III-4. *Если $\widetilde{\chi}_1^2, \widetilde{\chi}_2^2, \ldots, \widetilde{\chi}_k^2$ — независимые случайные величины, имеющие распределение χ^2 с числом степеней свободы, соответственно, $\nu_1, \nu_2, \ldots, \nu_k$, то случайная величина*

$$\widetilde{\chi}^2 = \widetilde{\chi}_1^2 + \widetilde{\chi}_2^2 + \ldots + \widetilde{\chi}_k^2$$

распределена также по закону χ^2 *с числом степеней свободы*

$$\nu = \nu_1 + \nu_2 + \ldots + \nu_k.$$

§ 8. Распределение Стьюдента

Наряду с нормальным и χ^2-распределением часто встречается иное распределение, являющееся комбинацией двух указанных. Пусть $\widetilde{u} \sim \sim N(0;1)$ и $\widetilde{\chi}^2$ — случайная величина, распределенная по закону χ^2 с числом степеней свободы ν. Если $\widetilde{u}$ и $\widetilde{\chi}^2$ независимы, то распределение отношения

$$\widetilde{t} = \frac{\widetilde{u}}{\sqrt{\widetilde{\chi}^2/\nu}}$$

называется распределением Стьюдента, или t-распределением. Единственный параметр этого распределения — число степеней свободы ν. Оно также табулировано (см. табл. III Приложения 1).

Распределение Стьюдента симметрично относительно оси ординат (его математическое ожидание равно нулю). Вид распределения показан на рис. 29. При достаточно больших ν (практически при $\nu > 30$) распределение Стьюдента аппроксимируется нормированным нормальным распределением.

§ 9. Распределение Снедекора–Фишера

Рассмотрим, наконец, еще одно важное распределение, встречающееся в биометрии. Пусть $\widetilde{\chi}_1^2$ и $\widetilde{\chi}_2^2$ — независимые случайные величины, имеющие распределение χ^2 с числом степеней свободы ν_1 и ν_2 соответственно. Тогда случайная величина

$$\widetilde{F} = \frac{\widetilde{\chi}_1^2/\nu_1}{\widetilde{\chi}_2^2/\nu_2}$$

имеет *распределение Снедекора–Фишера*, или F-распределение. Это распределение в отличие от χ^2- и t-распределений имеет два параметра: число степеней свободы числителя и число степеней свободы знаменателя. Из определения следует, что случайная величина $1/\widetilde{F}$ также имеет t-распределение, но с числом степеней свободы ν_1 и ν_2. Общий вид t-распределения показан на рис. 30. Оно табулировано (см. табл. IV Приложения 1).

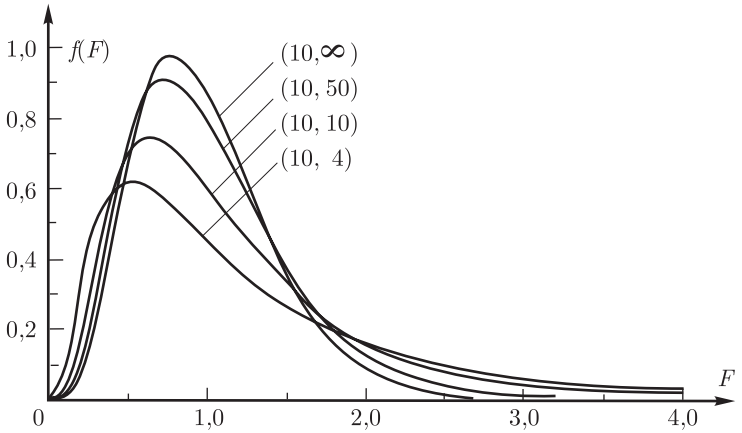


Рис. 30. Плотность F-распределения. В скобках указаны значения числа степеней свободы ν_1 и ν_2

§ 10. Взаимосвязи между различными распределениями

Выше неоднократно говорилось о предельных переходах между распределениями, об аппроксимации одного распределения другим. Было показано, что дискретные распределения, такие как биномиальное и пуассоновское, при определенных условиях достаточно точно аппроксимируются непрерывным нормальным распределением (§ 5) и что распределение Пуассона есть предельный случай биномиального распределения (гл. II, § 5).

Укажем еще два важных соотношения между дискретными (биномиальным и пуассоновским) и непрерывными (F и χ^2 соответственно) распределениями, которые выполняются точно.

1. Если случайная величина $\widetilde{x}$ имеет биномиальное распределение с параметрами p и n, то она связана со случайными величинами F_1 и F_2 соотношениями

$$P\{\widetilde{x} \leqslant k\} = 1 - P\left\{\frac{(k+1)\cdot \widetilde{F}_1}{(n-k)+(k+1)\cdot \widetilde{F}_1}\right\}$$

и

$$P\{\widetilde{x} \geqslant k\} = 1 - P\left\{\frac{k}{k+(n-k+1)\cdot \widetilde{F}_2}\right\},$$

где $\widetilde{F}_1$ — случайная величина, имеющая F-распределение с параметрами $\nu_1{=}2(k{+}1)$ и $\nu_2{=}2(n{-}k)$, а случайная величина $\widetilde{F}_2$ имеет F-распределение с параметрами $\nu_1 = 2(n-k+1)$ и $\nu_2 = 2k$.

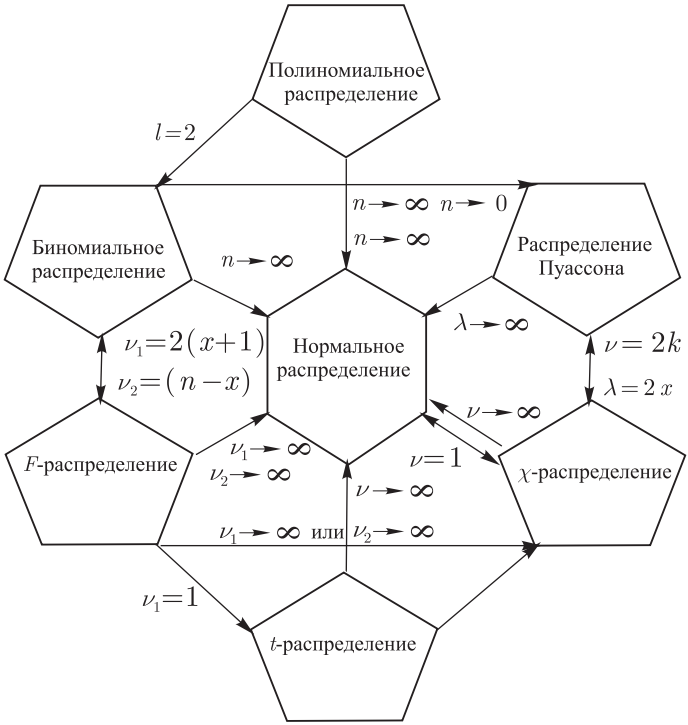


Рис. 31. Взаимосвязи между распределениями. Рядом со стрелками указаны условия предельных переходов или взаимно-однозначных соответствий

Отсюда следует, что для вычисления вероятностей $P\{x \leqslant k\}$ и $P\{x \geqslant k\}$ можно использовать значения функции F-распределения, которое, как уже говорилось, достаточно подробно табулировано (табл. IV Приложения 1).

2. Если случайная величина $\widetilde{x}$ подчиняется распределению Пуассона с параметром k, то она связана со случайными величинами $\widetilde{\chi}^2_1$ и $\widetilde{\chi}^2_2$

соотношениями

$$P\{\widetilde{x} \leqslant k\} = 1 - P\{\widetilde{\chi}_1^2 \leqslant 2\lambda\}$$

и

$$P\{\widetilde{x} \geqslant k\} = P\{\widetilde{\chi}_2^2 \leqslant 2\lambda\},$$

где $\widetilde{\chi}_1^2$ имеет распределение χ^2 с параметром $\nu = 2(k+1)$, а $\widetilde{\chi}_2^2$ имеет распределение χ^2 с параметром $\nu = 2k$. Это означает, что для вычисления указанных вероятностей можно использовать таблицы распределения χ^2 (табл. V Приложения 1). Указанные свойства будут использованы нами в дальнейшем (гл. V, § 5 и § 6).

Основные взаимосвязи между различными распределениями изображены на рис. 31. Видно, что нормальное распределение занимает особое положение, являясь пределом (в смысле теории функций) всех рассмотренных нами распределений.

Задачи

III-1. (Иллюстрация центральной предельной теоремы.) Постройте графики распределений трех случайных величин:

1) числа очков, выпадающих при подбрасывании одной игральной кости;

2) суммы очков на обеих костях при подбрасывании двух;

3) суммы при подбрасывании трех игральных костей.

III-2. Докажите, что для $\widetilde{x}_1$ и $\widetilde{x}_2$ — двух независимых случайных величин — $E(\widetilde{x}_1\widetilde{x}_2) = (E\widetilde{x}_1)(E\widetilde{x}_2)$.

III-3. Докажите, что $E\widetilde{y} = 0$ и $D\widetilde{y} = 1$, где y — нормированная случайная величина.

III-4. Пусть $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$. Найдите:

а) $\Phi(0)$;

б) $\Phi(-1)$;

в) $\Phi(2)$.

III-5. Пусть $\widetilde{u} \sim N(0;1)$. Найдите:

а) $P\{-3 < \widetilde{u} < 2\}$;

б) $P\{-1 < \widetilde{u} < \infty\}$;

в) $P\{|\widetilde{u}| \leqslant 1\}$;

г) $P\{|\widetilde{u}| \leqslant 2\}$;

д) $P\{|\widetilde{u}| \leqslant 3\}$;

е) $P\{|\widetilde{u}| \geqslant 3\}$.

III-6. Пусть $\widetilde{u} \sim N(0;1)$ и $P\{|\widetilde{u}| \leqslant \gamma\} = \beta$. Найдите γ при β, равном

а) 0,90;

б) 0,95;

в) 0,99.

III-7. Пусть $\widetilde{x} \sim N(1;2)$. Найдите $P\{-3 < \widetilde{x} < 7\}$.

III-8. Пусть $\widetilde{x} \sim N(4;2)$. Найдите $P\{|2\widetilde{x} - 7| > 2\}$.

III-9. Пусть $\widetilde{x}_1 \sim N(0;2)$ и $\widetilde{x}_2 \sim N(1;1)$. Найдите $P\{-4 < \widetilde{x}_1 - \widetilde{x}_2 < 8\}$.

III-10. Пусть $\widetilde{x}_1 \sim N(3;3)$ и $\widetilde{x}_2 \sim N(0;4)$. Найдите $P\{|\widetilde{x}_1 + 2\widetilde{x}_2 - 3| \geqslant 8\}$.

III-11. Пусть $\widetilde{x} \sim N(2;5)$ и $P\{|\widetilde{x}| \leqslant \gamma\}$. Найдите γ.

III-12. Докажите, что $E\widetilde{\chi}^2_{(\nu)} = \nu$, $D\widetilde{\chi}^2_{(\nu)} = 2\nu$. Указание: сначала докажите, что $E\widetilde{\chi}^2_{(1)} = 1$ и $D\widetilde{\chi}^2_{(1)} = 2$.

III-13. Пользуясь табл. V Приложения 1, найдите:

а) $P\{\widetilde{\chi}^2 \geqslant 27{,}9\}$, $\nu = 9$;

б) $P\{\widetilde{\chi}^2 < 0{,}004\}$, $\nu = 1$;

в) $P\{\widetilde{\chi}^2 \geqslant 5{,}23\}$, $\nu = 15$.

III-14. Пользуясь табл. III Приложения 1, найдите:

а) $P\{\widetilde{t} \geqslant 2{,}20\}$, $\nu = 11$;

б) $P\{|\widetilde{t}| \geqslant 2{,}20\}$, $\nu = 11$;

в) $P\{\widetilde{t} \leqslant \gamma\} = 0{,}95$, $\nu = 7$, $\gamma = ?$;

г) $P\{\widetilde{t} \leqslant \gamma\} = 0{,}99$, $\nu = 7$, $\gamma = ?$

III-15. Пользуясь табл. IV Приложения 1, найдите:

а) $P\{\widetilde{F} \geqslant 17{,}12\}$, $\nu_1 = 8$, $\nu_2 = 4$;

б) $P\{\widetilde{F} \geqslant 13{,}18\}$, $\nu_1 = 15$, $\nu_2 = 4$;

в) $P\{\widetilde{F} \geqslant 13{,}18\}$, $\nu_1 = 4$, $\nu_2 = 15$.

III-16. Пусть $f(x)$ — плотность распределения случайной величины $\widetilde{x}$, такая, что для нее существует точка симметрии x_0, т. е. для любого $a{>}0$ $f(x_0 + a) = f(x_0 - a)$. Тогда $Ex = x_0$, $\zeta = x_0$. Что можно сказать о нормальном распределении?

Глава IV

Статистические задачи в биологии и основные понятия математической статистики

«Если Эксперимент — Король Наук, то статистические Методы - его Телохранители».

М. Трайбус

Предыдущей главой завершено вероятностное введение в биометрию. Теперь мы приступим к рассмотрению основ математической статистики. Начнем с постановки некоторых наиболее часто встречающихся статистических задач, которые возникают в биологии при проведении количественных экспериментов и наблюдений. В процессе ознакомления с этими задачами будут сформулированы основные понятия математической статистики. Еще раз подчеркнем то, о чем уже шла речь во Введении: количественное исследование начинается с организации эксперимента, включающей и выбор адекватных методов, с помощью которых затем будет производиться анализ результатов.

§ 1. Генеральная совокупность и выборка

Понятия генеральной совокупности и выборки являются начальными, основополагающими понятиями математической статистики.

Рассмотрим урну, содержащую N шаров, из которых N_0 помечены цифрой 0, N_1 — цифрой 1, ..., N_r — цифрой r. Обозначим $p_i = N_i/N$ долю шаров, помеченных цифрой i: $p_i \geqslant 0, i = 0, \ldots, r$; $p_0 + \ldots + p_r = 1$. Из этой урны проведем n-кратный выбор наугад по одному шару с возвращением. В результате получим *выборку* объемом n шаров, содержащую n_i шаров, помеченных цифрой i: $i = 0, \ldots, r$; $n_i \geqslant 0$ (в случайной выборке некоторые n_i могут быть равны нулю!); $n_0 + \ldots + n_r = n$. Обозначим $p_i' = n_i/n$ долю

шаров в выборке, помеченных i: $p'_i \geqslant 0, i = 0, \ldots, r$; $p'_1 + \ldots + p'_r = 1$. Зная состав урны $(N_0, \ldots, N_r)$, можно найти вероятность того, что будет получена выборка данного состава — $P\{n_0, \ldots, n_r\}$.

По отношению к различным возможным выборкам урна, содержащая N шаров, является *генеральной совокупностью*. Понятия генеральной совокупности и выборки оказываются, таким образом, понятиями соотносительными: бессмысленно говорить о совокупности n шаров, что это выборка, без указания на то, откуда выбрана данная совокупность, из какой урны, из какой генеральной, общей совокупности; бессмысленно говорить, что урна данного состава есть генеральная, общая совокупность, без указания на то, для каких выборок эта совокупность является генеральной. Связью, создающей отношения между генеральной совокупностью и выборкой, служит процедура выбора шаров: в рассматриваемом случае это последовательный выбор шаров наугад с возвращением, который далее будет называться *простым случайным выбором*.

Пример IV-1. У 245 юношей, студентов университета, с точностью до 1 см измерялся рост — важнейший антропометрический показатель (см. табл. 3). Требуется охарактеризовать генеральную совокупность, из которой извлечена выборка.

Генеральная совокупность в этом примере — множество всех юношей — студентов данного университета, составляющее несколько тысяч человек. Выборка — те юноши, у которых измерялся рост; объем выборки n равен 245.

Главная задача, которую решает математическая статистика (ради чего, собственно говоря, эта наука и существует), заключается в том, чтобы на основании изучения выборки сделать выводы о свойствах генеральной совокупности. Выборка — лишь часть генеральной совокупности. Поэтому выводы о генеральной совокупности, которые делают на основании изучения выборки, могут иметь лишь вероятностный характер. Мы анализируем выборку для того, чтобы по ее свойствам вывести заключение о свойствах генеральной совокупности. При этом нас не интересует индивидуальная характеристика каждого обследуемого, в статистической задаче речь идет о массовой характеристике совокупности.

Если сначала, собирая материал, исследователь шел от генеральной совокупности к выборке, то цель анализа состоит в переходе от свойств выборки к свойствам генеральной совокупности. Последний шаг в статистической модели возможен только в том случае, если первый шаг в организации эксперимента был сделан правильно, если правильной была процедура взятия выборки. В этом учебном пособии мы рассматриваем лишь простейший

Таблица 3

Результаты антропометрического измерения юношей — студентов университета (к примеру IV-1)

Рост, см; выборка объема $n = 245$												
174	168	167	181	171	182	178	170	178	170	169	170	170
177	169	168	178	172	175	177	174	183	174	173	168	171
193	175	180	168	175	154	167	172	178	173	169	178	172
192	180	168	170	181	182	184	192	173	168	178	182	169
196	180	158	178	165	185	170	180	193	173	167	183	174
177	170	170	180	165	181	165	180	185	161	181	179	
182	168	172	180	185	164	186	184	170	174	178	169	
146	173	177	173	188	164	162	170	170	173	167	173	
175	160	178	175	185	156	170	187	170	170	170	170	
182	182	178	164	188	170	168	170	180	165	168	169	
181	170	180	175	170	173	170	183	174	174	192	176	
176	168	173	163	166	172	167	193	162	174	190	201	
181	169	177	176	166	169	180	185	180	173	172	174	
175	173	173	176	168	177	172	174	184	170	180	189	
178	162	172	168	192	178	165	162	187	178	178	170	
173	165	177	175	180	177	169	172	170	157	179	178	
180	165	173	165	178	167	172	167	183	160	190	178	
183	150	169	180	179	173	175	175	194	171	196	182	
173	158	178	178	177	181	180	190	185	160	175	183	
179	162	167	173	173	172	180	169	176	165	179	176	

метод взятия выборки — простой случайный выбор. Если в примере IV-1 действовать по аналогии с урновой схемой, то надо было бы «пронумеровать» всех студентов, затем случайно, например по жребию, извлечь последовательно с возвращением 245 номеров и измерить рост соответствующих студентов. Подробнее о такого рода процедуре говорится в § 3.

ПРИМЕР IV-2. Этот пример демонстрирует нарушение условий простого случайного выбора. Пусть требуется охарактеризовать рост студентов университета. В качестве выборки, изучение которой позволит сделать заключение о признаке в генеральной совокупности, предлагается взять баскетбольную команду университета. Абсурдность такого подхода очевидна.

Впервые систематическое внимание на приемы организации простого случайного выбора стал обращать английский математик и биолог Р. А. Фишер в 20-е годы нашего столетия при планировании и анализе полевых экспериментов на Ротамстедской сельскохозяйственной опытной стан-

ции. Был проведен ряд опытов, наглядно показавших, что процедура выбора, основанная на субъективном представлении исследователя о «типичности» выбираемых объектов, зачастую не обеспечивает выполнения условий простого случайного выбора. Один из этих опытов, описанных сотрудником Фишера Ф. Иейтсом в 1935 г., заключался в следующем.

Пример IV-III (Ф. Иейтс, по В. Н. Перегудову, 1961 г.). Производилось обследование высоты растений яровой пшеницы в три срока: 31 мая, 14 июня и 28 июня. Цель обследования — установить интенсивность роста растений. Брались 32 пробы, в каждой по 8 растений. Растения отбирались двумя способами:

а) случайным образом — от края делянки отступали на 25 см и измеряли ближайшее растение;

б) определялась высота типичного (с точки зрения исследователя) растения в том же рядке.

Результаты показали систематические различия между двумя способами формирования выборки (табл. 4).

Таблица 4

Средняя высота растений яровой пшеницы (см) при двух способах формирования выборки (к примеру IV-3)

Дата измерения	Случайная выборка	Отбор «типичного» образца	Отклонение
31 мая	47,49	50,82	+3,33
14 июня	76,56	78,29	+1,73
28 июня	118,84	116,12	−2,72

Из таблицы можно видеть, что в первый срок при низкой высоте растений наблюдатели стремятся ее «подтягивать»; в последний срок, когда растения достигали большой высоты, появляется склонность не доверять этому факту. Заметим, что средний прирост различается почти на 10%: $118{,}84 - 47{,}49 = 71{,}35$ см при случайном отборе и $116{,}12 - 50{,}82 = 65{,}30$ см при отборе типичного образца.

Субъективизм экспериментатора особенно четко проявляется в другом, модельном опыте.

Таблица 5

Средняя масса камней (унции[1]) при отборе серий по 20 камней (к примеру IV-4)

Повторность	Наблюдатель											
	1	2	3	4	5	6	7	8	9	10	11	12
1	1,9	2,4	2,4	1,9	2,2	2,3	2,4	1,6	2,2	2,6	2,4	2,4
2	1,8	3,0	2,4	2,0	2,7	2,6	2,6	2,0	2,2	2,2	2,4	3,0
3	1,7	2,4	2,1	2,0	3,1	2,8	2,5	2,0	2,2	3,1	1,8	2,4
Среднее	1,8	2,6	2,3	2,0	2,7	2,7	2,5	1,9	2,2	2,6	2,2	2,6

ПРИМЕР IV-4 (Ф. Иейтс, по В. Н. Перегудову, 1961 г.). На стол положили 1 200 камней различных размеров и массы. Было предложено 12 лицам (каждому трижды) отобрать 20 камней, «типичных» для этой совокупности. После каждого испытания камни возвращались обратно и перемешивались, так что все наблюдатели и каждый из них при повторном отборе находились в одинаковых условиях. Результаты приведены в табл. 5.

Средняя масса всех отобранных камней равна 2,34 унции, точная средняя масса всех камней — 1,91 унции, т. е., в целом, участники эксперимента увеличили среднюю массу камней примерно на 25 %. Обратите внимание, что только у 2 испытуемых из 12 средняя масса оказалась ниже истинной, у остальных 10 — выше. Более интересно, пожалуй, то, что испытуемые явно проявляют «определенные склонности»: первый, например, систематически занижает результаты, четвертый твердо держится на уровне среднего, девятый вообще не показывает изменчивости от опыта к опыту и т. д. Как заметил один статистик, человек — удивительно плохой инструмент для сознательного осуществления случайной выборки.

Правда, в целом ряде случаев, особенно в медицинских и сельскохозяйственных опытах, при проведении экспериментов в природе осуществить случайную выборку бывает не так просто. Это требует тщательной организации опыта.

Не следует думать, что в каждой биометрической задаче всегда четко и однозначно очерчена генеральная совокупность. Можно спорить, например, точно ли очерчена генеральная совокупность в примере IV-1: кто-то

[1] 1 унция равна 28,35 г.

будет толковать ее как совокупность юношей-студентов всех университетов такого-то региона, кто-то — всей страны, а кто-то — как совокупность студентов всех вузов. В рассматриваемом ниже примере IV-8, где объектами исследования являются белые крысы, вряд ли кто будет настаивать, что генеральную совокупность составляют разводимые сегодня во всех лабораториях мира все белые крысы. При изучении морфологии краба *Pachygrapsus crassipes* (см. пример IV-7) также, наверное, не идет речь о всех особях вида, биолог-профессионал введет какие-то ограничения на возраст животных, условия их обитания и т. д., и т. п. Постановка подобных вопросов, подчас очень важных, — дело биологов соответствующих специальностей. При построении статистической модели конкретизации понятия «генеральная совокупность» не требуется.

«Понятие бесконечной генеральной совокупности не является логически безупречным и необходимым. Для решения статистических задач нужна не сама генеральная совокупность, а лишь те или иные характеристики соответствующей функции распределения $F(x)$. С точки зрения теории вероятностей, выборка из бесконечной генеральной совокупности представляет собой наблюдаемые значения нескольких случайных величин, имеющих заданный закон распределения»[2].

Генеральная совокупность, таким образом, задается как некоторая генеральная случайная величина $\widetilde{x}$, имеющая функцию распределения $F(x)$, которая может быть известна или не известна исследователю. При таком подходе простой случайный выбор состоит в n-кратном повторении случайного испытания, математической моделью которого является случайная величина $\widetilde{x}$. Иными словами, простой случайный выбор дает n-мерную случайную величину $(\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n)$, каждая компонента $\widetilde{x}_i$ которой имеет функцию распределения $F_i(x)$, совпадающую с функцией распределения $F(x)$ случайной величины $\widetilde{x}$. Из определения простого случайного выбора следует независимость в совокупности случайных величин $\widetilde{x}_1, \ldots, \widetilde{x}_n$, что позволяет записать функцию их совместного распределения:

$$F(x_1, \ldots, x_n) = F(x_1) \cdot \ldots \cdot F(x_n).$$

При этом конкретные значения $x_1, \ldots, x_n$, регистрируемые экспериментатором, рассматриваются как реализации случайных величии $\widetilde{x}_1, \ldots, \widetilde{x}_n$.

Вернемся к примеру IV-1. Разобраться в 245 беспорядочно расположенных числах довольно трудно, поэтому упорядочим их по возрастанию

[2]Большев Л. Н. Генеральная совокупность : Математическая энциклопедия. — М.: Советская энциклопедия, 1977. — Т. 1. — С. 918.

выборочных значений — *вариант*:

$$146 < 150 < 154 \ldots 196 = 196 < 201.$$

Такую упорядоченную последовательность называют *ранжированным рядом*:

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)},$$

где новый индекс (в скобках) есть порядковый номер (ранг) варианты в этом ряду. Поскольку некоторые варианты имеют одинаковые, совпадающие значения (например, 158, 160 и т. д.), бывает удобнее представить данные в виде *вариационного ряда* — таблицы, один столбец которой содержит значения вариант x_i а в другом столбце числа n_i показывают, сколько раз варианта встречается в выборке (табл. 6). По ранжированному ряду можно построить функцию $F_n(x)$. По аналогии с генеральной функцией распределения $F(x)$ естественно назвать $F_n(x)$ *выборочной (эмпирической) функцией распределения*; для примера IV-1 она представлена графически на рис. 32.

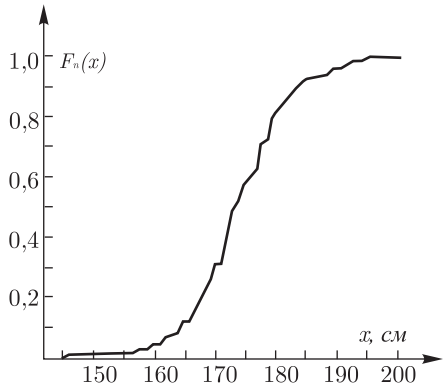


Рис. 32. Рост юношей — студентов университета. Эмпирическая функция распределения

Теперь, наконец, можно поставить вопрос, который, наверное, уже возник у вдумчивого читателя: на каком основании, собственно говоря, мы можем рассчитывать на то, что вариационный ряд, выборочная функция распределения $F_n(x)$ дают информацию о генеральной совокупности, т. е. функции распределения $F(x)$? Строгий ответ на этот вопрос содержится в замечательной теореме, доказанной советским математиком

В. И. Гливенко в 1933 г. После некоторых предварительных рассуждений мы сможем сформулировать теорему, правда, в более слабом виде, чем это сделал ее автор.

Таблица 6

Вариационный ряд роста студентов, $n = 245$ (по данным табл. 3)

Рост (x_i), см	Число студентов (n_i)	Рост (x_i), см	Число студентов (n_i)	Рост (x_i), см	Число студентов (n_i)
146	1	165	9	184	3
147	0	166	2	185	6
148	0	167	8	186	1
149	0	168	12	187	2
150	1	169	11	188	2
151	0	170	24	189	1
152	0	171	2	190	3
153	0	172	11	191	0
154	1	173	19	192	4
155	0	174	10	193	3
156	1	175	11	194	1
157	1	176	6	195	0
158	2	177	9	196	2
159	0	178	19	197	0
160	3	179	5	198	0
161	2	180	17	199	0
162	5	181	7	200	0
163	1	182	7	201	1
164	3	183	6		

Мы строили выборочную функцию распределения $F_n(x)$, используя конкретные числа $x_1, \ldots, x_n$. Поэтому в каждой точке x функция $F_n(x)$ дает числовое значение $y = F_n(x)$, зависящее от чисел $x_1, \ldots, x_n$ как от параметров: $y = y(x_1, \ldots, x_n)$. Однако выборку $(x_1, \ldots, x_n)$ мы договорились рассматривать в качестве реализации случайного вектора $(\widetilde{x}_1, \ldots, \widetilde{x}_n)$. Подставляя в функцию $y = y(x_1, \ldots, x_n)$ вместо числовых парамет-

ров $x_1, \ldots, x_n$ случайные величины $\widetilde{x}_1, \ldots, \widetilde{x}_n$, получаем вместо числа $y = y(x_1, \ldots, x_n)$ случайную величину $\widetilde{y} = \widetilde{y}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, для которой число $y = y(x_1, \ldots, x_n)$ есть просто одна из возможных реализаций. Итак, наряду с выборочной функцией $F_n(x)$ с числовыми параметрами $x_1, \ldots, x_n$ мы получаем *случайную выборочную функцию* $\widetilde{F}_n(x)$ со случайными параметрами $\widetilde{x}_1, \ldots, \widetilde{x}_n$. Случайная функция $\widetilde{F}_n(x)$ сопоставляет аргументу x уже не число, но случайную величину $\widetilde{F}_n(x)$.

Поскольку $\widetilde{F}_n(x)$ есть для каждого x некоторая случайная величина, то случайной величиной будет и наибольшее расхождение $\widetilde{D}_n$ между выборочной и генеральной функциями распределения:

$$\widetilde{D}_n = \sup \left| \widetilde{F}_n(x) - F(x) \right|.$$

Для случайной величины $\widetilde{D}_n$ можно говорить о вероятности события $\{\widetilde{D}_n < \varepsilon\}$, где $\varepsilon > 0$.

Теорема IV-1 (теорема В. И. Гливенко). *Пусть $\widetilde{F}_n(x)$ есть случайная выборочная функция распределения, построенная по случайной выборке $(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ из генеральной совокупности, описываемой функцией распределения $F(x)$. Тогда для любого $\varepsilon > 0$ в любой точке x имеет место соотношение*

$$P\{\widetilde{D}_n < \varepsilon\} \underset{n\to\infty}{\longrightarrow} 1,$$

где $\widetilde{D}_n = \sup \left| \widetilde{F}_n(x) - F(x) \right|$.

Про случайную выборочную функцию распределения $\widetilde{F}_n(x)$ говорят, что она *сходится по вероятности* в любой точке x к генеральной функции распределения $F(x)$.

Если функция $F(x)$ непрерывна, то можно найти предельное распределение величины $\widetilde{D}_n$. Соответствующая теорема принадлежит А. Н. Колмогорову (см. гл. VII, § 4) и занимает такое же положение в теории математической статистики, как центральная предельная теорема (§ 3 гл. III).

§ 2. Анализ одной выборки

Вернемся к примеру IV-1. Задача в примере сформулирована следующим образом: охарактеризуйте генеральную совокупность, из которой

извлечена выборка. Это значит, что нужно найти вид распределения генеральной случайной величины $\widetilde{x}$ и параметры этого распределения. Вначале надо сформулировать гипотезу о виде распределения, а затем проверить ее.

Любое предположение о виде распределения случайной величины или значении ее параметров называют *статистической гипотезой*. Каковы предпосылки для формулировки статистической гипотезы о виде распределения? Гипотеза формулируется на основании некоторых общих соображений о характере признака, структуре эксперимента; на основании аналогичных экспериментов, выполненных ранее; наконец, на основании анализа материалов данного эксперимента. Что мы имеем в примере IV-1? Рост студентов — это количественный признак, а выше говорилось, что количественные признаки часто распределены нормально (см. гл. III, § 3). Ранее приводился также пример почти идеального согласия распределения мужчин по росту с нормальным распределением (см. рис. 18). Обратимся, наконец, к материалам нашей выборки.

Полезно представить имеющиеся данные графически, построив *выборочное распределение*, являющееся аналогом плотности распределения $f(x)$. Для этого, однако, нельзя непосредственно воспользоваться вариационным рядом: число значений x_i для данного объема выборки n слишком велико, что видно хотя бы из обилия нулей и единиц в графе n_i табл. 6. Каким образом следует сгруппировать данные, объединяя соседние значения в классы? Какое выбрать число классов? Какие значения $x - i$ выбрать в качестве границ классов? На эти вопросы нельзя дать однозначный ответ. Часто пользуются эмпирическим правилом: оптимальное число классов лежит между величинами $l_{\min} = 1 + 3{,}22 \lg n$ и $l_{\max} = \sqrt{n}$.

На рис. 33 в виде гистограммы представлены результаты группировки данных примера IV-1 в десять классов: 146–151, 152–157 см и т. д. Середины классовых интервалов равны 148,5; 154,5 см и т. д. Высота столбиков на гистограмме пропорциональна относительной частоте классов в выборке. Можно видеть, что выборочное распределение куполообразное, одновершинное, более или менее симметричное. Следовательно, выборочное распределение дает самое грубое, глазомерное представление о виде распределения генеральной совокупности (ср. задачи IV-1, IV-3, IV-4). В случае примера IV-1, суммируя приведенные выше соображения и результаты, изображенные на рис. 33, можно сказать, что формулировка гипотезы о том, что признак «рост юношей — студентов университета» распределен нормально, представляется вполне разумной. Теперь можно перейти к следующему этапу решения задачи — проверке этой статистической гипотезы.

Проверка статистических гипотез — один из разделов математической статистики.

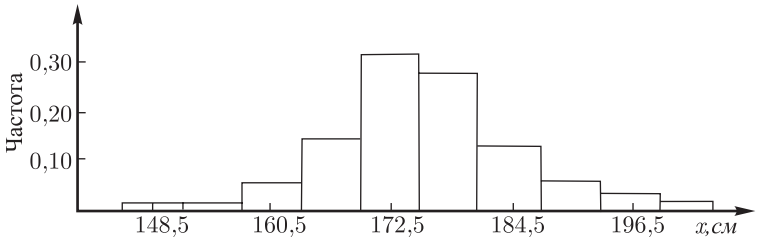


Рис. 33. Рост юношей — студентов университета. Выборочное распределение в виде гистограммы. Выборка разбита на 10 классов. Основание каждого прямоугольника — это длина классового интервала, высота — частота соответствующего класса; таким образом, площадь прямоугольника пропорциональна доле соответствующего класса в выборке

Проверяемая гипотеза называется *нулевой гипотезой* и обозначается H_0. В примере IV-1 H_0: случайная величина «рост студентов университета» имеет нормальное распределение. В результате проверки нулевой гипотезы может быть получено согласие с ней, и тогда нулевая гипотеза будет принята. Может, однако, оказаться, что эмпирические результаты не согласуются с гипотетическими. В этом случае нулевая гипотеза должна быть отвергнута и будет принята некоторая *альтернативная* (*конкурирующая*) *гипотеза* H_1. К такому исходу статистического анализа нужно быть готовым заранее, поэтому H_1 формулируется одновременно с H_0. Конкурирующая гипотеза H_1 в примере IV-1 гласит: случайная величина «рост студентов университета» не распределена нормально.

Для того чтобы выяснить, справедлива ли H_0, нужен какой-то *статистический критерий*, или *критерий значимости*. Построением критериев значимости для решения различных биометрических задач мы будем заниматься в последующих главах. Здесь же рассмотрим некоторые общие принципы построения критериев значимости и интерпретации полученных результатов.

Конструктивная формулировка нулевой гипотезы сопровождается построением некоторой случайной величины $\widetilde{g}$, являющейся функцией величин $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$. Функция $\widetilde{g}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ выбирается таким образом, чтобы закон ее распределения был полностью определен, т. е. заданы и вид распределения, и его параметры. Эта функция строится в предположении правильности H_0 и называется *статистикой критерия значимости*. Ее значение $g_{\text{эксп}}(x_1, \ldots, x_n)$, вычисленное по выборочным данным, рассматрива-

ется как единичная реализация g. Тогда оказывается возможным вычислить вероятность P одного из событий

$$P\{\widetilde{g} \geqslant g_{\text{эксп}}\},$$
$$P\{\widetilde{g} \leqslant g_{\text{эксп}}\}$$

или их объединения

$$P\{\widetilde{g} \leqslant g_{\text{эксп}}\} + P\{\widetilde{g} \geqslant g_{\text{эксп}}\};$$

какого именно — зависит от формулировки альтернативной гипотезы H_1.

Если P достаточно велика, больше некоторого α, то H_0 принимают; если P слишком мала, меньше α, то H_0 отклоняют и принимают альтернативную гипотезу. В этом смысл критерия значимости. Вероятность α называют *уровнем значимости статистического критерия*.

Выбор величины α — задача не математическая, но биометрическая. Шуточный пример Н. В. Лучника разъясняет суть дела. Некий изобретатель обращается с предложением к медикам. Он изобрел новое средство против рака, которое эффективно на самых последних, неизлечимых на сегодня стадиях заболевания, накануне смерти; приняв этот препарат, больной с вероятностью 0,95 поправляется, но с вероятностью 0,05 умирает. Можно ли в такой ситуации пренебречь вероятностью 0,05? Несомненно! Окрыленный успехом, изобретатель предлагает еще один препарат — теперь против гриппа, наносящего, как известно, огромный ущерб экономике. Больной, принявший этот препарат, с вероятностью 0,95 поправляется немедленно, но с вероятностью 0,05 умирает. Можно ли в такой ситуации пренебречь вероятностью 0,05? Разумеется, нет!

Таким образом, выбор уровня значимости определяется важностью биологических выводов, которые должен сделать экспериментатор. Принятие же решения (выбор α) в обычных, рядовых ситуациях основывается на всем опыте развития количественной биологии. Вначале биометрики брали $\alpha = 0,0027$, это так называемое «правило трех сигм»: можно пренебречь долей площади под нормальной кривой, удаленной по обе стороны от среднего значения более чем на три средних квадратичных отклонения (см. § 4 гл. III). Накопление материала в разных областях биологии показало, однако, что при этом слишком часто принимается неверная нулевая гипотеза. На следующем этапе брали уровень значимости $\alpha = 0,05$. Однако сейчас создается впечатление, что при этом слишком часто неправильно отвергают нулевую гипотезу. Поэтому в настоящее время многие биометрики склоняются к следующему правилу:

а) если $P > 0{,}05$, то принимается нулевая гипотеза;

б) если $P < 0{,}01$, то нулевая гипотеза отклоняется и принимается конкурирующая;

в) если $0{,}01 < P < 0{,}05$, то результат считается неопределенным.

В последнем случае обычно требуется проведение дополнительных опытов. Однако биолог вправе, учитывая всю совокупность знаний по изучаемому вопросу, все-таки принять то или иное решение. Он может это сделать и берет на себя ответственность за этот шаг, но должен помнить при этом, что уровень значимости критерия есть вероятность ошибочного отклонения нулевой гипотезы.

Обсуждая пример IV-1, мы сформулировали представление о проверке статистических гипотез и ввели некоторые важные понятия этого раздела математической статистики. Теперь необходимо подобрать к экспериментальным данным гипотетическое (теоретическое) нормальное распределение. Однако у нас нет никаких априорных соображений о значении параметров μ и σ этого распределения! И нет другого пути, кроме оценки параметров гипотетического нормального распределения по выборочным данным.

Оценка параметров распределения — другой раздел математической статистики.

Для оценки параметров распределений также используются *статистики*, т. е. случайные величины, являющиеся функциями от случайных величин $\widetilde{g}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, реализация которых $g(x_1, \ldots, x_n)$ наблюдается в эксперименте.

Для оценки одного и того же параметра генеральной совокупности могут быть использованы разные статистики. Например, оценкой $E\widetilde{x}$ могут быть и $\widetilde{m} = \frac{1}{n}\sum_{i-1}^{n} \widetilde{x}_i$ и $\frac{1}{2}(\widetilde{x}_{\max} - \widetilde{x}_{\min})$. Как правило, они не будут равны. Какую же статистику следует предпочесть?

Прежде всего, «хорошая» статистика $\widetilde{g}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ должна сходиться к «истинному» генеральному параметру η, для оценки которого она служит, т. е. для любого $\varepsilon > 0$ должно выполняться соотношение

$$P\{|\eta - \widetilde{g}(\widetilde{x}_1, \ldots, \widetilde{x}_n)| < \varepsilon\} \underset{n\to\infty}{\longrightarrow} 1.$$

Оценки, удовлетворяющие этому условию, называются *состоятельными*.

Состоятельность является довольно слабым требованием и к тому же описывает поведение случайной оценки $\widetilde{g}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ при безграничном

увеличении объема выборки. Для выборок конечного объема важной характеристикой является *несмещенность*, под которой понимают выполнение соотношения

$$E[\widetilde{g}(\widetilde{x}_1, \ldots, \widetilde{x}_n)] = \eta.$$

Две несмещенные оценки могут обладать разным рассеянием вокруг истинного значения параметра η. В качестве меры этого рассеяния естественно выбрать дисперсию $D[\widetilde{g}(\widetilde{x}_1, \ldots, \widetilde{x}_n)]$.

Отсюда следует определение: статистика $\widetilde{g}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ называется *эффективной*, если ее дисперсия минимальна в классе всех несмещенных оценок.

Найденные с помощью статистик оценки параметров являются *точечными оценками*, т. е. определяемыми одним числом. Как говорилось выше, статистика есть случайная величина, поэтому отдельное ее значение, полученное в опыте, как правило, не будет совпадать со значением параметра. Особенно большими могут быть различия между значениями статистик и параметра при малых выборках. Поэтому возникает вопрос: нельзя ли на основании выборочных данных указать интервал, заключающий в себе оцениваемый параметр генерального распределения? Речь идет, таким образом, о нахождении *интервальной оценки параметра*, т. е. оценки, определяемой двумя числами — концами интервала.

Выберем некоторую большую вероятность $1 - \alpha$, считая ее настолько близкой к единице, чтобы возможностью появления событий, вероятность которых меньше α, можно было практически пренебречь. Назовем эту вероятность $1 - \alpha$ *доверительной вероятностью*. Тогда интервал (a_1, a_2), построенный на основании выборочного значения статистики и покрывающий с вероятностью $1 - \alpha$ значение параметра генеральной совокупности, называется *доверительным интервалом*. Обычно в биологических исследованиях принимают $1 - \alpha = 0{,}95$. Однако в более ответственных ситуациях приходится брать $1 - \alpha = 0{,}99$ и даже $1 - \alpha = 0{,}999$.

Рассмотрим еще несколько примеров анализа одной выборки.

Пример IV-5. После действия колхицина на растения земляники получены ягоды, имеющие массу (г): 1,2; 1,3; 1,8; 1,4; 1,5; 3,3; 1,4; 1,5; 1,8; 1,7; 1,2; 1,3; 1,4; 1,8; 1,4. Каково среднее значение признака у земляники при воздействии колхицином?

Речь идет о количественном признаке. Однако объем выборки мал: $n = 15$. Анализ вида распределения здесь проводить трудно. Остается:

1) или предполагать, что распределение нормальное,

2) или считать распределение произвольным.

В первом случае нужно получить точечные оценки μ и σ, а также интервальную оценку μ (интервальная оценка σ в задаче не требуется!). Во втором случае можно найти точечную оценку медианы ζ и ее доверительный интервал. Способ решения этих задач будет указан в следующей главе.

ПРИМЕР IV-6 (данные Стьюдента, по Р. А. Фишеру [1958]). Производился подсчет дрожжевых клеток в счетной камере (табл. 7). Согласуется ли полученное распределение с распределением Пуассона?

Таблица 7

Распределение числа дрожжевых клеток по квадратам счетной камеры (к примеру IV-6)

Число		Число		Число	
дрожжевых клеток	квадратов счетной камеры	дрожжевых клеток	квадратов счетной камеры	дрожжевых клеток	квадратов счетной камеры
1	20	5	70	9	10
2	43	6	54	10	5
3	53	7	37	11	2
4	86	8	18	12	2

С какой целью проводятся эксперименты такого рода и почему представляет интерес решение указанной статистической задачи? В исследованиях по физиологии и генетике микроорганизмов, в различных производственных экспериментах зачастую встает вопрос об оценке числа клеток в культуре; это может потребоваться для характеристики темпа размножения клеток, для оценки частоты вновь возникающих мутаций и т. д., и т. п. Достаточно точное определение числа клеток возможно лишь в том случае, если можно учесть каждую клетку отдельно (не образуется комков клеток) и если клеточная культура гомогенна. Как это установить? Берется некоторый небольшой объем взвеси и помещается в счетную камеру. Нас интересует распределение дискретной (целочисленной) случайной величины: число клеток в квадрате счетной камеры может быть равно $0, 1, 2, \ldots$ Если по объему взвеси клетки распределены независимо и если учет числа клеток в квадратах счетной камеры ведется точно, то мы оказываемся в условиях задачи, рассмотренной в § 4 гл. II. Там было показано, что такая случайная величина должна быть распределена по закону Пуассона. Таким образом, решение задачи о согласии наблюдаемого распределения с распределением Пуассона дает ответ на вопрос о том, удовлетворительна ли техника про-

ведения эксперимента, производится ли подсчет числа клеток в культуре методически чисто. Разумеется, для полного ответа на вопрос необходимо провести исследование повторных проб разными лаборантами; однако это усложнение задачи мы обсудим несколько позже (см. гл. VII).

Итак, мы имеем эмпирическое распределение целочисленной случайной величины. Решение задачи сводится к нахождению оценки m параметра λ, вычислению гипотетического распределения и проверке согласия эмпирического распределения с гипотетическим. Можно найти и доверительный интервал для λ.

Пример IV-7. При рассадке черенков томатов с использованием кинетина из 2 417 черенков укоренился 561. Какова эффективность этого способа укоренения?

В качестве меры эффективности укоренения примем $561/2\,417 = 0{,}232$, или, в процентах, 23,2 %. Какова изменчивость выбранного показателя? Имеем качественный альтернативный признак «черенок укоренился» — «не укоренился»; если черенки для опыта выбирались более или менее одинаковыми по биологическим показателям и если все они одинаково обрабатывались кинетином, то вероятность укоренения одна и та же для всех черенков. Заметим, что если бы результаты опыта были приведены подробнее — материал был бы разбит на субвыборки по их размещению в теплице или по повторным во времени экспериментам, — это предположение можно было бы проверить! Естественно предположить, что укоренения разных черенков — события независимые. В таком случае мы оказываемся в условиях схемы Бернулли, приводящей к биномиальному распределению случайной величины «число укоренившихся черенков» (§ 5 гл. II). Один параметр распределения известен: $n = 2417$. Возникает задача точечной и интервальной оценки параметра p.

§ 3. Сравнение двух выборок

Перейдем к рассмотрению задач, возникающих при сравнении двух выборок.

Пример IV-8 (по Дж. У. Снедекору [1961]). Исследовалось влияние содержания белка в диете на прибавку массы белых крыс (табл. 8). Достоверна ли полученная в опыте разница между средними прибавками массы?

Начнем с обсуждения организации эксперимента. Целью исследования является изучение действия одного фактора — содержания белка в диете.

Таблица 8

Прибавка массы белых крыс в возрасте между 28 и 84 днями жизни при разном содержании белка в диете (к примеру IV-8)

Содержание белка в диете	Число животных	Прибавка массы, г	Средняя прибавка
Высокое	7	146, 119, 161, 113, 129, 83, 123	124,9
Низкое	7	70, 118, 101, 85, 107, 132, 94	101,0

Поэтому, естественно, экспериментатор стремится организовать опыт так, чтобы сравниваемые группы животных отличались только по исследуемому фактору; по всем же другим условиям опыта группы должны быть если не идентичны (это идеальная ситуация), то, по крайней мере, практически сходны. Наша книга — пособие по биометрии, поэтому «за кадром» остается огромная работа, которую проводит специалист, готовя такой эксперимент.

Решается вопрос о выборе материала: что означает термин «белые крысы»? Исходя из целей исследования и реальных возможностей, экспериментатор выбирает или животных определенной генетической линии, или гибридов первого поколения между линиями (что может иметь свои преимущества), или, наконец, просто «беспородных» животных. Необходимо так организовать разведение животных, чтобы к началу эксперимента было нужное число крыс определенного возраста. В примере не оговорено, велось ли исследование на самках или самцах, ограничивалась ли какими-то пределами масса животных в начале эксперимента (28 дней жизни) и т. п. Не описаны и условия эксперимента на интервале 28-й и 84-й дни жизни животных, которые также должны быть одинаковыми для обеих групп: нелепостью было бы содержание животных одной группы в индивидуальных клетках, а животных другой группы — по 3–4 в клетке. Подразумевается, наконец, что взвешивание животных в начале и в конце эксперимента проводится в соответствии с принятыми методическими требованиями.

Обсуждая биометрические задачи, мы всегда предполагаем наличие профессиональной культуры при проведении соответствующего эксперимента. Без этого просто нет смысла обсуждать биометрические задачи.

Однако, даже если эксперимент организован профессионально грамотно, возникает следующий принципиально важный вопрос. В нашем распоряжении 14 животных, равноценных во всех отношениях, т. е. любое животное с равными основаниями может быть взято в ту или другую экс-

периментальную группу (вариант опыта). Каким образом из 14 животных выбрать 7, которые будут получать диету с высоким содержанием белка, и 7, которые будут получать диету с низким содержанием белка? Или, другими словами: взяв первую попавшуюся из 14 крыс, в каком варианте опыта ее использовать?

Эта задача может быть решена следующим образом: животные распределяются по вариантам опыта случайно. Такая процедура носит название *рандомизации* (от англ. random — случайно). Современные представления о биологической изменчивости позволяют утверждать, что 14 крыс просто не могут быть идентичны по всем признакам и свойствам. Рандомизация, естественно, не приведет к тому, что животные, отнесенные к разным вариантам опыта, станут идентичными по всем показателям. Рандомизация являет собой в некотором смысле «механическую» процедуру, снимающую субъективизм определенного экспериментатора, определенные неконтролируемые предпочтения. При проведении рандомизации во всех экспериментах мы обеспечиваем выравнивание в среднем, избавляемся от систематических тенденций в формировании групп. Читателю, по-видимому, уже ясно, что вопрос о рандомизации, возникший при рассмотрении двух выборок, идентичен вопросу об обеспечении простого случайного выбора из генеральной совокупности. Это одна из основ математической статистики.

Как практически осуществляется рандомизация? Все 14 животных, отобранных для опыта, нумеруются в произвольном порядке: $1, 2, \ldots, 14$. Затем обращаются к таблицам равномерно распределенных случайных чисел (табл. I Приложения 1). Таблицы случайных чисел содержат несколько сотен тысяч цифр $0, 1, \ldots, 9$, каждая из которых представлена в одинаковом количестве, и все цифры хорошо перемешаны, т. е. рядом с любой произвольно выбранной цифрой равновероятно оказывается любая из $0, 1, \ldots, 9$. Начиная с любого места таблицы и двигаясь в произвольном направлении (по строке, столбцу, диагонали и т. д.), выбирают первые 7 двузначных чисел. Например, это $07, 08, 01, 05, 11, 02, 14$. Тогда для варианта опыта с высоким содержанием белка в диете возьмем животных №№ $1, 2, 5, 7, 8, 11, 14$, а оставшихся животных (№№$3, 4, 6, 9, 10, 12, 13$) возьмем для варианта с низким содержанием белка в диете. Таким образом, распределение животных по вариантам опыта осуществлено случайно, вне зависимости от желаний и склонностей экспериментатора.

В связи с рассмотрением рандомизации опишем два примера, отвлекшись на время от анализа примера IV-8. В одном из них отсутствие рандомизации чуть было не повлекло за собой признания открытием явного артефакта. Другой пример представляет собой редкий случай эффектив-

ной рандомизации в медицинских исследованиях, остроумно проведенной в виде «квазиэксперимента».

ПРИМЕР IV-9 (В. И. Корогодин). Изучалось влияние гамма-облучения дрожжей на интенсивность сбраживания сахара. Опыты ставили следующим образом. Навеску прессованных дрожжей делили на шесть равных частей, три облучали, а три оставляли в качестве контроля. Затем каждую пробу помещали в сахариметр — прибор для определения концентрации сахара. Всего использовали шесть сахариметров, совершенно новых, только что полученных. Сахариметры разместили на стенде, причем №№ $1, 3$ и 5 использовали для работы с облученными образцами, а №№ $2, 4$ и 6 — для работы с необлученными. Первый эксперимент показал, что облученные клетки в десять раз интенсивнее сбраживают сахар, чем необлученные! Пораженный столь необычным результатом, исследователь повторил опыт — то же самое. Третья повторность дала такие же результаты... Тогда, не поверив самому себе, исследователь применил другую методику регистрации брожения, — и эффект исчез... Тщательное обследование сахариметров показало, что три из них, всегда бывшие «контрольными», имели еле заметные трещинки в стекле, через которые улетучивался CO_2 — продукт брожения, а три «опытные» были совершенно исправными.

Превосходная воспроизводимость результатов всех последовательных опытов была воспроизводимостью артефакта. Если бы перед каждым опытом проводилась рандомизация, ошибка обнаружилась бы скорее: уже после второй повторности была бы установлена чрезвычайно большая изменчивость как в опытных, так и в контрольных группах, что свидетельствовало бы о каких-то дефектах методики. Ошибки удалось бы избежать, если бы все шесть сахариметров были тщательно выверены перед началом опытов. Только скрупулезность и трезвость исследователя избавили радиобиологию от лжеоткрытия.

ПРИМЕР IV-10 (Э. Ф. Казанец). На материалах травматологического института изучался вопрос об отдаленных последствиях черепно-мозговых травм. В архиве института были отобраны истории болезни лиц, попавших в автомобильную катастрофу 10–15 лет назад. Контрольная группа была представлена людьми, получившими какую-то травму, например перелом конечности; при этом записи в истории болезни полностью отрицали возможность черепно-мозговой травмы. «Опытная» группа, напротив, включала людей, перенесших черепно-мозговые травмы определенной категории. Все избранные для изучения лица по почте были вызваны в институт. Большинство из них откликнулось, и они были подвергнуты квалифицированному психоневрологическому обследованию.

Оказалось, что в группе, перенесшей 10–15 лет назад черепно-мозговую травму, намного чаще, чем в контрольной, встречаются шизофрения, эпилепсия и другая тяжелая патология. Это явно противоречило всей совокупности имеющихся данных об отдаленных последствиях черепно-мозговой травмы! Тогда автор, будучи очень аккуратным исследователем, повторил эксперимент: провел новую выборку, новые обследования. Итог: на сей раз та же тяжелая патология встречалась достоверно выше . . . в контрольной группе! Явный абсурд, показывающий наличие серьезных дефектов в организации эксперимента. Автор подверг детальному изучению все материалы и обнаружил, к своему удивлению, что лица, принадлежащие к контрольной и опытной группам, различаются отнюдь не только по характеру травмы, но и по ряду признаков, не имеющих на первый взгляд прямой связи с исследуемым вопросом: по соотношению полов, распределениям по возрасту, профессии и т. д., и т. п. Таким образом, наблюдаемые эффекты суть следствия каких-то неконтролируемых в эксперименте причин и условий.

И вот тут-то автору и удалось найти изящное решение задачи. Он стал отбирать в архиве не все истории болезни подряд, а исходя из определенной «модельной» ситуации: автомашина врезалась в группу людей, кто-то чисто случайно получил черепно-мозговую травму, а кто-то — другой вид повреждений. Остроумная организация «квазиэксперимента» 10–15 лет спустя: внешнее и явно нецеленаправленное воздействие потерявшего управление автомобиля является случайным по сути! Анализ этого «эксперимента», также повторенного, дал очень четкие результаты: в группе лиц, перенесших черепно-мозговую травму, несколько с большей частотой встречались относительно слабо выраженные психоневрологические отклонения.

Отметим одну особенность, весьма характерную для обоих вышеприведенных примеров. И в том, и в другом случаях исследователи активно использовали весь багаж знаний, ранее накопленных в соответствующей области. Они не спешили обработать — описать — опубликовать результаты, но действовали в высшей степени профессионально.

Вернемся теперь к примеру IV-8. Предположим, что продуманы все детали техники эксперимента, 14 животных разбиты на две группы. Какие статистические задачи возникнут после того, как эксперимент будет завершен? Обратите внимание: мы обсуждаем статистические задачи до начала эксперимента, поскольку это последний этап его планирования! Работа ведется с количественным признаком «приращение массы между 28 и 84 днями жизни». Объемы выборок невелики: $n_1 = n_2 = 7$, поэтому

вопрос о том, является ли распределение генеральной совокупности нормальным, вряд ли удастся решить, используя данные этого эксперимента.

Допустим, что общие соображения делают это предположение вполне правомочным. Допустим, что средние прибавки (средние арифметические) $m_1 = 124{,}9$ и $m_2 = 101{,}0$, можно использовать в качестве оценок средних значений μ_1 и μ_2 генеральных совокупностей (покажем это мы в следующей главе). Тогда, чтобы ответить на поставленный в примере IV-8 вопрос, нужно проверить нулевую гипотезу H_0: $\mu_1 = \mu_2$; альтернативная гипотеза H_1: $\mu_1 \neq \mu_2$, поскольку у нас нет никаких оснований полагать, например, что $\mu_1 > \mu_2$, но не $\mu_1 < \mu_2$. Может возникнуть вопрос, нельзя ли в этом примере поменять местами H_0 и H_1, т. е. H_0': $\mu_1 \neq \mu_2$; H_1': $\mu_1 = \mu_2$. Нет, нельзя! Дело в том, что нулевая гипотеза должна содержать число, однозначно задающее генеральную совокупность. В случае H_0: $\mu_1 = \mu_2$ имеем H_0: $\mu_1 - \mu_2 = 0$. А вот в случае H_0': $\mu_1 \neq \mu_2$ мы ничего не можем сказать о значении разности $\mu_1 - \mu_2$, эта гипотеза не конструктивна. Обратите внимание, что биолога в рассматриваемом примере интересует разность $\mu_1 - \mu_2$; вопрос же о сравнении σ_1^2 и σ_2^2 не ставится. Быть может, в данном случае это и оправдано, но еще Р. А. Фишер как-то заметил, что биолог склонен делать акцент на сравнении средних, пренебрегая часто сравнением изменчивости, т. е. дисперсий.

Если на уровне значимости α нулевая гипотеза будет принята, то говорят, что разница между средними привесами не значима (синонимы: статистически не достоверна, не существенна); если нулевая гипотеза будет отклонена и, следовательно, принята альтернативная гипотеза, говорят, что разница значима.

Если предположение о нормальности распределения генеральной совокупности не правомочно, приходится пользоваться другими методами, описанными в гл. VI.

Рассмотрим еще несколько задач, возникающих при сравнении двух выборок.

ПРИМЕР IV-11 (по Л. Д. Мешалкину [1963]). Исследование длины и ширины 139 черепов, найденных в Верхнем Египте и относимых к расе, жившей за 8 000 лет до нашей эры, показало, что выборочные дисперсии этих признаков равны 32,741 и 21,271 соответственно. Те же величины, выведенные на основании обследования 1 000 европейцев, оказались равными 37,958 и 25,553. Можно ли считать, что изменчивость изучаемых признаков сильнее выражена у современных европейцев?

В противоположность примеру IV-8 здесь задача исследования сосредоточена на изучении изменчивости. Предполагая нормальность распреде-

лений, сформулируем для каждого из признаков H_0: $\sigma_1^2 = \sigma_2^2$. Результат проверки этой нулевой гипотезы и даст ответ на вопрос, содержащийся в примере.

Пример IV-12 (по Ф. Б. Хатту, 1963 г.). Различаются ли две породы медоносной пчелы по устойчивости к американскому гнильцу — болезни, вызываемой *Bacillus larvae* (табл. 9)?

Таблица 9

Устойчивость к американскому гнильцу двух пород медоносной пчелы (к примеру IV-12)

Порода	Число зараженных личинок	Выжившие личинки	
		Число	%
Ван-Скоя	276	69	25,0
Шартрезская	395	185	46,8

Здесь мы имеем дело с качественным альтернативным признаком. Предполагая, что признак имеет биномиальное распределение, можно получить оценки h_1 параметра p_1 и h_2 параметра p_2. статистическая задача будет заключаться в проверке H_0: $p_1 = p_2$.

Пример IV-13. На чашку Петри с плотной питательной средой нанесен смыв с почвенного образца. На чашке выросли 42 колонии одного вида микроорганизмов и 71 колония другого вида. Есть ли основания полагать, что численности популяций этих двух видов различны?

Предположив, что признак имеет распределение Пуассона, используем $m_1 = 42$ как оценку λ_1 и $m_2 = 71$ как оценку λ_2. Задача заключается в проверке H_0: $\lambda_1 = \lambda_2$.

Пример IV-14 (по В. А. Елизарову, 1972 г.). Дважды (с интервалом в 10 лет) изучалось распределение больных туберкулезом по возрасту. Результаты приведены в табл. 10. Изменился ли возрастной состав больных?

Таблица 10

Распределение больных туберкулезом по возрасту (к примеру IV-14)

Год обследования	Возраст, годы						Всего
	< 1	$1-6$	$7-19$	$20-39$	$40-59$	$\geqslant 60$	
1954	6	42	181	468	296	40	1233
1964	12	61	164	233	260	62	792

Здесь речь идет о качественном признаке, подразделенном на шесть групп (так представлена группировка материала по количественному признаку). Задача сводится к сравнению двух полиномиальных распределений. Нулевая гипотеза заключается в том, что вероятности всех возрастных классов не изменились за 10 лет, так что, например, частоты h_{11} и h_{21} (первый индекс — принадлежность к первой или второй выборке; второй индекс — 1-й возрастной класс) являются оценками неизвестного значения параметра p_1; частоты h_{12} и h_{22} — оценками параметра p_2 и т. д. Альтернативная гипотеза: хотя бы один из параметров изменяется, т. е. имеем разные p_{1i} и p_{2i} хотя бы для одного из классов (см. гл. III, § 3).

§ 4. Сравнение нескольких выборок

Обсуждая следующий пример, мы покажем, чем отличается задача сравнения нескольких выборок от задачи сравнения двух выборок.

ПРИМЕР IV-15. На относительно однородном участке, разбитом на 20 делянок одинакового размера, были высеяны 4 сорта пшеницы, каждый в 5 повторностях. Полученные урожаи приведены в табл. 11.

Различаются ли урожаи разных сортов? Прежде всего, обратим внимание на организацию опыта. Проводится одновременная оценка (в одном эксперименте) четырех сортов. Если 20 делянок с точки зрения экспериментатора равноценны, это означает, что сорта должны размещаться по делянкам случайно, т. е. должна иметь место схема полной рандомизации. Экспериментатор нумерует подряд все делянки; первые пять двузначных цифр $\leqslant 20$, извлеченные из таблицы случайных чисел (см. табл. I Приложения 1), означают номера делянок, на которые будет высеваться сорт A; вторые пять — сорт B и т. д. (рис. 34, *а*). Обратите внимание, что номера

Таблица 11

Урожаи четырех сортов пшеницы (кг на делянку) (к примеру IV-15)

Сорт	Делянка					Средний
	1	2	3	4	5	урожай
A	32,3	34,0	34,3	35,0	36,5	34,42
B	33,3	33,0	36,3	36,8	34,5	34,78
C	30,8	34,3	35,3	32,3	35,8	33,70
D	29,3	26,0	29,8	28,0	28,8	28,38

делянок $1, 2, \ldots$ в табл. 11 — это просто порядковые номера, так что сравнение в табл. 11 по столбцам никакого смысла не имеет. Читателю ясно, что задача сравнения четырех сортов теряет всякий смысл, если, скажем, урожай сортов A и B учитывался при посеве в один год, а сортов C и D — в другой.

Средние урожаи в опыте являются оценками средних урожаев генеральных совокупностей (сортов в целом) $\mu_1, \mu_2, \ldots$ Необходимо ответить на вопрос: являются ли различия между выборочными средними случайными или действительно, по крайней мере, некоторые, генеральные средние разные? Другими словами, нулевая гипотеза H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$ против H_1: среднее хотя бы для одного сорта отличается от остальных, например $\mu_i \neq \mu_j$ $(i \neq j)$.

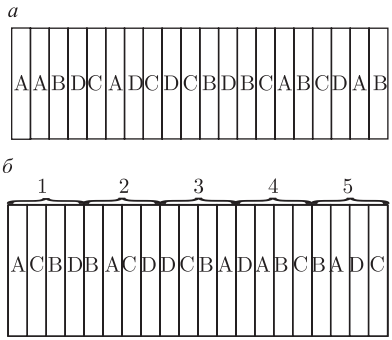


Рис. 34. Размещение 4 сортов, каждый в пяти повторностях, по 20 делянкам: *а* — полная рандомизация; *б* — случайные блоки

Можно ли свести эту задачу к нескольким задачам попарного сравнения сортов, например, сравнивать A и B, A и C, $\ldots$, C и D? Тогда бы мы оказались в условиях примера IV-8. Нет, нельзя. Такой подход не соответствует структуре эксперимента. Ведь экспериментатор работает одновременно с четырьмя выборками. Тем самым мы имеем совместное четырехмерное распределение, которое, как мы установили в гл. I, не сводится к своим маргинальным двумерным распределениям, возникающим при попарном сравнении результатов опыта.

В экспериментах типа рассматриваемого схема полной рандомизации имеет определенные недостатки. Посмотрите на рис. 34, *а*: участок из 20 делянок вытянут в длину, а на таком участке вполне возможны систематические изменения почвенных условий (влажности, химического состава почвы и т. д.). В этих условиях предпочтительно наложение частичного ограничения на рандомизацию. Разобьем участок на 5 последовательных блоков, каждый из которых включает 4 делянки, и будем проводить рандомизацию сортов в пределах каждого блока (рис. 34, *б*). При таком планировании опыта последующий статистический анализ позволяет учесть изменчивость между блоками, не имеющую никакого отношения к изменчивости между сортами. Если опыт проводился по методу случайных блоков, то столбцы в табл. 11 являются блоками, и именно в пределах каждого столбца ведется сравнение сортов. Решение задач этого типа дается в гл. VIII.

ПРИМЕР IV-16 (по В. П. Эфроимсону, 1964 г.). Изучались частоты групп крови системы $AB0$ у разных народов (табл. 12). Различаются ли эти распределения достоверно?

Задача, содержащаяся в данном примере, заключается в сравнении пяти выборочных полиномиальных распределений.

Обратите внимание, что объемы выборок разные. Нулевая гипотеза формулируется следующим образом. H_0: пять выборок извлечены из одной генеральной совокупности, имеющей полиномиальное распределение с параметрами p_1, p_2, p_3, p_4. Альтернативная гипотеза — H_1: хотя бы одна выборка извлечена из другой генеральной совокупности, т. е. хотя бы один параметр $p_{ij} \neq p_{ik}$. Решение этой задачи будет дано в гл. VII.

Перейдем теперь к рассмотрению биологических задач, приводящих к анализу статистических связей.

§ 5. Анализ статистических связей

Этот тип задач отличается от задач, описанных в предыдущих параграфах, тем, что у каждого исследуемого объекта (клетки, особи, популяции, сообщества) регистрируется сразу несколько признаков и оценивается наличие связи между признаками. В нашем учебном пособии мы рассмотрим случай лишь двух признаков. Методы решения задач, содержащихся в примерах IV-17 и IV-18, даются в гл. IX, а в примере IV-19 — в гл. VII.

ПРИМЕР IV-17 [Sokal, Rohlf, 1969]. С целью изучения морфологии краба *Pachygrapsus crassipes* у каждой из 12 отловленных особей определя-

Таблица 12

Частоты групп крови системы $AB0$ (в скобках — проценты от общего числа обследованных в данной выборке) (к примеру IV-16)

Выборка	Число людей, имеющих группу крови				Всего обследовано
	0	A	B	AB	
Китайцы (Кантон)	228 (45,6)	113 (22,6)	125 (25,0)	34 (6,8)	500
Индийцы	712 (30, 2)	577 (24,5)	877 (37,2)	191 (8,1)	2 357
Арабы	1 283 (44,0)	962 (33,0)	516 (17,7)	154 (5,3)	2 915
Бушмены	279 (83,0)	0 (0,0)	57 (17,0)	0 (0,0)	336
Коренные жители Австралии	327 (54,2)	243 (40,3)	23 (3,8)	10 (1,7)	603

лись масса жабр и масса тела (табл. 13). Скоррелированы ли исследуемые признаки (рис. 35)?

Каждая особь, таким образом, характеризуется парой чисел (x_{1i}, x_{2i}). Поэтому в качестве статистической модели мы можем рассматривать двумерные случайные величины $(\widetilde{x}_{1i}, \widetilde{x}_{2i})$. Мерой связи между признаками служит коэффициент корреляции ϱ, причем для некоторых важных в биологии случаев (в частности, для двумерного нормального распределения) $\varrho = 0$ означает, что случайные величины независимы. Таким образом, прежде всего нужно решить вопрос о правомочности предположения о двумерном нормальном распределении (этого предположения потребует и последующее использование статистических методов). Если, например, данное предположение правомочно, то необходимо оценить выборочный коэффициент корреляции r и проверить гипотезу, что коэффициент корреляции генеральной совокупности $\varrho = 0$. Если предположение о двумерном нор-

мальном распределении генеральной совокупности неправомочно, то оценивают ранговый коэффициент корреляции.

Таблица 13

Масса жабр и масса тела у краба *Pachygrapsus crassipes* (к примеру IV-17)

Номер особи	Масса		Номер особи	Масса	
	жабр, мг	тела, г		жабр, мг	тела, г
1	159	14,40	7	100	1,41
2	179	15,20	8	320	15,81
3	100	11,30	9	80	4,19
4	45	2,50	10	220	15,39
5	384	22,70	11	320	17,29
6	230	14,90	12	210	9,52

В предыдущих параграфах, когда у каждого объекта учитывался один признак, мы всякий раз приходили к двум типам статистических задач: оценке параметров распределений и сравнению параметров (или распределений). На примере IV-17 можно видеть, что к тем же типам задач мы приходим и при анализе систем двух признаков.

ПРИМЕР IV-18 (по М. Демерецу, 1943 г.). Изучалась зависимость между дозой рентгеновского облучения и частотой видимых мутаций у нейроспоры (табл. 14). Проанализируйте зависимость «доза — эффект».

Здесь речь идет об установлении вида функциональной зависимости $y = f(x)$. Один признак (доза облучения) произволен и обычно предполагается — точно задается экспериментатором, это независимая переменная, ее значения x_1. Другой признак — значения y_i случайных величин $\widetilde{y}_i$, наблюдаемые при фиксированных x_i. Отличие от корреляционной задачи, где оба признака варьировали вне воли экспериментатора, очевидно.

Приступая к регрессионным задачам, всегда следует начинать с построения графика. На рис. 36 видно, что предположение о линейной зависимости процента мутаций у нейроспоры от дозы облучения совершенно естественно (если бы зависимость отличалась от линейной, меняя масштаб по осям координат, мы попытались бы линеаризировать, «спрямить» ее, потому что это простейший вид функции и его проще всего анализировать). Задача заключается теперь в том, чтобы по данным выборки найти уравнение $y = a + bx$, где a и b — выборочные оценки параметров $\varkappa$ и β генеральной совокупности. Хотя из рис. 36 влияние облучения представля-

Таблица 14

Зависимость между дозой рентгеновского облучения и частотой мутаций у нейроспоры (к примеру IV-18)

Доза облучения (кР)	Частота мутаций (%)
0	1,2
0,1	1,3
0,5	2,0
1,5	3,7
3,0	7,4

ется «на глаз» очевидным, все-таки необходимо строго проверить гипотезу $\varrho = 0$.

Пример IV-19 (по Л. Д. Мешалкину [1963]). У 413 человек определялась односторонность в развитии рук по испытанию в поднимании тяжестей и глазная односторонность по общему астигматизму (табл. 15). Есть ли связь между этими признаками?

Таблица 15

Взаимосвязь между односторонностью в развитии рук и глаз (к примеру IV-19)

Односторонность в развитии рук	Глазная односторонность		
	левоглазые	обоеглазые	правоглазые
Леворукие (левши)	34	62	28
Обоерукие	27	28	20
Праворукие	57	105	52

По сути своей это корреляционная задача. Однако коэффициент корреляции введен лишь для непрерывных случайных величин, а в данном примере оба признака имеют полиномиальное распределение.

* *
*

Выше (§§ 2–5) мы рассмотрели основные типы статистических задач, возникающих в количественной биологии. Советуем читателю, прежде чем двигаться дальше в изучении предмета, еще раз просмотреть этот материал,

чтобы научиться распознавать, «в чем суть задачи». Вопрос «как решать задачу» будет рассматриваться в следующих главах.

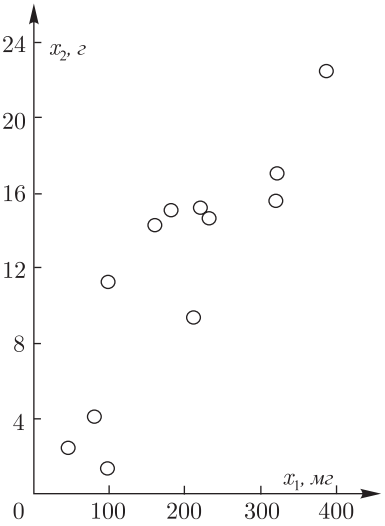


Рис. 35. Связь между массой жабр (x_1) и массой тела (x_2) у краба (пример IV-17)

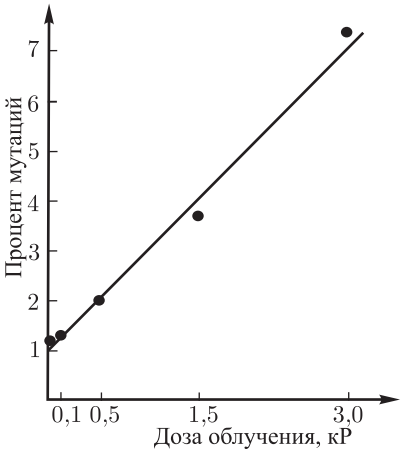


Рис. 36. Зависимость процента мутаций у нейроспоры от дозы рентгеновского облучения (пример IV-18)

На протяжении предыдущих глав неоднократно подчеркивалось, что статистический анализ в биологии начинается с выбора адекватной математико-статистической модели; по существу, это уже первый этап статистического анализа.

Следующий этап анализа — применение статистического аппарата для решения поставленной задачи к данным выборки. Как правило, соответствующий метод удается найти в учебниках и справочниках по статистике; нередко и модель строится уже под известный метод. Однако нельзя исключить, что в некоторых случаях определенная постановка биологической задачи потребует какой-то новой статистической разработки; обычно эту разработку может осуществить или специалист в области математической статистики, или статистически грамотный биолог в содружестве с математиком.

Наконец, последний этап анализа — биологическая интерпретация полученных статистических результатов. Для успешного осуществления этого этапа от биолога требуются и профессиональная биологическая подготовка, и знание основ математической статистики.

Задачи

IV-1 [Кендалл, Стьюарт, 1966]. Постройте график распределения по возрасту для смертности от менингита (табл. 16). Почему материал сгруппирован в классы с неравными интервалами?

Таблица 16

Смертность от менингита в Англии и Уэльсе в 1953 г[3]., $n = 414$ (к задаче IV-1)

Возраст (x_i), годы	Число смертей (n_i)	Возраст (x_i), годы	Число смертей (n_i)
0	162	35	5
1	29	40	11
2	11	45	17
3	6	50	21
4	0	55	32
5	10	60	23
10	8	65	22
15	6	70	15
20	4	75	11
25	3	80	10
30	5	85 и выше	3

IV-2 (данные С. Хсиа и др., по А. Пьяцца, 1981 г.). С целью оценить влияние родственных связей на интенсивность иммунного ответа было изучено распределение титров антител к вирусу цитомегалии в семьях меннонитов, проживающих в Индиане (табл. 17). Постройте и опишите график этого распределения. Попробуйте получить распределение, близкое к нор-

[3] 0 — обозначает возраст до 1 года, 1 — возраст от 1 года до 2 лет и т. д.

мальному, используя преобразования $1/x$, $\sqrt{x}$ и $\log x$ (в последнем случае обоснуйте выбор основания логарифма).

Таблица 17

Распределение титров антител к вирусу цитомегалии у меннонитов Индианы, $n = 380$ человек (к задаче IV-2)

Обратный титр (x_i)	Число людей (n_i)	Обратный титр (x_i)	Число людей (n_i)
2	23	16	110
4	53	32	60
8	90	64	34
		128	10

Таблица 18

Интенсивность облачности в Гринвиче для июля 1890–1904 гг., $n = 1715$ (к задаче IV-3)

Степень облачности (x_i), усл. ед.	Число наблюдений (n_i)	Степень облачности (x_i), усл. ед.	Число наблюдений (n_i)
0	320	6	55
1	129	7	65
2	74	8	90
3	68	9	148
4	45	10	676
5	45		

IV-3 (данные Г. Э. Пирс, по М. Кендаллу и А. Стьюарту [1966]). Постройте и опишите график распределения степени облачности в Гринвиче для июля (табл. 18).

IV-4 (данные Дж. В. Т. Метьюза, по К. Мардиа [1978]). Найдите наглядное графическое изображение для данных по ориентации немигрирующих английских крякв *Anas platyrhynchos* (табл. 19).

Таблица 19

Распределение значений угла исчезновения[4] английских крякв, $n = 734$ птиц (к задаче IV-4)

Направление (x_i)	Число птиц (n_i)	Направление (x_i)	Число птиц (n_i)
0°–	40	180°–	3
20–	22	200–	11
40–	20	220–	22
60–	9	240–	24
80–	6	260–	58
100–	3	280–	136
120–	3	300–	138
140–	1	320–	143
160–	6	340–	69

Таблица 20

Длина клещей у самцов уховертки на Фарнейских островах, $n = 584$ особи (к задаче IV-5)

Длина клещей (x_i), мм	Число особей (n_i)	Длина клещей (x_i), мм	Число особей (n_i)
3,0	64	6,5	42
3,5	125	7,0	90
4,0	52	7,5	68
4,5	7	8,0	44
5,0	12	8,5	8
5,5	24	9,0	6
6,0	42		

IV-5. «При своем посещении Фарнейских островов у берегов Нортумберлэнда Бэтсон обратил внимание на их богатство уховертками...» [Филипченко, 1978, с.103–104]. Воспроизведите график полученного Бэтсоном распределения длины клещей у самцов уховертки *Forficula auricularia* (табл. 20). Чем можно объяснить наблюдаемую картину?

[4]Угол исчезновения — это азимут той точки на горизонте, в которой выпущенная наблюдателем птица скрывается из виду. 0° — обозначает направление от 0° (север) до 20°, 20 — от 20° до 40° и т. д.

Таблица 21

Показания часов в витринах часовщиков, $n = 1\,000$ часов (к задаче IV-6)[5]

Показания часов (x_i)	Число наблюдений (n_i)	Показания часов (x_i)	Число наблюдений (n_i)
0–	77	6–	73
1–	81	7–	70
2–	95	8–	77
3–	86	9–	82
4–	98	10–	84
5–	90	11–	87

Таблица 22 Распределение суммы 4 случайных чисел, $n = 200$ (к задаче IV-7)

Сумма (x_i)	Число случаев (n_i)	Сумма (x_i)	Число случаев (n_i)
0–4	0	19–20	25
5–6	2	21–22	22
7–8	3	23–24	28
9–10	7	25–26	10
11–12	21	27–28	8
13–14	16	29–30	4
15–16	20	31–36	0
17–18	34		

IV-6. А. К. Эйткен (по Г. Крамеру [1975]) приводит распределение для показаний часов, выставленных в витринах часовщиков (табл. 21). Постройте график этого распределения. о чем он говорит?

IV-7. А. Хальд [1956] приводит два эмпирических распределения:

1) для суммы 4 случайных чисел (табл. 22) и

2) для суммы 16 случайных чисел (табл. 23).

Постройте и опишите графики этих распределений.

[5] 0 — обозначает промежуток времени от 0 до 1 ч, 1 — от 1 до 2 ч и т. д.

Таблица 23 Распределение суммы 16 случайных чисел, $n = 200$ (к задаче IV-77)

Сумма (x_i)	Число случаев (n_i)	Сумма (x_i)	Число случаев (n_i)
0–44	0	75–79	33
45–49	3	80–84	25
50–54	14	85–89	12
55–59	14	91–94	10
60–64	19	95–99	2
65–69	28	100–104	1
70–74	39	105–144	0

Таблица 24

Распределение числа наводнений в устье р. Невы по годам за 278 лет (1703–1980 гг.) (к задаче IV-12)

Число наводнений в году	Число лет при подъеме воды над ординаром	
	свыше 150 см	свыше 180 см
0	139	206
1	79	53
2	37	14
3	13	4
4	5	0
5	3	1
6	0	
7	1	
8	1	
Всего наводнений	242	98

IV-8 [Терентьев, Ростова, 1977]. Проведите графический анализ признака «размеры крипт» (в делениях окуляр-микрометра) в ободочной кишке крыс:

10	7	8	12	16	9	10	7	9	9	14	10	10
8	9	11	9	8	10	11	9	10	11	10	10	10
8	11	12	9	9	10	10	10	10	10	7	8	9
11	8	9	8	10	10	9	10	9	8	9	9	7
9	10	8	13	9	9	12	10	9	8	7	10	11
12	13	8	8	8	13	12	10	9	8	7	11	12
14	10	10	8	8	7	9	10	11	9	8	8	7
9	10	11	14	10	7	13	10	9	12	12	10	8
9	8	7	10									

IV-9. Сопоставьте примеры IV-8 и IV-15. Если вы располагаете методом анализа второго из них, что можно сказать о применимости этого метода к первому?

IV-10. Сопоставьте примеры IV-12, IV-14, IV-16 и IV-19. Сформулируйте для них H_0 одного типа. Если вы располагаете методом анализа примера IV-19, что можно сказать о его приложимости к остальным примерам?

IV-11. Можно ли использовать табл. 23 в качестве таблицы случайных чисел?

IV-12 (Р. А. Нежиховский, 1981 г.). В табл. 24 приведены статистические данные о наводнениях в устье р. Невы за 278 лет существования г. Ленинграда (с 1703 по 1980 г.). Постройте графики распределений и сравните их. Какому распределению подчиняются эти данные?

Глава V

Оценка параметров распределений

Выше (гл. IV, § 2) говорилось, что для оценки параметров распределений используются статистики, к которым предъявляются требования состоятельности, несмещенности иэффективности. Учитывая эти требования, а также то обстоятельство, что для оценки одного и того же параметра могут быть использованы разные статистики, естественно задать вопрос: с помощью каких методов следует выбирать их? Наиболее распространен метод максимального правдоподобия, предложенный Р. А. Фишером.

§ 1. Метод максимального правдоподобия

Пусть $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ — независимые случайные величины, каждая из которых имеет плотность распределения $f(x)$. Для целочисленных случайных величин мы рассматривали бы распределение вероятностей $p\{x\}$. Предполагается, что вид распределения $f(x)$ (или $p\{x\}$) известен; неизвестно только значение некоторого параметра Θ, который надлежит оценить по реализациям $x_1, x_2, \ldots, x_n$ этих случайных величин. Например, если $p\{x\}$ — распределение Пуассона, то Θ — параметр λ. Пусть $\psi(x)$ — плотность совместного распределения случайных величин $\widetilde{x}_1, \ldots, \widetilde{x}_n$. В дальнейшем будем иметь дело только с независимыми случайными величинами, поскольку это, как правило, соответствует структуре большинства биологических экспериментов. Для конкретных реализаций $x_1, x_2, \ldots, x_n$ функция $L(\Theta) = \psi(x_1, x_2, \ldots, x_n, \Theta)$ называется *функцией правдоподобия*.

Сущность метода максимального правдоподобия состоит в том, чтобы найти «наиболее правдоподобное» значение параметра Θ, т. е. такое, при котором значение функции правдоподобия максимально. При максимизации $L(\Theta)$ подразумевается, что значения $x_1, x_2, \ldots, x_n$ случайных величин $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ фиксированы, а переменной величиной является параметр Θ. Другими словами, максимум $L(\Theta)$ отыскивается в предположении, что $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ заменены их выборочными значениями. Так как удобнее

иметь дело не с величиной $L(\Theta)$, а с ее логарифмом, то требуемое значение Θ находят, решая относительно Θ *уравнение правдоподобия*:

$$\frac{d\ln L}{d\Theta} = 0.$$

Если распределение имеет несколько, например, k, параметров $\Theta_1, \Theta_2, \ldots$ $\ldots, \Theta_k$, необходимо приравнять нулю частные производные $\ln L$ по этим параметрам и решить систему:

$$\frac{\partial \ln L}{\partial \Theta_1};$$
$$\ldots$$
$$\frac{\partial \ln L}{\partial \Theta_k}.$$

Рассмотрим задачу оценивания параметров μ и σ^2 нормального распределения. В силу того, что плотность совместного распределения независимых случайных величин равна произведению их плотностей (см. гл. III, § 2), функция правдоподобия в этом случае имеет вид:

$$L(\Theta_1, \Theta_2) = (\sqrt{2\pi\Theta_2})^{-n} \exp\left[-\frac{1}{2\Theta_2}\sum_{i=1}^{n}(x_i - \Theta_1)^2\right]$$

и ее логарифм

$$\ln L = -\frac{n}{2}\ln(2\pi\Theta_2) - \frac{1}{2\Theta_2}\sum_{i=1}^{n}(x_i - \Theta_1)^2.$$

Дифференцируя это уравнение по Θ_1 и Θ_2 и приравнивая производную нулю, имеем:

$$\frac{\partial \ln L}{\partial \Theta_1} = -\frac{1}{2\Theta_2}\sum_{i=1}^{n} 2(x_i - \Theta_2) = 0;$$

$$\frac{\partial \ln L}{\partial \Theta_2} = -\frac{n}{2}\cdot\frac{1}{\Theta_2} + \frac{1}{2\Theta_2^2}\sum_{i=1}^{n}(x_i - \Theta_1)^2 = 0.$$

Из первого уравнения получаем:

$$\sum_{i=1}^{n}(x_i - \Theta_1) = \sum_{i=1}^{n} x_i - n\Theta_1 = 0.$$

Решая это уравнение относительно Θ_1, находим оценку максимального правдоподобия для среднего значения μ, которую обозначим m:

$$m = \Theta_1 = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Итак, для оценки среднего значения μ нормального распределения следует выбрать статистику

$$\widetilde{m} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{x}_i.$$

Чтобы найти оценку дисперсии σ^2, подставим во второе уравнение m вместо Θ_1. Разрешая это уравнение относительно Θ_2, получаем оценку максимального правдоподобия, которую обозначим

$$s_0^2 = \Theta_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2.$$

Таким образом, для оценки дисперсии σ^2 нормального распределения следует выбрать статистику

$$\widetilde{s}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{x}_i - \widetilde{m})^2.$$

Статистики, полученные методом максимального правдоподобия, обладают тремя замечательными свойствами: они асимптотически (т. е. при $n \to \infty$) распределены нормально, являются асимптотически несмещенными и асимптотически эффективными.

Читатель может самостоятельно показать, что оценкой максимального правдоподобия для параметра p биномиального распределения является

$$h = \frac{k}{n},$$

где n — общее число независимых испытаний, k — число испытаний, в которых произошло данное событие (задача V-2), и что оценкой максимального правдоподобия для параметра λ распределения Пуассона является

$$m = \frac{\sum_{i=1}^{l} x_i n_i}{n},$$

где x_i — наблюденное значение случайной величины; n_i показывает, сколько раз оно наблюдалось; $n = \sum_{i=1}^{l} n_i$ есть общее число наблюдений и l — число различных значений x_i (вариант) (задача V-2).

Методом максимального правдоподобия мы будем пользоваться в дальнейшем для нахождения оценок коэффициента корреляции p и коэффициента линейной регрессии β (см. гл. IX, § 2, § 6).

§ 2. Распределение выборочного среднего $\widetilde{m}$ и выборочной дисперсии $\widetilde{s}^2$ в случае нормально распределенной генеральной совокупности

Если $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ распределены нормально, то статистика $\widetilde{m}$ также имеет нормальное распределение, поскольку она является их линейной комбинацией (см. гл. III, § 3). При этом

$$E\widetilde{m} = E\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i\right) = \frac{1}{n}\sum_{i=1}^{n}E\widetilde{x}_i = \frac{1}{n}n\mu = \mu.$$

Найдем дисперсию $\widetilde{m}$:

$$D\widetilde{m} = D\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}D\widetilde{x}_i = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Это очень важный результат, показывающий еще одно ценное свойство нормального распределения; дисперсия выборочного среднего в n раз меньше дисперсии отдельных случайных величин, т. е. средние группируются вокруг центра распределения теснее, чем отдельные значения. Итак, справедлива следующая теорема.

Теорема V-1. *Если n независимых случайных величин $\widetilde{x}_1, \ldots, \widetilde{x}_n$ распределены нормально с параметрами μ и σ^2 каждая, то случайная величина*

$$\widetilde{m} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i$$

распределена также нормально с параметрами μ и σ^2/n.

В качестве статистики для оценки дисперсии в предыдущем параграфе была получена случайная величина

$$\widetilde{s}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{x}_i - \widetilde{m})^2.$$

Это тот самый случай, когда оценка максимального правдоподобия, асимптотически несмещенная, оказывается смещенной при конечном n (см. задачу V-4). Несмещенной статистикой является

$$\widetilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widetilde{x}_i - \widetilde{m})^2.$$

Величину $(n-1)$ называют *числом степеней свободы*, смысл ее заключается в следующем. Хотя все n случайных величин независимы, квадраты $(\widetilde{x}_i - \widetilde{m})^2$ независимыми не являются, так как разности $(\widetilde{x}_i - \widetilde{m})$ связаны уравнением

$$\sum_{i=1}^{n}(\widetilde{x}_i - \widetilde{m}) = 0.$$

Говорят, что на n независимых случайных величин наложена одна линейная связь, или одно линейное ограничение. Поэтому среди всех n величин $(\widetilde{x}_i - \widetilde{m})$ независимыми («свободными») являются любые $(n-1)$. Отсюда становится ясным происхождение термина «число степеней свободы».

Найдем теперь вид распределения $\widetilde{s}^2$. Перейдем от независимых случайных величин $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ к случайным величинам $\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_n$ по формулам:

$$\widetilde{y}_1 = \frac{1}{\sqrt{1\cdot 2}}(\widetilde{x}_1 - \widetilde{x}_2);$$

$$\widetilde{y}_2 = \frac{1}{\sqrt{2\cdot 3}}(\widetilde{x}_1 + \widetilde{x}_2 - \widetilde{x}_3);$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\widetilde{y}_{n-1} = \frac{1}{\sqrt{(n-1)n}}[\widetilde{x}_1 + \widetilde{x}_2 + \ldots + \widetilde{x}_{n-1} - (n-1)\widetilde{x}_n];$$

$$\widetilde{y}_n = \frac{1}{\sqrt{n}}(\widetilde{x}_1 + \widetilde{x}_2 + \ldots + \widetilde{x}_n) = \sqrt{n}\cdot\widetilde{m}.$$

Так как матрица, составленная из коэффициентов этого преобразования,

$$C = \begin{pmatrix} \frac{1}{\sqrt{1\cdot 2}} & -\frac{1}{\sqrt{1\cdot 2}} & 0 & \dots & 0 \\ \frac{1}{\sqrt{2\cdot 3}} & \frac{1}{\sqrt{2\cdot 3}} & -\frac{2}{\sqrt{2\cdot 3}} & \dots & 0 \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ \frac{1}{\sqrt{(n-1)n}} & \frac{1}{\sqrt{(n-1)n}} & \frac{1}{\sqrt{(n-1)n}} & \dots & \frac{(n-1)}{\sqrt{(n-1)n}} \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \dots & \frac{1}{\sqrt{n}} \end{pmatrix}$$

удовлетворяет соотношению $\boldsymbol{CC}^{\mathrm{т}} = \boldsymbol{I}$, где $\boldsymbol{I}$ — единичная матрица порядка n, а т — знак транспонирования, очевидно, что случайные величины $\widetilde{y}_1, \widetilde{y}_2, \dots, \widetilde{y}_n$ получены ортогональным преобразованием случайных величин $\widetilde{x}_1, \widetilde{x}_2, \dots, \widetilde{x}_n$. Из курса линейной алгебры известно, что ортогональные преобразования сохраняют длины векторов, т. е.

$$\sum_{i=1}^{n} \widetilde{x}_i^2 = \sum_{i=1}^{n} \widetilde{y}_i^2.$$

Поскольку случайные величины $\widetilde{y}_1, \widetilde{y}_2, \dots, \widetilde{y}_n$ являются линейными комбинациями нормально распределенных случайных величин $\widetilde{x}_1, \widetilde{x}_2, \dots, \widetilde{x}_n$, то $\widetilde{y}_i$ также распределены нормально, причем читатель может убедиться, что

$$E\widetilde{y}_i = 0; \quad D\widetilde{y}_i = \sigma^2 (i = 1, 2, \dots, n) \quad \text{и} \quad \operatorname{Cov}(\widetilde{y}_i, \widetilde{y}_j) = 0 (i \neq j),$$

т. е. случайные величины $\widetilde{y}_i$ независимы (см. гл. III, § 6).

Из представления статистики $\widetilde{s}^2$ получаем:

$$(n-1)\widetilde{s}^2 = \sum_{i=1}^{n} (\widetilde{x}_i - \widetilde{m})^2 = \sum_{i=1}^{n} \widetilde{x}_i^2 - n\widetilde{m}^2.$$

Но $\widetilde{y}_n = \sqrt{n} \cdot \widetilde{m}$, поэтому

$$(n-1)\widetilde{s}^2 = \sum_{i=1}^{n} \widetilde{y}_i^2 - \widetilde{y}_n^2 = \sum_{i=1}^{n-1} \widetilde{y}_i^2.$$

Поделим теперь обе части уравнения на σ^2:

$$\frac{(n-1)\widetilde{s}^2}{\sigma^2} = \sum_{i=1}^{n-1} \left(\frac{\widetilde{y}_i}{\sigma^2} \right)^2.$$

Случайные величины $\widetilde{y}_i/\sigma$ $(i = 1, 2, \ldots, n-1)$ имеют нормированное нормальное распределение каждая, они независимы, поэтому сумма их квадратов имеет распределение χ^2 с числом степеней свободы $\nu = n - 1$. Итак, доказана следующая теорема.

Теорема V-2. *Если $\widetilde{x}_1, \ldots, \widetilde{x}_n$ — независимые нормально распределенные случайные величины с параметрами μ и σ^2 каждая, то*

$$\frac{(n-1)\widetilde{s}^2}{\sigma^2} \sim \chi^2(\nu),$$

где $\nu = n - 1$.

Докажем, наконец, еще одну очень важную теорему.

Теорема V-3. *Если $\widetilde{x}_1, \ldots, \widetilde{x}_n$ — независимые нормально распределенные случайные величины с параметрами μ и σ^2 каждая, то статистики $\widetilde{m}$ и $\widetilde{s}^2$ независимы.*

Доказательство:
Выше мы получили

$$\widetilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}\widetilde{y}_i^2; \quad \widetilde{m} = \frac{1}{\sqrt{n}}\widetilde{y}_n.$$

Случайные величины $\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_n$ независимы, откуда и следует независимость статистик $\widetilde{m}$ и σ^2. ■

В заключение параграфа укажем, что в биометрии в качестве показателя изменчивости часто вычисляют *коэффициент вариации*

$$v = \frac{s}{m}.$$

Удобство коэффициента вариации состоит в том, что эта величина безразмерная. Таким образом, с его помощью возможно сравнение изменчивости признаков, значения которых измерены в разных единицах.

§ 3. Доверительный интервал для среднего значения μ нормального распределения

В § 1 была указана точечная оценка μ. Задачей этого параграфа является нахождение интервала (a_1, a_2), в пределах которого с доверительной

вероятностью $(1-\alpha)$ лежит значение параметра μ:

$$P\{a_1 < \mu < a_2\} = 1 - \alpha.$$

Полностью определен закон распределения случайной величины:

$$\frac{\widetilde{m} - \mu}{\sigma/\sqrt{n}} \sim N(0;1).$$

Трудность, однако, заключается в том, что в биологических задачах практически никогда не известно значение параметра σ^2, его можно лишь оценить по выборке с помощью статистики $\widetilde{s}^2$. Другими словами, требуется найти закон распределения случайной величины

$$\frac{\widetilde{m} - \mu}{\widetilde{s}\sqrt{n}}.$$

Результат был указан в 1908 г. английским исследователем В. Госсетом, опубликовавшим свою работу за подписью Стьюдент (student — студент), но строго был получен в 1925 г. Р. А. Фишером. Покажем, что эта случайная величина имеет t-распределение Стьюдента, которое было определено ранее (см. гл. III, § 8).

Так как

$$\frac{\widetilde{m} - \mu}{\sigma/\sqrt{n}} = \widetilde{u} \sim N(0;1),$$

а в предыдущем параграфе было показано, что

$$\frac{(n-1\widetilde{s}^2)}{\sigma^2} = \widetilde{\chi}^2 \sim \chi^2(\nu),$$

где $\nu = n-1$, причем эти две случайные величины независимы, то

$$\frac{\widetilde{u}}{\sqrt{\widetilde{\chi}^2/\nu}} = \frac{(\widetilde{m}-\mu)\sqrt{n}\sigma\sqrt{n-1}}{\sigma\sqrt{n-1}\widetilde{s}} = \frac{\widetilde{m}-\mu}{\widetilde{s}/\sqrt{n}} = \widetilde{t} \sim t(\nu), \quad \nu = n-1.$$

Перейдем к рассмотрению выражения

$$P\left\{t_1 < \frac{\widetilde{m}-\mu}{\widetilde{s}/\sqrt{n}} < t_2\right\} = 1 - \alpha.$$

Какие дополнительные соображения необходимы для выбора t_1 и t_2? Случайная величина $\widetilde{x}$ генеральной совокупности распределена нормально, симметрично относительно μ. Также симметрично (относительно нуля) t-распределение. Исследуемая выборка случайна, т. е. равновозможны значения $m < \mu$ и $m > \mu$. Поэтому представляется естественным строить симметричный относительно μ доверительный интервал.

Введем следующее обозначение, которым в дальнейшем будем часто пользоваться: x_a есть значение случайной величины $\widetilde{x}$, такое, что $P\{\widetilde{x} \geqslant x_a\} = a$. В нашем случае верхней границей искомого доверительного интервала будет $t_{\alpha/2}$, т. е. такое значение случайной величины t при $\nu = n-1$, что $P\{\widetilde{t} \geqslant t_{\alpha/2}\} = \alpha/2$. Тогда симметрично расположенной нижней границей будет $t_{1-\alpha/2}$, т. е. такое значение случайной величины t при $\nu = n-1$, что $P\{\widetilde{t} \geqslant t_{1-\alpha/2}\} = 1-\alpha/2$ (рис. 37). Так как t-распределение симметрично относительно нуля, то очевидно, что $t_{1-\alpha/2} = -t_{\alpha/2}$. Поэтому приведенное выше неравенство мы запишем в виде:

$$\left|\frac{\widetilde{m}-\mu}{\widetilde{s}\sqrt{n}}\right| < t_{\alpha/2}.$$

Другими словами, найти границы доверительного интервала для μ означает найти пределы интегрирования плотности t-распределения:

$$\int_{-t_{\alpha/2}}^{t_{\alpha/2}} f(x)dx = 1-\alpha.$$

Это можно сделать с помощью табл. III Приложения 1.

Разрешая полученное неравенство относительно μ, имеем

$$\widetilde{m} - t_{\alpha/2}\cdot\widetilde{s}/\sqrt{n} < \mu < \widetilde{m} + t_{\alpha/2}\cdot\widetilde{s}/\sqrt{n}.$$

Для конкретной выборки объема n значения статистик $\widetilde{m}$ и $\widetilde{s}$ будут равны m и s, тогда доверительный интервал для μ с доверительной вероятностью $(1-\alpha)$ вычисляется по формуле

$$m - t_{\alpha/2}\cdot s/\sqrt{n} < \mu < m + t_{\alpha/2}\cdot s/\sqrt{n}.$$

Поскольку при больших n (практически $n > 30$) t-распределение хорошо аппроксимируется нормированным нормальным, то доверительный интервал приобретает вид:

$$m - u_{\alpha/2}\cdot s/\sqrt{n} < \mu < m + u_{\alpha/2}\cdot s/\sqrt{n},$$

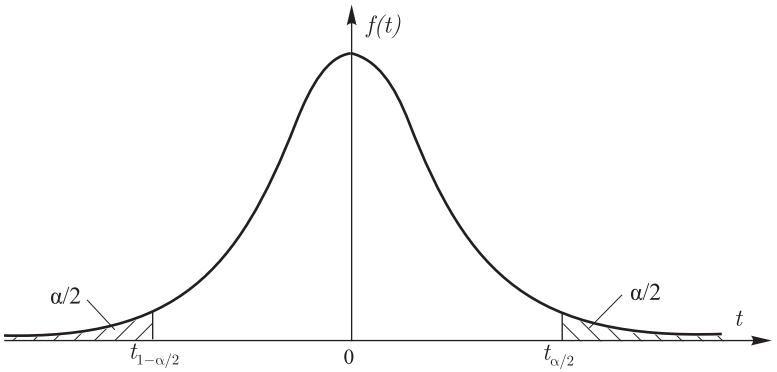


Рис. 37. К построению доверительного интервала для среднего значения μ нормального распределения

где $u_{\alpha/2}$ — такое значение случайной величины $\widetilde{u} \sim N(0;1)$, что $P\{\widetilde{u} \geqslant \geqslant u_{\alpha/2}\} = \alpha/2$, и оно в отличие от $t_{\alpha/2}$ не зависит от объема выборки (см. последнюю строку табл. III Приложения 1). Величину

$$s_m = \frac{s}{\sqrt{n}}$$

называют *ошибкой среднего* и, представляя результаты обработки экспериментальных данных, часто пишут $m \pm s_m$. Так, запись $101{,}8 \pm 0{,}16$ означает, например, что среднее $m = 101{,}8$, а ошибка среднего $s_m = 0{,}16$.

ПРИМЕР V-1. У растений нового сорта тетраплоидной ржи измерена длина междоузлий (см): 7,2; 7,1; 7,0; 6,8; 6,6; 6,8; 7,2; 7,1; 7,4; 7,0; 7,2; 7,3; 7,1; 7,1; 7,2; 7,3; 7,1; 7,0; 6,8. Требуется оценить среднее значение и изменчивость признака.

Объем выборки $n = 19$. Выборочное среднее есть

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{19}(7{,}2 + 7{,}1 + \ldots + 6{,}8) = 7{,}068 = 7{,}07.$$

Необходимо сделать замечание о точности вычислений: среднее вычисляется с точностью, на порядок большей, чем точность измерений; среднее квадратичное отклонение — с точностью, на порядок большей, чем точ-

ность вычисления среднего. Напомним правило округления: если округляемая цифра <5, то она отбрасывается; если $\geqslant 5$, то в предыдущий разряд добавляется единица. Сумма квадратов отклонений

$$\sum_{i=1}^{n}(x_i - m)^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 =$$

$$= 7,2^2 + 7,1^2 + \ldots + 6,8^2 - \frac{1}{19}134,3^2 = 0,741.$$

Выборочная дисперсия

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{18}0,741 = 0,041$$

и среднее квадратичное отклонение $s = 0{,}203$. Коэффициент вариации $v = s/m = 0{,}203/7{,}07 = 0{,}029$ очень мал, всего 2,9%. Ошибка среднего $s_m = s/\sqrt{n} = 0{,}203/\sqrt{19} = 0{,}047$. Таким образом, имеем $7{,}07 \pm 0{,}047$. Построим теперь 95%-ный доверительный интервал для среднего значения. Так как $\nu = n - 1 = 18$, то значение $t_{0,025} = 2{,}10$ (см. табл. III Приложения 1):

$$7{,}07 - 2{,}10 - 0{,}047 < \mu < 7{,}07 + 2{,}10 - 0{,}047,$$

т. е. $6{,}97 < \mu < 7{,}17$.

§ 4. Доверительный интервал для дисперсии σ^2 нормального распределения

Случайная величина $(n-1)\widetilde{s}^2/\sigma^2$ имеет распределение χ^2 с числом степеней свободы $\nu = n - 1$ (см. § 2). Поэтому решение данной задачи оказывается простым. Необходимо найти границы (a_1, a_2) доверительного интервала для σ^2 из условия

$$P\left\{a_1 < \frac{(n-1)\widetilde{s}^2}{\sigma^2} < a_2\right\} = 1 - \alpha.$$

Распределение χ^2 несимметрично, поэтому границы доверительного интервала выбирают симметричными по вероятностям — площадям, отсекаемым на хвостах распределения (рис. 38). Имеем

$$\chi^2_{1-\alpha/2} < \frac{(n-1)\widetilde{s}^2}{\sigma^2} < \chi^2_{\alpha/2},$$

где $\chi^2_{1-\alpha/2}$ и $\chi^2_{\alpha/2}$ суть такие значения случайной величины χ^2 при $\nu = n-1$, что $P\{\widetilde{\chi}^2 \geqslant \chi^2_{1-\alpha/2}\} = 1 - \alpha/2$, $\widetilde{P}\{\chi^2 \geqslant \chi_{\alpha/2}\} = \alpha/2$ и $\chi^2_{1-\alpha/2} \neq \chi_{\alpha/2}$. Разрешая неравенство относительно σ^2 и подставляя вместо $\widetilde{s}^2$ значение выборочной дисперсии s^2, получаем искомый доверительный интервал:

$$\frac{s^2\nu}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{s^2\nu}{\chi^2_{1-\alpha/2}}.$$

Значения знаменателей находят из табл. V Приложения 1.

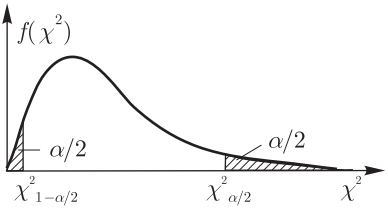


Рис. 38. Границы доверительного интервала для σ^2 нормального распределения выбирают симметричными по вероятностям, отсекаемым на хвостах χ^2-распределения

ПРИМЕР V-2. Пусть выборочная дисперсия нормально распределенной совокупности $s^2 = 0{,}0349$ при $\nu = 29$. Требуется построить 95 %-ный доверительный интервал для σ^2.

По табл. V Приложения 1 при $\nu = 29$ находим $\chi^2_{0{,}025} = 45{,}722$ и $\chi^2_{0{,}975} = 16{,}047$, отсюда

$$\frac{0.0349 \cdot 29}{45{,}7} < \sigma^2 < \frac{0{,}0349 \cdot 29}{16{,}0}.$$

Итак, с доверительной вероятностью 0,95 значение дисперсии генеральной совокупности лежит в интервале $0{,}0221 < \sigma^2 < 0{,}0631$. В случае больших выборок $(n > 30)$ для нахождения доверительного интервала для параметра σ^2 нормального распределения можно использовать следующее выражение, которое мы приводим без вывода:

$$s^2\left[1 - u_{\alpha/2}\sqrt{\frac{2}{n}} + \frac{2}{3n}(u^2_{\alpha/2} - 1)\right] < \sigma^2 < \\ < s^2\left[1 + u_{\alpha/2}\sqrt{\frac{2}{n}} + \frac{2}{3n}(u^2_{\alpha/2} - 1)\right],$$

где $u_{\alpha/2}$ — такое значение $\widetilde{u} \sim N(0;1)$, что $P\{\widetilde{u} \geqslant u_{\alpha/2}\} = \alpha/2$ (см. последнюю строку в табл. III Приложения 1).

§ 5. Доверительный интервал для параметра p биномиального распределения

Ранее отмечалось, что биномиальное распределение может быть представлено через F-распределение (см. гл. III, § 10). Это обстоятельство используется для нахождения доверительного интервала параметра p. Приведем без вывода точные формулы для нахождения верхней $p_в$ и нижней $p_н$ границ доверительного интервала. С доверительной вероятностью $(1-\alpha)$ параметр p лежит в пределах $p_н < p < p_в$:

$$p_в = \frac{\nu_1 \cdot F_{\alpha/2}}{\nu_2 \cdot \nu_1 \cdot F_{\alpha/2}},$$

где $F_{\alpha/2}$ — такое значение случайной величины $\widetilde{F}$ при $\nu_1 = 2(k+1)$, $\nu_2 = 2(n-k)$, что $P\{\widetilde{F} \geqslant F_{\alpha/2}\} = \alpha/2$;

$$p_н = \frac{\nu_2}{\nu_2 + \nu_1 \cdot F_{\alpha/2}},$$

где $F_{\alpha/2}$ — значение $\widetilde{F}$ при $\nu_1 = 2(n-k+1)$, $\nu_2 = 2k$, такое, что $P\{\widetilde{F} \geqslant F_{\alpha/2}\} = \alpha/2$.

Пример V-3. При проверке некоторого лекарственного препарата на $n = 17$ обезьянах у $k = 3$ животных наблюдались побочные эффекты. Указать 95%-ный доверительный интервал для доли животных, дающих побочные эффекты.

Точечная оценка p равна $h = k/n = 3/17 = 0{,}176$, или 17,6%. Для нахождения верхней границы 95 %-ного интервала берем из табл. IVб Приложения 1 $F_{0{,}025}(\nu_1 = 8; \nu_2 = 28) = 2{,}69$, тогда

$$p_в = \frac{4 \cdot 26{,}9}{14 + 4 \cdot 2{,}69} = 0{,}435.$$

Для нахождения нижней границы из той же таблицы получаем $F_{0{,}025}(\nu_1 = 30; \nu_2 = 6) = 5{,}07$, тогда

$$p_н = \frac{3}{3 + 15 \cdot 5{,}07} = 0{,}033.$$

Таким образом, 95 %-ный доверительный интервал оказывается очень широким — от 3,8 до 43,5%, что связано, конечно, с малым объемом

выборки, обусловленным, по-видимому, дороговизной опытов на обезьянах.

Для нахождения доверительных интервалов параметра p в биологии широко используется аппроксимация биномиального распределения нормальным. С доверительной вероятностью $(1-\alpha)$

$$h - \frac{1}{2n} - u_{\alpha/2}\cdot\sqrt{\frac{h(n-1)}{n}} < p < h + \frac{1}{2n} + u_{\alpha/2}\cdot\sqrt{\frac{h(n-1)}{n}},$$

где $h = k/n$ (ср. гл. III, § 5). Подчеркнем, что эта аппроксимация хороша лишь при довольно жестких требованиях: необходимо, чтобы $nh(1-h)>25$. Если, например, $h = 1-h = 0{,}5$, то необходимо $n > 100$, а для $h = 0{,}1$ требуется $n > 277$.

Когда частоту h умножают на 100, выражая в процентах, то величину $\sqrt{h(n-1)/n}\cdot 100\%$ называют ошибкой процента (аналогично ошибке среднего). Распространена также форма записи $h \pm \sqrt{h(n-1)/n}$, например, $0{,}528 \pm 0{,}015$.

ПРИМЕР V-4. Среди отловленных в природной популяции мух Drosophila melanogaster оказалось 539 самцов и 570 самок. Какова частота встречаемости самцов в популяции?

Общее число наблюдений $n = 539 + 570 = 1\,109$. Частота самцов $h = 539/1\,109 = 0{,}486$. Поскольку $nh(l-h) = 207{,}5 > 25$, возможна аппроксимация биномиального распределения нормальным: $\sqrt{h(1-h)/n} = = 0{,}486 - 0{,}514/1\,109 = 0{,}015$. Построим 95 %-ный доверительный интервал для частоты самцов p в популяции: $0{,}486 - 1{,}96 - 0{,}015 < p < 0{,}486 + + 1{,}96 - 0{,}015$. Таким образом, частота самцов в выборке составляет 48,6 % и частота самцов в популяции с вероятностью 0,95 лежит в интервале от 45,7 до 51,69 %. Аппроксимация нормальным распределением здесь очень хороша: нахождение точных границ дает 45,7 и 51,6 %!

§ 6. Доверительный интервал для параметра λ распределения Пуассона

Ранее отмечалось, что распределение Пуассона может быть представлено через распределение χ^2 (см. гл. III, § 10). Это обстоятельство используется для нахождения доверительного интервала параметра λ. Приведем без вывода точные формулы для нахождения границ доверительного интервала. С доверительной вероятностью $(1-\alpha)$ параметр λ лежит в пределах

$\lambda_н < \lambda < \lambda_в$:

$$\lambda_в = \frac{1}{2}\chi^2_{\alpha/2},$$

где $\chi^2_{\alpha/2}$ — такое значение случайной величины χ^2 при $\nu = 2(k+1)$, что $P\{\widetilde{\chi}^2 \geqslant \chi^2_{\alpha/2}\} = \alpha/2$;

$$\lambda_н = \frac{1}{2}\chi^2_{1-\alpha/2},$$

где $\chi^2_{1-\alpha/2}$ — такое значение χ^2 при $\nu = 2k$, что $P\{\chi^2 \geqslant \chi^2_{1-\alpha/2}\} = 1-\alpha/2$.

Пример V-5 (К. А. Браунли [1977]). На самолете, покидающем сборочный цех, не хватает одной заклепки $(k = 1)$. Если предположить, что число недостающих заклепок на одном самолете распределено по закону Пуассона, то 99 %-ный доверительный интервал для λ (т. е. для среднего числа недостающих заклепок на один самолет во всей партии самолетов) находим как

$$\lambda_в = \frac{1}{2}\chi^2_{0{,}005}(\nu = 4) = 14{,}9/2 = 7{,}45;$$

$$\lambda_н = \frac{1}{2}\chi^2_{0{,}995}(\nu = 2) = 0{,}010/2 = 0{,}005.$$

Значения $\chi^2_{0{,}005}$ и $\chi^2_{0{,}995}$ берутся из табл. V Приложения 1.

Для нахождения доверительного интервала нередко пользуются аппроксимацией распределения Пуассона нормальным; с доверительной вероятностью $(1-\alpha)$.

Условие аппроксимации довольно жесткое: $mn > 25$.

Пример V-6. Получено распределение островков Лангерганса по отдельным квадратам ткани поджелудочной железы *Macaca rhesus* (табл. 25).

Предположив, что наблюдаемое распределение описывается распределением Пуассона, оценим параметр λ.

Найдем выборочное среднее:

$$m = \frac{1}{n}\sum_{i=1}^{l} x_i n_i = (0 \cdot 327 + 1 \cdot 340 + \ldots + 6 \cdot 1)/900 = 904/900 = 1{,}00.$$

Так как $mn = 900 > 25$, возможна аппроксимация нормальным распределением: $\sqrt{m/n} = \sqrt{1{,}00/900} = 0{,}033$, и 95 %-ный доверительный

Таблица 25

Распределение островков Лангерганса в поджелудочной железе макаки резус, $n = 900$ (к примеру V-6)

Число островков Лангерганса (x_i)	Число квадратов ткани (n_i)
0	327
1	340
2	160
3	53
4	16
5	3
6	1

интервал будет $1{,}00 - 1{,}96 \cdot 0{,}033 < \lambda < 1{,}00 + 1{,}96 \cdot 0{,}033$, т. е. с вероятностью 0,95 параметр пуассоновского распределения находится в пределах от 0,94 до 1,06.

§ 7. Оценка медианы неизвестного распределения

Несмотря на широкое распространение нормального распределения в биологии, все многообразие распределений признаков отнюдь не сводится к нему. Хотя и для не нормально распределенных признаков распределение выборочного среднего $\widetilde{m}$ нередко нормально, его дисперсия обычно неизвестна. Когда у исследователя есть сомнения в нормальности распределения генеральной совокупности, производят выборочную оценку другого параметра положения — медианы ζ (см. гл. III, § 1).

Пусть $\widetilde{x}$ — непрерывная случайная величина с неизвестной плотностью распределения $f(x)$ и ζ — медиана распределения. Значения независимых наблюдений $x_1, x_2, \ldots, x_n$ — случайная выборка из генеральной совокупности. Расположим все n значений в порядке возрастания:

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)},$$

т. е. ранжируем их. В полученном ранжированном ряду новый индекс (в скобках) означает порядковый номер, т. е. *ранг* соответствующего наблюдения. Исходя из определения медианы ζ, естественно определить выборочную медиану Z как значение наблюдения, имеющего средний ранг $\frac{1}{2}(n+1)$.

Отсюда при нечетном n получим $Z = x_{\left(\frac{n+1}{2}\right)}$, при четном n имеем

$$Z = \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}\right].$$

Таким образом, в качестве статистики для точечной оценки медианы ζ берется случайная величина

$$\widetilde{Z} = \widetilde{x}_{\left(\frac{n+1}{2}\right)} \quad \text{или} \quad \widetilde{Z} = \frac{1}{2}\left[\widetilde{x}_{\left(\frac{n}{2}\right)} + \widetilde{x}_{\left(\frac{n+1}{2}\right)}\right].$$

Можно получить распределение $\widetilde{Z}$, если известна плотность $f(x)$. При $n \to \infty$ это распределение аппроксимируется нормальным со средним значением ζ и дисперсией $0{,}25 \cdot \frac{1}{n}[f(\zeta)]^{-2}$, где $f(\zeta)$ — значение плотности распределения $f(x)$ в точке ζ. Отсюда следует, что если $\widetilde{x} \sim N(\mu; \sigma^2)$, то асимптотически $\widetilde{Z} \sim N(\mu; \frac{\pi}{2n}\sigma^2)$. Таким образом, в случае нормального распределения дисперсия выборочной медианы $\widetilde{Z}$ в $\pi/2$ раз больше дисперсии выборочного среднего $\widetilde{m}$ (см. § 2). Однако, как правило, если $\widetilde{x}$ распределена не нормально, плотность $f(x)$ в биологических исследованиях не известна. Поэтому для медианы ζ неизвестного распределения приходится строить непараметрический доверительный интервал. Перейдем к рассмотрению такой задачи.

Поскольку медиана ζ делит площадь под кривой распределения на две равные части, а выборка производится случайным образом, то вероятность получить выборочное значение, меньшее ζ, равна вероятности получить выборочное значение, большее ζ. Это соответствует случайному извлечению шара из урны, содержащей равное количество белых и черных шаров. Поэтому если объем выборки n известен, то число выборочных значений, меньших (больших) ζ, является случайной величиной $\widetilde{k}$, имеющей биномиальное распределение с параметрами $p = 0{,}5$ и n. Последовательность рассуждений, приводящих к построению доверительного интервала для ζ с доверительной вероятностью $(1-\alpha)$, продемонстрируем для конкретного значения n.

Пусть $n = 9$. В табл. 26 указаны значения k, соответствующие всем возможным вариантам расположения выборочных наблюдений относительно

Таблица 26

К нахождению доверительного интервала для медианы неизвестного распределения при объеме выборки $n = 9$

Число k выборочных значений, меньших ζ	Расположение выборочных значений относительно ζ	$P\{\widetilde{k} = k\}$
0	$\zeta \leqslant x_{(1)} \leqslant \ldots \leqslant x_{(9)}$	0,002
1	$x_{(1)} \leqslant \zeta \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(9)}$	0,018
2	$x_{(1)} \leqslant x_{(2)}\zeta \leqslant x_{(3)} \leqslant \ldots \leqslant x_{(9)}$	0,070
3	$x_{(1)} \leqslant x_{(2)} \leqslant x_{(3)}\zeta \leqslant x_{(4)} \leqslant \ldots \leqslant x_{(9)}$	0,164
4	$x_{(1)} \leqslant \ldots \leqslant x_{(4)}\zeta \leqslant x_{(5)} \leqslant \ldots \leqslant x_{(9)}$	0.24
5	$x_{(1)} \leqslant \ldots \leqslant x_{(5)}\zeta \leqslant x_{(6)} \leqslant \ldots \leqslant x_{(9)}$	0,246
6	$x_{(1)} \leqslant \ldots \leqslant x_{(6)}\zeta \leqslant x_{(7)} \leqslant x_{(8)} \leqslant x_{(9)}$	0,164
7	$x_{(1)} \leqslant \ldots \leqslant x_{(7)}\zeta \leqslant x_{(8)} \leqslant x_{(9)}$	0,070
8	$x_{(1)} \leqslant \ldots \leqslant x_{(8)}\zeta \leqslant x_{(9)}$	0,018
9	$x_{(1)} \leqslant \ldots \leqslant x_{(9)}\zeta$	0,002

медианы и вероятности таких событий. Выберем доверительную вероятность $(1 - \alpha) = 0,95$. Исходя из симметрии биномиального распределения при $p = 0{,}5$, будем строить симметричный доверительный интервал, «отсекая» от концов распределения по $\alpha/2 = 0{,}025$. Начнем с отыскания нижней границы доверительного интервала. Вероятность того, что ни одно из выборочных значений не попало левее ζ, т. е. $k = 0$, очень мала: $0{,}002 < \alpha/2$; это означает, что по крайней мере первое значение в ранжированном ряду может быть взято в качестве нижней границы медианы. Вероятность того, что ни одного наблюдения не попало или одно попало левее ζ, т. е. $k \leqslant 1$, также мала: $(0{,}002 + 0{,}018) < \alpha/2$; это означает, что по крайней мере второе значение в ранжированном ряду может быть взято в качестве нижней границы медианы. Однако уже следующий шаг $k \leqslant 2$ дает $(0{,}002 + + 0{,}018 + 0{,}070) > \alpha/2$. Поэтому в качестве нижней границы 95 %-ного доверительного интервала для ζ принимаем значение наблюдения, номер (ранг) которого равен двум: $x_{(2)}$. Проводя подобные рассуждения, найдем и верхнюю границу: $x_{(8)}$.

Читатель может убедиться, что доверительный интервал при $(1 - \alpha) = = 0{,}99$ равен $x_{(1)} < \zeta < x_{(9)}$, а вопрос о величине доверительного интервала при $(1 - \alpha) = 0{,}999$ для выборки объема $n = 9$ смысла не имеет.

При решении биометрических задач нет необходимости выполнять каждый раз такого рода достаточно трудоемкие построения. В табл. VII Приложения 1 приведены объемы выборок, для которых при заданной доверительной вероятности $(1-\alpha)$ указаны номера $(b=k+1)$ ранжированных наблюдений, являющихся нижними границами доверительного интервала для ζ. Номер наблюдения, являющегося верхней границей, по симметрии равен $(n-b+1)$. При больших n (в табл. VII при $n>75$) применяют аппроксимацию биномиального распределения нормальным:

$$b=\frac{1}{2}n-u_{\alpha/2}\cdot\frac{1}{2}\sqrt{n}+\frac{1}{2},$$
$$n-b+1=\frac{1}{2}n+u_{\alpha/2}\cdot\frac{1}{2}\sqrt{n}+\frac{1}{2},$$

где $u_{\alpha/2}$, как и раньше, такое значение нормированной нормальной случайной величины $u\sim N(0;1)$, что $P\{\widetilde{u}\geqslant u_{\alpha/2}\}=\alpha/2$.

Пример V-7 (Дж. Вайнберг, Дж. Шумекер, 1979 г.). Определяли время прохождения лабиринта крысами, $n=17$. Получены следующие (уже упорядоченные) значения (c): 12, 13, 13, 14, 14, 15, 15, 16, 16, 16, 17, 17, 18, 18, 19, 19, 20. Оценим медиану неизвестного распределения.

Средний ранг $(n+1)/2=(17+1)/2=9$, следовательно, точечная оценка медианы есть $Z=x_{(9)}=16$. Найдем 95 %-ный доверительный интервал для ζ: из табл. VII Приложения 1 находим $b=5$, т. е. нижняя граница есть $x_{(5)}=14$; соответственно, $n-b+1=13$, т. е. верхняя граница есть $x_{(13)}=18$. Итак, $14<\zeta<18$ с доверительной вероятностью $1-\alpha=0{,}95$.

Задачи

V-1. Найдите оценку максимального правдоподобия параметра p биномиального распределения.

V-2. Найдите оценку максимального правдоподобия параметра λ в распределении Пуассона.

V-3. Найдите оценки максимального правдоподобия параметров p_1, p_2, $\ldots$, p_{l-1} полиномиального распределения.

V-4. Покажите, что оценка максимального правдоподобия дисперсии нормального распределения $\widetilde{s}_0^2$ является смещенной. Покажите, что несмещенной оценкой является статистика

$$\widetilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widetilde{x}_i - \widetilde{m})^2.$$

V-5. Укажите границы доверительного интервала, если выбрана доверительная вероятность, равная единице.

V-6. Проведите статистический анализ примера IV-1.

V-7. Проведите статистический анализ примера IV-5.

V-8. Проведите статистический анализ примера IV-7.

V-9. На 1 800 обследованных зарегистрировано 72 больных диабетом. Оцените процент заболеваемости диабетом в данном районе.

V-10. На чашке Петри выросли 42 колонии. Предполагая, что это единственная реализация пуассоновской случайной величины, оцените параметр λ.

V-11. Постройте доверительный интервал для среднего квадратичного отклонения σ нормального распределения.

V-12. На практике встречаются ситуации, когда вместо параметра p биномиального распределения требуется оценить математическое ожидание np биномиальной случайной величины $\widetilde{x}$. Постройте соответствующий доверительный интервал.

V-13. Постройте доверительный интервал для величины $n\lambda$, где n — объем выборки из распределения Пуассона, а λ — его параметр.

Глава VI

Сравнение параметров двух распределений

В этой главе будут рассмотрены задачи, сводящиеся к сравнению параметров двух распределений. Речь идет, как правило, о сравнении выборочных оценок, поскольку в биологии мы обычно не можем выдвинуть содержательной гипотезы о значении параметров генеральной совокупности μ, σ^2, λ, p. Исключение составляют задачи анализа расщеплений в генетике, где гипотетические значения параметров биномиального или полиномиального распределений задаются правилами Г. Менделя (см. гл. VII).

§ 1. Сравнение дисперсии σ_1^2 и σ_2^2 двух независимых нормальных распределений

Пусть $\widetilde{x}_1$ и $\widetilde{x}_2$ — независимые нормально распределенные случайные величины; σ_1^2 и σ_2^2 — соответственно, их дисперсии. Пусть s_1^2 и s_2^2 — выборочные дисперсии. Требуется проверить нулевую гипотезу о равенстве дисперсий H_0: $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Покажем, что если H_0 верна, то случайная величина $\widetilde{s}_1^2/\widetilde{s}_2^2$ имеет F-распределение с числом степеней свободы для числителя $\nu_1 = n_1 - 1$ и для знаменателя $\nu_2 = n_2 - 1$.

Действительно, согласно гл. V, § 2,

$$\frac{(n_1-1)\widetilde{s}_1^2}{\sigma^2} = \widetilde{\chi}_1^2, \quad \nu_1 = n_1 - 1 \quad \text{и} \quad \frac{(n_2-1)\widetilde{s}_2^2}{\sigma^2} = \widetilde{\chi}_2^2, \quad \nu_2 = n_2 - 1,$$

причем эти случайные величины независимы. Отсюда следует, что

$$\widetilde{s}_1^2 = \frac{\sigma^2}{\nu_1}\widetilde{\chi}_1^2 \quad \text{и} \quad \widetilde{s}_2^2 = \frac{\sigma^2}{\nu_2}\widetilde{\chi}_2^2.$$

Взяв отношение случайных величин $\widetilde{s}_1^2$ и $\widetilde{s}_2^2$, убеждаемся, согласно § 9 гл. III, что это отношение имеет F-распределение:

$$\frac{\widetilde{s}_1^2}{\widetilde{s}_2^2} = \frac{\widetilde{\chi}_1^2/\nu_1}{\widetilde{\chi}_2^2/\nu_2} = \widetilde{F} \sim F(\nu_1, \nu_2).$$

Знание закона распределения отношения $\dfrac{\widetilde{s}_1^2}{\widetilde{s}_2^2}$ (когда $\sigma_1^2 = \sigma_2^2$) позволяет построить F-критерий (критерий Фишера) для проверки нулевой гипотезы H_0: $\sigma_1^2 = \sigma_2^2$ по выборочному значению[1] $F_{\text{эксп}} = s_1^2/s_2^2$.

Пусть верна нулевая гипотеза $\sigma_1^2 = \sigma_2^2$. Тогда случайная величина $\widetilde{F} = \widetilde{s}_1^2/\widetilde{s}_2^2$ с вероятностью $(1-\alpha)$ принимает значение в интервале $[F_{1-\alpha/2}(\nu_1, \nu_2), F_{\alpha/2}(\nu_1, \nu_2)]$, где $F_{1-\alpha/2}(\nu_1, \nu_2)$ и $F_{\alpha/2}(\nu_1, \nu_2)$ суть такие значения случайной величины F при ν_1 и ν_2, что $P\{\widetilde{F} \geqslant F_{1-\alpha/2}(\nu_1, \nu_2)\} = 1 - \alpha/2$ и $P\{\widetilde{F} \geqslant F_{\alpha/2}(\nu_1, \nu_2)\} = \alpha/2$, причем $F_{1-\alpha/2}(\nu_1, \nu_2) < 1$, а $F_{\alpha/2}(\nu_1, \nu_2) > 1$ (рис. 39). Мы проводим конкретный эксперимент: извлекаем две независимые выборочные дисперсии s_1^2, s_2^2 и их отношение $F_{\text{эксп}} = s_1^2/s_2^2$. Число $F_{\text{эксп}}$ есть единичная реализация случайной величины $\widetilde{F}$ с параметрами ν_1, ν_2.

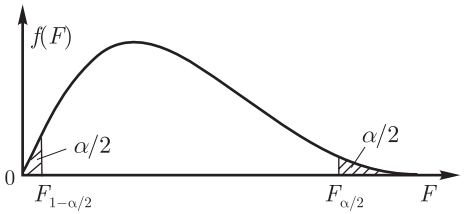


Рис. 39. Двусторонний F-критерий. Сумма заштрихованных площадей равна α

Если оказалось, что $F_{1-\alpha/2}(\nu_1, \nu_2) < F_{\text{эксп}} < F_{\alpha/2}(\nu_1, \nu_2)$, то на уровне значимости α нулевая гипотеза принимается. Если же $F_{\text{эксп}} \leqslant F_{1-\alpha/2}(\nu_1, \nu_2)$ или $F_{\text{эксп}} \geqslant F_{\alpha/2}(\nu_1, \nu_2)$, то на уровне значимости α нулевая гипотеза отвергается, и мы говорим, что $\sigma_1^2 \neq \sigma_2^2$. Покажем, однако, что в действительности нет необходимости сравнивать число $F_{\text{эксп}}$ с двумя табличными значениями: $F_{1-\alpha/2}(\nu_1, \nu_2)$ и $F_{\alpha/2}(\nu_1, \nu_2)$.

[1] Для значений статистик, вычисленных по выборочным (экспериментальным) данным, мы всегда будем использовать индекс «эксп».

Выше (§ 7 гл. III) отмечалось, что если $\widetilde{F} \sim F(\nu_1, \nu_2)$, то $1/\widetilde{F} \sim \sim F(\nu_1, \nu_2)$. Следовательно, имеем

$$\alpha/2 = P\{\widetilde{F} < F_{1-\alpha/2}(\nu_1, \nu_2)\} = P\{1/\widetilde{F} > F_{1-\alpha/2}(\nu_1, \nu_2)].$$

С другой стороны, по указанному свойству

$$P\{1/\widetilde{F} \geqslant F_{\alpha/2}(\nu_1, \nu_2)\} = \alpha/2.$$

Сопоставляя два последних выражения, получим

$$1/F_{1-\alpha/2}(\nu_1, \nu_2) = F_{\alpha/2}(\nu_1, \nu_2).$$

Отметим, что и $F_{\alpha/2}(\nu_1, \nu_2)$, и $F_{\alpha/2}(\nu_2, \nu_1)$ больше единицы. Из полученного следует, что двойное неравенство

$$F_{1-\alpha/2}(\nu_1, \nu_2) < F_{\text{эксп}} < F_{\alpha/2}(\nu_1, \nu_2)$$

равносильно одновременному выполнению двух неравенств:

$$F_{\text{эксп}} < F_{\alpha/2}(\nu_1, \nu_2) \quad \text{и} \quad 1/F_{\text{эксп}} < F_{\alpha/2}(\nu_2, \nu_1).$$

Условимся всегда брать в числителе бóльшую из сравниваемых дисперсий; допустим, что $s_1^2 > s_2^1$. Тогда $F_{\text{эксп}} = s_1^2/s_2^2 > 1$ и второе из указанных неравенств автоматически выполняется. Следовательно, для проверки нулевой гипотезы достаточно использовать первое из неравенств и потому иметь таблицы F-распределения только для значений $F_{\text{эксп}}$, больших единицы (см. табл. IV Приложения 1).

Таким образом, процедура проверки гипотезы о равенстве двух дисперсий сводится к следующему. По выборочным дисперсиям вычисляется величина $F_{\text{эксп}} = s_1^2/s_2^2$, причем в качестве s_1^2 всегда берется бóльшая дисперсия, а в качестве s_2^2 — меньшая. Вычисленное значение $F_{\text{эксп}}$ сравнивается с табличным $F_{\alpha/2}(\nu_1, \nu_2)$. На уровне значимости α нулевая гипотеза принимается, если $F_{\text{эксп}} < F_{\alpha/2}(\nu_1, \nu_2)$; нулевая гипотеза отвергается, если $F_{\text{эксп}} \geqslant F_{\alpha/2}(\nu_1, \nu_2)$.

Пример VI-1 [Урбах, 1975]. Два сорта пшеницы имеют почти одинаковую среднюю урожайность за 9 лет: $m_1 = 20{,}4$ ц/га и $m_2 = 20{,}3$ ц/га, но один из них как будто более подвержен влиянию изменений погодных условий: $s_1^2 = 16{,}9$; $s_2^2 = 4{,}92$. Требуется проверить гипотезу H_0: $\sigma_1^2 = \sigma_2^2$.

Выберем уровень значимости $\alpha = 0{,}05$. Вычислим $F_{\text{эксп}} = s_1^2/s_2^2 = = 16{,}9/4{,}92 = 3{,}44$; $\nu_1 = \nu_2 = 8$. Обратимся к табл. IV*б* Приложения 1,

где приведены критические значения $F_{0,025}(\nu_1,\nu_2)$; имеем $F_{0,025}(8,8) = 4{,}43$. Поскольку $F_{\text{эксп}}$ меньше табличного, различие дисперсий статистически незначимо, и нулевая гипотеза не отвергается. Таким образом, имеющиеся данные не свидетельствуют о том, что урожайность первого сорта пшеницы более изменчива от года к году.

§ 2. Сравнение средних значений μ_1 и μ_2 двух независимых нормальных распределений

Пусть $\widetilde{x}_1$ и $\widetilde{x}_2$ — независимые нормально распределенные случайные величины; μ_1 и μ_2 — их средние значения, а σ_1^2 и σ_2^2 — соответственно, дисперсии; m_1 и m_2 — выборочные средние. Требуется проверить нулевую гипотезу о равенстве средних H_0: $\mu_1 = \mu_2 = \mu$.

Рассмотрим случай, когда $\sigma_1^2 = \sigma_2^2 = \sigma^2$, т. е. когда s_1^2 и s_2^2 являются выборочными оценками одной и той же величины σ^2.

Определим средневзвешенную (обобщенную) оценку дисперсии σ^2 следующим образом:

$$\widetilde{s}^2 \frac{\nu_1 \widetilde{s}_1^2 + \nu_2 \widetilde{s}_2^2}{\nu_1 + \nu_2},$$

где $\nu_1 = n_1 - 1$; $\nu_2 = n_2 - 1$. Покажем, что если H_0 верна, то в предположении равенства дисперсий случайная величина $\dfrac{(\widetilde{m}_1 - \widetilde{m}_2)}{(\widetilde{s}\sqrt{1/n_1 + 1/n_2})}$ имеет t-распределение с $\nu = n_1 + n_2 - 2$ степенями свободы.

Действительно, $\widetilde{m}_1$ и $\widetilde{m}_2$ имеют нормальное распределение с дисперсиями σ_1^2/n_1 и σ_2^2/n_2 соответственно. Если $\sigma_1^2 = \sigma_2^2 = \sigma^2$ и верна нулевая гипотеза H_0: $\mu_1 = \mu_2 = \mu$, то случайная величина $(\widetilde{m}_1 - \widetilde{m}_2)$ имеет нормальное распределение с нулевым математическим ожиданием

$$E(\widetilde{m}_1 - \widetilde{m}_2) = \mu_1 - \mu_2 = 0$$

и дисперсией

$$D(\widetilde{m}_1 - \widetilde{m}_2) = D\widetilde{m}_1 + D\widetilde{m}_2 = \sigma_1^2/n_1 + \sigma_2^2/n_2 = \sigma^2(1/n_1 + 1/n_2).$$

Следовательно, нормированная случайная величина

$$\widetilde{u} = \frac{(\widetilde{m}_1 - \widetilde{m}_2) - E(\widetilde{m}_1 - \widetilde{m}_2)}{\sqrt{D(\widetilde{m}_1 - \widetilde{m}_2)}} = \frac{\widetilde{m}_1 - \widetilde{m}_2}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0;1).$$

Далее $\nu_1\widetilde{s}_1^2/\sigma^2 \sim \chi^2(\nu_1)$ и $\nu_2\widetilde{s}_2^2/\sigma^2 \sim \chi^2(\nu_2)$. Поскольку выборки независимы, новая случайная величина $\nu\widetilde{s}^2/\sigma^2 = \nu_1\widetilde{s}_1^2/\sigma^2 + \nu_2\widetilde{s}_2^2/\sigma^2$ в силу аддитивности хи-квадрата также имеет χ^2-распределение: $\nu\widetilde{s}^2/\sigma^2 \sim \chi^2(\nu)$, $\nu = \nu_1 + \nu_2 = n_1 + n_2 - 2$ (см. гл. III, § 7). Окончательно имеем

$$\frac{\widetilde{m}_1 - \widetilde{m}_2}{\widetilde{s}\sqrt{1/n_1 + 1/n_2}} = \frac{\widetilde{m}_1 - \widetilde{m}_2}{\sigma\sqrt{1/n_1 + 1/n_2}} \div \frac{\widetilde{s}}{\sigma} = \frac{\widetilde{u}}{\sqrt{\widetilde{\chi}^2/\nu}} = \widetilde{t} \sim t(\nu).$$

Знание закона распределения величины $\dfrac{(\widetilde{m}_1 - \widetilde{m_2})}{\widetilde{s}\sqrt{1/n_1 + 1/n_2}}$ позволяет построить *t-критерий* (*критерий Стьюдента*) для проверки нулевой гипотезы H_0: $\mu_1 = \mu_2$. При справедливости нулевой гипотезы $\widetilde{t} \sim t(\nu)$. С вероятностью $(1 - \alpha)$ случайная величина $\widetilde{t}$ принимает значение в интервале $[-t_{\alpha/2}(\nu), t_{\alpha/2}(\nu)]$, где, как обычно, $t_{\alpha/2}(\nu)$ есть такое значение случайной величины $\widetilde{t}$, что $P\{\widetilde{t} \geqslant t_{\alpha/2}(\nu)\} = P\{\widetilde{t} \leqslant -t_{\alpha/2}(\nu)\} = \alpha/2$ (см. рис. 37). Мы проводим конкретный эксперимент: извлекаем две выборки, вычисляем выборочные средние и дисперсии и затем величину

$$t_{\text{эксп}} = \frac{m_1 - m_2}{s\sqrt{1/n_1 + 1/n_2}}.$$

Число $t_{\text{эксп}}$ есть единичная реализация случайной величины $\widetilde{t}$. Если оказывается, что $-t_{\alpha/2}(\nu) < t_{\text{эксп}} < t_{\alpha/2}(\nu)$ или $|t_{\text{эксп}}| < t_{\alpha/2}(\nu)$, то на уровне значимости α нулевая гипотеза принимается. Если же $t_{\text{эксп}} \leqslant -t_{\alpha/2}(\nu)$ или $t_{\text{эксп}} \geqslant t_{\alpha/2}(\nu)$, иными словами, если $|t_{\text{эксп}}| \geqslant t_{\alpha/2}(\nu)$, то на уровне значимости α нулевая гипотеза отклоняется, и мы говорим, что $\mu_1 \neq \mu_2$. Процедура проверки гипотезы о равенстве двух средних значений сводится к следующему. Выборочные дисперсии s_1^2 и s_2^2 сравниваются с помощью критерия F. Если гипотеза H_0: $\sigma_1^2 = \sigma_2^2$ принимается, то по выборочным дисперсиям вычисляется выборочное значение

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

По выборочным средним вычисляется значение $|t_{\text{эксп}}| = \dfrac{|m_1 - m_2|}{s\sqrt{1/n_1 + 1/n_2}}$, которое сравнивается с $t_{\alpha/2}$ $(\nu = n_1 + n_2 - 2)$.

ПРИМЕР VI-2. В табл. 27 приведены результаты эксперимента по действию яда в двух дозах (A и B) на комнатную муху. Определялось время реакции: сколько минут проходит от момента соприкосновения мухи с ядом до момента наступления паралича (муха падает). Является ли наблюдаемое различие во времени реакции статистически значимым?

Таблица 27

Время наступления паралича (мин) у комнатной мухи при воздействии двумя дозами яда (к примеру VI-2)

Доза A		Доза B	
x_{1i}	$y_{1i} = \lg x_{1i}$	x_{2i}	$y_{2i} = \lg x_{2i}$
3	0,48	2	0,30
5	0,70	5	0,70
5	0,70	5	0,70
7	0,85	7	0,85
9	0,95	8	0,90
9	0,95	9	0,95
10	1,00	14	1,15
12	1,08	18	1,26
20	1,30	24	1,38
24	1,38	26	1,42
24	1,38	26	1,42
34	1,53	34	1,53
43	1,63	37	1,57
46	1,66	42	1,62
58	1,76	90	1,95
140	2,15		

Обширные материалы по такого рода признакам показывают, что время реакции не имеет нормального распределения. Однако нормальное распределение можно получить, взяв логарифм времени реакции: $y = \lg x$. Поэтому будем работать с этим новым признаком. Имеем $n_1 = 16$; $n_2 = 15$. Выборочные средние $m_1 = 1{,}219$ и $m_2 = 1{,}180$. Необходимо сравнить их на 5 %-ном уровне значимости. Выборочные дисперсии $s_1^2 = 0{,}2074$ и $s_2^2 = 0{,}1928$ не различаются: $F_{\text{эксп}} 1{,}08 < F_{0{,}025}(15{,}14) = 2{,}95$. Поэтому вычисляем общую по эксперименту выборочную дисперсию $s_2 = [(16 -$

$-1)-0{,}2074+(15-1)\cdot 0{,}1928]/(16+15-2)=0{,}2004$. Вычисляем $|t_{\text{эксп}}|=$ $=|1{,}219-1{,}180|/(0{,}2004\sqrt{1/16+1/15})=0{,}24$.

По табл. III Приложения 1 находим, что эта величина много меньше $t_{0{,}025}(\nu=29)=2{,}05$. Таким образом, нулевая гипотеза не отвергается: различия между эффектами доз A и B статистически незначимы.

Особо следует рассмотреть случай неравных дисперсий. Если F-критерий выявил, что $\sigma_1^2\neq\sigma_2^2$, то величина $\nu s^2/\sigma^2$ уже не имеет χ^2-распределения. Однако специальное исследование показало, что случайная величина

$$\frac{\widetilde{m_1}-\widetilde{m}_2}{\sqrt{\widetilde{s}^2_{1/n_1}+\widetilde{s}^2_{2/n_2}}}$$

аппроксимируется t-распределением, но с числом степеней свободы ν, которое находят из соотношения

$$\frac{1}{\nu}=\frac{C^2}{\nu_1}+\frac{(1-C)^2}{\nu_2},$$

где

$$C=\frac{s_1^2/n_1}{s_1^2/n_1+s_2^2/n_2},$$

т. е. ν лежит между меньшим из двух чисел $\nu_1=n_1-1$ и $\nu_2=n_2-1$ и их суммой $\nu_1+\nu_2=n_1+n_2+2$.

Пример VI-3. В двух популяциях нивяника обыкновенного измерялась высота растений в сантиметрах. Получены следующие результаты. Популяция I: 48, 45, 50, 44, 42, 46, 49, 45, 41, 47, 44, 39, 41, 48, 52, 45, 46, 49. Популяция II: 38, 50, 35, 33, 45, 31, 54, 39, 39, 43, 47, 40, 35, 42, 45, 38, 35, 39. Значимы ли различия между популяциями?

Объемы выборок $n_1=n_2=18$. Выборочные средние $m_1=45{,}6$ и $m_2=40{,}4$. Выборочные дисперсии $s_1^2=11{,}90$ и $s_2^2=36{,}50$. Вычисляем $F_{\text{эксп}}=36{,}50/11{,}90=3{,}07$. Табл. IV Приложения 1 не содержит $\nu_1=\nu_2=$ $=17$. Ближайшее значение $F_{0{,}025}(15{,}16)=2{,}79$. Ясно, что тем более $F_{\text{эксп}}>$ $>F_{0{,}025}(17,17)$, т. е. нулевая гипотеза о равенстве дисперсий отвергается на 5 %-ном уровне значимости. Поэтому нельзя вычислять общую для обеих выборок дисперсию.

Находим

$$|t_{\text{эксп}}|=\frac{|45{,}6-40{,}4|}{\sqrt{\dfrac{11{,}90}{18}+\dfrac{36{,}50}{18}}}=3{,}17.$$

Вычисляем число степеней свободы: $C = 0{,}66/(0{,}66 + 2{,}03) = 0{,}245$; $\frac{1}{\nu} = \frac{0{,}245^2}{17} + \frac{0{,}759^2}{17} = 0{,}0374$. Отсюда $\nu = 26{,}7$. Табл. III Приложения 1 не содержит значений $t_{\alpha/2}$ для $\nu = 27$. Однако $|t_{\text{эксп}}| > t_{0{,}005}(26) = 2{,}78$, поэтому нулевая гипотеза о равенстве средних отвергается на 1 %-ном уровне значимости.

Таким образом, популяции различаются и по средней высоте растений, и по изменчивости этого признака.

§ 3. Сравнение средних значений двух зависимых нормальных распределений (парные наблюдения)

В биологии довольно часто встречаются ситуации, когда выборки зависимы. Например, испытания двух лекарственных препаратов могут проводиться таким образом, что каждый препарат испытывается на одном и том же животном. При этом, естественно, должны приниматься меры предосторожности, чтобы действия препаратов не перекрывались. Учитывая индивидуальность реакции разных особей, такой способ сравнения препаратов должен оказаться и более эффективным, поскольку нас интересует не сравнение реакции разных особей, а сравнение эффективности препаратов.

Пусть распределение признака в одном варианте опыта задается случайной величиной $\widetilde{x}_1$, а в другом — $\widetilde{x}_2$. Возьмем случайную величину, являющуюся их разностью: $\widetilde{d} = \widetilde{x}_1 - \widetilde{x}_2$, и будем полагать, что $\widetilde{d}$ распределена нормально. Из генеральной совокупности разностей $\widetilde{d}_i = \widetilde{x}_{1i} - \widetilde{x}_{2i}$ произведем выборку объема n независимых пар наблюдений $(x_{11} - x_{21})$, $(x_{12} - x_{22})$, …, $(x_{1n} - x_{2n})$. Если согласно нулевой гипотезе, требующей проверки, средние значения μ_1 и μ_2 не различаются, то $\widetilde{d} \sim N(0; \sigma_d^2)$, где значение σ_d^2 неизвестно, но можно получить лишь ее выборочное значение s_d^2. Проведем рассуждения, аналогичные рассуждениям предыдущего параграфа. В качестве статистики для оценки разности средних $(\mu_1 - \mu_2)$ возьмем

$$\widetilde{m}_d = \frac{1}{n} \sum_{i=1}^{n} \widetilde{d}_i.$$

Очевидно, если справедлива нулевая гипотеза, то

$$\widetilde{m}_d \sim N(0; \sigma_d^2/n)$$

и, следовательно,

$$\frac{\widetilde{m}_d}{\sigma_d/\sqrt{n}} \sim N(0;1).$$

Поскольку

$$\frac{(n-1)\widetilde{s}_d^2}{\sigma_d^2} \sim \chi^2\nu, \quad \nu = n-1,$$

то

$$\frac{\widetilde{m}_d}{s_d/\sqrt{n}} = \frac{\widetilde{m}_d}{s_d/\sqrt{n}} \div \frac{\widetilde{s}_d}{\sigma_d} = \widetilde{t} \sim t(\nu), \quad \nu = n-1.$$

Таблица 28

Содержание крахмала (усл. ед.) в картофеле (к примеру VI-4)

Номер клубня (i)	Содержание крахмала		Разность
	метод I (x_{1i})	метод II (x_{2i})	$(d_i = x_{1i} - x_{2i})$
1	21,7	21,5	0,2
2	18,7	18,7	0,0
3	18,3	18,3	0,0
4	17,5	17,4	0,1
5	18,5	18,3	0,2
6	15,6	15,4	0,2
7	17,0	16,7	0,3
8	16,6	16,9	−0,3
9	14,0	13,9	0,1
10	17,2	17,0	0,2
11	21,7	21,4	0,3
12	18,6	18,6	0,0
13	17,9	18,0	−0,1
14	17,7	17,6	0,1
15	18,3	18,5	−0,2
16	16,6	16,5	0,1

Таким образом, мы приходим к следующей процедуре сравнения парных наблюдений, которая иногда называется *парным t-критерием.* По выборочным данным вычисляем разности $d_i = x_{1i} - x_{2i}$, среднюю раз-

ность $m_d = \frac{1}{n}\sum_{i=1}^{n} d_i$ (учетом знаков d_i!), выборочную дисперсию $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - m)^2$ и, наконец, $t_{\text{эксп}} = \frac{m_d}{s_d/\sqrt{n}}$ и $\nu = n - 1$.

ПРИМЕР VI-4 [Хальд, 1956]. Сравнивались два метода определения крахмала в картофеле (табл. 28). Были взяты 16 клубней с изменяющимся в широких пределах содержанием крахмала и к каждому клубню применялись оба метода. Есть ли статистические различия между методами?

Естественно, при сравнении методов не должны учитываться различия в содержании крахмала в разных клубнях. Поэтому рассмотрим разности $d_i = x_{1i} - x_{2i}$. Средняя разность $m_d = 0{,}075$, $s_d^2 = 0{,}0287$ при $\nu = 15$ и $t_{\text{эксп}} = \frac{0{,}075}{0{,}169/4} = 1{,}78$, по табл. III Приложения 1 $t_{0{,}025}(15) = 2{,}13$. Таким образом, нет оснований отклонять нулевую гипотезу об одинаковой чувствительности обоих методов на уровне значимости $\alpha = 0{,}05$.

§ 4. Сравнение параметров p_1 и p_2 двух биномиальных распределений

Необходимо сравнить выборочные параметры $h_1 = k_1/n_1$ и $h_2 = k_2/n_2$, т. е. проверить H_0: $p_1 = p_2 = p$. В предположении правильности нулевой гипотезы объединим данные обеих выборок: $k_1 + k_2 = k$ и $n_1 + n_2 = n$ и вычислим $h = k/n$. Здесь оказывается эффективной аппроксимация биномиального распределения нормальным

$$\widetilde{u} = \frac{|\widetilde{h}_1 - \widetilde{h}_2| - \frac{1}{2n}}{\sqrt{\widetilde{n}(1-\widetilde{h})(1/n_1 + 1/n_2)}} \sim N(0;1).$$

Формула применима при условии, что $n_i h \geqslant 5$ и $n_i(1-h) \geqslant 5$, $i = 1, 2$. Это позволяет, вычисляя $u_{\text{эксп}}$, построить критерий значимости, основанный на нормальном распределении, который мы будем для простоты называть *u-критерием*. Если $|u_{\text{эксп}}| < u_{\alpha/2}$, то нулевая гипотеза принимается; если $|u_{\text{эксп}}| \geqslant u_{\alpha/2}$ — отвергается на уровне значимости α.

ПРИМЕР VI-5. Решим задачу, сформулированную в примере IV-12 (см. табл. 9). Находим $h_1 = 69/276 = 0{,}250$, $h_2 = 185/395 = 0{,}468$ и (в предположении H_0: $p_1 = p_2$) $h = 254/671 = 0{,}379$. Поскольку $n_1 h = 104{,}6$; $n_1(1-h) = 171{,}4$; $n_2 h = 149{,}7$; $n_2(1-h) = 245{,}3$, — все больше 5,

то возможно использование u-критерия: $u_{\text{эксп}} = -0{,}218/0{,}0380 = -5{,}73$. Поправка на дискретность $1/(2 \cdot 671) = 0{,}00075$ ничтожна и ею можно пренебречь. Поскольку $|u_{\text{эксп}}| > u_{0{,}0005} = 3{,}27$, то порода Ван-Скоя более устойчива к заболеванию на 0,1%-ном уровне значимости.

Пример VI-6 [Бейли, 1962]. В табл. 29 приведены результаты лечения редкого заболевания двумя разными методами. Что можно сказать об относительной эффективности этих методов?

Условия аппроксимации биномиального распределения нормальным здесь не выполняются, поэтому приходится прибегать к другому способу статистического анализа — *точному критерию Фишера*.

Таблица 29

Результаты лечения редкого заболевания двумя методами (к примеру VI-6)

Метод	Число больных		Итого
	выздоровевших	не имевших улучшения	
А	4	1	5
Б	0	4	4
Итого	4	5	9

Таблица 30

К построению точного критерия Фишера

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

Рассмотрим таблицу 2×2 в общем виде (табл. 30) и проследим, как приходит к соответствующему критерию сам Р. А. Фишер:

«Пусть p есть вероятность какого-либо события, тогда вероятность того, что оно произойдет a раз в $(a+b)$ независимых испытаниях, определяется биномиальной формулой

$$\frac{(a+b)!}{a!b!} p^a q^b,$$

где $q = 1 - p$. Вероятность того, что в $(c+d)$ испытаниях оно произойдет

c раз, есть

$$\frac{(c+d)!}{c!d!}p^c q^d.$$

Следовательно, вероятность наблюдать в таблице 2×2 численности a, b, c и d равна произведению

$$\frac{(a+b)!(c+d)!}{a!b!c!d!}p^{a+c}q^{b+d}$$

и в общем случае она должна быть неизвестной, если неизвестно p. Однако неизвестный множитель, содержащий p и q, будет одним и тем же для всех таблиц, имеющих одинаковые маргинальные численности $a+c$, $b+d$, $a+b$, $c+d$. Так что вероятность любого из возможных наборов наблюдений, имеющих одинаковые маргинальные численности, оказывается пропорциональной величине

$$\frac{1}{a!b!c!d!},$$

каким бы ни было значение p, или, другими словами, для любых совокупностей, в которых четыре численности находятся в пропорциональном отношении. Тогда можно найти, что сумма величин $\frac{1}{a!b!c!d!}$ для всех выборок, имеющих одинаковые маргинальные численности, равна

$$\frac{n!}{(a+b)!(c+d)!(a+c)!(b+d)!},$$

где $n = a+b+c+d$. Таким образом, для данных маргинальных численностей вероятность любого наблюдаемого набора входящих значений равна

$$\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!} \cdot \frac{1}{a!b!c!d!} \text{ »}$$

[Фишер, 1958].

Вернемся теперь к примеру VI-6. В соответствии с проведенными рассуждениями, в табл. 31 приведены все возможные варианты таблиц 2×2 при постоянстве сумм по строкам и столбцам, соответствующие им вероятности, а в последней строке — также соответствующие им разности эффективности методов лечения А и Б.

Вероятность получить наблюдаемый результат случайно (при равной эффективности методов А и Б) равна, таким образом, $5/126 = 3{,}97\,\%$.

Таблица 31

К построению точного критерия Фишера (данные табл. 29)

Варианты таблиц 2×2	$\begin{array}{c\|c} 4 & 1 \\ \hline 0 & 4 \end{array}$	$\begin{array}{c\|c} 3 & 2 \\ \hline 1 & 3 \end{array}$	$\begin{array}{c\|c} 2 & 3 \\ \hline 2 & 2 \end{array}$	$\begin{array}{c\|c} 1 & 4 \\ \hline 3 & 1 \end{array}$	$\begin{array}{c\|c} 0 & 5 \\ \hline 4 & 0 \end{array}$
Вероятность	$\frac{5}{126}$	$\frac{40}{126}$	$\frac{60}{126}$	$\frac{20}{126}$	$\frac{1}{126}$
А $-$ Б, %	$+80$	$+35$	-10	-55	-100

Поскольку ранее не было оговорено, что метод А не может быть хуже Б, то мы должны учесть и вероятность еще большего отклонения, пусть оно и имеет другой знак: $1/126 = 0{,}79\,\%$. Таким образом, вероятность получить случайно наблюдаемое или еще большее отклонение от нулевой гипотезы равна $6/126 = 4{,}76\,\%$. Другими словами, лекарство А более эффективно на $5\,\%$-ном уровне значимости.

На практике нет необходимости проводить подобные вычисления, достаточно обратиться к табл. VI Приложения 1. Находим в первом столбце структуру, отвечающую нашему эксперименту: $\begin{array}{c\|c} a & 1 \\ \hline 0 & 4 \end{array}$, двигаясь по этой строке, видим, что $P\{a \geqslant 4\} \leqslant 0{,}05$, как и в результате прямых вычислений. Если бы в нашем эксперименте было $a = 8$, это означало бы $P\{a \geqslant 8\} \leqslant 0{,}01$. В случае $a = 3$ имели бы $P\{a \geqslant 3\} > 0{,}05$, т. е. нулевая гипотеза должна была бы быть принята.

§ 5. Сравнение параметров λ_1 и λ_2 распределений Пауссона

Пусть $\widetilde{x}_1$ и $\widetilde{x}_2$ имеют распределение Пуассона с параметрами λ_1 и λ_2. По двум независимым выборкам объема n_1 и n_2 получаем оценки параметров λ_i: m_1 и m_2. При довольно мягких для экспериментатора условиях, а именно $m_1 + m_2 \geqslant 5$, возможна следующая аппроксимация распределения

Пуассона нормальным:

$$\widetilde{u} = \frac{\widetilde{m}_1 - \widetilde{m}_2}{\widetilde{m}_1/n_1 + \widetilde{m}_2/n_2} \sim N(0;1).$$

Это позволяет проверить H_0: $\lambda_1 = \lambda_2$. Подобно предыдущему параграфу получаем критерий, основанный на нормальном распределении: если $|u_{\text{эксп}}| < \alpha/2$, то нулевая гипотеза принимается; если $|u_{\text{эксп}}| \geqslant \alpha/2$, то нулевая гипотеза отвергается на уровне значимости α.

ПРИМЕР VI-7. Решим задачу, сформулированную в примере IV-13.

Имеем $n_1 = n_2 = 1$; $m_1 = 42$, $m_2 = 71$. Условие $m_1 + m_2 \geqslant 5$ выполняется. Поэтому $|u_{\text{эксп}}| = |42 - 71|/\sqrt{42 + 71} = 2{,}73$. Так как $|u_{\text{эксп}}| > u_{0{,}005} = 2{,}58$, то H_0: $\lambda_1 = \lambda_2$ отвергается на 1 %-ном уровне значимости.

§ 6. О сравнении параметров неизвестных распределений

Если распределение изучаемого признака не нормально или неизвестно, то для сравнения параметров предпочтительно применение *непараметрических критериев*, свободных от предположения о виде распределения.

Среди множества непараметрических критериев предпочтение отдается тем, эффективность которых в случае выборок из нормального распределения сравнима с эффективностью параметрических критериев типа t-критерия для сравнения средних или F-критерия для сравнения дисперсий и остается высокой в случае выборок из не нормальных распределений. Под эффективностью критерия подразумевается его чувствительность к отклонениям от проверяемой нулевой гипотезы.

К сожалению, существующие непараметрические критерии для сравнения параметров рассеяния значительно уступают в эффективности F-критерию. Поэтому мы рассмотрим лишь два критерия для сравнения параметров положения, которые по своей эффективности способны конкурировать с t-критерием: критерий Вилкоксона–Манна–Уитни для случая двух независимых выборок и парный критерий Вилкоксона для случая парных наблюдений.

§ 7. Сравнение параметров положения двух неизвестных независимых распределений

Пусть $(x_{11}, \ldots, x_{1n_1})$ и $(x_{21}, \ldots, x_{2n_2})$ — две независимые выборки объема n_1 и n_2 соответственно и случайные величины $\widetilde{x}_{1i}$ и $\widetilde{x}_{2i}$ имеют непрерывные функции распределения $F_1(x)$ и $F_2(x)$. С помощью критерия Вилкоксона–Манна–Уитни можно проверить гипотезу H_0: $\tau_1 = \tau_2$, где τ_1 и τ_2 — некие параметры положения сравниваемых распределений, не обязательно задаваемые в явном виде. При этом генеральные совокупности, которым принадлежат выборки, не обязательно должны характеризоваться одинаковыми параметрами рассеяния δ_1 и δ_2 (рис. 40, *а*).

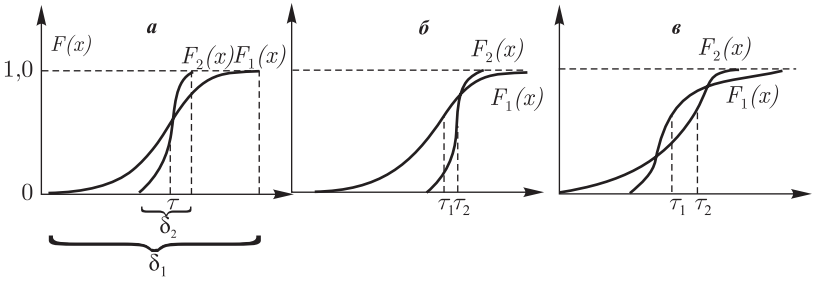


Рис. 40. К каким отклонениям от гипотезы H_0: $\tau_1 = \tau_2$ чувствителен критерий Вилкоксона–Манна–Уитни?

а — функции распределения $F_1(x)$ и $F_2(x)$ имеют равные параметры положения $\tau_1 = \tau_2 = \tau$, но различные параметры рассеяния $\sigma_1 \neq \sigma_2$; гипотеза H_0 принимается;

б — $F_1(x)$ стохастически больше $F_2(x)$; соответственно, $\tau_1 > \tau_2$; гипотеза H_0 отвергается;

в — параметры положения различаются: $\tau_1 > \tau_2$, однако $F_2(x)$ стохастически не отличается от $F_1(x)$, и критерий Вилкоксона–Манна–Уитни может не отвергнуть H_0

В качестве альтернативы подразумевается H_1: $\tau_1 \neq \tau_2$, и если, например, $\tau_2 > \tau_1$, то говорят, что $F_1(x)$ стохастически больше $F_2(x)$ (рис. 40, *б*). Именно к таким альтернативам, когда для большинства значений x $F_1(x) > F_2(x)$ или $F_1(x) < F_2(x)$, чувствителен описываемый критерий. Критерий этот, однако, оказывается малочувствительным к аль-

тернативам типа изображенной на рис. 40, *в*, когда одна из функций распределения попеременно то больше, то меньше другой.

Обратимся к статистике t-критерия (см. § 2). В ней в качестве естественной меры различия двух параметров положения μ_1 и μ_2 используется разность выборочных средних $(\widetilde{m}_1 - \widetilde{m}_2)$. Заметим, что эту разность можно представить в виде

$$\widetilde{m}_1 - \widetilde{m}_2 = \frac{1}{n_1}\sum_{i=1}^{n_1}\widetilde{x}_{1i} - \frac{1}{n_2}\sum_{j=1}^{n_2}\widetilde{x}_{2j} = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\widetilde{x}_{1i} - \widetilde{x}_{2j}),$$

т. е. разность двух средних равна среднему всех $n_1 n_2$ возможных разностей $(\widetilde{x}_{1i} - \widetilde{x}_{2j})$. Это наводит на мысль построить меру различия двух параметров положения τ_1 и τ_2, которая не будет зависеть от вида сравниваемых распределений, если вместо разностей $(\widetilde{x}_{1i} - \widetilde{x}_{2j})$ использовать их знаки. Для нахождения знаков разностей удобно использовать прямоугольную матрицу сравнений, в клетках которой отмечаются знаки (sign) соответствующих разностей (табл. 32).

Таблица 32

Матрица сравнений к построению критерия Вилкоксона–Манна–Уитни

x_{1i} \ x_{2j}	$\widetilde{x}_{21}$	$\widetilde{x}_{22}$	$\dots$	$\widetilde{x}_{2n_2}$
$\widetilde{x}_{11}$	$\mathrm{sign}(\widetilde{x}_{11} - \widetilde{x}_{21})$	$\mathrm{sign}(\widetilde{x}_{11} - \widetilde{x}_{22})$	$\dots$	$\mathrm{sign}(\widetilde{x}_{11} - \widetilde{x}_{2n_2})$
$\widetilde{x}_{12}$	$\mathrm{sign}(\widetilde{x}_{12} - \widetilde{x}_{21})$	$\mathrm{sign}(\widetilde{x}_{12} - \widetilde{x}_{22})$	$\dots$	$\mathrm{sign}(\widetilde{x}_{12} - \widetilde{x}_{2n_2})$
$\cdot$	$\cdot$	$\cdot$	$\dots$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\dots$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\dots$	$\cdot$
$\widetilde{x}_{1n_1}$	$\mathrm{sign}(\widetilde{x}_{1n_1} - \widetilde{x}_{21})$	$\mathrm{sign}(\widetilde{x}_{1n_1} - \widetilde{x}_{22})$	$\dots$	$\mathrm{sign}(\widetilde{x}_{1n_1} - \widetilde{x}_{2n_2})$

Положительным разностям $(\widetilde{x}_{1i} - \widetilde{x}_{2j}) > 0$ припишем знак «+», отрицательным $(\widetilde{x}_{1i} - \widetilde{x}_{2j}) < 0$ — знак «−». Введем две статистики: $\widetilde{U}^+$ — число всех положительных разностей, т. е. число знаков «+» в матрице, и $\widetilde{U}^-$ — число всех отрицательных разностей, т. е. число знаков «−» в той же матрице. Соотношение этих статистик оказывается достаточно эффективной мерой различия параметров положения τ_1 и τ_2 двух независимых совокупностей. Действительно, если $\tau_1 = \tau_2 = \tau$ (см. рис. 40, *а*), то $\widetilde{U}^+ = \widetilde{U}^-$,

т. е. должно быть поровну положительных и отрицательных разностей. Если же $\tau_2 > \tau_1$ (см. рис. 40, *б*), то $\widetilde{U}^+ < \widetilde{U}^-$, т. е. число отрицательных разностей будет больше, чем положительных. Таким образом, мы сводим задачу сравнения параметров положения двух непрерывных распределений к сравнению двух выборочных дискретных случайных величин $\widetilde{U}^+$ и $\widetilde{U}^-$.

Статистика $\widetilde{U}^+$ (или $\widetilde{U}^-$) называется *статистикой Манна–Уитни*, а соответствующий критерий — *критерием Вилкоксона–Манна–Уитни*. Ф. Вилкоксон в 1945 г. одним из первых разработал этот критерий, основанный на замене выборочных значений их рангами. Вслед за ним в 1947 г. Х. Б. Манн и Д. Р. Уитни предложили статистику $\widetilde{U}^+$ и показали, что она идентична (с точностью до константы) ранговой статистике Вилкоксона. Статистика Манна–Уитни есть мера различия таких параметров положения τ_1 и τ_2, разность которых $\triangle = \tau_1 - \tau_2$ оценивается статистикой $\triangle = \mathrm{med}(\widetilde{x}_{1i} - \widetilde{x}_{2j})$, являющейся медианой всех $n_1 n_2$ разностей $(x_{1i} - \widetilde{x}_{2j})$.

Пример VI-8 (Н. Б. Петров, 1972 г.). Нуклеотидный состав ДНК может быть использован в качестве таксономического признака. Сравним, например, содержание ГЦ-пар нуклеотидов в ДНК беспозвоночных двух родов: Cancer и Drosophila (табл. 33).

Таблица 33

Содержание ГЦ-пар в ДНК беспозвоночных (к примеру VI-8)

Виды рода *Cancer*	x_{1i}, мол. %	Виды рода *Drosophila*	x_{2j}, мол. %
C. antennaris	*38,4*	*D. virilis*	*40,0*
C. borealis	*40,2*	*D. melanogaster*	*41,2*
C. gracilis	*38,4*	*D. simulans*	*42,5*
C. irroratus	*40,0*	*D. funebris*	*38,5*
C. magister	*39,4*		
C. productus	*39,0*		
C. oregonensis	*40,5*		
C. pagurus	*38,0*		

Составим матрицу сравнений (табл. 34). Можно видеть, что на практике встречаются нулевые разности: $(x_{1i} - x_{2j}) = 0$. В таких случаях говорят, что имеет место *совпадение*, т. е. $x_{1i} = x_{2j}$.

Таблица 34 Матрица сравнений для данных примера VI-8.

x_{1i}	x_{2j}			
	40,0	41,2	41,5	38,5
38,4	$-$	$-$	$-$	$-$
40,2	$+$	$-$	$-$	$+$
33,4	$-$	$-$	$-$	$-$
40,0	$\pm$	$-$	$-$	$+$
39,4	$-$	$-$	$-$	$+$
39,0	$-$	$-$	$-$	$+$
40,5	$+$	$-$	$-$	$+$
38,0	$-$	$-$	$-$	$-$

Нулевые разности будем обозначать знаком «$\pm$». Искомое выборочное значение $U^{+}_{\text{эксп}}$ статистики $\widetilde{U}^{+}$ находим суммированием количества знаков «$+$» и половины количества знаков «$\pm$»:

$$U^{+}_{\text{эксп}} = n(+) + \frac{1}{2}n(\pm) = 7 + 0{,}5 = 7{,}5.$$

Соответственно, выборочное значение $U^{-}_{\text{эксп}}$ статистики $\widetilde{U}^{-}$ есть сумма количества знаков «$-$» и половины количества знаков «$\pm$»:

$$U^{-}_{\text{эксп}} = n(-) + \frac{1}{2}n(\pm) = 24 + 0{,}5 = 24{,}5.$$

Очевидно, что для данных объемов выборок n_1 и n_2 имеет место соотношение $U^{+}_{\text{эксп}} + U^{-}_{\text{эксп}} = n_1 n_2 = 32$.

Матрица сравнений может принимать различные «значения»: от матрицы с одними плюсами (когда все x_{1i} больше всех x_{2j}) до матрицы с одними минусами (когда все x_{1i} меньше всех x_{2j}). Общее число всех возможных конфигураций такой матрицы есть $(n_1+n_2)!/n_1!n_2!$, т. е. равно общему числу всех возможных сочетаний из n_1 значений x_{1i} и n_2 значений x_{2j}. Как видим, ситуация равносильна подсчету числа всех возможных способов извлечения (без возвращения) n_1 шаров из урны, содержащей $N = n_1 + n_2$ шаров.

Если обе сравниваемые совокупности имеют одно и то же распределение, то все возможные конфигурации матрицы сравнения будут равновероятны, и вероятность любой из них равна $n_1!n_2!/(n_1 + n_2)!$. Другими словами, если имеет место равенство параметров положения $\tau_1 = \tau_2$, то

для статистики $\widetilde{U}^+$ (или $\widetilde{U}^-$) можно получить распределение вероятностей, которое зависит только от n_1 и n_2. Для конкретного рассматриваемого нами случая ($n_1 = 8, n_2 = 4$) вид этого распределения представлен на рис. 41. Здесь на левой оси ординат отложены значения a_i, т. е. сколько раз $\widetilde{U}^+$ принимает значения U_i^+, а на правой оси ординат отложены соответствующие им вероятности $p_i(U^+) = \dfrac{a_i(n_1+n_2)!}{n_1!n_2!} = \dfrac{a_1(8+4)!}{8!4!} = \dfrac{a_i}{495}$. На оси абсцисс указаны значения U_i^+.

Можно видеть, что случайная величина $\widetilde{U}^+$ пробегает значения от 0 до $n_1n_2 = 32$; ее распределение симметрично относительно среднего значения $n_1n_2/2 = 16$, которое при H_0, очевидно, равно математическому ожиданию случайных величин $\widetilde{U}^+$ и $\widetilde{U}^-$:

$$E\widetilde{U}^+ = E\widetilde{U}^- = \frac{1}{2}n_1n_2.$$

Заштрихованные области («хвосты» распределения) соответствуют вероятностям $P\{\widetilde{U}^+ \leqslant 4\} = 0{,}024$ и $P\{\widetilde{U}^+ \geqslant 28\} = 0{,}024$. Следовательно, при $n_1 = 8$ и $n_2 = 4$ значения $U^+ = 4$ и $U^- = 28$ являются критическими для уровня значимости $\alpha = 2 \cdot 0{,}024 \approx 0{,}05$.

В силу того, что для фиксированных объемов выборок n_1 и n_2 между значениями $U^+_{\text{эксп}}$ и $U^-_{\text{эксп}}$ имеет место однозначное соответствие, для построения критерия достаточно использовать любое одно из них; договорились брать меньшее, т. е. $U_{\text{эксп}} = \min(U^+_{\text{эксп}}, U^-_{\text{эксп}})$. Поэтому таблицы критических значений (табл. VIII Приложения 1) содержат лишь значения $U_{\alpha/2}(n_1, n_2)$, соответствующие левому хвосту распределения статистики Манна–Уитни, такие, что $P\{\widetilde{U} \leqslant U_{\alpha/2}(n_1, n_2)\} = \alpha/2$[2]. Таким образом, если $U_{\text{эксп}} \leqslant U_{\alpha/2}(n_1, n_2)$, то гипотезу H_0 отвергают при уровне значимости α; если $U_{\text{эксп}} > U_{\alpha/2}(n_1, n_2)$, гипотезу H_0 принимают.

В нашем примере $U_{\text{эксп}} = 7,5$ больше табличного значения $U_{0,025}(4, 8) = 4$, и поэтому имеющийся экспериментальный материал не дает оснований сомневаться в справедливости гипотезы об отсутствии различий в нуклеотидном составе ДНК двух родов беспозвоночных на уровне значимости $\alpha = 0{,}05$.

Распределение на рис. 41 напоминает нормальное и отражает реальное свойство распределения статистики $\widetilde{U}$ стремиться к нормальному. Это

[2] Заметим, что эти таблицы отличаются от таблиц распределений t, F и χ^2, в которых значения $c_{\alpha/2}$ соответствуют правому хвосту распределения, т. е. $P\{\widetilde{x} \geqslant x_{\alpha/2}\} = \alpha/2$ (см. §§ 1–2 в гл. V, §§3–6).

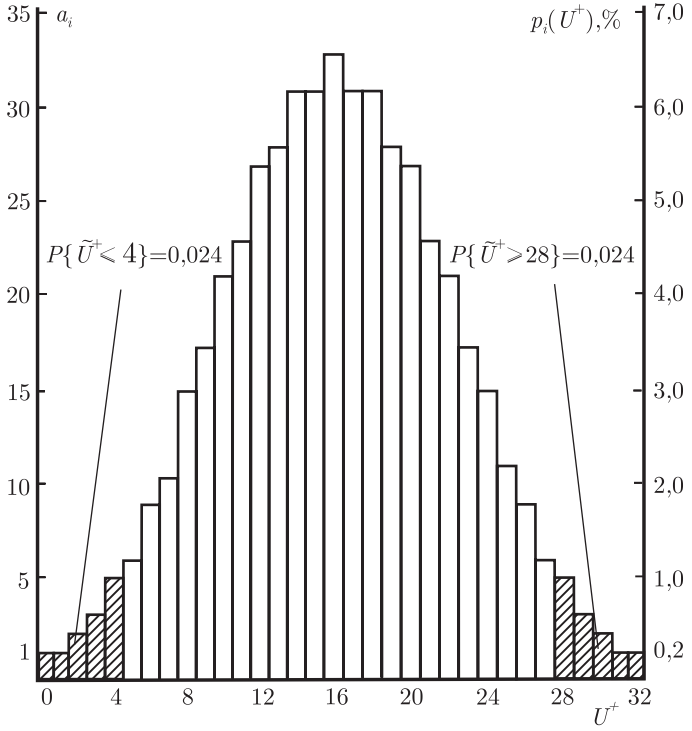


Рис. 41. Распределение статистики $\widetilde{U}$ Манна–Уитни при $n_1 = 8$ и $n_2 = 4$. Для наглядности дискретные значения представлены в виде единичных отрезков на оси абсцисс. Сумма заштрихованных площадей есть $\alpha = 0{,}048 \approx 0{,}05$

означает, что нет надобности в обширных таблицах и для больших объемов выборок (на практике уже для $n_1 \geqslant n_2 \geqslant 8$) можно использовать нормальную аппроксимацию

$$\widetilde{u} = \frac{\widetilde{U} - E\widetilde{U}}{\sqrt{D\widetilde{U}}} \sim N(0;1),$$

где

$$E\widetilde{U} = \frac{1}{2}n_1 n_2 \quad \text{и} \quad D\widetilde{U} = \frac{1}{12}n_1 n_2 (n_1 + n_2 + 1).$$

Если имеют место совпадения, то статистика $\widetilde{U}$ перестает быть свободной от распределения (она становится зависимой от неизвестного распределения совпадающих значений), и критерий становится приближенным. При этом изменяется (уменьшается) дисперсия $D\widetilde{U}$, выражение для которой приобретает вид

$$D\widetilde{U} = \frac{1}{12}n_1n_2(n_1+n_2+1) - \frac{n_1n_2}{12(n_1+n_2)(n_1+n_2+1)}\sum_{i=1}^{g}(c_i^3 - c_i),$$

где g — число групп совпадений и c_i — число совпадающих значений в i-й группе. Совпадениями являются только те случаи, когда совпадающие значения принадлежат разным выборкам, а группу образуют одинаковые по величине совпадающие значения. Например, в матрице число групп совпадений $g = 2$, а именно группа «двоек», в которой число совпадающих значений $c_1 = 2$, и группа «девяток», в которой $c_2 = 3$; «четверки» не являются совпадениями.

x_{1i} \ x_{2j}	2	4	4	9
2	$\pm$	$-$	$-$	$-$
9	$+$	$+$	$+$	$\pm$
9	$+$	$+$	$+$	$\pm$

Таким образом, при достаточно больших объемах выборок n_1 и n_2 гипотеза H_0: $\tau_1 = \tau_2$ о равенстве параметров положения двух неизвестных распределений становится равносильной гипотезе H_0: $\widetilde{U} \sim N(E\widetilde{U}, D\widetilde{U})$, которую можно проверить с помощью u-критерия. Для этого надо вычислить значение

$$|u_{\text{эксп}}| = \frac{|U_{\text{эксп}} - E\widetilde{U}|}{\sqrt{D\widetilde{U}}}$$

и сравнить его с табличным значением $u_{\alpha/2}$.

§ 8. Сравнение параметров положения двух неизвестных распределений в случае парных наблюдений

Числитель статистики парного t-критерия для сравнения парных наблюдений $\widetilde{m}_d = \frac{1}{n}\sum_{i=1}^{n}\widetilde{d}_i$ (см. § 3) можно привести к следующему виду:

$$2\widetilde{m}_d = \frac{2}{n(n+1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(\widetilde{d}_i + \widetilde{d}_j), \quad i \leqslant j,$$

т. е. удвоенное среднее равно среднему всех $\frac{1}{2}n(n+1)$ возможных сумм $(\widetilde{d}_i + \widetilde{d}_j), i \leqslant j$. Поскольку величины d_i являются разностями парных величин $\widetilde{x}_{1i}$ и $\widetilde{x}_{2j}$ и могут принимать как положительные, так и отрицательные значения, суммы $(\widetilde{d}_i + \widetilde{d}_j)$ также могут быть положительными или отрицательными. Если теперь вместо сумм $(\widetilde{d}_i + \widetilde{d}_j)$ использовать их знаки, то число положительных $\widetilde{W}^+$ или отрицательных $\widetilde{W}^-$ сумм $(\widetilde{d}_i + \widetilde{d}_j), i \leqslant j$, может служить статистикой непараметрического критерия для сравнения парных наблюдений, носящего название *парного критерия Вилкоксона*.

Для нахождения сумм $(\widetilde{d}_i + \widetilde{d}_j), i \leqslant j$, удобно использовать треугольную матрицу сравнений, в клетках которой отмечают знаки соответствующих сумм: $\operatorname{sign}(\widetilde{d}_i + \widetilde{d}_j), i \leqslant j$ (табл. 35). Естественно приписать положительным суммам $(\widetilde{d}_i + \widetilde{d}_j) > 0$ знак «+», а отрицательным $(\widetilde{d}_i + \widetilde{d}_j) < 0$ — знак «−». Тогда статистика $\widetilde{W}^+$ есть число плюсов в такой матрице, а статистика $\widetilde{W}^-$ — число минусов. Их соотношение может служить критерием при сравнении параметров положения парных наблюдений τ_1 и τ_2. Действительно, если $\tau_1 = \tau_2$, то $\widetilde{W}^+ = \widetilde{W}^-$, т. е. число плюсов и минусов должно быть одинаковым. Если же, например, $\tau_1 > \tau_2$, то $\widetilde{W}^+ < \widetilde{W}^-$, т. е. минусов будет больше, чем плюсов. Статистика парного критерия Вилкоксона $\widetilde{W}^+$ есть мера различия таких параметров положения совокупности парных наблюдений τ_1 и τ_2, разность которых Δ оценивается статистикой $\widetilde{\Delta} = \operatorname{med}\left[\frac{1}{2}(\widetilde{d}_i + \widetilde{d}_j)\right], i \leqslant j$, являющейся медианой всех $\frac{1}{2}n(n+1)$ полусумм $\frac{1}{2}(\widetilde{d}_i + \widetilde{d}_j)$. Соответственно, парный критерий Вилкоксона проверяет нулевую гипотезу о равенстве нулю этой медианы.

ПРИМЕР VI-9 (Ю. Н. Стройков и др., 1981 г.). У 14 крыс замеряли суммарно-пороговый показатель, т. е. значения напряжения электрическо-

Таблица 35

Матрица сравнений к построению парного критерия Вилкоксона

d_1	$\operatorname{sign}(\widetilde{d}_1+\widetilde{d}_1)$			
$\widetilde{d}_2$	$\operatorname{sign}(\widetilde{d}_2+\widetilde{d}_1)$	$\operatorname{sign}(\widetilde{d}_2+\widetilde{d}_2)$		
$\cdot$	$\cdot$	$\dots$		
$\cdot$	$\cdot$	$\dots$		
$\cdot$	$\cdot$	$\dots$		
$\widetilde{d}_n$	$\operatorname{sign}(\widetilde{d}_n+\widetilde{d}_1)$	$\operatorname{sign}(\widetilde{d}_n+\widetilde{d}_2)$	$\dots$	$\operatorname{sign}(\widetilde{d}_n+\widetilde{d}_n)$
	$\widetilde{d}_1$	$\widetilde{d}_2$	$\dots$	$\widetilde{d}_n$

го тока в вольтах, при котором вызывается судорожное сокращение мышц. В табл. 36 приведены полученные значения показателя до (x_{1i}) и после (x_{2j}) ингаляции метилового эфира нитроуксусной кислоты.

Таблица 36

Значение суммарно-порогового показателя (B) у крыс до и после воздействия метиловым эфиром нитроуксусной кислоты (к примеру VI-9)

Номер животного (i)	x_{1i}	x_{2i}	d_i
1	5,2	6,5	−1,3
2	5,8	5,1	+0,7
3	5,7	5,5	+0,2
4	4,0	4,4	−0,4
5	5,3	6,0	−0,7
6	4,8	5,3	−0,5
7	4,5	5,5	−1,0
8	4,7	5,5	−0,8
9	6,0	6,6	−0,6
10	4,8	6,0	−1,2
11	4,0	4,0	0,0
12	5,1	5,1	0,0
13	4,5	4,5	0,0
14	4,0	4,0	0,0

Из табл. 36 видно, что на практике возможны нулевые разности $d_i = 0$. Для рассматриваемого парного критерия они не несут никакой информации и их просто отбрасывают. При этом изменяется только объем выборки,

Таблица 37

Матрица сравнений для данных примера VI-9

$-1,3$	$-$									
$+0,7$	$-$	$+$								
$+0,2$	$-$	$+$	$+$							
$-0,4$	$-$	$+$	$-$	$-$						
$-0,7$	$-$	$\pm$	$-$	$-$	$-$					
$-0,5$	$-$	$+$	$-$	$-$	$-$	$-$				
$-1,0$	$-$	$-$	$-$	$-$	$-$	$-$	$-$			
$-0,8$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$		
$-0,6$	$-$	$+$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	
$-1,2$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$
d_{1i} / d_{2i}	$-1,3$	$+0,7$	$+0,2$	$-0,4$	$-0,7$	$-0,5$	$-1,0$	$-0,8$	$-0,6$	$-1,2$

который в данном случае становится фактически равным $N = n - n(0) = 10$, где $n(0)$ — количество нулевых разностей.

Составим матрицу сравнений (табл. 37). В случаях, когда имеют место совпадения, т. е. когда $d_i = -d_j$, получающимся нулевым суммам $(d_i + d_j = 0)$ будем приписывать знак «±», тогда искомое выборочное значение $W^+_{\text{эксп}}$ статистики $\widetilde{W^+}$ есть сумма количества знаков «+» и половины количества знаков «±»: $W^+_{\text{эксп}} = n(+) + n(\pm)1/2 = 6 + 0,5 = 6,5$; соответственно, $W^-_{\text{эксп}} = n(-) + n(\pm)1/2 = 48 + 0,5 = 48,5$. Очевидно, что для данного объема выборки N имеет место соотношение $W^+_{\text{эксп}} + W^-_{\text{эксп}} = \frac{1}{2}N(N+1) = 55$. Поэтому для построений критерия используют меньшее из этих значений, т. е. $W_{\text{эксп}} = \min(W^+_{\text{эксп}}, W^-_{\text{эксп}}) = 6,5$.

Чтобы оценить значимость полученного результата $W_{\text{эксп}} = 6{,}5$, нужно знать распределение статистики $\widetilde{W}$. Это распределение моделируется извлечением N шаров из урны, содержащей равное число шаров двух типов, помеченных, например, знаками «+» и «−». Количество матриц только с плюсами равно $\frac{N!}{N!0!}$, с одним минусом $\frac{N!}{(N-1)!1!}$, и т. д., всего возможно

$$\frac{N!}{N!0!} + \frac{N!}{(N-1)!1!} + \ldots + \frac{N!}{0!N!} = (1+1)^N = 2^N$$

вариантов матриц (в примере VI-9 это $2^{10} = 1024$). Если справедлива нулевая гипотеза, то все возможные конфигурации матрицы равновероятны

и вероятность любой из них равна $(1/2)^N$. Следовательно, можно получить распределение статистики $\widetilde{W}$, которое зависит только от $N = n - n(0)$. Для нашего примера $(N = 10)$ это распределение показано на рис. § 8. Заштрихованные области соответствуют вероятностям $P\{\widetilde{W} \leqslant 8\} = 0{,}024$. Это означает, что при $N = 10$ значения $W = 47$ и $W = 8$ являются критическими для уровня значимости $\alpha = 2 \cdot 0,024 \approx 0{,}05$.

Для статистики $\widetilde{W}$ парного критерия Вилкоксона имеются таблицы критических значений $W_{\alpha/2}(N)$, таких, что $P\{\widetilde{W} \leqslant W_{\alpha/2}(N)\} = \alpha/2$ (табл. IX Приложения 1). Как обычно, использование таблиц сводится к тому, что найденное по выборке значение $W_{\text{эксп}}$ сравнивается с табличным $W_{\alpha/2}(N)$. Если $W_{\text{эксп}} \leqslant W_{\alpha/2}(N)$, гипотезу H_0 отвергают на уровне значимости α; в противном случае ее принимают.

В данном примере $W_{\text{эксп}} = 6{,}5 < W_{0{,}025}(10) = 8$, и мы склоняемся к выводу, что влияние примененного воздействия на нервную деятельность крыс статистически значимо при $\alpha = 0{,}05$.

Вернемся к рис. 41. Можно видеть, что $\widetilde{W}$ принимает значения от 0 до $\frac{1}{2}N(N+1) = 55$, ее распределение симметрично относительно среднего значения $\frac{1}{4}N(N+1) = 27{,}5$, которое при H_0, очевидно, равно математическому ожиданию случайной величины $\widetilde{W}$:

$$E\widetilde{W} = \frac{1}{4}N(N+1).$$

Ее дисперсия есть

$$D\widetilde{W} = \frac{1}{24}N(N+1)(2N+1).$$

При наличии совпадений критерий становится приближенным и выражение для дисперсии принимает вид

$$D\widetilde{W} = \frac{1}{24}N(N+1)(2N+1) - \frac{1}{48}\sum_{i=1}^{g}(c_i^3 - c_i),$$

где g — число групп совпадений, а c_i — число совпадающих значений в i-й группе. Совпадениями являются только те случаи, когда $d_i = -d_j$, т. е. когда разности одинаковы по абсолютной величине, но противоположны по знаку.

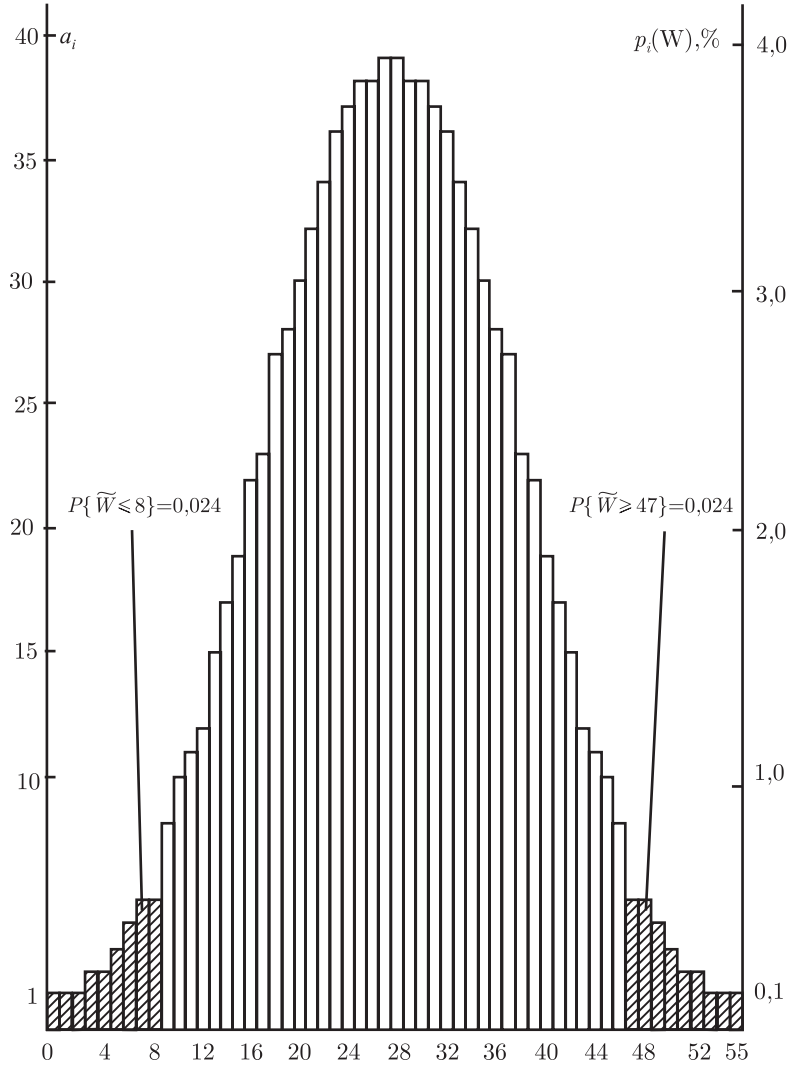


Рис. 42. Распределение статистики $\widetilde{W}$ парного критерия Вилкоксона при $N = 10$. Для наглядности дискретные значения представлены в виде единичных отрезков на оси абсцисс. Сумма заштрихованных площадей есть $\alpha = 0{,}048 \approx 0{,}05$

При больших N (практически при $N \geqslant 25$) величина

$$\widetilde{u} = \frac{\widetilde{W} - E\widetilde{W}}{\sqrt{D\widetilde{W}}}$$

имеет приблизительно нормальное распределение $N(0;1)$.

Таким образом, при достаточно большом объеме выборки гипотеза H_0: $\tau_1 = \tau_2$ о равенстве параметров положения для парных наблюдений из неизвестной генеральной совокупности становится равносильной гипотезе H_0: $\widetilde{W} \sim N(E\widetilde{W}, D\widetilde{W})$, которую можно проверить с помощью u-критерия. Для этого надо вычислить значение

$$|u_{\text{эксп}}| = \frac{|W_{\text{эксп}} - E\widetilde{W}|}{\sqrt{D\widetilde{W}}}$$

и сравнить его с критическим значением $u_{\alpha/2}$.

Задачи

VI-1 [Тьюки, 1981]. В 1942 г. Дж. Билл сосчитал число особей насекомого *Phlegethontius quinquemaculata* на делянку после применения различных способов борьбы с ним. Результаты для двух способов выглядят так:

способ I: 0, 1, 7, 2, 3, 1, 2, 1, 3, 0, 1, 4;
способ II: 3, 5, 3, 5, 3, 6, 1, 1, 3, 2, 6, 4.

Число особей на делянку имеет распределение Пуассона; какое преобразование данных следует применить (см. гл. VIII, § 6)? Какой из способов борьбы с насекомым более эффективен? Примените критерий Стьюдента и критерий Вилкоксона–Манна–Уитни и сравните полученные выводы.

VI-2 [Снедекор, 1961]. Фермер тратил на удобрение поля A по одному фунту стерлингов на акр, а поля Б — по два фунта. Чистый доход на акр в течение 5 лет с этих полей (не учитывая стоимости удобрений) приведен в табл. 38. Имеет ли смысл продолжать применять усиленное удобрение?

VI-3. Как вы будете обеспечивать простой случайный выбор (проводить рандомизацию) в примере VI-6?

VI-4 [ван дер Варден, 1960]. С 1946 по 1951 г. в медицинской клинике Цюрихского университета для лечения последствий тромбоза — образования сгустков крови в кровеносных сосудах — 252 раза применялись антикоагулянты. Из 252 пациентов умерли 14. С 1937 по 1945 г. антикоагулянты вообще не применялись. Оказалось, что из 205 пациентов, лечение антикоагулянтами которым было бы не противопоказано, умерли 74. Можно ли благотворность действия антикоагулянтов считать статистически значимой?

Таблица 38

Чистый доход на акр в течение 5 лет (к задаче VI-2)

Год	Поле	
	А	Б
1	17,0	18,0
2	14,0	16,5
3	21,0	24,0
4	18,5	19,0
5	22,0	25,0

VI-5. Проведите статистический анализ данных из примера IV-8.

VI-6. Проведите статистический анализ данных из примера IV-11.

VI-7 (Н. В. Железнова, 1980 г.). Для производства вирусных вакцин необходимо подавить инфекционную активность вирусных частиц, не изменив их иммунологических свойств. Орудием для такого «нежного убийства» является ультрафиолетовое излучение. С этой целью изучена фоточувствительность двух штаммов вируса гриппа и получены следующие данные (см/10^{15} квант):

штамм «Техас-77» (12 опытов): 0,23, 0,27, 0,28, 0,28, 0,29, 0,29, 0,30, 0,31, 0,31, 0,32, 0,33, 0,34;

штамм «СССР-77» (5 опытов): 0,34, 0,36, 0,42, 0,43, 0,50.

Различаются ли штаммы по чувствительности к УФ-свету?

VI-8 (А. П. Пуговкин, 1982 г.). Изучали регуляцию тонуса кровеносных сосудов. Для этого подопытное животное (кошку) подключали к системе искусственного кровообращения, замеряли артериальное давление (миллиметры ртутного столба) — x_{1i}, затем прерывали и вновь возобновля-

ли кровоток, повторно замеряя давление — x_{2i} (табл. 39). Можно ли говорить о восстановлении артериального давления?

Таблица 39

Артериальное давление у кошки (к задаче VI-8)

Номер животного (i)	x_{1i}	x_{2i}	Номер животного (i)	x_{1i}	x_{2i}
1	80	85	10	90	85
2	90	77	11	90	85
3	90	95	12	90	88
4	90	70	13	80	70
5	90	80	14	80	65
6	100	85	15	90	90
7	70	75	16	90	92
8	70	70	17	90	110
9	70	70	18	90	110

Таблица 40

Дополнительные часы сна у десяти пациентов, вызванные действием двух снотворных средств A и B (к задаче VI-9)

Пациент (i)	Средство A (x_{1i})	Средство B (x_{2i})	Разность $(d_i = x_{1i} - x_{2i})$
1	1,9	0,7	1,2
2	0,8	−1,6	2,4
3	1,1	−0,2	1,3
4	0,1	−1,2	1,3
5	−0,1	−0,1	0,0
6	4,4	3,4	1,0
7	5,5	3,7	1,8
8	1,6	0,8	0,8
9	4,6	0,0	4,6
10	3,4	2,0	1,4

VI-9 [Фишер, 1958; Крамер, 1975]. В табл. 40 приведены данные, заимствованные из классической работы Стьюдента 1908 г. о t-распределении.

По этим данным требуется проверить, существует ли значимая разница между действием двух снотворных средств A и B, если предположить, что разность между дополнительными часами сна, вызванными действием этих двух средств, распределена нормально.

VI-10. Используя известное соответствие между t-распределением и F-распределением (см. § 10 гл. III), постройте статистику F-критерия для сравнения средних значений μ_1 и μ_2 двух нормальных распределений, когда $\sigma_1^2 = \sigma_2^2$.

VI-11 [Большев, Смирнов, 1982]. В группе из 25 человек 20 человек были подвергнуты действию противогриппозной сыворотки, и в течение шести месяцев из них лишь 6 человек заболели гриппом. Остальные пятеро от вакцинации отказались, и среди них наблюдались четыре случая заболевания гриппом. Убеждают ли эти результаты в благотворном действии испытываемой сыворотки? Удовлетворительна ли такая постановка эксперимента?

Глава VII

Сравнение распределений

В этой главе мы рассмотрим сначала задачу проверки согласия эмпирического распределения с полностью определенным гипотетическим. Данная задача, встречающаяся при анализе расщеплений в генетике, позволяет аккуратно ввести критерий χ^2. Далее обсуждаются задачи проверки согласия эмпирического распределения с гипотетическим дискретным, параметры которого оцениваются по выборке, и сравнения нескольких выборочных дискретных распределений. Эти задачи также сводятся к применению критерия χ^2. Затем обсуждается задача согласия эмпирического распределения с гипотетическим непрерывным нормальным распределением. В этом случае мы предпочитаем использовать негруппированные данные и применять непараметрический критерий Колмогорова.

§ 1. Согласие выборочного распределения с полностью определенным дискретным распределением

Пример VII-1 (Г. Мендель, 1865 г.). При скрещивании двух чистых линий гороха, одна из которых имеет желтые семена с морщинистой поверхностью, а другая — зеленые с гладкой поверхностью, в первом поколении получены растения, все горошины на которых были желтыми и гладкими. Во втором поколении, полученном в результате самоопыления, обнаружено расщепление: 315 горошин были желтыми гладкими, 101 — желтой морщинистой, 108 — зелеными гладкими и 32 — зелеными морщинистыми. Общее число независимых наблюдений равно $315 + \ldots + 32 = 556$.

Формулируется простейшая гипотеза: различия между исходными линиями определяются двумя генами, которые комбинируются независимо. Один ген, определяющий различия по окраске семян, имеет доминантную аллель, обусловливающую зеленую окраску, и рецессивную аллель, обусловливающую желтую окраску. Другой ген, определяющий различия по форме семян, имеет доминантную аллель, обусловливающую гладкую форму, и рецессивную аллель, обусловливающую морщинистую форму.

Если эта гипотеза верна, то во втором поколении горошины желтые гладкие должны появляться с вероятностью 9/16, желтые морщинистые и зеленые гладкие — соответственно, с вероятностями 3/16 и 3/16 и, наконец, зеленые морщинистые — с вероятностью 1/16. Это позволяет вычислить ожидаемые численности всех классов семян.

Таким образом, в эксперименте получены эмпирические численности классов 315: 101: 108: 32; на основании законов Менделя найдены гипотетические (ожидаемые) численности 312,75: 104,25: 104,25: 34,75. Требуется проверить согласие эмпирических соотношений с гипотетическими. Заметим, что менделевская гипотеза приводит к полностью определенному распределению. В самом деле, проведено $n = 556$ независимых испытаний, каждое из которых имеет четыре несовместных исхода с постоянными от испытания к испытанию вероятностями $p_1 = 9/16$, $p_2 = p_3 = 3/16$, $p_4 = 1/16$, причем $p_1 + p_2 + p_3 + p_4 = 1$, т. е. мы имеем полиномиальное распределение $(9/16 + 3/16 + 3/16 + 1/16)^{556}$.

Сформулируем теперь этот тип задачи в общем виде.

Пусть результатом некоторого испытания может быть один из l различных попарно несовместных исходов $A_1, A_2, \ldots, A_l$, так что событие $U_{i=1}^{l} A_i$ есть достоверное событие. Вероятность исхода A_i обозначим через p_i и будем считать ее известной. Производится n независимых испытаний, в k_1 из них мы наблюдаем событие A_1, в $k_2 - A_2$ и т. д. Общее число всех испытаний $n = \sum_{i=1}^{l} k_i$. Исходя из гипотезы, что вероятность каждого события A_i есть p_i, можно вычислить, сколько раз в n испытаниях ожидается каждый исход эксперимента: при $np_1, np_2, \ldots, np_l$. Необходимо проверить согласие ожидаемых (гипотетических) численностей классов np_i с наблюдаемыми численностями k_i.

Для того чтобы решить эту задачу, нужно предложить меру расхождения между гипотетическими и эмпирическими данными. Представляется естественным рассмотреть отклонения наблюдаемых численностей от ожидаемых суммарно по всем классам $A_1, A_2, \ldots, A_l$. Однако $\sum_{i=1}^{l}(k_i - np_i) = 0$. По аналогии с выборочной дисперсией можно попытаться рассмотреть сумму квадратов отклонений $\sum_{i=1}^{l}(k_i - np_i)^2$. Но эта величина, очевидно, будет зависеть и от числа классов l, и от объема выборки n. Поэтому должны быть введены некоторые коэффициенты (нормирующие множители): $\sum_{i=1}^{l} c_i(k_i - np_i)^2$. Что выбрать в качестве c_i? К. Пирсон предложил брать $c_i = 1/np_i$. Тогда мера расхождения между гипотетическими

и эмпирическими данными (или мера согласия между ними) принимает вид

$$\frac{(k_i - np_i)^2}{np - i}.$$

«Хорошая» эта мера или «плохая»? Ответ зависит от того, удастся ли найти распределение предложенной статистики и на его основе построить критерий значимости. Попытаемся это сделать.

Для упрощения задачи будем рассматривать частный случай полиномиального испытания — биномиальное, в котором возможны лишь два несовместных исхода A_1 и $A_2 = \overline{A_1}$. Вероятность $P\{A_1\}$ исхода A_1 обозначим p_1, $P\{A_1\} = p_2$, так что $p_1 + p_2 = 1$. Производится n независимых испытаний, в каждом из которых $p_i = \text{const}$, $i = 1, 2$, причем в k_1 испытаниях происходит событие A_1, в k_2 испытаниях — событие A_2 и $k_1 + k_2 = n$.

Тогда предложенная мера согласия принимает вид

$$\frac{(k_1 - np_1)^2}{np_1} + \frac{(k_2 - np_2)^2}{np_2}.$$

Отклонения $(k_1 - np_1)$ и $(k_2 - np_2)$ линейно связаны:

$$(k_1 - np_1) + (k_2 - np_2) = (k_1 + k_2) - n(p_1 + p_2) = 0$$

и

$$(k_1 - np_1) = -(k_2 - np_2).$$

Тогда

$$\frac{(k_1 - np_1)^2}{np_1} + \frac{(k_2 - np_2)^2}{np_2} = (k_1 - np_1)^2 \left(\frac{1}{np_1} + \frac{1}{np_2} \right) =$$

$$= (k_1 - np_1)^2 \frac{p_2 + p_1}{np_1 p_2} = \frac{(k_1 - np_1)^2}{np_1(1 - p_1)}.$$

Уменьшив числитель по абсолютной величине на 0,5 (поправка на дискретность) и рассматривая $\widetilde{k}_1$ как случайную величину, распределенную биномиально, получим случайную величину

$$\widetilde{u} = \frac{|\widetilde{k}_1 - np_1| - 0{,}5}{\sqrt{np_1(1 - p_1)}},$$

которая, согласно результату, приведенному в § 5 гл. III, при $n \to \infty$ будет иметь нормированное нормальное распределение $N(0;1)$. Тогда, по определению (гл. III, § 7), случайная величина

$$\frac{(|\widetilde{k}_1 - np_1| - 0,5)^2}{np_1(1-p_1)} = \sum_{i=1}^{2} \frac{(|\widetilde{k}_i - np_1| - 0,5)^2}{np_i} \sim \chi^2(\nu), \quad \nu = 1.$$

Этот результат можно обобщить на полиномиальное распределение, при этом поправка на дискретность (0,5) не используется. Совокупность l отклонений

$$\frac{\widetilde{k}_1 - np_1}{\sqrt{np_1(1-p_1)}}, \quad \frac{\widetilde{k}_2 - np_2}{\sqrt{np_2(1-p_2)}}, \quad \ldots, \frac{\widetilde{k}_l - np_l}{\sqrt{np_l(1-p_l)}}$$

можно представить в виде $(l-1)$ независимой случайной величины $\widetilde{u}_i \sim N(0;1)$. Тогда

$$\widetilde{\chi}^2 = \sum_{i=1}^{l} \frac{(\widetilde{k}_i - np_i)^2}{np_i} \sim \chi^2(\nu), \quad \nu = l-1.$$

Поскольку закон распределения статистики, предложенной К. Пирсоном, установлен, можно построить *критерий значимости* χ^2 (*критерий Пирсона*). Сформулируем нулевую гипотезу H_0: имеет место согласие между гипотетическим и эмпирическим распределениями; альтернатива H_1 заключается в том, что эмпирическое распределение не описывается гипотетическим.

При правильности нулевой гипотезы значения $\chi^2_{\text{эксп}}$, вычисленные по экспериментальным (выборочным) данным, не должны превосходить критическое значение $\chi^2_\alpha(\nu)$, такое, что $P\{\widetilde{\chi}^2 \geqslant \chi^2_\alpha(\nu)\} = \alpha$. Поэтому, если $\chi_{\text{эксп}} < \chi^2_\alpha$, т. е. $P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{эксп}}\} > \alpha$, то нулевая гипотеза принимается на уровне значимости α. Если $\chi^2_{\text{эксп}} \geqslant \chi^2_\alpha(\nu)$, т. е. $P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{эксп}}\} \leqslant \alpha$, то нулевая гипотеза отклоняется и принимается альтернативная.

Так обычно поступают, когда речь идет о мере согласия между какими-то распределениями. Такой способ построения критерия χ^2 приводит к односторонему по своей сути критерию (рис. 43). Односторонним он является потому, что статистика критерия не учитывает знаков отклонений «наблюдаемых» от «ожидаемых», т. е. знаков разностей $(k_i - np_i)$.

Если $\chi^2_{\text{эксп}} < \chi^2_\alpha(\nu)$, то при описании экспериментальных результатов принято указывать не просто $P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{эксп}}\} > \alpha$, но более точное значение

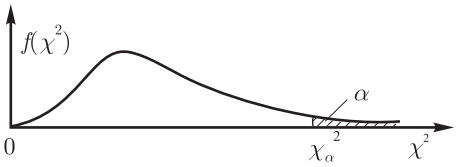


Рис. 43. Односторонний критерий χ^2

вероятности, что можно сделать, пользуясь табл. V Приложения 1. В таблице приведены и очень большие значения $P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{табл}}\}$ и даже такие, как 0,975 и 0,995. Это связано с обстоятельством, на которое впервые обратил внимание Р. А. Фишер. Если вычисленное значение $\chi^2_{\text{эксп}}$ очень мало, то это приводит к очень большим значениям $P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{эксп}}\}$ и свидетельствует об очень большой вероятности согласия гипотезы с экспериментальными данными. Однако в то же время это означает, что вероятность случайного отклонения экспериментальных данных от гипотетических значений $P\{\widetilde{\chi}^2 \leqslant \chi^2_{\text{эксп}}\} = 1 - P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{эксп}}\}$ очень мала, т. е. получено слишком хорошее согласие с гипотезой H_0. Разумеется, такие значения изредка могут появляться и случайно. Экспериментатору нужно это обстоятельство просто иметь в виду и при систематическом получении малых значений $\chi^2_{\text{эксп}}$ еще раз просмотреть структуру и технику эксперимента с целью обнаружения факторов, искусственно снижающих случайную изменчивость. Таковыми могут быть, в частности, субъективизм в оценках, необоснованное отсеивание (выбраковка) «нежелательных» результатов и т. п. Мы пришли к выражению

$$\sum_{i=1}^{l} \frac{(\widetilde{k}_i - np_i)^2}{np_i} \sim \chi^2(\nu), \quad \nu = l - 1,$$

в предположении, что $n \to \infty$. Возникает вопрос, при каких условиях на практике эта аппроксимация правомочна. Поскольку в знаменателе стоят np_i, дело сводится к минимально допустимым значениям ожидаемых численностей. Точного решения этой задачи нет, однако ее всестороннее обсуждение привело к эмпирическому правилу, которым и пользуются в биометрии: минимальное ожидаемое не должно быть меньше 5. Если ожидаемое оказывается слишком малым, приходится объединять соседние (обычно крайние) классы, соответственно, уменьшая число степеней свободы. Эта процедура, естественно, влечет за собой уменьшение чувствительности критерия.

Ряд исследователей все более склоняется к точке зрения, что требование минимального ожидаемого меньше 5 является слишком жестким и что допустимое минимальное ожидаемое, вообще говоря, зависит от числа степеней свободы. Широкое распространение получила следующая рекомендация: если число степеней свободы больше 6, то одно из ожидаемых может снижаться даже до 1/2. При числе степеней свободы, равном 60 и более, и при малых ожидаемых критерий χ^2 очень надежен. При двух степенях свободы ожидаемое может снижаться до 2. Только при одной степени свободы нужно соблюдать осторожность и требовать, чтобы ожидаемое было не меньше 4.

Распределение χ^2 — это непрерывное распределение. Мы же рассматриваем задачу о согласии с полиномиальным, т. е. дискретным, распределением. Однако при соблюдении указанных выше ограничений на минимальное ожидаемое дискретность не приводит к сколько-нибудь существенным погрешностям, за одним исключением: когда число степеней свободы равно единице. Для этого случая Ф. Иейтсом была указана очень простая поправка, дающая удовлетворительные результаты: абсолютные значения разностей между наблюдаемыми и ожидаемыми необходимо уменьшить на 0,5.

ПРИМЕР VII-1 (продолжение). Ожидаемые численности уже были вычислены выше. Имеем

$$\chi^2_{\text{эксп}} = \frac{(315 - 312{,}75)^2}{312{,}75} + \ldots + \frac{(32 - 34{,}75)^2}{34{,}75} = 0{,}47.$$

По табл. V Приложения 1 при $\nu = l - 1 = 3$ находим $\chi^2_{0{,}05}(3) = 7{,}81$, которое больше $\chi^2_{\text{эксп}}$. Таким образом, $P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{эксп}}\} > 0{,}05$, следовательно, наблюдаемое расщепление не противоречит выдвинутой гипотезе при уровне значимости $\alpha = 0{,}05$.

Отметим, что статистика критерия Пирсона может быть преобразована к виду

$$\chi^2 = \frac{1}{n} \sum_{i=1}^{l} \frac{k_i^2}{p_i} - n,$$

который более удобен для вычислений в случае, когда p_i кратны целым числам.

§ 2. Согласие выборочного распределения с дискретным распределением, параметры которого оцениваются по выборке

Обратимся к данным задачи II-11. Сопоставление эмпирического и гипотетического распределений графически показывает их явное расхождение. Решим эту задачу точно с помощью критерия χ^2. Отличие ее от задачи, обсуждаемой в § 1, заключается в том, что нам неизвестно точное значение параметра гипотетического распределения Пуассона λ, мы можем вычислить лишь его оценку $m = 0,205$.

Можно показать, что при проверке согласия с распределением, параметры которого оцениваются по выборке, мы также имеем распределение χ^2, но с числом степеней свободы, меньшим на число оцениваемых параметров. Например, в рассматриваемом случае проверки согласия с распределением Пуассона число степеней свободы будет равно $(l-1)-1$. Если проверять согласие с нормальным распределением, среднее значение μ и дисперсия σ^2 которого оцениваются по выборке, то $\nu = (l-1)-2$.

Пример VII-2 (продолжение задачи II-11). Ожидаемые численности для клеток имеющих более двух поврежденных хромосом очень малы (см. решение задачи). Поэтому необходимо объединить численности четырех последних классов тогда $l = 3$ и $\nu = 3-1-1 = 1$. Вычисляем значение

$$\chi_{\text{эксп}} = \frac{(877-814\,7)^2}{814\,7} + \frac{(63-167\,0)^2}{167\,0} + \frac{(60-18\,3)^2}{18\,3} = 164\,55$$

которое существенно превышает критическое табличное значение $\chi^2_{0\,001}(1) = 10\,8$. Следовательно $P\{\chi^2 \geqslant \chi^2_{\text{эксп}} = 164\,55\} \ll 0\,001$. Таким образом нулевая гипотеза о согласии наблюдаемого распределения поврежденных хромосом по клеткам с распределением Пуассона должна быть несомненно отвергнута.

§ 3. Сравнение нескольких дискретных распределений (критерий однородности)

Пример VII-3 (Г. И. Арнаутова, 1982 г.). Природные популяции *Primula sibthorpii* полиморфны по окраске цветка. Встречаются растения с четырьмя типами окраски: белые цветки (неокрашенные), светло-сиреневые, сиреневые, фиолетовые. В табл. 41 приведены распределения этих растений по окраске цветков в трех популяциях Дагестана.

Таблица 41

Распределение растений *P. sibthorpii* по окраске цветка в популяциях Дагестана (к примеру VII-3)

Популяция	Окраска цветка				Всего
	белая	светло-сиреневая	сиреневая	фиолетовая	
Дылым	1 529	177	211	45	1 962
Сыртыч	725	360	348	89	1 522
Маджалис	182	354	252	99	887
Всего	2 436	891	811	233	4 371

Необходимо решить, различаются ли популяции по частотам растений с разной окраской или наблюдаемые различия между популяциями носят случайный характер, определяемый объемом выборки. Другими словами, совпадают ли распределения генеральных совокупностей, из которых извлечены выборки?

Сформулируем задачу в общем виде. Пусть $n_{1\cdot}$ (читается: «эн один с точкой») объектов первой выборки разбиты по какому-то признаку на l классов и пусть $k_{11}, k_{12}, \ldots, k_{1l}$ — численности объектов в этих классах. Следующие $n_{2\cdot}$ объектов второй выборки разбиты на те же l классов и $k_{21}, k_{22}, \ldots, k_{2l}$ — численности соответствующих классов, и т. д. Наконец, численности в l классах для n_r, объектов r-й выборки равны $k_{r1}, k_{r2}, \ldots, k_{rl}$. Всего получено rl чисел, которые расположим в прямоугольную таблицу (табл. 42). Справа приведены суммы по строкам (выборкам), внизу — по столбцам (классам признаков), тогда точка в индексе при $n_{\cdot i}$ указывает на суммирование по всем выборкам для числа особей, имеющих признак, соответствующий i-му классу.

Таким образом, имеем r выборочных полиномиальных распределений, и нужно проверить, могут ли параметры $p_1, p_2, \ldots, p_l$ быть одинаковыми для всех строк. Наилучшими выборочными оценками значений p_i являются частоты классов, определенные по всем выборкам (см. гл. V, § 1):

$$h_1 = \frac{n_{\cdot 1}}{n}, \quad h_2 = \frac{n_{\cdot 2}}{n}, \ldots h_l = \frac{n_{\cdot l}}{n},$$

где h_i — выборочная оценка вероятности p_i. Поскольку известен объем каждой выборки, теперь можно найти ожидаемые численности классов. Например, для первой выборки они будут равны $n_{\cdot 1}h_1; n_{\cdot 1}h_2; \ldots; n_{\cdot 1}h_l$. Точно так же поступаем для второй и т. д. выборки и, наконец, для r-й выборки.

Таблица 42

Таблица сопряженности $r \times l$ к построению критерия однородности

k_{11}	k_{12}	$\dots$	k_{1i}	$\dots$	k_{1l}	$n_{1\cdot}$
k_{21}	k_{22}	$\dots$	k_{2i}	$\dots$	k_{2l}	$n_{2\cdot}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
k_{j1}	k_{j2}	$\dots$	k_{ji}	$\dots$	k_{jl}	$n_{j\cdot}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
k_{r1}	k_{r2}	$\dots$	k_{ri}	$\dots$	k_{rl}	$n_{r\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot i}$	$\dots$	$n_{\cdot l}$	n

Вычитая из каждого наблюдаемого соответствующее ожидаемое, вновь приходим к прямоугольной таблице отклонений $(k_{ji} - n_{j\cdot}h_i)$ (табл. 43). Суммы по строкам и суммы по столбцам в этой таблице, естественно, равны нулю, что следует использовать для проверки правильности вычислений.

Таблица 43

Таблица отклонений к построению критерия однородности

$k_{ji} - n_{j\cdot}h_i$					
$k_{11} - n_{1\cdot}h_1$	$k_{12} - n_{1\cdot}h_2$	$\dots$	$k_{1i} - n_{1\cdot}h_i$	$\dots$	$k_{1l} - n_{1\cdot}h_l$
$k_{21} - n_{2\cdot}h_1$	$k_{22} - n_{2\cdot}h_2$	$\dots$	$k_{2i} - n_{2\cdot}h_i$	$\dots$	$k_{2l} - n_{2\cdot}h_l$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$k_{j1} - n_{j\cdot}h_1$	$k_{j2} - n_{j\cdot}h_2$	$\dots$	$k_{ji} - n_{j\cdot}h_i$	$\dots$	$k_{jl} - n_{j\cdot}h_l$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$k_{r1} - n_{r\cdot}h_1$	$k_{r2} - n_{r\cdot}h_2$	$\dots$	$k_{ri} - n_{r\cdot}h_i$	$\dots$	$k_{rl} - n_{r\cdot}h_l$

Аналогично рассуждениям § 1 можно показать, что сумму квадратов разностей между наблюдаемыми и ожидаемыми, деленных на соответствующие ожидаемые

$$\frac{(k_{11} - n_{1\cdot}h_1)^2}{n_{1\cdot}h_1} + \frac{(k_{12} - n_{1\cdot}h_2)^2}{n_{1\cdot}h_2} + \ldots + \frac{(k_{rl} - n_{r\cdot}h_l)^2}{n_{r\cdot}h_l},$$

можно рассматривать как реализацию случайной величины $\widetilde{\chi}^2$ с числом степеней свободы $\nu = (r-1)(l-1)$. Действительно, всего наблюдается rl значений численностей. Объем каждой выборки фиксирован, т. е. на эти rl наблюдений наложено r линейных связей:

$$k_{j1} + k_{j2} + \ldots + k_{jl} = n_j.$$

Кроме того, было оценено $(l-1)$ параметров p_i (число оцененных параметров равно $l-1$, а не l, поскольку $p_1+p_2+\ldots+p_l=1$). Тогда число степеней свободы будет равно

$$\nu = rl - r - (l-1) = (r-1)(l-1),$$

т. е. произведению числа строк без единицы на число столбцов без единицы. Заметим, что $n_{j\cdot}h_i = n_{j\cdot}n_{i\cdot}/n$, т. е. для того чтобы найти любое ожидаемое, необходимо произведение сумм по соответствующим строкам и столбцам разделить на общую сумму наблюдений.

ПРИМЕР VII-3 (продолжение). Прежде всего вычислим ожидаемые: для численности 1 529 имеем $2\,436 - 1\,962/4\,371 = 1\,093{,}4$; для 177 имеем $891 \cdot 962/4\,371 = 399{,}9$ и т. д. и, наконец, для 99 имеем $233 - 887/4371 = 47{,}3$. Все они больше 5, следовательно, можно применять критерий χ^2. Тогда

$$\chi^2_{\text{эксп}} = \frac{(1\,529-1093{,}4)^2}{1\,093{,}4} + \frac{(17-399{,}9)^2}{399{,}9} + \ldots + \frac{(99-47{,}3)^2}{47{,}3} = 905{,}14.$$

По табл. V Приложения 1 при $\nu = (3-1)(4-1) = 6$ находим $\chi^2_{0{,}001} = 22{,}5$, которое существенно меньше $\chi^2_{\text{эксп}}$. Следовательно, $P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{эксп}} = 905{,}14\} \ll 0{,}001$. Таким образом, распределения растений по окраске цветка в разных популяциях, несомненно, различны.

Заметим, что более удобной для вычислений является формула

$$\chi^2 = n\left(\sum_{j=1}^{r}\frac{1}{n_{j\cdot}}\sum_{i=1}^{l}\frac{k_{ij}^2}{n_{\cdot i}} - 1\right),$$

согласно которой сначала для каждой строки вычисляют значения $k_{ij}^2/n_{\cdot i}$, суммируют их по i и каждую j-ю сумму делят на $n_{i\cdot}$.

Рассмотрим в заключение этого параграфа вопрос о суммировании значений статистики хи-квадрат и вопрос об аппроксимации χ^2-распределения нормальным.

ПРИМЕР VII-4 [Бейли, 1962]. При отработке методики подсчета дрожжевых клеток в счетной камере четырьмя лаборантами были проведены независимые эксперименты, в которых исследовалось распределение клеток по квадратам счетной камеры. Если суспензии клеток приготовлены правильно, т. е. каждая клетка отделена от другой, так что не образуется

комков, то эмпирическое распределение не должно отличаться от распределения Пуассона (см. гл. II, § 4). Результаты статистического анализа этих экспериментов приведены в табл. 44.

Таблица 44

Проверка согласия распределения дрожжевых клеток по квадратам счетной камеры с распределением Пуассона (к примеру VII-4)

Лаборант	$\chi^2_{\text{эксп}}$	ν	$P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{эксп}}\} \approx$
А	17,32	13	0,1
Б	14,71	10	0,1
В	24,45	17	0,1
Г	14,88	9	0,05
Сводные данные	71,36	49	0,025

Как видно, ни в одном случае нет оснований отклонить нулевую гипотезу, т. е. следует вывод о согласии эмпирических распределений с распределением Пуассона. Правда, обращает на себя внимание близость всех значений $P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{эксп}}\}$ к уровням значимости $\alpha = 0{,}10$ и $\alpha = 0{,}05$.

Однако такой анализ не использует всей информации, содержащейся в эксперименте: анализировалась работа каждого лаборанта индивидуально. Поскольку экспериментаторы работали независимо, можно продолжить анализ, объединив результаты. Вспомним теорему сложения для распределения χ^2 (см. гл. III, § 7) и на ее основании просуммируем все значения $\chi_{\text{эксп}}$ и числа ν степеней свободы. Для суммарных данных (нижняя строка в табл. 44) $P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{эксп}}\} < 0{,}025$, т. е. при 2,5 %-ном уровне значимости нулевая гипотеза должна быть отклонена. Если кто-то сочтет выбранный уровень значимости слишком мягким, следует провести дополнительные эксперименты, ведь речь идет об отработке методики! Полученные же результаты позволяют заподозрить, что у всех лаборантов имеют место какие-то небольшие, но систематические погрешности (возможно, одинаковые) в приготовлении суспензии клеток.

Не во всех статистических таблицах содержатся критические значения χ^2_α при 49 степенях свободы. Однако при $\nu > 30$ можно воспользоваться аппроксимацией распределения χ^2 нормальным (см. гл. III, § 7). Тогда

$$u_{\text{эксп}} = \sqrt{2\chi^2} - \sqrt{2\nu - 1} = \sqrt{2 \cdot 71{,}36} - \sqrt{2 \cdot 49 - 1} = 2{,}10,$$

что меньше $u_{0,025} = 1{,}96$. Таким образом, $P\{\widetilde{u} \geqslant u_{\text{эксп}}\} < 0{,}025$, т. е. имеем тот же результат (ведь аппроксимируется распределение статистики одностороннего по самой своей сути критерия!).

§ 4. Согласие выборочного распределения с нормальным

Удобным способом проверки согласия выборочного распределения с гипотетическим непрерывным распределением является применение так называемой вероятностной бумаги (сетки). Наибольшее распространение получила вероятностная бумага для проверки нормальности выборочного распределения. *Нормальная вероятностная бумага* — это специальным образом сконструированная координатная сетка с таким выбором масштаба по оси ординат, что график функции нормального распределения изображается прямой линией (рис. 44). По оси абсцисс масштаб берется равномерным, а по оси ординат — неравномерным, вероятностным, «выпрямляющим» график функции нормального распределения (ср. с рис. 17). Нормальный вероятностный масштаб симметричен относительно точки 0,5 (50 %); вверх и вниз от этой точки цена деления резко уменьшается. В пределах от 20 до 80 % цена деления составляет 2 %; в пределах от 1 до 20 % и от 80 до 99 % цена деления равна 1 % и за этими пределами — 0,1 %.

В биологических исследованиях практически не возникает задача сравнения выборочного распределения с полностью определенным нормальным. Обычно выборочное распределение сравнивается с гипотетическим нормальным, параметры которого оцениваются по той же выборке. Использование вероятностной бумаги в этом случае заключается в следующем.

Пусть по выборке объема n из нормально распределенной генеральной совокупности вычислены выборочное среднее m и выборочная дисперсия s^2. По оси абсцисс на вероятностной бумаге откладывают значения признака, охватывающие весь диапазон выборочных наблюдений. На графике отмечают три точки, имеющие координаты $(x_1 = m, y_1 = 50\,\%)$, $(x_2 = m + 2s, y_2 = 97{,}72\,\%)$ и $(x_3 = m - 2s, y_3 = 2{,}28\,\%)$. Если все технические операции были выполнены верно, то эти три точки должны лежать на одной прямой, которая соответствует функции гипотетического нормального распределения $F_0(x)$ с параметрами $\mu_0 = m$, $\sigma_0^2 = s^2$. Теперь необходимо построить в том же вероятностном масштабе выборочную функцию распределения $F_n(x)$. Для этого производится ранжировка выборочных значений — находят $x_{(i)}$. Ожидаемые для нормального распределения значения $y_{(i)} = E[\widetilde{F}_n(x)]$ с вполне достаточной точностью можно вычислить по фор-

муле $y_{(i)} = \dfrac{i - 0{,}3175}{n + 0{,}3650}$, где i — номер (ранг) наблюдения $x_{(i)}$ и n — объем выборки. После этого на график наносят точки с координатами $(x_{(i)}, y_{(i)})$.

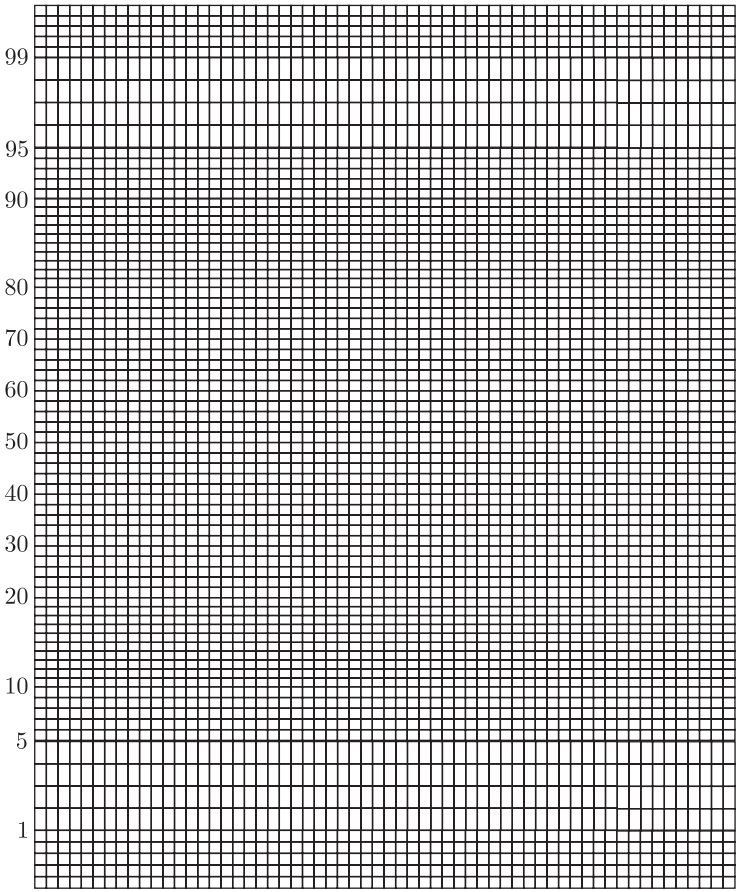


Рис. 44. Нормальная вероятностная бумага

Пример VII-5 (М. В. Падкина, 1978 г.). Определялась константа Михаэлиса–Ментен $K_M(10^{-3}\,M)$ для кислой фосфатазы дрожжей

Saccharomyces cerevisiae:

$$\begin{array}{cccccccccc} 2\,17 & 1\,92 & 2\,20 & 2\,08 & 1\,85 & 2\,00 & 2\,08 & 2\,00 & 2\,12 \\ 1\,96 & 1\,92 & 2\,00 & 1\,83 & 2\,00 & 2\,08 & 1\,92 & 1\,75 & 1\,96 \end{array}$$

Согласуется ли выборочное распределение с нормальным?

Объем выборки $n = 18$, выборочное среднее $m = 1{,}99$, выборочное среднее квадратичное отклонение $s = 0{,}118$. Абсциссы трех точек для построения графика $F_0(x)$ суть $x_1 = 1{,}99$, $x_2 = m + 2s = 2{,}23$ и $x_3 = m - 2s = l{,}75$. Ранжируем все 18 наблюдений, получая i и $x_{(i)}$ и вычисляем $y_{(i)}$ (табл. 45). Наносим точки $(x_{(i)}, y_{(i)})$ на вероятностную бумагу.

Таблица 45

К проверке нормальности выборочного распределения (к примеру VII-5)

i	$x_{(i)}$	$y_{(i)}$, %
1	1,75	3,7
2	1,83	9,2
3	1,85	14,6
4 — 6	1,92	20,1 — 30,9
7 — 8	1,96	36,4 — 41,8
9 — 12	2,00	47,3 — 63,6
13 — 15	2,08	69,1 — 80,0
16	2,12	85,5
17	2,17	90,8
18	2, 20	96,3

Теперь возникает вопрос о согласии эмпирического распределения с гипотетическим. Нередко, когда видны систематические различия, достаточно глазомерного сравнения. Однако в общем случае, конечно, необходимо иметь количественную меру согласия распределений и соответствующий критерий значимости.

Такой критерий был найден А. Н. Колмогоровым в 1933 г. Пусть $F(x)$ — произвольная, но полностью определенная функция распределения непрерывной случайной величины $\widetilde{x}$ и $\widetilde{F}_n(x)$ — случайная выборочная функция распределения. В качестве статистики для критерия согласия между эмпирическим и гипотетическим распределениями А. Н. Колмогоров выбрал случайную величину $\widetilde{D}_n = \sup|\widetilde{F}_n(x) - F(x)|$ (см. § 1, гл. IV). Было показано, что $P\{\sqrt{n}\widetilde{D}_n < \lambda\} \to K(\lambda)$, $\lambda > 0$, при $n \to \infty$, где $K(\lambda)$ — функция

распределения Колмогорова, для которой найдено аналитическое выражение. Лишь 45 лет спустя (в 1978 г.) Г. А. Несененко и Ю. Н. Тюрину удалось найти асимптотическое распределение статистики критерия Колмогорова для случая, когда параметры $F(x)$ оцениваются по выборке. В частности, было получено решение для нормального распределения с неизвестными параметрами μ и σ. На практике достаточно знать два критических значения статистики $\beta_{0,05} = 0{,}895$ и $\beta_{0,01} = 1{,}035$, с которыми сравнивается величина, вычисленная по экспериментальным данным:

$$\beta_{\text{эксп}} = D_n(\sqrt{n} - 0{,}01 + 0{,}85/\sqrt{n}).$$

Итак, проверяется гипотеза H_0: $E[\widetilde{F}_n(x)] - F_0(x) = 0$ о согласии эмпирической функции распределения $F_n(x)$ с гипотетической функцией нормального распределения $F_0(x)$, параметры μ_0 и σ_0^2 которой оцениваются по выборке ($\mu_0 = m, \sigma_0^2 = s^2$). Нулевая гипотеза принимается на уровне значимости α, если $\beta_{\text{эксп}} < \beta_\alpha$, и отклоняется, если $\beta_{\text{эксп}} > \beta_\alpha$, где β_α такое, что $P\{\widetilde{\beta} \geqslant \beta_\alpha\} = \alpha$. Критерий Колмогорова в изложенном виде применим и для малых выборок.

Пример VII.4 (продолжение). Выборочное значение D_n можно достаточно точно установить по графику на вероятностной бумаге. В данном примере максимальное по абсолютной величине отклонение выборочной функции распределения от гипотетической принадлежит значению $x_{(12)} = 2{,}00$ и составляет $D_n = 0{,}106$ (рис. 45). Находим значение $\beta_{\text{эксп}} = 0{,}106(\sqrt{18} - 0{,}01 + 0{,}85/\sqrt{18}) = 0{,}470$. Поскольку $\beta_{\text{эксп}} < \beta_{0,05}$, то на уровне значимости $\alpha = 0{,}05$ нет оснований отклонять гипотезу о нормальности выборочного распределения. Следует, однако, заметить, что для более надежных выводов требуется анализ выборки большего объема.

Задачи

VII-1 (В. П. Эфроимсон, 1956 г.). В табл. 46 представлены результаты опытов Г. Менделя и его последователей по изучению наследования одного признака — окраски семядолей — у гороха. Согласуются ли результаты разных авторов друг с другом и с ожидаемым расщеплением 3: 1?

VII-2. Проверьте согласие с нормальным распределением данных примера IV-1 (см. табл. 3 и 6).

VII-3. Проведите статистический анализ данных из примера IV-6 (см. табл. 7).

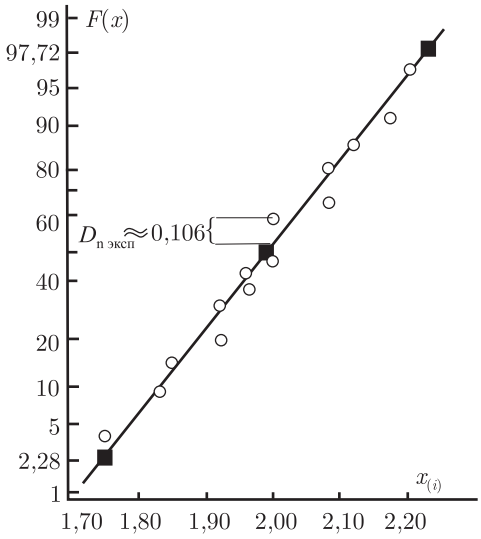


Рис. 45. Проверка нормальности распределения с помощью вероятностной бумаги и критерия Колмогорова. Указано наблюдаемое значение статистики D_n для данных примера VII-5

VII-4. Проведите статистический анализ данных из примера IV-12 (см. табл. 9).

VII-5. Проведите статистический анализ данных из примера IV-14 (см. табл. 10).

VII-6. Проведите статистический анализ данных из примера IV-19 (см. табл. 15).

VII-7. Чему равно число степеней свободы при проверке согласия с биномиальным распределением при неизвестном значении p?

VII-8 (Дж. Э. Юл, М. Дж. Кендэл, 1960 г.). Число рождений в Англии и Уэльсе по месяцам в 1941 г.: январь — 50 159, февраль — 45 885, март — 50 819, апрель — 49 070, май — 50 771, июнь — 46 788, июль — 49 395, август — 50 443, сентябрь — 51 562, октябрь — 50 224, ноябрь — 47 168, декабрь — 50 529 (всего зарегистрировано 592 813 рождений). Выявляют ли эти данные наличие некоторой сезонности в рождениях?

Таблица 46

Расщепление по признаку окраски семядолей у гороха (к задаче VII-1)

Автор	Расщепление во втором поколении $Aa \times Aa$	
	число желтых (A)	число зеленых (a)
Мендель	6 022	2 001
Корренс	1394	453
	1 012	344
	225	70
Чермак	3 580	1 190
	3 000	959
Херст	1 310	445
Бэтсон	11902	3 903
Локк	1438	514
	3 085	1 008
	2 400	850
Дэрбишир	109 060	36 186
	1 089	354
	5 662	1856
Уайт	1647	543
Всего	152 823	50 676

VII-9 (Н. В. Тимофеев-Ресовский, К. Пэтау, 1943 г.). На протяжении ряда лет проводились эксперименты по изучению радиационно-индуцированного мутационного процесса у дрозофилы. Особенно большие материалы накопились при облучении самцов дозой 3000 Р: всего было проанализировано 29600 хромосом! Эта выборка включала 148 последовательных субвыборок по 200 хромосом в каждой), полученных в хронологическом порядке на протяжении семи лет. Полученное авторами распределение представлено в табл. 47.

Если условия эксперимента были воспроизводимы, то вероятность наступления события «мутация» должна быть постоянной на протяжении всего опыта. В этом случае должно быть получено согласие с биномиальным распределением. Проверьте это.

VII-10. Почему при построении графика выборочной функции нормального распределения на вероятностной бумаге (см. рис. 45) берутся три, а не две точки? Почему избранным при этом трем значениям x_i: m, $(m+2s)$ и $(m-2s)$ — соответствуют значения y_i: 50, 97,72 и 2,28 %?

VII-11. Проверьте гипотезу о равномерном распределении для показаний часов, выставленных в витринах часовщиков (см. задачу IV-6, табл. 21).

Таблица 47

Результаты рентгеновского облучения самцов дрозофилы дозой 3 кР (к задаче VII-9)

Число хромосом	Число субвыборок	Число хромосом	Число субвыборок
6	1	19	9
7	2	20	8
8	0	21	6
9	2	22	9
10	5	23	6
11	7	24	5
12	3	25	6
13	9	26	2
14	8	27	2
15	9	28	1
16	10	29	0
17	17	30	1
18	20		

VII-12. Проверьте, согласуются ли с нормальным два выборочных распределения из примера VI-2 (см. табл. 27). Действительно ли преобразование $\lg x$ в данном случае является нормализующим?

VII-13. Для данных задачи VI-1 графически, с помощью нормальной вероятностной бумаги, проследите, как изменяется вид двух выборочных функций распределения после нормализующего преобразования.

VII-14. Проверьте выдвинутое Стьюдентом предположение о нормальности распределения для разности между дополнительными часами сна, вызванными двумя снотворными (см. задачу VI-9, табл. 40).

VII-15. Ч. Дарвин[1] приводит результаты многолетних наблюдений различных авторов за соотношением полов у домашних животных (табл. 48). Согласуются ли эти данные друг с другом? Можно ли считать, что соотношение полов у домашних животных соответствует предполагаемому 1: 1?

[1] Дарвин Ч. Соч. т. 5. Происхождение человека и половой отбор. — М.: Изд-во АН СССР, 1953.

Таблица 48 Соотношение полов у домашних животных (к задаче VII-15).

Животные	Число самцов	Число самок
Лошади	12 763	12 797
Собаки	3 605	3 273
Овцы	29 478	30 172
Коровы	477	505
Куры	487	614

Таблица 49

Распределение личинок мясной мухи по теплоустойчивости при $42\,^{\circ}\mathrm{C}$ (к задаче VII-18)

x_i, мин	n_i	x_i, мин	n_i	x_i, мин	n_i
9	3	20	10	31	5
10	6	21	7	32	6
11	6	22	11	33	2
12	7	23	9	34	5
13	6	24	7	35	3
14	8	25	7	36	2
15	19	26	7	37	2
16	7	27	5	38	2
17	12	28	6	46	1
18	14	29	3	48	1
19	32	30	3		

VII-16. Для эмпирического распределения 26 306 бросаний 12 игральных костей (см. задачу II-9, табл. 1) проверьте гипотезу о согласии его с биномиальным распределением, параметр которого равен $p = 1/3$. Если гипотеза будет отклонена, проверьте согласие с гипотетическим биномиальным распределением, параметр которого оцените по выборке.

VII-17. Для двух распределений числа наводнений в устье р. Невы (см. задачу IV-12, табл. 24) проверьте гипотезу о согласии их с распределением Пуассона.

VII-18 (Б. П. Ушаков, 1981 г.). Согласуется ли с нормальным распределением распределение личинок мухи *Calliphora erythrocephalia* по теп-

лоустойчивости мышц при 42°С (табл. 49)? Согласуется ли распределение логарифма теплоустойчивости с нормальным распределением?

Глава VIII

Сравнение параметров нескольких распределений

Сравнение параметров нескольких распределений проводится с помощью *дисперсионного анализа*, которому и посвящена бóльшая часть этой главы. Однако дисперсионный анализ позволяет решать и другую важную задачу, не рассматриваемую в элементарном курсе биометрии, — разложение дисперсии на компоненты; мы сочли необходимым упомянуть об этом в § 4. Последний параграф посвящен изложению непараметрического метода дисперсионного анализа с помощью критерия Крускала–Уоллиса.

§ 1. Однофакторный дисперсионный анализ: модель *I*

Биологическая задача сравнения параметров нескольких нормальных распределений сформулирована в примере IV-15. Метод решения подобного рода задач был найден Р. А. Фишером и носит название *дисперсионного анализа*. Однако если постановка задачи такая, как в примере IV-15, где речь идет о сравнении нескольких выборочных средних, то становится непонятным, почему метод получил название дисперсионный анализ или, при дословном переводе английского analysis of variance, анализ дисперсии? Тем не менее, здесь все верно. Сравнение средних производится путем анализа (разложения на компоненты) общей дисперсии.

Пусть k независимых выборок объема n каждая извлечены из нормально распределенных генеральных совокупностей с равными дисперсиями: $\sigma_1^2{=}\sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$. Иметь равное число наблюдений в каждой выборке желательно для простоты организации эксперимента. Такой план эксперимента называется *сбалансированным* в отличие от *несбалансированного* плана с выборками неравного объема. Каждое наблюдение x_{ij}, где i — номер выборки (сорта в примере IV-15), а j — номер повторности (делянки в примере IV-15), будем рассматривать как реализацию случайной величины

$$\widetilde{x}_{ij} = \mu_i + \widetilde{e}_{ij},$$

где μ_i — среднее значение i-й генеральной совокупности; $\widetilde{e}_{ij} \sim N(0;\sigma^2)$ — независимые случайные величины, имеющие нормальное распределение с параметрами 0 и σ^2.

Случайная величина $\widetilde{e}_{ij}$ называется *случайной ошибкой* (случайным отклонением) или просто «ошибкой». Она характеризует изменчивость внутри i-й совокупности (обусловленную, например, случайными различиями в составе почвы, влажности, качестве семян, индивидуальном развитии растений и т. п.).

Таким образом, мы рассматриваем однофакторный дисперсионный анализ, где фактор в нашем случае — это сорт, он представлен k уровнями (число сортов).

Отметим, что в сформулированной задаче сравнения k сортов μ_i — не случайные величины, а некоторые константы; такая линейная статистическая модель называется *моделью с фиксированными эффектами*, или *моделью I*. В этой модели нулевая гипотеза принимает вид

$$H_0\colon \mu_1 = \mu_2 = \ldots = \mu_k = \mu.$$

Модель I дисперсионного анализа можно записать и по-иному:

$$\widetilde{x}_{ij} = \mu + \alpha_i + \widetilde{e}_{ij},$$

где $\mu = \frac{1}{k}\sum_{i=1}^{k} \mu_k$ — общее среднее для всех k генеральных совокупностей; $\alpha_i = \mu_i - \mu$ — отклонение среднего значения i-й генеральной совокупности от общего среднего μ; очевидно, что $\sum_{i=1}^{k} \alpha_i = 0$; $e_{ij} \sim N(0;\sigma^2)$ имеет прежний смысл случайной ошибки.

В этих обозначениях нулевая гипотеза принимает вид

$$H_0\colon \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0.$$

Используя метод максимального правдоподобия, можно показать, что несмещенными и эффективными оценками параметров модели являются статистики:

$$\text{для} \quad \mu - \widetilde{m} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n} \widetilde{x}_{ij};$$

$$\text{для} \quad \mu_i - \widetilde{m}_i = \frac{1}{n}\sum_{j=1}^{n} \widetilde{x}_{ij};$$

$$\text{для} \quad \alpha_i - \widetilde{m}_i - \widetilde{m},$$

где $N = nk$.

Рассмотрим тождество

$$\widetilde{x}_{ij} = \widetilde{m} + (\widetilde{m}_i - \widetilde{m}) + (\widetilde{x}_{ij} - \widetilde{m}_i).$$

Перенесем $\widetilde{m}$ в левую часть:

$$\widetilde{x}_{ij} - \widetilde{m} = (\widetilde{m}_i - \widetilde{m}) + (\widetilde{x}_{ij} - \widetilde{m}_i),$$

возведем обе части в квадрат:

$$(\widetilde{x}_{ij} - \widetilde{m})^2 = (\widetilde{m}_i - \widetilde{m})^2 + 2(\widetilde{m}_i - \widetilde{m})(\widetilde{x}_{ij} - \widetilde{m}_i) + (\widetilde{x}_{ij} - \widetilde{m}_i)^2$$

и просуммируем обе части уравнения по j:

$$\sum_{j=1}^{n}(\widetilde{x}_{ij} - \widetilde{m})^2 = n(\widetilde{m}_i - \widetilde{m})^2 +$$

$$+2(\widetilde{m}_i - \widetilde{m})\sum_{j=1}^{n}(\widetilde{x}_{ij} - \widetilde{m}_i) + \sum_{j=1}^{n}(\widetilde{x}_{ij} - \widetilde{m}_i)^2.$$

Однако $\widetilde{m}_i = \frac{1}{n}(\widetilde{x}_{i1} + \widetilde{x}_{i2} + \ldots + \widetilde{x}_{in})$, поэтому $\sum_{j=1}^{n}(\widetilde{x}_{ij} - \widetilde{m}_i) = 0$, т. е. второй член в правой части уравнения равен нулю. Просуммируем обе части уравнения по i:

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(\widetilde{x}_{ij} - \widetilde{m})^2 = \sum_{i=1}^{k} n(\widetilde{m}_i - \widetilde{m})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(\widetilde{x}_{ij} - \widetilde{m}_i)^2.$$

В левой части данного уравнения мы имеем сумму квадратов отклонений случайных величин $\widetilde{x}_{ij}$ от общего для всех выборок среднего $\widetilde{m}$. Это так называемая *общая*, или *полная, сумма квадратов*, которую мы будем обозначать $\widetilde{SS}_t$.

Первый член в правой части — сумма квадратов отклонений выборочных средних каждой группы (уровня, выборки) $\widetilde{m}_i$ от общего выборочного среднего $\widetilde{m}$. Это *сумма квадратов между группами* ($\widetilde{SS}_b$).

Наконец, второй член в правой части — сумма квадратов отклонений $\widetilde{x}_{ij}$ от среднего своей группы $\widetilde{m}_i$. Эту величину называют *суммой квадратов внутри групп* ($\widetilde{SS}_w$).

Итак, полная сумма квадратов разложена на две аддитивные составляющие — суммы квадратов между группами и внутри групп:

$$\widetilde{SS}_t = \widetilde{SS}_b + \widetilde{SS}_w.$$

Каждую составляющую из суммы квадратов необходимо теперь связать с соответствующим числом степеней свободы.

Общая сумма квадратов $\widetilde{SS}_t$ основана на $N = nk$ наблюдениях, на которые наложена одна линейная связь: при вычислении $\widetilde{SS}_t$ использована оценка общего среднего m, поэтому $\nu_1 = N - 1$. Аналогично $\nu_b = k - 1$ и $\nu_w = (n-1)k = N - k$. Заметим, что

$$\nu_t = \nu_b + \nu_w.$$

Сумма квадратов, деленная на соответствующее число степеней свободы, называется *средним квадратом*[1]:

$$\widetilde{MS}_b = \frac{\widetilde{SS}_b}{\nu_b}; \quad \widetilde{MS}_w = \frac{\widetilde{SS}_w}{\nu_w}.$$

В нашей модели предполагалось, что внутригрупповые выборочные дисперсии суть оценки σ^2. Если это так, то внутригрупповой средний квадрат есть взвешенная по отдельным группам оценка σ^2:

$$\widetilde{MS}_w = \frac{\sum_{j=1}^n (\widetilde{x}_{1j} - \widetilde{m}_1)^2 + \sum_{j=1}^n (\widetilde{x}_{2j} - \widetilde{m}_2)^2 + \ldots + \sum_{j=1}^n (\widetilde{x}_{kj} - \widetilde{m}_k)^2}{(n-1) + (n-1) + \ldots + (n-1)} =$$

$$= \frac{n-1}{N-1} \sum_{i=1}^k \widetilde{s}_i^2,$$

где $\widetilde{s}_i^2$ — дисперсия i-й выборки (группы). Поскольку эта оценка является несмещенной, то $E(\widetilde{MS}_w) = \sigma^2$.

Можно показать, что

$$E(\widetilde{MS}_b) = \sigma^2 + \frac{1}{n} \sum_{i=1}^n n(\mu_i - \mu)^2.$$

[1] Для сумм квадратов и средних квадратов традиционно используются двухбуквенные обозначения, являющиеся сокращениями английских выражений sum of squares (SS) и mean square (MS).

Следовательно, если нулевая гипотеза верна, то $E(\widetilde{MS}_b) = \sigma^2 = E(\widetilde{MS}_w)$. Если же нулевая гипотеза неверна, то $E(\widetilde{MS}_b) > E(\widetilde{MS}_w)$.

Таким образом, проверка нулевой гипотезы сводится к рассмотрению соотношения $\widetilde{MS}_b/\widetilde{MS}_w$. Поскольку $E(\widetilde{MS}_b) \geqslant \sigma^2$, а $E(\widetilde{MS}_w) = \sigma^2$, то отношение $\widetilde{MS}_b/\widetilde{MS}_w$ может принимать значение меньше единицы лишь в двух случаях:

1) случайно, вследствие слишком малых отклонений выборочных средних $\widetilde{m}_i$ от общего среднего значения μ;

2) при неадекватности рассматриваемой модели экспериментальным данным.

Мы уже неоднократно пользовались тем, что случайная величина $\dfrac{(n-1)\widetilde{s}^2}{\sigma^2} \sim \chi^2(\nu)$, $\nu = n-1$. Для $\widetilde{s}_i^2$ это можно представить в виде

$$\frac{(n-1)\widetilde{s}_i^2}{\sigma^2} = \sum_{j=1}^{n} \frac{(\widetilde{x}_{ij} - \widetilde{m}_i)^2}{\sigma^2} \sim \chi^2(\nu), \quad \nu = n-1.$$

Тогда для $\widetilde{SS}_w$ имеем $\widetilde{\chi}_w^2 = \dfrac{\widetilde{SS}_w}{\sigma^2} \sim \chi^2(\nu_w)$, $\nu_w = N-k$, или

$$\frac{\widetilde{\chi}_w^2}{\nu_w} = \frac{\widetilde{MS}_w}{\sigma^2}.$$

Обратимся теперь к $\widetilde{MS}_b$. Имеем k случайных величин $\widetilde{m} \sim N(\mu; \sigma/\sqrt{n})$ и $\widetilde{m} = \frac{1}{k}\sum_{i=1}^{k} \widetilde{m}_i$. Тогда

$$\widetilde{\chi}_b^2 = \frac{(\widetilde{m}_1 - \widetilde{m})^2}{\sigma^2/n} + \frac{(\widetilde{m}_2 - \widetilde{m})^2}{\sigma^2/n} + \ldots + \frac{(\widetilde{m}_k - \widetilde{m})^2}{\sigma^2/n} =$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{k} n(\widetilde{m}_i - \widetilde{m})^2 \sim \chi^2(\nu_b), \quad \nu_b = k-1,$$

или $\widetilde{\chi}_b^2 = \widetilde{SS}_b/\sigma^2 \sim \chi^2(\nu_b)$, $\nu_b = k-1$, и, поделив обе части на число степеней свободы ν_b,

$$\widetilde{\chi}_b^2/\nu_b = \widetilde{MS}_b/\sigma^2.$$

Итак,

$$\frac{\widetilde{MS}_b}{\widetilde{MS}_w} = \frac{\widetilde{\chi}_b^2/\nu_b}{\widetilde{\chi}_w^2/\nu_w} \sim F(\nu_b, \nu_w), \quad \nu_b = k-1, \quad \nu_w = N-k,$$

поскольку можно показать, что $\widetilde{\chi}_b^2$ и $\widetilde{\chi}_w^2$ суть независимые случайные величины.

На этой основе можно построить *критерий значимости* F для сравнения средних значений нескольких независимых нормальных распределений (с равными дисперсиями). Если нулевая гипотеза верна, т. е. $\mu_1 = \mu_2 = \ldots = \mu_k = \mu$, то статистика критерия $\widetilde{F} = \dfrac{\widetilde{MS}_b}{\widetilde{MS}_w}$ будет иметь F-распределение с параметрами $\nu_1 = \nu_b = k-1$ и $\nu_2 = \nu_w = N-k$. Поскольку, как было показано выше, $E(\widetilde{MS}_b) \geqslant E(\widetilde{MS}_w)$, критерий является односторонним. Вычислив значение

$$F_{\text{эксп}} = \frac{MS_b}{MS_w},$$

его сравнивают с критическим значением $F_\alpha(\nu_1, \nu_2)$, $\nu_1 = \nu_b$, $\nu_2 = \nu_w$, таким, что $P\{\widetilde{F} \leqslant F_\alpha(\nu_1, \nu_2)\} = \alpha$. Если $F_{\text{эксп}} < F_\alpha(\nu_1, \nu_2)$, то нет оснований для отклонения гипотезы H_0 и различия между средними признаются незначимыми на уровне α. Если $F_{\text{эксп}} \geqslant F_\alpha(\nu_1, \nu_2)$, то нулевая гипотеза отклоняется, т. е., по крайней мере, одно из k сравниваемых средних значимо отличается от остальных.

Подведем итог приведенным выше рассуждениям.

Сформулировав в статистических терминах задачу сравнения средних значений, возникающую в биологическом эксперименте, мы построили линейную статистическую модель, оговорив условия ее выполнения и указав статистики для оценки параметров модели. Далее было проведено разложение полной суммы квадратов и полного числа степеней свободы. Для получаемых затем межгруппового и внутригруппового средних квадратов указаны их математические ожидания; при этом оказалось, что при правильности нулевой гипотезы величина $\dfrac{MS_b}{MS_w}$, вычисляемая по экспериментальным данным, является реализацией случайной величины, имеющей F-распределение. Это позволило построить критерий значимости F.

Таким образом, разлагая общую дисперсию на компоненты и сравнивая их, оказывается возможным решить задачу сравнения k выборочных средних.

§ 2. Техника вычислений в однофакторном дисперсионном анализе

Вычисления в схеме однофакторного дисперсионного анализа удобно проводить в следующей последовательности:

1) общее число наблюдений $N = nk$, где k — число групп (выборок, уровней), n — число наблюдений в каждой группе;

2) полная сумма наблюдаемых значений $G = \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}$;

3) корректирующий член $C = G^2/N$;

4) полная сумма квадратов отклонений $SS_t = \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}^2 - C$;

5) сумма квадратов между группами

$$SS_b = \frac{1}{n} \sum_{i=1}^{k} \left(\sum_{j=1}^{n} x_{ij} \right)^2 - C;$$

6) сумма квадратов внутри групп $SS_w = SS_t - SS_b$;

7) числа степеней свободы $\nu_t = N - 1$, $\nu_b = k - 1$, $\nu_w = N - k$;

8) средний квадрат между группами $MS_b = SS_b/\nu_b$;

9) средний квадрат внутри групп $MS_w = SS_w/\nu_w$;

10) экспериментальное значение статистики F-критерия $F_{\text{эксп}} = MS_b/MS_w$.

Результаты однофакторного дисперсионного анализа удобно представлять в виде таблицы (табл. 50).

Проведем вычисления для данных примера IV-15:

1) $N = 5 \times 4 = 20$;

2) $G = 656{,}4$;

3) $C = 21\,543{,}05$;

4) $SS_t = 21\,725{,}22 - 21\,543{,}05 = 182{,}17$;

Таблица 50

Таблица однофакторного дисперсионного анализа с фиксированными эффектами (модель I)

Изменчивость	Сумма квадратов	Число степеней свободы	Средний квадрат	$F_{\text{эксп}}$
Между группами	SS_b	$k-1$	MS_b	MS_b/MS_w
Внутри групп	SS_w	$N-k$	MS_w	–
Полная	SS_t	$N-1$	–	–

5) $SS_b = 21\,677{,}50 - 21\,543{,}05 = 134{,}45$;

6) $SS_w = 182{,}17 - 134{,}45 = 47{,}72$;

7) $\nu_b = 4-1 = 3$; $\nu_w = 20-4 = 16$; $\nu_t = 20-1 = 19$;

8) $MS_b = 134{,}45/3 = 44{,}82$;

9) $MS_w = 47{,}72/16 = 2{,}98$;

10) $F_{\text{эксп}} = 44{,}82/2{,}98 = 15{,}03$.

Результаты анализа представлены в табл. 51.

Таблица 51

Результаты дисперсионного анализа для данных примера IV-15

Изменчивость	Сумма квадратов	Число степеней свободы	Средний квадрат	$F_{\text{эксп}}$
Между группами	134,45	3	44,82	
Внутри групп	47,72	16	2,98	15,03
Полная (общая)	182,17	19	–	

Далее находим, что $F_{\text{эксп}} > F_{0{,}001}(3;16) = 9{,}0$ (см. табл. IV∂ Приложения 1).

Таким образом, $P\{\widetilde{F} \geqslant F_{\text{эксп}}\} < 0{,}001$ и нулевая гипотеза отвергается: различия между сортами по урожайности статистически высоко значимы.

Если число наблюдений в группах разное (несбалансированный план эксперимента), то это не вносит существенных изменений в схему однофакторного дисперсионного анализа.

Мы не будем рассматривать этот случай подробно, укажем лишь, что в вычислениях изменяются только два пункта:

1) общее число наблюдений: $N = \sum_{i=1}^{k} n_i$;

2) сумма квадратов между группами

$$SS_b = \sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{j=1}^{n_i} x_{ij} \right)^2 - C,$$

где n_i — объем i-й выборки.

§ 3. Множественные сравнения

Сравнение средних значений $\mu_1, \mu_2, \ldots, \mu_k$ нескольких (k) независимых нормальных распределений не ограничивается простой констатацией факта: есть различие (совокупности неоднородны) или нет различия (совокупности однородны). Желательно знать, какие конкретно из анализируемых совокупностей различаются своими средними значениями. Существуют разнообразные методы множественных сравнений для выявления таких различий.

Мы рассмотрим следующую двушаговую процедуру, которая одновременно и проста, и достаточно эффективна. На первом этапе проводится дисперсионный анализ (модель *I*), как описано в § 1. И только после того, как на основании дисперсионного анализа сделан вывод о неоднородности сравниваемых совокупностей (при заданном уровне значимости α), переходят ко второму этапу, на котором проводится попарное сравнение средних значений всех k совокупностей с помощью t-критерия (при том же уровне значимости α).

Вспомним, что знаменатель статистики t-критерия для сравнения средних значений двух $(k = 2)$ независимых нормальных распределений содержит обобщенную (средневзвешенную) оценку дисперсии σ^2 вида

$$\widetilde{s}^2 = \frac{\nu_1 \widetilde{s}_1^2 + \nu_2 \widetilde{s}_2^2}{\nu_1 + \nu_2}$$

(см. гл. VI, § 2). Естественным образом эта оценочная статистика обобщается на случай нескольких $(k > 2)$ независимых выборок и принимает вид

$$\frac{\sum_{i=1}^{k} \nu_i \widetilde{s}_i^2}{\sum_{i=1}^{k} \nu_i} = \frac{\sum_{i=1}^{k} \nu_i \widetilde{s}_i^2}{N - k} = \widetilde{MS}_w,$$

т. е. равна статистике $\widetilde{MS}_w$, которая в модели I дисперсионного анализа называется средним квадратом внутри групп и является взвешенной по k группам (выборкам) оценкой дисперсии σ^2. Тогда статистика

$$\widetilde{t}_{ij} = \frac{\widetilde{m}_i - \widetilde{m}_j}{\sqrt{\widetilde{MS}_w\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim t(\nu), \quad \nu = N - k, \quad i \neq j,$$

т. е. имеет t-распределение с параметром $\nu = N - k$. Очевидно, что $\widetilde{t}_{ij}$ может служить статистикой критерия для попарного сравнения средних значений μ_i и μ_j $(i \neq j)$ нескольких независимых нормальных распределений. Подчеркнем, что результат этот справедлив только для случая равенства дисперсии сравниваемых распределений $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$.

Такой критерий часто называют *множественным t-критерием*. Подчеркнем еще раз, что он применяется только на втором этапе обсуждаемой двушаговой процедуры, т. е. после того как дисперсионный анализ выявил неоднородность сравниваемых совокупностей. Лишь когда $k = 2$, т. е. в случае двух независимых нормальных распределений, процедура сравнения является одношаговой: статистики F-критерия и критерия t_{ij} становятся идентичными (см. задачи VI-10 и VIII-3)!

Обратимся вновь к примеру IV-15. Поскольку дисперсионный анализ выявил существенные различия между сортами (при $\alpha = 0{,}001$), встает задача множественных сравнений, т. е. задача проверки новой гипотезы H_0: $\mu_i = \mu_j$ $i \neq j$. Для этого вычисляем значения

$$t_{ij(\text{эксп})} = \frac{|m_i - m_j|}{\sqrt{\widetilde{MS}_w(1/n_i + 1/n_j)}}, \quad i \neq j, \quad \nu = N - k.$$

В примере IV-15 объемы всех выборок одинаковы (сбалансированный план эксперимента). В таком случае

$$t_{ij(\text{эксп})} = \frac{|m_i - m_j|}{\sqrt{NS_w 2/n}}.$$

Находим $\sqrt{NS_w 2/n} = \sqrt{2{,}98} \cdot \sqrt{2/5} = 1{,}09$. Полученные значения $t_{ij(\text{эксп})}$ приведены в табл. 52.

Согласно табл. III Приложения 1, $t_{0{,}0005}(16) = 4{,}02$. В табл. 52 три значения $t_{ij(\text{эксп})}$) превышают данное табличное, а именно: t_{AD}, t_{BD} и t_{CD}.

Таблица 52

Матрица множественных сравнений для данных примера IV-15 (приведены значения m_i и $t_{ij(\text{эксп})}$)

Сорт		A	B	C
	m_i	34,42	34,78	33,70
B	34,78	0,33		
C	33,70	0,78	0,99	
D	28,38	5,54	5,87	4,88

Следовательно, только низкоурожайный сорт D ($m_D = 28{,}38$) отличается на 0,1 %-ном уровне значимости от трех остальных сортов.

§ 4. Понятие о модели II дисперсионного анализа

Исследователя часто не интересуют сравнительные оценки тех конкретных уровней фактора, которые изучаются в определенном эксперименте. Такие ситуации очень часты в популяционной биологии.

Пример VIII-1. Изучалась изменчивость размеров листа у дуба скального *Quercus petraea* на северо-западном Кавказе. По маршрутному ходу со 163 деревьев брали по 5 листьев. Изменчивость листьев внутри деревьев — это метамерная изменчивость. Изменчивость листьев на разных деревьях связана с их генетическими различиями и различиями в условиях произрастания. Таким образом, в этой схеме изучается один фактор (деревья), представленный $k = 163$ уровнями (группами) с $n = 5$ повторными наблюдениями. При этом нас совершенно не интересует сравнение выборочных средних конкретных 163 деревьев.

Исследователь рассматривает эти уровни как случайную выборку из практически бесконечного множества уровней изучаемого фактора A. Его интересует, насколько общая изменчивость признака определяется изменчивостью данного фактора A. Другими словами, разлагая общую дисперсию на составляющие

$$\sigma^2 = \sigma_a^2 + \sigma_e^2,$$

в конечном итоге нужно узнать:

1) отличается ли σ_a^2 от нуля;

2) если отличается, то оценить долю (силу) влияния изучаемого фактора в общей дисперсии, т. е. оценить величину $\sigma_a^2/\sigma^2 = \rho_w$, которая называется *коэффициентом внутриклассовой корреляции*. Это *модель II* дисперсионного анализа, или *модель со случайными эффектами*. При этом меняется смысл исходной линейной модели

$$\widetilde{x}_{ij} = \mu + \widetilde{a}_i + \widetilde{e}_{ij}.$$

Здесь μ — общее среднее (как и в модели I); $\widetilde{a}_i \sim N(0;\sigma_a^2)$ $(i = 1,2,\ldots,k)$ — независимые нормально распределенные случайные величины со средними 0 и дисперсиями σ_a^2 для всех i; $\widetilde{e}_{ij} \sim N(0;\sigma_a^2)$ $(i = 1,2,\ldots,k, j = 1,2,\ldots,n)$ — тоже независимые нормально распределенные случайные величины со средними 0 и дисперсиями σ_e^2.

В модели предполагается также независимость случайных величин $\widetilde{a}_i$ и $\widetilde{e}_{ij}$.

Метод разложения общей дисперсии остается тем же самым. Вычисляются статистики $\widetilde{SS}_t$, $\widetilde{SS}_b$ и $\widetilde{SS}_w$. В качестве оценок внутригрупповой и межгрупповой дисперсий используются статистики $\widetilde{MS}_w$ и $\widetilde{MS}_b$. Проверяется гипотеза H_0: $\sigma_a^2 = 0$. Можно показать, что

$$E(\widetilde{MS}_w) = \sigma_e^2 \quad \text{и} \quad E(\widetilde{MS}_b) = \sigma_e^2 + l\sigma_a^2,$$

где

$$l = \frac{1}{k-1}\left(N - \frac{1}{N}\sum_{i=1}^{k} n_i^2\right).$$

В случае равного числа наблюдений во всех группах ясно, что $l = n$.

Следовательно, если нулевая гипотеза верна, то $E(\widetilde{MS}_b) = E(\widetilde{MS}_w)$, если же гипотеза неверна, то $E(\widetilde{MS}_b) > E(\widetilde{MS}_w)$. Тогда очевидно, что, как и в модели I, проверка нулевой гипотезы сводится к рассмотрению отношения $\widetilde{MS}_b/\widetilde{MS}_w$, которое при справедливости H_0 имеет F-распределение:

$$\widetilde{MS}_b/\widetilde{MS}_w \sim F(\nu_1,\nu_2), \quad \nu_1 = \nu_b = k-1, \quad \nu_2 = \nu_b = N-k.$$

Таким образом, в двух разных моделях дисперсионного анализа для проверки по существу одной и той же гипотезы можно использовать статистику $\widetilde{MS}_b/\widetilde{MS}_w$ и соответствующий односторонний F-критерий.

Если на основании этого критерия влияние фактора A признано статистически значимым, то представляет интерес оценить долю влияния

изучаемого фактора. Для этого используются оценки соответствующих дисперсий:

$$\text{для} \quad \sigma_e^2 - \widetilde{MS}_b = \widetilde{s}_e^2;$$

$$\text{для} \quad \sigma_a^2 - \frac{1}{l}(\widetilde{MS}_b - \widetilde{MS}_w) = \widetilde{s}_a^2.$$

Тогда доля влияния фактора A оценивается (по Р. А. Фишеру) с помощью выборочного коэффициента внутриклассовой (внутригрупповой) корреляции

$$\widetilde{r}_w = \frac{\widetilde{s}_a^2}{(\widetilde{s}_a^2 + \widetilde{s}_e^2)}.$$

Эта статистика есть точечная оценка коэффициента внутриклассовой корреляции ρ_w в генеральной совокупности. Интервальная оценка (для равной или близкой численности групп) с доверительной вероятностью $1-\alpha$ может быть найдена как

$$\frac{F_{\text{эксп}} - F''_{\text{табл}}}{F_{\text{эксп}} + (k-1)\cdot F''_{\text{табл}}} < \rho_w < \frac{F_{\text{эксп}} - 1/F'_{\text{табл}}}{F_{\text{эксп}} + 1/[(k-1)\cdot F'_{\text{табл}}]},$$

где $F'_{\text{эксп}} = F_{\alpha/2}(N-k; k-1)$; $F''_{\text{табл}} = F_{\alpha/2}(k-1; N-k)$.

Пример VIII-2 [Снедекор, 1961]. Изучалась изменчивость размера тела у одной из пород свиней. От 5 свиноматок взято по 4 поросенка. Измеряли длину поросят по достижении ими 90 кг (табл. 53).

Таблица 53

Длина поросят (см), происходящих от 5 свиноматок (к примеру VIII-2)

Свиноматка (i)	Длина поросят (x_{ij})				Средняя длина (m_i)
1	93,5	94,5	96,0	89,5	93,38
2	91,0	99,0	96,0	93,5	95,63
3	94,0	91,0	93,0	92,0	92,50
4	88,0	88,0	87,0	90,0	88,23
5	93,0	91,0	94,0	90,0	92,00

Вычисляем:

1) $N = 4 \times 5 = 20$;

2) $G = 1847{,}0$.

3) $C = 170570{,}45$;

4) $SS_t = 170747{,}00 - 170570{,}45 = 176{,}55$;

5) $SS_b = 170685{,}38 - 170570{,}45 = 114{,}93$;

6) $SS_w = 176{,}55 - 114{,}93 = 61{,}62$;

7) $\nu_b = 5 - 1 = 4$; $\quad \nu_w = 20 - 5 = 15$; $\quad \nu_t = 20 - 1 = 19$;

8) $MS_b = 114{,}93/4 = 28{,}73$;

9) $MS_w = 61{,}62/15 = 4{,}11$;

10) $F_{\text{эксп}} = 28{,}73/4{,}11 = 6{,}99$; $F_{\text{эксп}} > F_{0{,}005}(4{,}15) \approx 5{,}8$ (см. табл. IV*г* Приложения 1).

Таким образом, $P\{\widetilde{F} \geqslant F_{\text{эксп}}\} < 0{,}005$, т.
t,e. различия между свиноматками значимы. Поэтому имеет смысл оценка компонент дисперсии $s_e^2 = 4{,}11$, $s_a^2 = \dfrac{28{,}73 - 4{,}11}{5} = 4{,}92$ и коэффициента внутриклассовой корреляции (доли влияния свиноматок) $r_w = \dfrac{4{,}92}{4{,}92 + 4{,}11} = 0{,}54$.

Построим, наконец, 90%-ный доверительный интервал для ρ_w: $F'_{\text{табл}} = F_{0{,}005}(15; 4)$; $F''_{\text{табл}} = F_{0{,}005}(4; 15) = 3{,}06$;

$$\frac{6{,}99 - 3{,}06}{6{,}99 + 3{,}06 \times 4} < \rho_w < \frac{6{,}99 - \dfrac{1}{5{,}86}}{6{,}99 + \dfrac{1}{5{,}86 \times 4}};$$

$$0{,}20 < \rho_w < 0{,}97.$$

Полученный результат показателен: доверительный интервал для ρ_w очень велик, поэтому не следует никогда полагаться только на точечную оценку!

§ 5. Понятие о двухфакторном дисперсионном анализе

Пример VIII-3 [Снедекор, 1961]. У трех видов цитрусовых деревьев было определено при трех условиях затенения отношение листовой поверхности к сухой массе листьев (табл. 54). Значимы ли различия по изучаемому признаку:

1) между видами цитрусовых,

2) при разной степени затенения?

Таблица 54

Отношение листовой поверхности к сухой массе листьев у цитрусовых (к примеру VIII-3).

Степень затенения	Вид		
	апельсин	грейпфрут	мандарин
На солнце	112	90	123
Частичное затенение	86	76	89
В тени	80	62	81

Решение поставленной задачи требует применения двухфакторного дисперсионного анализа. Один исследуемый фактор здесь — вид цитрусовых A, другой — условия выращивания B. Все три вида испытываются в одних и тех же для каждого вида условиях, т. е. каждому уровню одного фактора соответствует каждый и один и тот же уровень другого фактора. Это случай полной, или перекрестной, классификации. По самой постановке сформулированного в примере VIII-3 вопроса по обоим факторам следует рассматривать модели с фиксированными эффектами (модель I). Своеобразие этого примера заключается в том, что в каждой ячейке табл. 54 содержится только одно наблюдение.

Двухфакторный дисперсионный анализ с единственным наблюдением в ячейке (без повторностей) имеет важное применение в планировании эксперимента.

Пример VIII-4. В примере IV-15 проводилась полная рандомизация четырех сортов, каждый в 5 повторностях, по делянкам (см. табл. 11, рис. 34, a). Однако найти относительно однородный участок такого размера, чтобы его можно было разбить на 20 делянок, довольно трудно. Всегда

будут оставаться сомнения, нет ли в его пределах градиента каких-то почвенных условий, подчас сложно установимого. Гораздо проще поступить следующим образом: разбить участок на пять блоков с 4 делянками (см. рис. 34, *б*) и проводить рандомизацию сортов в пределах каждого блока, а при анализе учесть влияние второго фактора — «блоки». При полной рандомизации сравнение столбцов одного номера (для разных сортов) в табл. 11 смысла не имеет, это просто порядковый номер повторности. При выделении блоков, однако, такая процедура становится осмысленной: сравниваются урожаи разных сортов в пределах блока. Если один фактор (сорта в нашем примере) имеет k уровней, а другой (блоки) — n уровней, то математические ожидания средних квадратов имеют вид

$$E(\widetilde{MS}_w) = \sigma_e^2,$$

$$E(\widetilde{MS}_b) = \sigma_e^2 + \frac{1}{k-1} n \sum_{i=1}^{k} (\mu_i - \mu)^2;$$

$$E\widetilde{s}_a^2 = \sigma_e^2 + \frac{1}{n-1} k \sum_{j=1}^{n} (\mu_j - \mu)^2$$

и схема вычислении несколько изменяется:

1) (1—5) то же, что и в однофакторном анализе;

2) сумма квадратов между блоками

$$SS_a = \frac{1}{k} \sum_{j=1}^{n} \left(\sum_{i=1}^{k} x_{ij} \right)^2 - C;$$

3) $\nu_b = k - 1$; $\nu_a = n - 1$; $\nu_w = (k-1)(n-1)$; $\nu_t = N - 1$;

4) (8—10) то же, что и в однофакторном анализе;

5) сумма квадратов ошибки

$$SS_w = SS_t - SS_b - SS_a;$$

6) $F_{\text{эксп}} = s_a^2 / MS_w$.

Таблица 55

Таблица двухфакторного дисперсионного анализа с перекрестной классификацией

Изменчивость	Сумма квадратов	Число степеней свободы	Средний квадрат	$F_{\text{эксп}}$
Блоки	SS_a	$n-1$	s_a^2	s_a^2/MS_w
Сорта	SS_b	$k-1$	MS_b	MS_b/MS_w
Ошибка	SS_w	$(k-1)(n-1)$	MS_w	
Общая	SS_t	$N-1$		

Результаты двухфакторного дисперсионного анализа удобно представлять в виде таблицы (табл. 55).

Организация опыта с выделением блоков приводит в примере VIII-4 к следующему анализу:

1) $SS_a = 21\,564{,}51 - 21\,543{,}05 = 21{,}46$.

2) $\nu_b = 4-1=3$; $\nu_a = 5-1=4$; $\nu_w = 3\times 4 = 12$; $\nu_t = 20-1=19$.

3) $SS_w = 182{,}17 - 134{,}45 - 21{,}46 = 26{,}26$.

4) $F_{\text{эксп}} = 2{,}45$.

Таким образом, в нашем примере (табл. 56) различия между блоками оказались незначимыми.

Гораздо более интересная ситуация возникает в двухфакторном дисперсионном анализе, когда в каждой ячейке мы имеем несколько повторных наблюдений.

Таблица 56

Результаты дисперсионного анализа для данных примера VIII-4

Изменчивость	Сумма квадратов	Число степеней свободы	Средний квадрат	$F_{\text{эксп}}$	$P\{\widetilde{F} \geqslant F_{\text{эксп}}\}$
Блоки	21,46	4	5,36	2,45	>0,05
Сорта	134,45	3	44,82	20,47	$\ll$0,01
Ошибка	26,26	12	2,19	—	—
Общая	182,17	19	—	—	—

ПРИМЕР VIII-5 [Гласс, Стэнли, 1976]. Изучалась эффективность двух различных методов (A) и двух различных условий (B) обучения геометрии.

48 школьных классов были случайно разделены на 4 группы, так что для каждой группы из 12 классов применялся один из методов обучения в одном из условий. В конце полугодия исследователь провел одну и ту же контрольную работу по геометрии в каждом классе. Каждый класс был охарактеризован по некоторому критерию успеваемости (табл. 57), баллы округлены до целых.

Таблица 57

Исследование эффективности разных методов и условий преподавания геометрии

Метод (A)	Условия (B)					
	лекция в классе			программированное обучение		
Традиционный	2	4	4	9	10	10
	5	6	6	12	13	13
	6	7	7	14	14	14
	7	8	10	15	16	17
Современный	10	10	11	21	22	22
	13	13	13	25	26	30
	14	14	15	31	32	32
	16	17	17	33	34	35

В этом случае оказывается возможным оценить влияние не только *главных эффектов* A и B, но и их *взаимодействия* AB. Наличие взаимодействия означает, что эффекты разных факторов не просто суммируются, но при некоторых комбинациях уровней факторов оказываются больше (или меньше) аддитивных. Поясним это упрощенным примером. Пусть лечебный эффект сульфамидного препарата равен 3, а антибиотика — 5. Аддитивный эффект (отсутствие взаимодействия) заключается в том, что при совместном применении сульфамидного препарата и антибиотика лечебный эффект будет равен $3+5=8$. Взаимодействие же (положительное) может выразиться, например, в мультипликативной реакции: $3\times 5=15$. Возможность обнаружения взаимодействий является принципиальным достижением, полученным Р. А. Фишером. статистическая модель двухфакторного комплекса с повторностями выглядит следующим образом:

$$\widetilde{x}_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\widetilde{e}_{ijk},$$

где μ — общее среднее; α_i — отклонения от μ, обусловленные действием фактора A; β_i — отклонения от μ, обусловленные действием фактора B; $(\alpha\beta)_{ij}$ — отклонения от аддитивного эффекта (взаимодействие); $\widetilde{e}_{ijk} \sim N(0;\sigma^2)$ — «ошибки», независимые случайные величины, распределенные нормально со средним 0 и дисперсией σ^2.

И вновь очень важными являются предположения:

$$\sum_{i=1}^{k} \alpha_i = 0; \quad \sum_{i=1}^{k} (\alpha\beta)_{ij} = 0 \quad \text{для всех} \quad i$$

$$\text{и} \quad \sum_{j=1}^{l} \beta_j = 0; \quad \sum_{i=1}^{k} (\alpha\beta)_{ij} = 0 \quad \text{для всех} \quad j.$$

Двухфакторные комплексы с повторностями могут быть трех типов:

1) равномерные (с равным числом наблюдений в каждой ячейке);

2) пропорциональные и

3) непропорциональные (с произвольным, разным числом наблюдений в ячейке).

Таблица 58

Пропорциональный комплекс двухфакторного дисперсионного анализа

Факторы	Уровни j / i	B		$\sum_{j=1}^{2} n_{ij}$
		1	2	
A	1	$n_{11} = 2$	$n_{12} = 4$	$n_{1\cdot} = 6$
	2	$n_{21} = 3$	$n_{22} = 6$	$n_{2\cdot} = 9$
	3	$n_{31} = 2$	$n_{32} = 4$	$n_{3\cdot} = 6$
$\sum_{i=1}^{3} n_{ij}$		$n_{\cdot 1} = 7$	$n_{\cdot 2} = 14$	$n = 21$

Пример пропорционального комплекса показан в табл. 58. *Пропорциональным* называется такой комплекс, для численности любой ячейки

которого выполняется условие

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}.$$

И теоретически, и технически хорошо разработаны равномерные комплексы. Ситуация несколько осложняется для пропорциональных и становится совсем сложной для непропорциональных комплексов, здесь возможно применение лишь приближенных и весьма трудоемких с вычислительной точки зрения методов. Поэтому всегда следует стремиться при планировании эксперимента к равномерному или, в крайнем случае, к пропорциональному комплексу. Более того, если комплекс оказывается все-таки непропорциональным, выгоднее исключить (случайным образом) часть наблюдений (5–10 %), чем вести его полный анализ.

Если оба фактора имеют фиксированные уровни, то это будет модель I двухфакторного анализа; если уровни обоих факторов случайные, то речь идет о модели II. Возможен и третий вариант: уровни одного фактора фиксированные, другого — случайные; это *смешанная модель*.

До сих пор мы говорили о схеме полной классификации. Очень часто, однако, приходится иметь дело с *иерархической классификацией*.

ПРИМЕР VIII-6. В примере VIII-2 была приведена часть материалов более обширного эксперимента. В действительности 3 хряка скрещивались каждый с 5 свиноматками, причем ни один хряк не скрещивался с одной и той же свиноматкой и в потомстве каждой матки отдельно учитывались признаки четырех поросят. Таким образом, дисперсионный комплекс был двухфакторным иерархическим, его структура представлена на рис. 46.

§ 6. Несогласованность с моделью дисперсионного анализа и преобразование данных

Напомним, что в модели однофакторного дисперсионного анализа с фиксированными эффектами делаются следующие предположения:

1) $\widetilde{x}_{ij} = \mu + \alpha_i + \widetilde{e}_{ij}$, причем $\sum_{i=1}^{k} \alpha_i = 0$. «Нелинейность» экспериментальной ситуации может заключаться, например, в том, что эффект данного уровня неодинаков для всех изучаемых объектов, т. е., другими словами, имеет место внутригрупповая гетерогенность;

2) $\widetilde{e}_{ij}$ — независимые нормально распределенные случайные величины с одной и той же дисперсией σ^2. Нарушение условия независимости

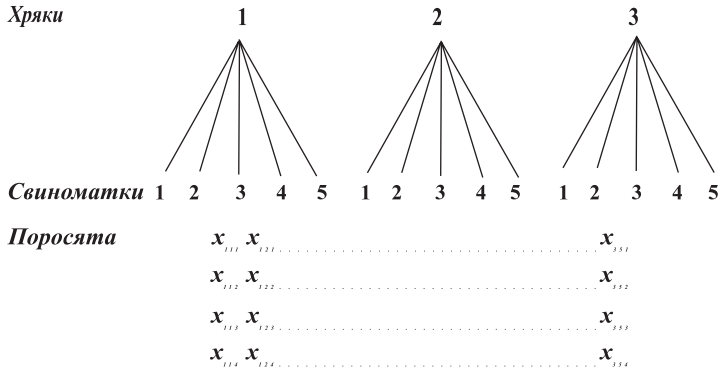


Рис. 46. Схема двухфакторной иерархической классификации (к примеру VIII-6)

может возникать, в частности, при проведении повторных наблюдений над одними и теми же объектами. Закон распределения, например, нормальность, является, как правило, свойством признака. Нередки ситуации, когда определенный уровень фактора влияет не только на групповую среднюю, но и на дисперсию, и тогда условие равенства дисперсий во всех группах будет нарушаться.

Существует ряд методов, с помощью которых можно проверить, выполняются ли требования модели в конкретном материале, подлежащем изучению с помощью дисперсионного анализа.

Поскольку вопрос о границах применимости дисперсионного анализа очень важен, отметим, что этот вопрос неоднократно изучался и продолжает интенсивно изучаться специалистами по математической статистике. Правда, исследование осуществляется, как правило, не аналитически (на необходимом уровне строгости), а (вследствие чрезвычайной сложности возникающих задач) путем проведения статистических экспериментов, часто с использованием электронно-вычислительных машин.

Если попытаться совсем кратко (и огрубленно) суммировать имеющиеся на сегодняшний день результаты, то можно сказать следующее. Аппарат дисперсионного анализа вполне применим, если распределение куполообразно и не резко асимметрично, а групповые дисперсии различаются в 1,5–2,5 раза. По-видимому, допустимы и значительно бóльшие отклонения, но при условии равенства числа наблюдений в группах в однофакторном анализе и равномерности двухфакторного комплекса.

Нередко еще до начала эксперимента у биолога могут быть соображения о виде распределения изучаемого признака. И если, например, предполагаются биномиальное или пуассоновское распределения, для которых среднее и дисперсия связаны, то, конечно, не следует надеяться на устойчивость модели дисперсионного анализа. В этих и аналогичных случаях надо заранее предусмотреть и осуществить преобразование шкалы измерений, которое дает распределение, по крайней мере, близкое к нормальному, и стабилизирует дисперсию.

В случае распределения Пуассона обычно используется преобразование $\sqrt{x}$ или, для малых значений x, преобразование $\sqrt{x+c}$, где оптимальным является $c = 0{,}386$.

В случае биномиального распределения обычно берется $2\arcsin\sqrt{h}$, где h — частота события (дисперсия стабилизируется для значений h от 0,05 до 0,95).

Для асимметричных, вытянутых вправо распределений, часто встречающихся, особенно, в физиологии, целесообразно преобразование $\lg x$.

Вообще говоря, если экспериментатор проводит обширные опыты на протяжении достаточно длительного времени, работая с какими-то определенными признаками, всегда полезно не ограничиваться вычислением статистических оценок параметров, но попытаться построить соответствующее распределение и исследовать его более подробно.

§ 7. Сравнение параметров положения нескольких неизвестных распределений

Для сравнения параметров положения нескольких генеральных совокупностей, имеющих неизвестное распределение, предложено несколько непараметрических критериев. Здесь мы рассмотрим один из них — критерий Крускала–Уоллиса, эффективность которого сравнима с эффективностью F-критерия.

Подобно тому, как F-критерий в однофакторном дисперсионном анализе является обобщением двухвыборочного t-критерия, критерий Крускала–Уоллиса есть обобщение критерия Вилкоксона–Манна–Уитни на случай сравнения параметров положения нескольких совокупностей с неизвестными распределениями.

Пусть $(x_{11}, \ldots, x_{1n_1}), (x_{21}, \ldots, x_{2n_2}), \ldots, (x_{k1}, \ldots, x_{kn_k})$ — независимые выборки из совокупностей, каждая из которых имеет непрерывную функцию распределения $F_i(x)$, $i = 1, \ldots, k$. Проверяется гипотеза H_0: $\tau_1 = \ldots = \tau_k = \tau$ о равенстве параметров положения этих k совокупностей.

Альтернатива H_1: $\tau_i \neq \tau$ подразумевает, что, по крайней мере, одна из совокупностей отличается от всех других параметром положения. Ситуация (за исключением предположений о нормальности распределений и о равенстве внутригрупповых дисперсий) аналогична модели I однофакторного дисперсионного анализа. Поскольку распределения сравниваемых совокупностей неизвестны, гипотезу H_0 следует проверять с помощью непараметрического критерия.

В качестве меры различия параметров положения в критерии Крускала–Уоллиса по существу выступает статистика $\widetilde{U}$ Манна-Уитни. Проиллюстрируем это примером.

Пример VIII-7 (М. В. Падкина, 1978 г.). У четырех штаммов дрожжей *Saccharomyces cerevisiae* определяли константу Михаэлиса–Ментен K_M ($10^{-3}M$) для кислой фосфатазы. Для каждого штамма проведено 3–4 повторных опыта по выделению фермента и определению его активности. Результаты представлены в табл. 59. Различаются ли штаммы по этому показателю?

Таблица 59

Значения K_M $(10^{-3}M)$ для четырех штаммов дрожжей *Saccharomyces cerevisiae* (к примеру VIII-7)

Номер штамма			
1	2	3	4
2,17	2,08	1,66	1,52
1,92	2,00	1,20	1,66
2,20	2,12	1,61	1,59
	1,92		1,47

Имеем 4 независимые выборки ($k = 4$), объемы которых, соответственно, $n_1 = 3$, $n_2 = 4$, $n_3 = 3$, $n_4 = 4$; общее число наблюдений $N = \sum_{i=1}^{k} n_i = 14$. Составим матрицу сравнений (табл. 60).

Каждый столбец этой матрицы представляет собой сравнение одной из выборок со всеми остальными. В двух нижних строках приведены значения U_i^+ и U_i^-:

$$U_i^+ = T_i^+ + \frac{1}{2}T_i^{\pm} \quad \text{и} \quad U_i^- = T_i^- + \frac{1}{2}T_i^{\pm},$$

Таблица 60

Матрица сравнений к построению критерия Крускала–Уоллиса (см. пример VIII-7)

x_{ij}	2,17	1,92	2,20	2,08	2,00	2,12	1,92	1,66	1,20	1,61	1,52	1,66	1,59	1,47
2,17				+	+	+	+	+	+	+	+	+	+	+
1,92				−	−	−	±	+	+	+	+	+	+	+
2,20				+	+	+	+	+	+	+	+	+	+	+
2,08	−	+	−					+	+	+	+	+	−	+
2,00	−	+	−					+	+	+	+	+	+	+
2,12	−	+	−					+	+	+	+	+	+	+
1,92	−	±	−					+	+	+	+	+	+	+
1,66	−	−	−	−	−	−	−				+	±	+	+
1,20	−	−	−	−	−	−	−				−	−	−	−
1,61	−	−	−	−	−	−	−				+	−	+	+
1,52	−	−	−	−	−	−	−	−	+	−				
1,66	−	−	−	−	−	−	−	±	+	+				
1,59	−	−	−	−	−	−	−	−	+	−				
1,47	−	−	−	−	−	−	−	−	+	−				
U_i^+	3,5			8,5				26,5			34,5			
U_i^-	29,5			31,5				6,5			5,5			

где T_i^+, T_i^-, $T_i^\pm$ — число знаков «+», «−», «±» в i-м столбце матрицы соответственно. (Используя соотношение $U_i^+ U_i^- = n_i(N - n_i)$, проверьте правильность вычислений.) Как и в случае критерия Вилкоксона–Манна–Уитни, статистики $\widetilde{U}_i^+$ и $\widetilde{U}_i^-$ равноправны, и мы будем использовать только одну из них — U_i^+, обозначая ее просто $\widetilde{U}_i$.

Статистика критерия Крускала–Уоллиса конструируется следующим образом. Построим нормированную случайную величину

$$\widetilde{u}_i = \frac{\widetilde{U}_i - E\widetilde{U}_i}{\sqrt{D\widetilde{U}_i}} \sim N(0;1),$$

распределение которой, как известно (гл. VI, § 7), уже при малых значениях n_i хорошо аппроксимируется нормированным нормальным распределением. Ее математическое ожидание есть при H_0

$$E\widetilde{U}_i = \frac{1}{2} n_i(N - n_i).$$

В качестве статистики $D\widetilde{U}_i$, как и в случае параметрического дисперсионного анализа, используется внутригрупповая дисперсия.

Можно показать, что

$$D\widetilde{U}_i = \frac{1}{12} n_i (N^2 - 1).$$

Если справедлива H_0, то распределение случайной величины $\widetilde{u}_i^2$ при $n_i \to \infty$ (практически при $n_i \geqslant 5$) есть распределение $\chi^2(\nu)$, $\nu = 1$:

$$\widetilde{u}_i^2 = \frac{(\widetilde{U}_i - E\widetilde{U}_i)^2}{D\widetilde{U}_i} \sim \chi^2(\nu), \quad \nu = 1.$$

Статистика Крускала–Уоллиса, по существу, есть сумма таких величин $\widetilde{H}' = \sum_{i=1}^{k} \widetilde{u}_i^2$. Очевидно, что только $(k-1)$ слагаемых зависимы, поэтому если величину $\widetilde{H}'$ усреднить по k и помножить на $(k-1)$, то полученная величина будет примерно распределена как $\widetilde{\chi}^2$ с числом степеней свободы $\nu = k - 1$:

$$\widetilde{H} = \frac{\widetilde{H}'(k-1)}{k} \sim \chi^2(\nu), \quad \nu = k - 1.$$

Аппроксимация удовлетворительна при объемах выборок $n_i \geqslant 5$ или числе выборок $k \geqslant 3$. Статистика $\widetilde{H}$ называется *статистикой Крускала–Уоллиса*. После соответствующих подстановок она принимает вид

$$\widetilde{H} = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{1}{n_i} |\widetilde{U}_i - E\widetilde{U}_i|^2.$$

Если имеются совпадения, то вычисляется статистика $\widetilde{H}^*$ с поправкой:

$$\widetilde{H}^* = \frac{\widetilde{H}}{1 - \widetilde{A}}; \quad \widetilde{A} = \frac{1}{1 - \widetilde{A}} \sum_{j=1}^{g} (\widetilde{c}_j^3 - \widetilde{c}_j),$$

где g — число групп совпадений; $\widetilde{c}_i$ — число совпадающих значений в j-й группе.

Подставляя вместо $\widetilde{U}_i$ и $\widetilde{c}_j$ наблюдаемые значения U_i и c_j, получим $H^*_{\text{эксп}}$, сравнение которого со значениями $\chi^2_\alpha(k-1)$ позволяет делать выводы о справедливости гипотезы H_0. Таким образом, мы получили критерий Крускала–Уоллиса.

ПРИМЕР VIII-7 (продолжение). Имеем $g = 2, c_1 = 2$ и $c_2 = 2$, поправка на совпадения $A = 0{,}0044$.

$$H_{\text{эксп}} = \frac{12}{14 \cdot 15}\frac{(3{,}5 - 16{,}5)^2}{3} + \frac{(8{,}5 - 20)^2}{4} + \frac{(26{,}5 - 16{,}5)^2}{3} + \\ + \frac{(34{,}5 - 20)^2}{4} = 10{,}02 \quad \text{и} \quad H^*_{\text{эксп}} = \frac{10{,}02}{1 - 0{,}0044} = 10{,}06.$$

Критические значения (табл. V Приложения 1) $H_{0{,}05}(3) = 7{,}81$ и $H_{0{,}01}(3) = 11{,}3$. Следовательно, $0{,}01 < P\{\widetilde{H}^* \geqslant H^*_{\text{эксп}}\} < 0{,}05$, т. е. нулевая гипотеза отклоняется на 5 %-ном уровне значимости.

Когда условия аппроксимации не выполняются, находят точное распределение статистики H Крускала–Уоллиса. Оно выводится из предположения, что при H_0 все возможные варианты расположения знаков «+» и «−» в матрице сравнений равновероятны, их общее число есть $N!/(n_1!n_2!\ldots!n_k!)$, вероятность каждого из них равна $n_1!n_2!\ldots!n_k!/N!$. Подсчитав прямым перебором число разных значений H, можно получить распределение статистики $\widetilde{H}$ (см. табл. X Приложения 1).

Пусть в примере VIII-7 сравниваются только первые три выборки. Тогда $U_1 = 3{,}5, U_2 = 8{,}5$ и $U_3 = 21$; $E\widetilde{U}_1 = E\widetilde{U}_3 = 10{,}5$ и $E\widetilde{U}_2 = E\widetilde{U}_4 = 12$; $H_{\text{эксп}} = 6{,}12$. Согласно табл. X Приложения 1, $H_{0{,}05}(3;3;4) = 5{,}73$ и $H_{0{,}01}(3;3;4) = 6{,}75$. Следовательно, $0{,}01 < P\{\widetilde{H} \geqslant H_{\text{эксп}}\} < 0{,}05$.

Только после того, как критерий Крускала–Уоллиса выявил неоднородность сравниваемых совокупностей в целом, можно перейти к следующему этапу статистического анализа — множественным сравнениям. Для этого все k выборок сравниваем попарно с помощью критерия Вилкоксона–Манна–Уитни.

ПРИМЕР VIII-7 (окончание). Экспериментальные значения $U_{ij(\text{эксп})}$ статистики Манна–Уитни, полученные при попарном сравнении четырех штаммов дрожжей, приведены в табл. 61.

Объемы первой и третьей выборок слишком малы, чтобы говорить о существенности значений $U_{31} = 0, U_{41} = 0$ и $U_{32} = 0$. Только значение $U_{42} = 0$, которое равно критическому $U_{0{,}025}(4,4)$ (см. табл. VIII Приложения 1), свидетельствует о различиях между параметрами положения второй и четвертой совокупностей.

Итак, при уровне значимости $\alpha = 0{,}05$ кислые фосфатазы второго и четвертого штаммов дрожжей существенно различаются по сродству к субстрату (по K_M).

Таблица 61

Матрица множественных сравнений для данных примера VIII-7 (приведены значения $U_{ij\,(\text{эксп})}$)

Номер штамма	1	2	3
2	3,5		
3	0	0	
4	0	0	5,5

Задачи

VIII-1. Изучали процент гемоглобина у кур разных пород (табл. 62).

Таблица 62

Процент гемоглобина у кур разных пород (к задаче VIII-1)

Порода	Процент гемоглобина					
Итальянские	87	92	86	91	90	93
Куропатчатые	91	90	88	89		
Минорки	85	82	85	86		
Бентамки	82	82	85	83	81	

Одинаков ли этот показатель у разных пород?

Какую модель дисперсионного анализа вы выберете: модель I или модель II? В задаче речь идет о «проценте»; следует ли пользоваться преобразованием $2\arcsin\sqrt{h}$? Проводя дисперсионный анализ полных данных, попробуйте выровнять число наблюдений, отбросив по таблице случайных чисел два наблюдения в первой и одно в четвертой группах.

VIII-2 [Bliss, 1967]. Определялась концентрация билирубина в сыворотке крови восьми здоровых мужчин. Предварительный анализ показал, что дисперсии признака внутри групп выравниваются, если использовать преобразование $\lg x$. Поэтому в табл. 63 приведены уже преобразованные значения. Достоверны ли и насколько велики в общей изменчивости различия между разными людьми по содержанию билирубина?

Таблица 63

Концентрация билирубина в сыворотке крови (10^{-1} мг/мл) (к задаче VIII-2)

Индивид							
1	2	3	4	5	6	7	8
0,36	0,43	0,61	0,99	0,79	0,72	0,92	1,18
0,15	0,53	0,61	0,83	0,91	0,83	1,18	1,11
0,61	0,61	0,85	0,83	0,83	0,74	0,98	0,68
0,43	0,53	0,51	0,83	0,83	1,09	1,16	0,98
0,43	0,96	0,74	0,79	0,88	0,88	0,95	0,83
0,74	0,85	0,79	0,88	1,01	0,91	0,98	1,11
0,53	0,74	0,96	1,00	0,92	1,06	1,04	1,09
0,82	0,51	0,74	0,88	0,87	0,91	1,04	1,11

VIII-3. Покажите, что в случае двух выборок ($k = 2$) статистика F-критерия для дисперсионного анализа идентична статистике t-критерия (ср. задачу VI-10).

VIII-4. С четырех участков почвы был произведен высев бактерий на 10 чашек Петри с питательной средой, где каждая бактерия дает начало колонии. Число колоний, образовавшихся на каждой из чашек, представлено в табл. 64. Различается ли число бактерий в почве сравниваемых участков? Нужно ли преобразование данных и какое?

Таблица 64

Число колоний на чашках Петри (к задаче VIII-4)

Участок почвы	Число колоний на чашках									
	1	2	3	4	5	6	7	8	9	10
1	7	4	8	10	10	7	16	11	7	12
2	5	10	9	4	7	5	1	11	12	15
3	6	7	9	10	15	14	12	12	4	7
4	7	7	11	10	8	8	12	7	12	8

VIII-5 [Снедекор, 1961]. В табл. 65 приведены частоты растений сои, пораженных раком стебля. Проведите дисперсионный анализ этих данных. Каков закон распределения изучаемого признака?

Таблица 65 Процент растений сои, пораженных раком стебля (к задаче VIII-5)

Блок	Сорт					
	A	*B*	*C*	*D*	*E*	*F*
1	19,3	10,1	25,2	14,0	3,3	3,1
2	29,2	34,7	36,5	30,2	35,8	9,6
3	1,0	14,0	23,4	7,2	1,1	1,0
4	6,4	5,6	12,9	8,9	2,0	1,0

VIII-6 [Снедекор, 1961]. Определялась численность четырех видов планктона в пробах (табл. 66). Проведите дисперсионный анализ этих данных. Почему здесь необходимо и что дает преобразование $\lg x$?

Таблица 66

Численность четырех видов планктона (к задаче VIII-6)

Проба	Вид			
	I	II	III	IV
1	895	1 520	43 300	11 000
2	540	1 610	32 800	8 600
3	1 020	1 900	28 800	8 200
4	470	1 350	34 600	9 830
5	428	9 80	27 800	760
6	620	1 710	32 800	9 650
7	760	1 930	28 100	8 900

VIII-7. Дарвин в течение многих лет экспериментировал с растениями. Результаты одного из его опытов представлены в табл. 67. По поводу этих данных он писал: «Так как измерению было подвергнуто небольшое число растений от перекрестного опыления и самоопыления, то мне было чрезвычайно важно узнать, в какой степени достоверны полученные средние величины. Поэтому я попросил мистера Гальтона, который имел большой опыт в статистических исследованиях, посмотреть некоторые из моих таблиц измерений».

В ответном письме Ф. Гальтон сообщал, что он согласен с выводом Дарвина о превосходстве перекрестноопыленных растений над самоопыленными растениями, но что точная цифровая оценка этого превосход-

Таблица 67

Zea mays (молодые растения)

Цифры (в дюймах) расположены							
так, как они приведены у м-ра Дарвина			согласно их величине				
			в отдельных горшках		в одном ряду		
	Перекрестноопыленные	Самоопыленные	Перекрестноопыленные	Самоопыленные	Перекрестноопыленные	Самоопыленные	Разница
I	II	III	IV	V	VI	VII	VIII
Горшок I	$23\frac{4}{8}$ 12 21	$17\frac{3}{8}$ $20\frac{3}{8}$ 20	$23\frac{4}{8}$ 21 12	$20\frac{3}{8}$ 20 $17\frac{3}{8}$	$23\frac{4}{8}$ $23\frac{2}{8}$ 23	$20\frac{3}{8}$ 20 20	$-3\frac{1}{8}$ $-3\frac{2}{8}$ -3
Горшок II	22 $19\frac{1}{8}$ $21\frac{4}{8}$	20 $18\frac{3}{8}$ $18\frac{5}{8}$	22 $21\frac{4}{8}$ $19\frac{1}{8}$	20 $18\frac{5}{8}$ $18\frac{3}{8}$	$22\frac{1}{8}$ $22\frac{1}{8}$ 22	$18\frac{5}{8}$ $18\frac{5}{8}$ $18\frac{3}{8}$	$-3\frac{4}{8}$ $-3\frac{4}{8}$ $-3\frac{5}{8}$
Горшок III	$22\frac{1}{8}$ $20\frac{3}{8}$ $18\frac{2}{8}$ $21\frac{5}{8}$ $23\frac{2}{8}$	$18\frac{5}{8}$ $15\frac{2}{8}$ $16\frac{4}{8}$ 18 $16\frac{2}{8}$	$23\frac{3}{8}$ $22\frac{1}{8}$ $21\frac{5}{8}$ $20\frac{3}{8}$ $18\frac{2}{8}$	$18\frac{5}{8}$ 18 $16\frac{4}{8}$ $16\frac{2}{8}$ $15\frac{2}{8}$	$21\frac{5}{8}$ $21\frac{4}{8}$ 21 21 $20\frac{3}{8}$	18 18 18 $17\frac{3}{8}$ $16\frac{4}{8}$	$-3\frac{5}{8}$ $-3\frac{4}{8}$ -3 $-3\frac{5}{8}$ $-3\frac{7}{8}$
Горшок IV	21 $22\frac{1}{8}$ 23 12	18 $12\frac{6}{8}$ $15\frac{4}{8}$ 18	23 $22\frac{1}{8}$ 21 12	18 18 $15\frac{4}{8}$ $12\frac{6}{8}$	$19\frac{1}{8}$ $18\frac{2}{8}$ 12 12	$16\frac{2}{8}$ $15\frac{4}{8}$ $15\frac{2}{8}$ $12\frac{6}{8}$	$-2\frac{7}{8}$ $-2\frac{6}{8}$ $+3\frac{2}{8}$ $+0\frac{6}{8}$

ства представляется невозможной; «. . . трудность заключается в том, что мы не знаем точного закона, которому подчиняется ряд; . . . наблюдения . . . слишком малочисленны. . . ».

Проведите статистический анализ результатов этого эксперимента. Прокомментируйте следующие строки из письма Ф. Гальтона:

«Данные наблюдений в том виде, в каком я получил их, показаны в столбцах II и III (см. табл. 67), где в них на первый взгляд, несомненно, нельзя усмотреть какой-либо правильности. Но как только мы расположим их согласно их величине, так, как это сделано в столбцах IV и V, дело существенно изменится. Мы видим теперь, за немногими исключениями, что в каждом горшке наиболее высокое растение на той стороне, где посажены перекрестноопыленные растения, превосходит наиболее высокое растение на той стороне, где посажены самоопыленные растения, что второе по высоте растение превосходит соответствующее ему второе, третье превосходит третье и т. д. Из пятнадцати случаев, приведенных в таблице, имеется лишь два исключения из этого правила. Мы можем поэтому с уверенностью утверждать, что ряд перекрестноопыленных растений всегда будет обнаруживать превосходство над рядом самоопыленных растений в пределах тех условий, при которых производился настоящий опыт»[2].

[2] Дарвин Ч. Соч. Т.6. Перекрестное опыление и самоопыление. — М.; Л.: Изд-во АН СССР, 1950. — С. 272–275.

Глава IX

Анализ статистических связей

В этой главе будут рассмотрены задачи линейной регрессии, линейной корреляции, оценки рангового коэффициента корреляции (при ненормальности двумерного распределения), а также задача анализа таблиц сопряженности для качественных признаков, по сути являющаяся корреляционной задачей.

§ 1. Модель линейной регрессии

Вспомним пример IV-18. В этом примере в отличие от ранее рассмотренных мы имеем две переменные: независимую, или свободную, переменную x (доза облучения) и зависимую, или связанную, переменную $\widetilde{y}$ (частота мутаций). При этом важно, что x есть величина неслучайная, ее значения регистрируются (фиксируются) точно, без ошибок, а $\widetilde{y}$ является случайной величиной. Нас прежде всего интересует установление функциональной зависимости $\widetilde{y}$ от x вида $y = f(x)$, наглядно показанной на рис. 36. Уравнение $y = f(x)$ называется *уравнением регрессии y на x*.

Вообще говоря, $\widetilde{y}$ может зависеть от нескольких переменных:

$$x^{(1)}, x^{(2)}, \ldots, x^{(n)}, \text{ т. е. } \quad y = f(x^{(1)}, x^{(2)}, \ldots, x^{(n)}).$$

Мы будем рассматривать лишь случай зависимости от одной переменной. Но и в этом случае вид функции $f(x)$ может быть любым и в общем случае описываться полиномом

$$y = f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n = \sum_{i=0}^{n} a_i x_i.$$

Мы будем рассматривать лишь *линейную регрессию*

$$y = \varkappa + \beta x.$$

Не слишком ли узким является круг обсуждаемых вопросов? Нет. Во-первых, зависимости, близкие к линейным, нередко встречаются в биологии. Во-вторых, преобразуя масштабы по осям координат, многие типы зависимости можно достаточно точно свести к линейным (см. § 4, табл. 69 и рис. 52).

В последнем уравнении параметр β называют *коэффициентом линейной регрессии; $\varkappa$ — свободным членом*. При $x = 0$, $y = \varkappa$, т. е. $\varkappa$ — значение y в точке пересечения линии регрессии с осью ординат. При $\varkappa = 0$ $\beta = y/x$, т. е. β — тангенс угла наклона линии регрессии к оси абсцисс (рис. 47).

Мы будем пользоваться в дальнейшем и другой формой записи уравнения линейной регрессии:

$$y = \varkappa_0 + \beta(x - m_x),$$

где $m_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ и n — число наблюдений. Тогда $y = (\varkappa_0 - \beta m_x) + \beta x$ и $\varkappa = \varkappa_0 - \beta m_x$.

Этот вид записи оказывается удобным, поскольку можно получить независимые статистические оценки коэффициента регрессии и свободного члена.

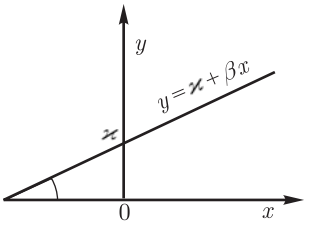


Рис. 47. Геометрическая интерпретация параметров $\varkappa$ и β в уравнении линейной регрессии

В дальнейшем греческими буквами $\varkappa$ и β будем обозначать параметры генеральной совокупности (константы), а латинскими a и b — выборочные оценки этих параметров.

Пусть x — неслучайная (фиксированная) переменная, значения которой $x_1, x_1, \ldots, x_n$ точно, без ошибки, задаются (или известны). Значениям x_i соответствуют значения случайных величин $\widetilde{y}_i$, независимых и нормально распределенных с параметрами μ_i и σ^2. Будем предполагать, что для любого μ и x имеет место линейная зависимость

$$\mu = \varkappa + \beta\varkappa = \varkappa_0 + \beta(x - m_x).$$

Таким образом, дисперсии $D\widetilde{y}_i$ предполагаются равными, а μ_i лежат на прямой (рис. 48).

Итак, статистическая задача заключается прежде всего в том, чтобы оценить параметры $\varkappa_0$, β и σ^2, т. е. необходимо подобрать прямую регрессии y по x:

$$\hat{y} = a_0 + b(x - m_x),$$

наилучшим образом описывающую (оценивающую) зависимость $\widetilde{y}$ от x в генеральной совокупности.

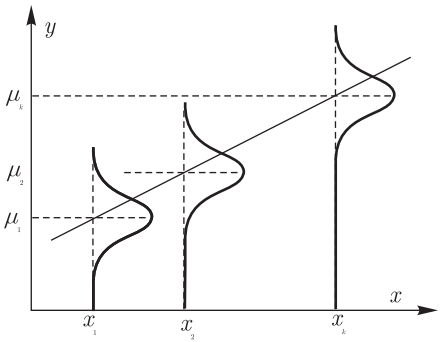


Рис. 48. Модель линейной регрессии

§ 2. Оценка параметров модели линейной регрессии

Для оценки параметров линии регрессии используется метод наименьших квадратов, который, по существу, совпадает с методом максимального правдоподобия в случае нормально распределенных наблюдений (см. гл. V, § 1). Поэтому статистики, получаемые методом наименьших квадратов, оказываются несмещенными и эффективными. Суть метода заключается в минимизации сумм квадратов отклонений наблюдаемых значений y_i от гипотетических (теоретических) $\hat{y}_i$:

$$R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - a_0 - b(x_i - m_x)]^2.$$

Чтобы найти a_0 и b, продифференцируем R по a_0 и по b, приравняв производные нулю:

$$\frac{\partial R}{\partial a_0} = -2 \sum_{i=1}^{n} [y_i - a_0 - b(x_i - m_x)]^2 = 0;$$

$$\frac{\partial R}{\partial b} = -2 \sum_{i=1}^{n} [y_i - a_0 - b(x_i - m_x)](x_i - m_x) = 0.$$

Отсюда имеем:

$$\sum_{i=1}^{n} y_i - na_0 - b\sum_{i=1}^{n}(x_i - m_x) = 0;$$
$$\sum_{i=1}^{n} y_i(x_i - m_x) - a_0\sum_{i=1}^{n}(x_i - m_x) - b\sum_{i=1}^{n}(x_i - m_x)^2 = 0.$$

Так как $\sum_{i=1}^{n} y_i(x_i - m_x) = 0$, то $a_0\frac{1}{n}\sum_{i=1}^{n} y_i = m_y$, тогда $a = m_y + bm_x$. Коэффициент линейной регрессии равен

$$b = \frac{\sum_{i=1}^{n}(x_i - m_x)y_i}{\sum_{i=1}^{n}(x_i - m_x)^2}.$$

В последнем выражении числитель можно видоизменить:

$$\sum_{i=1}^{n}(x_i - m_x)y_i = \sum_{i=1}^{n}(x_i - m_x)y_i - m_y\sum_{i=1}^{n}(x_i - m_x) =$$
$$= \sum_{i=1}^{n}(x_i - m_x)(y_i - m_y).$$

Таким образом, статистиками для оценивания $\varkappa$, $\varkappa_0$ и β являются

$$\widetilde{a}_0 = \widetilde{m}_y;$$
$$\widetilde{a} = \widetilde{m}_y - \widetilde{b}m_x;$$
$$\widetilde{b}\frac{\sum_{i=1}^{n}(x_i - m_x)(\widetilde{y}_i - \widetilde{m})}{\sum_{i=1}^{n}(x_i - m_x)^2}.$$

Обычно в биологических задачах представляет интерес статистический анализ коэффициента линейной регрессии. Коэффициент линейной регрессии β — это константа, которая оценивается с помощью статистики, являющейся случайной величиной. Поэтому выборочный коэффициент b есть реализация случайной величины $\widetilde{b}$, закон и параметры распределения которой требуется найти.

Поскольку $\widetilde{b}$ есть линейная комбинация нормально распределенных величин, то она распределена нормально со средним значением $E\widetilde{b} = \beta$ (проверьте это утверждение!), так как метод наименьших квадратов дает несмещенные оценки. Найдем дисперсию $D\widetilde{b}$, вспомнив, что $D\widetilde{y}_i = \sigma^2$:

$$D\widetilde{b} = D\left[\frac{\sum_{i=1}^n (x_i - m_x)\widetilde{y}_i}{\sum_{i=1}^n (x_i - m_x)^2}\right] =$$

$$= \frac{D[(x_1 - m_x)\widetilde{y}_1 + (x_2 - m_x)\widetilde{y}_2 + \ldots + (x_n - m_x)\widetilde{y}_n]}{\left[\sum_{i=1}^n (x_i - m_x)^2\right]^2} =$$

$$= \frac{\sigma^2 \sum_{i=1}^n (x_i - m_x)^2}{\left[\sum_{i=1}^n (x_i - m_x)^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - m_x)^2}.$$

Найдем теперь статистику для оценивания σ^2.

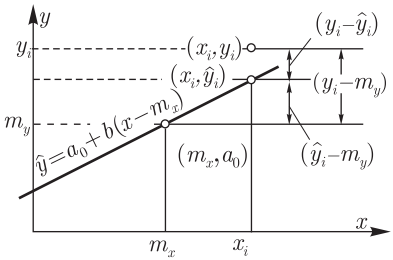


Рис. 49. Из чего слагается отклонение $(y_i - m_y)$?

Выборочное значение s^2 (оценка σ^2) связано с рассеянием полученных в эксперименте значений y_i вокруг значений $\hat{y}_i$, лежащих на линии регрессии $\hat{y} = a_0 + b(x - m_x)$, параметры которой были оценены по методу наименьших квадратов. Поэтому статистикой для оценивания σ^2 является случайная величина

$$\widetilde{s}^2 = \frac{1}{n-2} \sum_{i=1}^n (\widetilde{y}_i - \hat{y}_i)^2.$$

Число степеней свободы здесь меньше числа наблюдений на два, так как при вычислении линии регрессии на систему были наложены два линейных ограничения:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)=0;$$
$$\sum_{i=1}^{n}(x_i-m_x)=0.$$

Как удобно вычислять значение $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ по экспериментальным данным? Обратимся к рис. 49. Здесь показаны линия регрессии $\hat{y}=a_0+$ $+b(x-m_x)$, проходящая через точки (m_x,a_0) и $(x_i,\hat{y}_i)$, и значение (x_i,y_i), наблюдаемое в эксперименте. Можно видеть, что отклонение точки (x_i,y_i) от среднего арифметического m_y состоит из двух частей: отклонения y_i от значения $\hat{y}_i$, лежащего на линии регрессии, и отклонения $\hat{y}_i$ от среднего $m_y=a_0$. Рассмотрим тождество

$$\widetilde{y}_i-\widetilde{m}_y=(\widetilde{y}_i-\hat{y}_i)+(\hat{y}_i+\widetilde{m}_y).$$

Возведем в квадрат обе его части:

$$(\widetilde{y}_i-\widetilde{m}_y)^2=[(\widetilde{y}_i-\hat{y}_i)+(\hat{y}_i+\widetilde{m}_y)]^2.$$

Суммируя по i, получим

$$\sum_{i=1}^{n}(\widetilde{y}_i-\widetilde{m}_y)^2=\sum_{i=1}^{n}(\widetilde{y}_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i+\widetilde{m}_y)^2,$$

так как сумма попарных произведений равна нулю (см. гл. VIII, § 1).

Сумма квадратов, обусловленная регрессией, может быть найдена как

$$\sum_{i=1}^{n}(\hat{y}_i-\widetilde{m}_y)^2=\widetilde{b}^2\sum_{i=1}^{n}(x_i-m_x)^2=\widetilde{b}\sum_{i=1}^{n}(x_i-m_x)(\widetilde{y}_i-\widetilde{m}_y)=$$
$$=\frac{\left[\sum_{i=1}^{n}(x_i-m_x)(\widetilde{y}_i-\widetilde{m}_y)\right]^2}{\sum_{i=1}^{n}(x_i-m_x)^2}.$$

Таким образом, интересующее нас рассеяние экспериментальных точек вокруг линии регрессии («ошибка»), называемое суммой квадратов остатков,

или остатком, вычисляется как

$$\sum_{i=1}^{n}(\widetilde{y}_i-\hat{y}_i)^2=\sum_{i=1}^{n}(\widetilde{y}_i-\widetilde{m}_y)^2-\frac{\left[\sum_{i=1}^{n}(x_i-m_x)(\widetilde{y}_i-\widetilde{m}_y)\right]^2}{\sum_{i=1}^{n}(x_i-m_x)^2}$$

и статистика $\widetilde{s}^2$ для оценивания σ^2 принимает вид:

$$\widetilde{s}^2=\frac{1}{n-2}\left\{\sum_{i=1}^{n}(\widetilde{y}_i-\widetilde{m}_y)^2-\frac{\left[\sum_{i=1}^{n}(x_i-m_x)(\widetilde{y}_i-\widetilde{m}_y)\right]^2}{\sum_{i=1}^{n}(x_i-m_x)^2}\right\}.$$

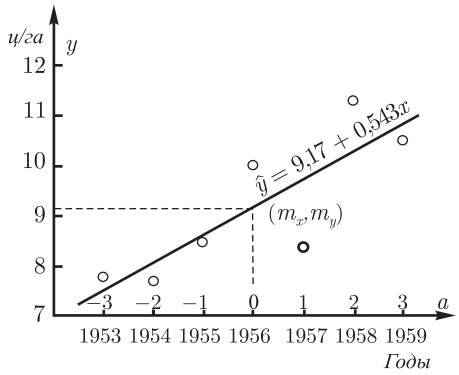


Рис. 50. Результат регрессионного анализа для данных примера IX-1

Тогда статистикой для оценивания дисперсии $D\widetilde{b}$ выборочного коэффициента линейной регрессии $\widetilde{b}$ является, очевидно,

$$\widetilde{s}_b^2=\frac{\widetilde{s}^2}{\sum_{i=1}^{n}(x_i-m_x)}.$$

Величина $\widetilde{s}_b$ называется *стандартной ошибкой выборочного коэффициента линейной регрессии* $\widetilde{b}$.

Для практических расчетов удобны формулы

$$b = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2};$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^n y_i^2 - \frac{1}{n}(\sum_{i=1}^n y_i)^2 - \right.$$

$$\left. - \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}\right];$$

$$s_b^2 = \frac{s^2}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}$$

(см. § 1 гл. II, задачи II-3 и IX-6).

Пример IX-1 [Рокицкий, 1973]. Фактическая урожайность культур на одной ферме в последовательные годы была следующей (ц/га):

Год	1953	1954	1955	1956	1957	1958	1939
Урожайность	7,8	7,7	8,5	10,0	8,4	11,3	10,5

Значимо ли увеличение урожайности?

Число наблюдений $n = 7$. Кодируя годы через $-3, -2, \ldots, +3$, получаем $\sum_{i=1}^7 x_i = 0$, $m_x = 0$; $\sum_{i=1}^7 y_i = 64{,}2$, $m_y = 9{,}17$; $\sum_{i=1}^7 x_i^2 = 28$; $\sum_{i=1}^7 y_i^2 = 600{,}88$; $\sum_{i=1}^7 x_i y_i = 15{,}2$. Тогда

$$\sum_{i=1}^7 (x_i - m_x)^2 = 28;$$

$$\sum_{i=1}^7 (y_i - m_y)^2 = 600{,}88 - \frac{1}{7} \times 64{,}2^2 = 12{,}07;$$

$$\sum_{i=1}^7 (x_i - m_x)(y_i - m_y) = 15{,}2.$$

Далее получаем $b = 15{,}2/28 = 0{,}543$; $a = 9{,}17$. Уравнение линейной регрессии $\hat{y} = 9{,}17 + 0{,}543x$ (в конкретных примерах обычно пишут для упрощения y вместо $\hat{y}$). Оценка σ^2 равна $s^2 = 1/5(12{,}07 - 15{,}2/28) = 2{,}305$, оценка $D\widetilde{b}$ равна $s_b^2 = 2{,}305/28 = 0{,}0832$ и оценка «ошибки» коэффициента линейной регрессии $s_b = 0{,}288$.

Представление исходных данных и результата их анализа на одном графике позволяет наглядно убедиться, насколько они согласуются друг с другом (рис. 50). В данном примере разброс экспериментальных точек вокруг линии регрессии достаточно велик и найденное уравнение регрессии отражает, по-видимому, лишь общую тенденцию к возрастанию в сложной картине изменения урожайности по годам.

§ 3. Проверка гипотезы о значении коэффициента линейной регрессии

После того как получено уравнение линейной регрессии $\hat{y} = a + bx$ и вычислена оценка дисперсии коэффициента регрессии s_b^2, возникает задача сравнения b с некоторым гипотетическим значением β. Чаще всего, когда просто требуется ответить на вопрос о достоверности применяемого воздействия, проверяется H_0: $\beta = 0$. Или, другими словами, достоверно ли угол наклона линии регрессии отличается от нуля? (Заметим, что H_0: $\beta = 0$ не единственная возможная гипотеза; см., например, задачу IX-7.)

Решение задачи оказывается простым. Точно так же, как в гл. V, было показано, что

$$\frac{\widetilde{b} - \beta}{\widetilde{s}_b} \sim t(\nu), \quad \nu = n - 1,$$

можно показать, что

$$\frac{\widetilde{b} - \beta}{\widetilde{s}_b} \sim t(\nu), \quad \nu = n - 2.$$

Поэтому, вычислив

$$t_{\text{эксп}} = \frac{|b - \beta|}{s_b},$$

остается сравнить полученное значение $t_{\text{эксп}}$ с табличным $t_{\alpha/2}(\nu)$, $\nu = n-2$ (см. табл. III Приложения 1).

Вернемся к примеру IX-1:

$$t_{\text{эксп}} = \frac{0{,}543}{0{,}288} = 1{,}89, \quad \nu = 5,$$

что дает $0{,}05 < P\{|t| \geqslant t_{\text{эксп}}\} < 0{,}10$, т. е. увеличение урожая от года к году значимо лишь на 10 %-ном уровне значимости.

§ 4. Сравнение двух коэффициентов линейной регрессии β_1 и β_2

Пример IX-2 [Фишер, 1958]. В табл. 68 приведены десятичные логарифмы объемов, занимаемых по мере роста клетками морской водоросли. Над культурой A наблюдения велись девять дней, над культурой B — восемь дней. Различаются ли кривые роста достоверно? Построив график, убеждаемся, что в принятой шкале измерений в обоих случаях зависимость $\lg V$ от x близка к линейной.

Таблица 68

Рост двух культур морской водоросли[1] (к примеру IX-2)

Дни учета (x_i)	Культура	
	A, $y_{1i} = \lg V_{Ai}$	B, $y_{2i} = \lg V_{Bi}$
1	3,59	3,54
2	3,82	3,83
3	4,17	4,35
4	4,53	4,83
5	4,96	4,91
6	5,16	5,30
7	5,50	5,57
8	5,60	6,04
9	6,09	

Таким образом, мы пришли к сравнению двух коэффициентов линейной регрессии β_1 и β_2, когда σ^2 неизвестна. Эта задача ничем не отличается от задачи сравнения двух средних μ_1 и μ_2 нормально распределенных совокупностей (см. гл. V).

[1] V — объем, занимаемый клетками.

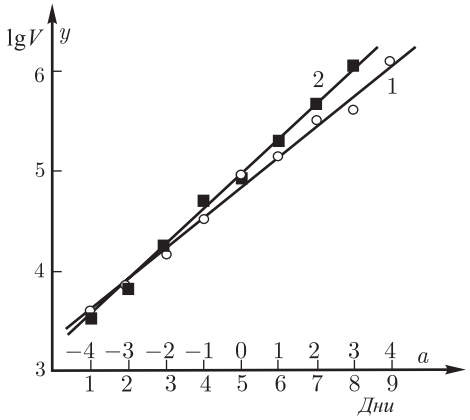


Рис. 51. Результаты регрессионного анализа для данных примера IX-2: 1 — линия регрессии для культуры A; 2 — то же для культуры B

Если $\sigma_1^2 = \sigma_2^2 = \sigma$, то вычисляем

$$t_{\text{эксп}} = \frac{|b_1 - b_2|}{s\sqrt{1/\sum_{i=1}^{n_1}(x_{1i} - m_{x_1})^2 + \sum_{i=1}^{n_2}(x_{2i} - m_{x_2})^2}}$$

при $\nu = n_1 + n_2 - 4$, где

$$s^2 = \frac{(n_1 - 2)s_1^2 + (n_2 - 2)s_2^2}{n_1 + n_2 - 4}.$$

Если $\sigma_1^2 \neq \sigma_2^2$, то, как и при сравнении выборочных средних, вычисляют s_1^2 и s_2^2, а затем

$$t_{\text{эксп}} = \frac{|b_1 - b_2|}{\sqrt{s_{b_1}^2 + s_{b_2}^2}}$$

при ν, определенном из

$$\frac{1}{\nu} = \frac{c^2}{\nu_1} + \frac{(1 - c^2)}{\nu_2},$$

где

$$c = \frac{s_{b_1}^2}{s_{b_1}^2 + s_{b_2}^2}.$$

Дадим решение примера IX-2.

Культура A: $n_1 = 9$; кодируя дни учета $-4, -3, \ldots, +4$, получаем $\sum_{i=1}^{9} x_{1i} = 0$; $m_{x_1} = 0$; $\sum_{i=1}^{9} x_{1i}^2 = 60$; $\sum_{i=1}^{9} y_{1i} = 43{,}42$; $m_{y1} = 4{,}824$; $\sum_{i=1}^{9} y_{1i}^2 = 214{,}8884$; $\sum_{i=1}^{9} x_{1i}y_{1i}^2 = 18{,}63$. Тогда $\sum_{i=1}^{9}(x_{1i} - m_{x_1})^2 = 60$; $\sum_{i=1}^{9}(y_{1i} - m_{x_1})^2 = 5{,}4496$; $\sum_{i=1}^{9}(x_{1i} - m_{x_1})^2 \times (y_{1i} - m_{x_1})^2 = 18{,}63$. Далее получаем $b_1 = 0{,}3105$; $a_1 = 4{,}82$. Уравнение регрессии: $\widehat{y}^{(1)} = 4{,}82 + 0{,}311x$. Оценка дисперсии σ_1^2 равна $s_1^2 = 0{,}0479$.

Культура B: $n_2 = 8$; $\sum_{i=1}^{9} x_{2i} = -4$; $m_{2i} = -0{,}5$; $m_{x_2} = -0{,}5$; $\sum_{i=1}^{8} x_{2i}^2 = 44$; $\sum_{i=1}^{8} y_{2i} = 38{,}41$; $\sum_{i=1}^{8} y_{2i} = 4{,}796$; $\sum_{i=1}^{8} y_{2i}^2 = 189{,}1565$; $\sum_{i=1}^{8} x_{2i}y_{2i} = -4{,}62$. Тогда $\sum_{i=1}^{8}(x_{2i} - m_{x_2}) = 42$; $\sum_{i=1}^{8}(y_{2i} - m_{y_2}) = 5{,}1436$; $\sum_{i=1}^{8}(x_{2i} - m_{x_2})(y_{2i} - m_{y_2}) = 14{,}565$. Далее получаем $b_2 = 0{,}3468$; $a_2 = 4{,}970$. Уравнение регрессии $\hat{y}^{(2)} = 4{,}97 + 0{,}347x$. Оценка дисперсии σ_2^2 равна $s_2^2 = 0{,}0154$.

Сравнение выборочных дисперсий дает $F_{\text{эксп}} = s_1^2/s_2^2 = 3{,}10$ и, согласно табл. IV Приложения 1, $P\{\widetilde{F} \geqslant F_{\text{эксп}}\} > 0{,}1$, т. е. принимается гипотеза H_0: $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Поэтому

$$s^2 = 7 \cdot 0{,}0479 + 6 \cdot 0{,}0154/13 = 0{,}0329;$$

$$t_{\text{эксп}} = -\frac{|0{,}311 - 0{,}347|}{\sqrt{0{,}0329(1/60 + 1/42)}} = 0{,}99; \quad \nu = 13;$$

$$P\{\widetilde{t} \geqslant t_{\text{эксп}}\} > 0{,}05.$$

Таким образом, кривые роста культур A и B не различаются. Результаты анализа наглядно представлены на рис. 51.

В примере IX-2 линейная зависимость от времени роста культуры наблюдалась не для регистрируемого признака — объема, занятого культурой, а для логарифма объема. Очень часто нелинейные зависимости, подчас достаточно сложные, могут быть линеаризованы простыми преобразованиями.

Некоторые типы таких преобразований представлены в табл. 69. На рис. 52 показано, как выглядят исходные кривые.

В связи с этим необходимо сделать два замечания. Во-первых, удовлетворительные результаты могут быть получены путем применения к одним и тем же исходным данным разных преобразований. Поэтому при выборе вида преобразования следует использовать все возможные сведения о предполагаемом виде зависимости. Во-вторых, в отношении преобразованных

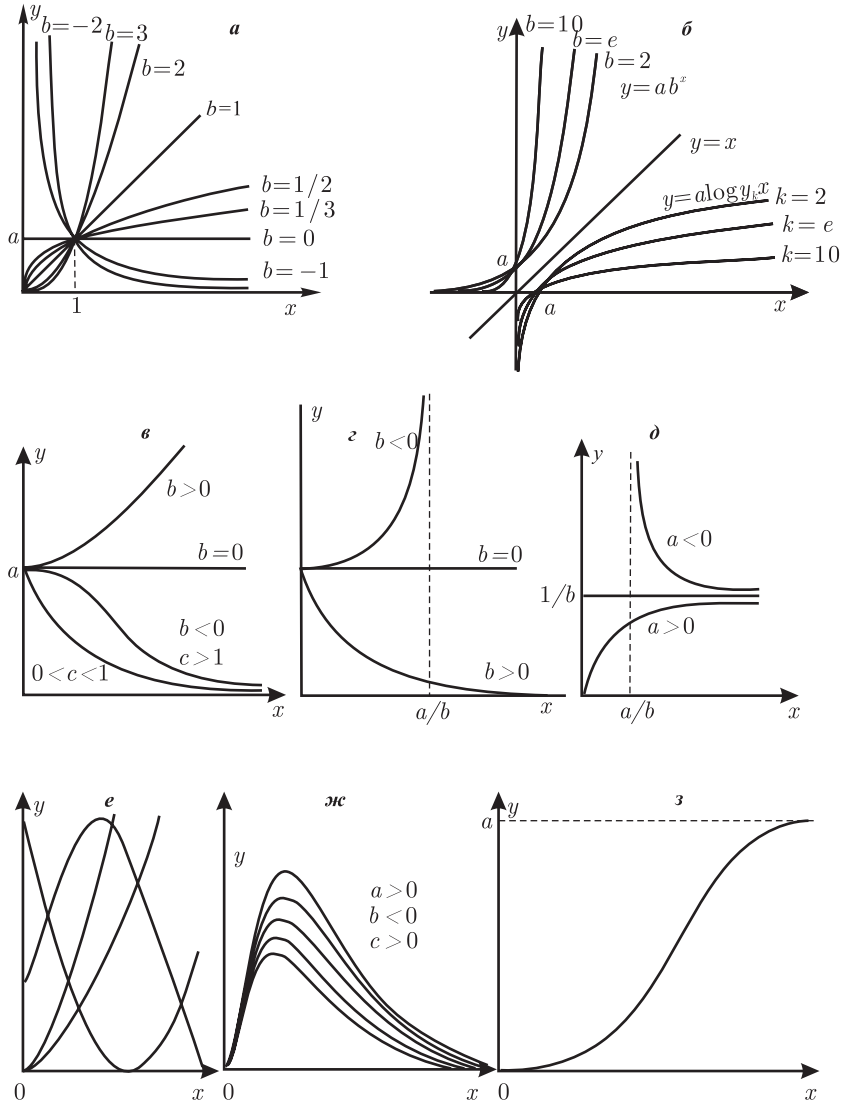


Рис. 52. Вид функций, методы линеаризации которых указаны в табл. 69

Таблица 69

Преобразования, используемые для линеаризации некоторых функций

Вид функции		Прямоугольные координаты[2]	
рис. 52	формула	абсцисса	ордината
а	$y = ax^b$	$\log x$	$\log y$
	$b = 2$	x^2	y
	$b = 3$	x^3	y
	$b = 1/2$	$\sqrt{x}$	y
	$b = 1/3$	$\sqrt[3]{x}$	y
	$b = -1$	$1/x$	y
	$b = -2$	$1/x^2$	y
б	$y = ab^x$	x	$\log y$
	$y = a \log x$	$\log x$	y
в	$y = ae^{bxc}$	x^c	$\log y$
г	$y = \dfrac{1}{a + bx}$	x	$1/y$
д	$y = \dfrac{x}{a + bx}$	x	x/y
е	$y = a + b + cx^2$	x	$(y - y_0)/(x - x_0)$
ж	$y = \dfrac{x}{a + bx + cx^2}$	$x - x_0$	$(x - x_0)/(y - y_0)$
з	$y = \dfrac{1}{1 + be^{-cx}}$	x	$\log[(a/y) - 1]$

величин должны выполняться все условия модели, обсуждавшиеся ранее (см. § 1).

[2] (x_0, y_0) — координаты произвольной точки на эмпирической кривой.

§ 5. Сопоставление регрессионной и корреляционной задач

ПРИМЕР IX-3. Оценим связь между массой жабр (x_1) и массой тела (x_2) у краба *Pachygrapsus crassipes* по данным примера IV-7.

Это типичный пример корреляционной задачи: у одного объекта регистрируются значения двух количественных признаков и ставится вопрос о наличии связи между значениями признаков в генеральной совокупности.

Модель линейной корреляции между двумя признаками основана на двумерном нормальном распределении, причем коэффициент корреляции ϱ является одним из параметров этого распределения (см. гл. III, § 6). Подчеркнем, что вычисление выборочного коэффициента корреляции по формулам, которые мы ниже получим, возможно для любого распределения, но статистические оценки и выводы, основанные на этом, справедливы только в предположении двумерного нормального распределения.

Сопоставление примеров IX-1 и IX-3 ясно показывает различия между регрессионной и корреляционной задачами. В регрессионной задаче только одна из двух величин $(\widetilde{y})$ является случайной, в корреляционной задаче обе величины ($\widetilde{x}_1$ и $\widetilde{x}_2$) — случайные.

В формальном (математическом) отношении регрессии и корреляция тесно связаны между собой. Мы не будем обсуждать эти вопросы, фиксируя внимание на принципиальных (особенно в биологических приложениях) различиях в структуре моделей регрессионного и корреляционного анализа.

§ 6. Оценка коэффициента корреляции ϱ

Найдем оценки m_1, m_2, s_1^2, s_2^2 и r параметров μ_1, μ_2, σ_1^2, σ_2^2 и ϱ двумерного нормального распределения по данным двумерной выборки $(\widetilde{x}_{11}, \widetilde{x}_{21}), (\widetilde{x}_{12}, \widetilde{x}_{22}), \ldots, (\widetilde{x}_{1n}, \widetilde{x}_{2n})$ с помощью метода максимального правдоподобия. Будем, как и раньше, искать максимум логарифма функции правдоподобия:

$$\ln L(\mu_1, \mu_2, \sigma_1, \sigma_2, \varrho) = -n \ln(2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}) -$$

$$-\frac{1}{2(1-\varrho^2)}\left[\frac{\sum_{i=1}^{n}(x_{1i}-\mu_1)^2}{\sigma_1^2} - 2\varrho\frac{\sum_{i=1}^{n}(x_{1i}-\mu_1)^2(x_{2i}-\mu_2)^2}{\sigma_1^2\sigma_2^2} +\right.$$

$$+\frac{\sum_{i=1}^{n}(x_{2i}-\mu_2)^2}{\sigma_2^2}\Bigg].$$

Дифференцируя $\ln L$ по $\mu_1, \mu_2, \sigma_1, \sigma_2$ и ϱ, приравнивая производные нулю, подставляя, наконец, вместо значений параметров их оценки m_1, m_2, s_1, s_2 и r, читатель может получить в результате решения уравнений правдоподобия следующие статистики для оценки всех пяти параметров двумерного нормального распределения:

$$\widetilde{m}_1=\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_{1i};\quad \widetilde{m}_2=\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_{2i};$$

$$\widetilde{s}_1=\frac{1}{n}\sum_{i=1}^{n}(\widetilde{x}_{1i}-\widetilde{m}_1)^2;\quad \widetilde{s}_2=\frac{1}{n}\sum_{i=1}^{n}(\widetilde{x}_{2i}-\widetilde{m}_2)^2;$$

$$\widetilde{r}=\frac{\sum_{i=1}^{n}(\widetilde{x}_{1i}-\widetilde{m}_1)(\widetilde{x}_{2i}-\widetilde{m}_2)}{\sqrt{\sum_{i=1}^{n}\widetilde{x}_{1i}-\widetilde{m}_1)^2\sum_{i=1}^{n}\widetilde{x}_{2i}-\widetilde{m}_2)^2}}.$$

Величина $\widetilde{r}$, служащая точечной оценкой параметра ϱ двумерного нормального распределения, называется *выборочным коэффициентом линейной корреляции*, или *коэффициентом корреляции Пирсона*.

Удобной для вычисления значений r является формула

$$r=\frac{\sum_{i=1}^{n}x_{1i}x_{2i}-\frac{1}{n}\sum_{i=1}^{n}x_{1i}\sum_{i=1}^{n}x_{2i}}{\sqrt{\left[\sum_{i=1}^{n}x_{1i}^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_{1i}\right)^2\right]\left[\sum_{i=1}^{n}x_{2i}^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_{2i}\right)^2\right]}}$$

(см. § 2 и задачу IX-6).

Вернемся к примеру IX-3. Глазомерная оценка данных, представленных на рис. 35, убеждает, что зависимость между $\widetilde{x}_1$ и $\widetilde{x}_2$ можно считать

линейной, т. е. линеаризующее преобразование данных не требуется (ср. задачу IX-3). Тогда имеем $n = 12$; $\sum_{i=1}^{12} x_{1i} = 2$

$t,347$; $\sum_{i=1}^{12} x_{2i} = 144{,}57$; $\sum_{i=1}^{12} x_{1i}^2 = 583$

$t,403$; $\sum_{i=1}^{12} x_{2i}^2 = 2$

$t,204{,}185$; $\sum_{i=1}^{n} x_{1i}x_{2i} = 34$

$t,837{,}10$. Отсюда $\sum_{i=1}^{12} (x_{1i} - m_1)^2 = 124$

$t,368{,}917$; $\sum_{i=1}^{12} (x_{2i} - m_2)^2 = 462{,}478$; $\sum_{i=1}^{12} (x_{1i} - m_1)^2 (x_{2i} - m_2)^2 = 6$

$t,561{,}618$ и

$$r = \frac{6 \qquad t,561{,}618}{124t,368{,}917 \times 462{,}478} = 0{,}87.$$

Полученное значение $r = 0{,}87$ свидетельствует, по-видимому, о сильной положительной корреляции между массой жабр и массой тела у краба.

Следующий необходимый этап анализа — установление статистической значимости полученного результата, что сводится, по существу, к проверке гипотезы о независимости двух нормально распределенных случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$.

§ 7. Проверка гипотезы о независимости двух нормальных распределений

Распределение статистики $\widetilde{r}$ (r-распределение) не выражается в элементарных функциях, тем не менее оно достаточно подробно изучено. В общем случае оно резко отличается от нормального и сходится к нему чрезвычайно медленно (неразумно пользоваться нормальной аппроксимацией при $n < 500$!).

Можно показать, что случайная величина $\widetilde{r}/\sqrt{1 - \widetilde{r}^2}$ является комбинацией трех независимых случайных величин: $\widetilde{u}\ N(0;1)$; $\widetilde{\chi}_1^2 \sim \chi^2(\nu_1), \nu_1 =$

$= n-1$ и $\widetilde{\chi}_2^2 \sim \chi^2(\nu_2), \nu_2 = n-2, -$

$$\frac{\widetilde{r}}{\sqrt{1-\widetilde{r}^2}} = \frac{\widetilde{u}+\varrho\sqrt{\widetilde{\chi}_1^2}}{\sqrt{\widetilde{\chi}_2^2}}.$$

Поэтому в частном, но важном для практики случае, когда $\varrho = 0$, имеем

$$\frac{\widetilde{r}}{\sqrt{1-\widetilde{r}^2}} = \frac{\widetilde{u}}{\sqrt{\widetilde{\chi}^2}},$$

где $\widetilde{\chi}^2 \sim \chi^2(\nu)$, $\nu = n-2$. Помножив обе части на $\sqrt{\nu} = \sqrt{n-2}$, получаем

$$\widetilde{t} = \frac{\widetilde{r}\sqrt{n-2}}{\sqrt{1-\widetilde{r}^2}} = \frac{\widetilde{u}}{\sqrt{\widetilde{\chi}^{2/\nu}}} \sim t(\nu), \quad \nu = n-2.$$

В таком случае можно проверить гипотезу H_0: $\varrho = 0$, т. е. проверить, являются ли два нормальных распределения независимыми (см. § 6 гл. III). Для этого требуется найти значение

$$t_{\text{эксп}} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$$

и сравнить его с критическим значением $t_{\alpha/2}(\nu)$, $\nu = n-2$.

Указанное свойство позволяет вычислить критические значения

$$r_{\alpha/2}(\nu) = \frac{t_{\alpha/2}(\nu)}{\sqrt{t_{\alpha/2}(\nu)+n-2}}, \quad \nu = n-2,$$

такие, что $P\{\widetilde{r} \geqslant r_{\alpha/2}(\nu)\} = \alpha/2$. Эти значения приведены в табл. IX Приложения 1. Таким образом, для проверки гипотезы о независимости достаточно сравнить экспериментальное значение $|r_{\text{эксп}}|$ с табличным $r_{\alpha/2}(\nu)$, $\nu = n-2$.

Проверим значимость результата в примере IX-3. Имеем $r = 0{,}87$, $\nu = 12-2 = 10$; находим

$$t_{\text{эксп}} = \frac{0{,}87\sqrt{12-2}}{\sqrt{1-0{,}87^2}} = 5{,}58.$$

Согласно табл. III Приложения 1 $t_{0{,}0005}(10) = 4{,}59$; следовательно, $P\{|\widetilde{t}| \geqslant t_{\text{эксп}}\} < 0{,}001$, т. е. между массой жабр и массой тела у краба имеется высокозначимая положительная корреляция.

По табл. IX Приложения 1 $r_{0{,}0005}(10) = 0{,}82$; следовательно, $P\{|\widetilde{r}| \geqslant r_{\text{эксп}}\} < 0{,}001$, т. е. мы приходим к тому же выводу.

§ 8. Сравнение двух коэффициентов корреляции ϱ_1 и ϱ_2

Р. А. Фишер в 1921 г. указал замечательное нормализующее и стабилизирующее дисперсию преобразование случайной величины $\widetilde{r}$:

$$\widetilde{z} = \frac{1}{2}\ln\frac{1+\widetilde{r}}{1-\widetilde{r}},$$

которое не зависит ни от ϱ, ни от n. Если $n > 50$, то распределение случайной величины $\widetilde{z}$ близко к нормальному со средним и дисперсией:

$$E\widetilde{z} = \frac{1}{2}\ln\frac{1+\varrho}{1-\varrho} \quad \text{и} \quad D\widetilde{z} = \frac{1}{n-3}.$$

Таким образом, нормированная случайная величина

$$\widetilde{u} = \frac{\widetilde{z} - E\widetilde{z}}{\sqrt{D\widetilde{z}}} \sim N(0;1).$$

Для перевода значений r в z и обратно используют табл. XII Приложения 1.

С помощью r-преобразования можно проверять гипотезу H_0: $\varrho_1 = \varrho_2$ о равенстве коэффициентов корреляции двух двумерных нормальных распределений. Если эта гипотеза верна, то случайная величина

$$\widetilde{u} = \frac{\widetilde{z}_1 - \widetilde{z}_2}{\sqrt{\dfrac{1}{n_1-3} + \dfrac{1}{n_2-3}}} \sim N(0;1)$$

и вычисленное по выборке для проверки H_0 значение $|u_{\text{эксп}}|$ сравнивают с $u_{\alpha/2}$.

При $n < 50$ такой критерий становится консервативным, т. е. малочувствительным к отклонениям от нулевой гипотезы. В связи с этим Г. Хотеллинг, изучая возможности улучшения r-преобразования, в 1953 г. нашел, что распределение случайной величины

$$\widetilde{z}^* = \widetilde{z} - \frac{(3\widetilde{z} + \widetilde{r})}{4n}$$

более близко к нормальному, чем распределение $\widetilde{z}$, при этом

$$D\widetilde{z}^* = \frac{1}{n-1}.$$

Следовательно, если верна гипотеза H_0: $\varrho_1 = \varrho_2$, то

$$\widetilde{u} = \frac{\widetilde{z}_1^* - \widetilde{z}_2^*}{\sqrt{\dfrac{1}{n_1 - 1} + \dfrac{1}{n_2 - 1}}} \sim N(0; 1).$$

Таким образом, сравнивая $|u_{\text{эксп}}|$ и $u_{\alpha/2}$, можно проверять H_0. Критерий пригоден для малых выборок ($n_1 \geqslant n_2 \geqslant 10$).

Пример IX.4 (Г. И. Арнаутова, 1982 г.). У *Primula macro calyx* изучалась связь между числом семян на коробочку и массой семени. Для длинностолбчатых растений (Д-форма) $r_1 = -0{,}41$ при $n_1 = 26$; для короткостолбчатых (К-форма) $r_2 = -0{,}51$ при $n_2 = 20$. Существенны ли различия в силе связи между этими признаками у Д- и К-форм?

По табл. XII*а* Приложения 1 находим $z_1 = 0{,}44$ и $z_2 = 0{,}56$;

$$|u_{\text{эксп}}| = \frac{|0{,}44 - 0{,}56|}{\sqrt{\dfrac{1}{23} + \dfrac{1}{17}}} = 0{,}38,$$

следовательно, $P\{|\widetilde{u} \geqslant u_{\text{эксп}}|\} \gg 0{,}10$. (Преобразование Г. Хотеллинга в данном случае не вносит существенных изменений.) Таким образом, нет оснований считать, что сила связи между числом семян на коробочку и массой семени различна у Д- и К-форм.

§ 9. Ранговая корреляция

Пусть $(x_1, y_1), \ldots, (x_n, y_n)$ — выборка объема n из двумерной генеральной совокупности с непрерывной функцией распределения $F(x, y)$. $G(x)$ и $H(y)$ — маргинальные функции распределения x и y соответственно. Требуется проверить гипотезу H_0: $F(x, y) = G(x) \cdot H(y)$ о независимости случайных величин x и y.

Если функция распределения $F(x, y)$ не имеет двумерного нормального распределения, то для проверки гипотезы H_0 необходимо предпочесть непараметрический критерий. Непараметрический критерий следует предпочесть и тогда, когда в качестве альтернативы ожидается нелинейная корреляционная зависимость. Непараметрическую статистику для оценки связи можно получить, если в статистике коэффициента корреляции Пирсона $\widetilde{r}$

заменить случайные величины $\widetilde{x}_i$ и $\widetilde{y}_i$ их рангами $R(\widetilde{x}_i)$ и $Q(\widetilde{y}_i)$, которые будем далее обозначать просто $\widetilde{R}_i$ и $\widetilde{Q}_i$; если

$$\widetilde{r} = \frac{\sum\limits_{i=1}^{n}(\widetilde{x}_i - \widetilde{m}_x)(\widetilde{y}_i - \widetilde{m}_y)}{\sqrt{\sum\limits_{i=1}^{n}(\widetilde{x}_i - \widetilde{m}_x)^2 \sum\limits_{i=1}^{n}(\widetilde{y}_i - \widetilde{m}_y)^2}}, \text{ то}$$

$$\widetilde{r}_S = \frac{\sum\limits_{i=1}^{n}(\widetilde{R}_i - \overline{R})(\widetilde{Q}_i - \overline{Q})}{\sqrt{\sum\limits_{i=1}^{n}(\widetilde{R}_i - \overline{R})^2 \sum\limits_{i=1}^{n}(\widetilde{Q}_i - \overline{Q})^2}}.$$

Символами $\overline{R}$ и $\overline{Q}$ обозначены средние ранги, которые равны $\overline{R} = \overline{Q} = \dfrac{1}{2(n+1)}$ (см. задачу II-4). Напомним также, что

$$\sum_{i=1}^{n}(\widetilde{R}_i - \overline{R})^2 = \sum_{i=1}^{n}(\widetilde{Q}_i - \overline{Q})^2 = \frac{1}{12}n(n^2-1) \quad \text{(см. задачу II-5).}$$

Подставив эти значения в выражение для $\widetilde{r}_S$, получаем

$$\widetilde{r}_S = \frac{12}{n(n^2-1)}\sum_{i=1}^{n}\left(\widetilde{R}_i - \frac{n+1}{2}\right)\left(\widetilde{Q}_i - \frac{n+1}{2}\right),$$

что легко преобразуется в выражение, более удобное для вычислений:

$$\widetilde{r}_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(\widetilde{R}_i - \widetilde{Q}_i)^2 = 1 - \frac{6\widetilde{D}}{n(n^2-1)},$$

где $\widetilde{D} = \sum\limits_{i=1}^{n}(\widetilde{R}_i - \widetilde{Q}_i)^2$. Эта статистика носит название статистики *коэффициента ранговой корреляции Спирмена*. Впервые она была предложена американским психологом Ч. Э. Спирменом в 1904 г. Величина $\widetilde{r}_S$ (как и $\widetilde{r}$)

может принимать значения от -1 до $+1$ и является мерой коррелированности рангов $\widetilde{R}_i$ и $\widetilde{Q}_i$.

Подобно многим другим непараметрическим статистикам, $\widetilde{r}_S$ основана на комбинаторике всех возможных упорядочений рангов $\widetilde{R}_i$ относительно $\widetilde{Q}_i$. Распределение ее можно вывести, используя следующую урновую схему. В двух урнах находятся по n шаров, пронумерованных от 1 до n. Шары извлекаются наугад парами, по одному из каждой урны. Пусть $\widetilde{R}_i$ — номер i-го шара из первой урны, а $\widetilde{Q}_i$ — номер i-го шара из второй урны. $\widetilde{R}_i$ может принимать значения $R_i = 1, \ldots, n$ и $\widetilde{Q}_i$ также может принимать значения $Q_i = 1, \ldots, n$. Нас интересует распределение случайной величины $\widetilde{D} = \sum_{i=1}^{n} (\widetilde{R}_i - \widetilde{Q}_i)^2$, которая есть сумма квадратов разностей для всех n пар номеров. Чтобы найти это распределение, расположим пары случайных величин $(\widetilde{R}_i, \widetilde{Q}_i)$ в порядке возрастания значений R_i:

$$\begin{array}{l} R_i\colon\ 1,\quad 2,\quad \ldots,\ n-1,\quad n; \\ Q_i\colon\ Q_1,\ Q_2,\ \ldots,\ Q_{n-1},\ Q_n. \end{array}$$

Очевидно, что если Q_1 приняло любое из n возможных значений, то Q_2 может принять любое из оставшихся $(n-1)$ значений и т. д., Q_{n-1} может принять любое из двух оставшихся значений, наконец, Q_n принимает единственное, оставшееся последним значение. Таким образом, всего возможно $n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1 = n!$ равновероятных наборов (перестановок) значений Q_i. Поскольку значения, которые может принимать случайная величина $\widetilde{R}_i$, не зависят от значений случайной величины $\widetilde{Q}_i$, то столько же возможно наборов значений и для пар $(\widetilde{R}_i, \widetilde{Q}_i)$. Следовательно, вероятность каждого такого набора значений равна $1/n!$, если $\widetilde{R}_i$ и $\widetilde{Q}_i$ являются независимыми. Чтобы найти распределение случайной величины $\widetilde{D}$, нужно подсчитать, сколько раз (a_i) сумма квадратов разностей $\sum_{i=1}^{n} (\widetilde{R}_i - \widetilde{Q}_i)^2$ принимает значение $\widetilde{D}$, и тогда вероятности $P\{\widetilde{D} = D\} = a_j/n!$ задают распределение случайной величины $\widetilde{D}$. Очевидно, что распределение $\widetilde{r}_S$ с точностью до константы совпадает с распределением $\widetilde{D}$. Распределение $\widetilde{r}_S$ затабулировано (табл. XIII Приложения 1).

Найдем математическое ожидание $\widetilde{r}_S$. Для этого сначала найдем математическое ожидание $\widetilde{D}$, которое можно представить

$$\widetilde{D} = \sum_{i=1}^{n} \widetilde{R}_i^2 - 2\sum_{i=1}^{n} \widetilde{R}_i \widetilde{Q}_i + \sum_{i=1}^{n} \widetilde{Q}_i^2.$$

Так как $(\widetilde{R}_1, \ldots, \widetilde{R}_n)$ и $(\widetilde{Q}_1, \ldots, \widetilde{Q}_n)$ суть перестановки целых чисел от 1 до n, то согласно результату в задаче II-5

$$\sum_{i=1}^{n} \widetilde{R}_i^2 = \sum_{i=1}^{n} \widetilde{Q}_i^2 = \frac{1}{6} n(n+1)(2n+1).$$

Следовательно,

$$\widetilde{D} = \frac{1}{3} n(n+1)(2n+1) - 2\sum_{i=1}^{n} \widetilde{R}_i \widetilde{Q}_i.$$

Тогда математическое ожидание этой величины есть

$$E\widetilde{D} = \frac{1}{3} n(n+1)(2n+1) - 2E\left(\sum_{i=1}^{n} \widetilde{R}_i \widetilde{Q}_i\right),$$

где, очевидно,

$$E\left(\sum_{i=1}^{n} \widetilde{R}_i \widetilde{Q}_i\right) = nE\widetilde{R}_i \cdot E\widetilde{Q}_i = \frac{1}{4} n(n+1)^2,$$

так что

$$E\widetilde{D} = \frac{1}{3} n(n+1)(2n+1) - \frac{1}{2} n(n+1)^2 = \frac{1}{6} n(n^2-1).$$

В итоге, возвращаясь к величине $\widetilde{r}_S$, получаем

$$E\widetilde{r}_S = E\left[1 - \frac{6\widetilde{D}}{n(n^2-1)}\right] = 0.$$

Таким образом, статистика $\widetilde{r}_S$ имеет все свойства коэффициента корреляции и проверка гипотезы H_0 о независимости случайных величин сводится

к проверке гипотезы H_0: $\varrho_S = 0$ о равенстве нулю такого параметра ϱ_S двумерной генеральной совокупности, оценкой которого служит статистика $\widetilde{r}_S$.

Далее можно показать, что дисперсия статистики $\widetilde{r}_S$ равна

$$d\widetilde{r}_S = \frac{1}{n-1}$$

и нормированная случайная величина $\widetilde{u}$ при $n \to \infty$

$$\widetilde{u} = \frac{\widetilde{r}_S - E\widetilde{r}_S}{\sqrt{D\widetilde{r}_S}} = \frac{\widetilde{r}_S - 0}{\frac{1}{\sqrt{n-1}}} = \widetilde{r}_S\sqrt{n-1} \sim N(0;1).$$

Практически эта аппроксимация удовлетворительна при $n > 30$. При $n \geqslant 15$ распределение $\widetilde{r}_S$ хорошо аппроксимируется t-распределением с числом степеней свободы $\nu = n-2$:

$$\widetilde{t} = \widetilde{r}_S\sqrt{\frac{n-2}{1-\widetilde{r}_S^2}} \sim t(\nu), \quad \nu = n-2.$$

Если среди значений x_i и (или) y_i имеются совпадающие значения, то им приписывают усредненные ранги и статистику $\widetilde{u}$ вычисляют с поправками $\widetilde{T}_x$ и $\widetilde{T}_y$ на совпадения:

$$\widetilde{u}^* = \frac{\left[\widetilde{r}_S - \frac{1}{2}(\widetilde{T}_x + \widetilde{T}_y)\right]\sqrt{n-1}}{(1-\widetilde{T}_x)(1-\widetilde{T}_y)},$$

где

$$\widetilde{T}_x = \frac{1}{n(n^2-1)}\sum_{j=1}^{k}(\widetilde{c}_j^2 - \widetilde{c}_j) \quad \text{и} \quad \widetilde{T}_y = \frac{1}{n(n^2-1)}\sum_{g=1}^{l}(\widetilde{b}_g^2 - \widetilde{b}_g).$$

Здесь, в свою очередь, k — число групп совпадающих значений x_i; c_j — число совпадающих значений x_i в j-й группе; l — число совпадающих значений y_i; b_g — число совпадающих значений y_i в g-й группе.

Пример IX-5 (А. П. Виноградов, 1961 г., А. Ленинджер, 1974 г.).
В табл. 70 приведено содержание наиболее распространенных химических элементов (ат. %) в литосфере (x_i) и в организме человека (y_i). Связаны ли эти показатели между собой?

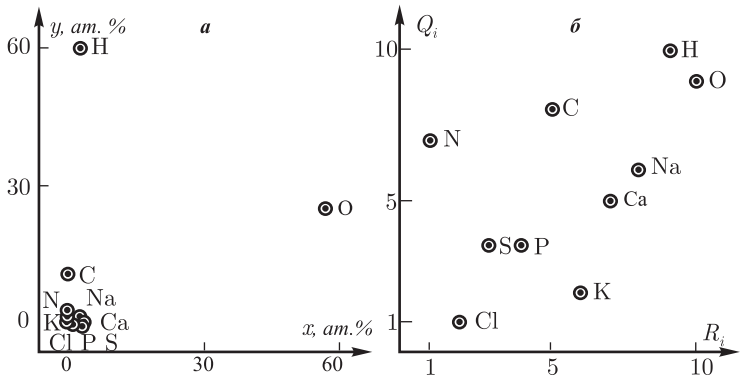


Рис. 53. Содержание химических элементов в литосфере и в организме человека: *а* — исходные данные; *б* — ранжированные данные (пример IX-5)

Прежде чем приступать к вычислениям, полезно представить данные графически (рис. 53, *а*). Тогда становится очевидным, что статистическую связь между $\widetilde{x}_i$ и $\widetilde{y}_i$ (если она имеется) вряд ли можно назвать линейной, и потому для проверки гипотезы об их независимости применение коэффициента ранговой корреляции Спирмена представляется более оправданным, нежели коэффициента корреляции Пирсона. Действительно, связь между изучаемыми величинами становится гораздо наглядней, если предварительно ранжировать выборочные значения (рис. 53, *б*).

Значения рангов R_i и Q_i показаны в табл. 70. Двум совпавшим значениям $y_{(3)} = y(4) = 0{,}13$ приписываем усредненный ранг $Q_3 = Q_4 = (3 + 4)/2 = 3{,}5$. Очевидно, что сумма разностей $\sum_{i=1}^{n}(R_i - Q_i)$ должна быть равна нулю.

Находим сумму $D = \sum_{i=1}^{1} 0(R_i - Q_i)^2 = 72{,}5$ и вычисляем

$$r_S = 1 - \frac{6 \cdot 72{,}5}{10(10^2 - 1)} = 0{,}56.$$

Чтобы оценить достоверность наблюдаемого значения рангового коэффициента корреляции, необходимо знать распределение статистики $\widetilde{r}_S$ в случае, когда $\varrho_s = 0$. Для $n = 10$ это распределение показано на рис. 54. Оно получено перебором $n! = 101 = 3\,628\,800$ всех возможных перестановок для пар чисел $R_i = 1, \ldots, 10$ и $Q_i = 1, \ldots, 10$ и подсчетом числа встречаемости каждого из возможных значений r_S. Заштрихованные области

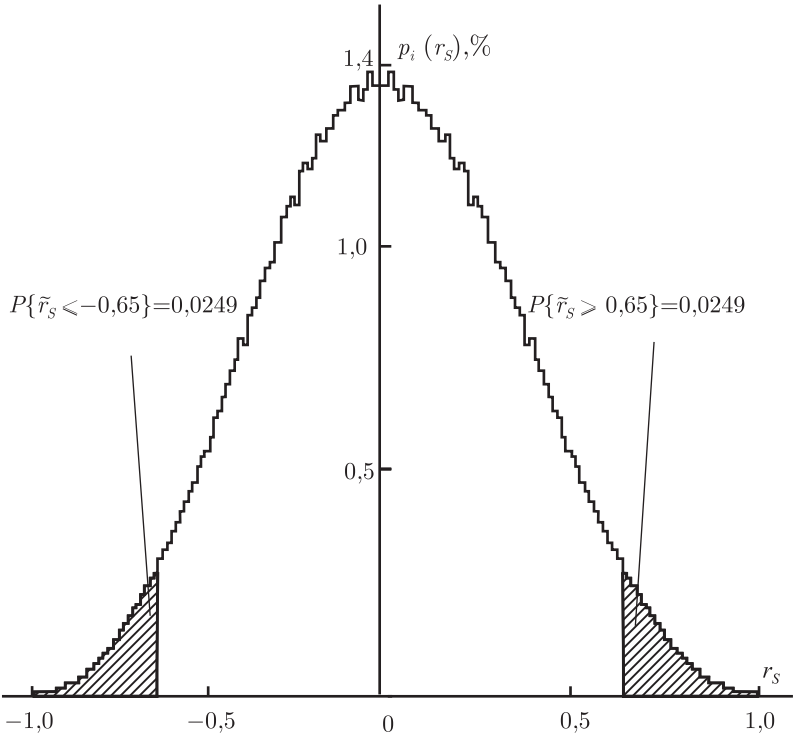


Рис. 54. Распределение статистики $\widetilde{r}_S$ коэффициента корреляции Спирмена при $n = 10$. Сумма заштрихованных площадей есть $\alpha = 0{,}0498 \approx 0{,}05$

на хвостах распределения соответствуют вероятностям $P\{r_S \leqslant -0{,}65\} = 0{,}0249$ и $P\{\widetilde{r}_S \geqslant 0{,}65\} = 0{,}0249$. Найденное нами значение не попадает ни в одну из этих областей, поэтому нет оснований отклонять нулевую гипотезу на уровне значимости $\alpha = 0{,}05$.

Поскольку распределение $\widetilde{r}_S$ симметрично относительно нуля, то в таблицах приводят лишь положительные значения $r_{S(\alpha)}$, т. е. такие, для которых $P\{|\widetilde{r}_S| \geqslant r_{S(\alpha)}(n)\} = \alpha$. В нашем случае $|r_S| = r_{S(0{,}05)}(10) = 0{,}56$, и поэтому наблюдаемую корреляцию можно признать значимой лишь на уровне $\alpha = 0{,}10$.

Таблица 70

Содержание химических элементов (ат. %) в литосфере (x_i) и организме человека (y_i) (к примеру IX-5)

Элемент	x_i	y_i	R_i	Q_i	$R_i - Q_i$	$(R_i - Q_i)^2$
H	3,0	60,3	9	10	-1	1
O	58,0	25,5	10	9	1	1
C	0,15	10,5	5	8	-3	9
N	0,025	2,42	1	7	-6	36
Na	2,4	0,73	8	6	2	4
Ca	2,0	0,23	7	5	2	4
P	0,05	0,13	4	3,5	0,5	0,25
S	0,03	0,13	3	3,5	$-0,5$	0,25
K	1,4	0,036	6	0	4	16
Cl	0,026	0,032	2	1	1	1

§ 10. Критерий χ^2 как критерий независимости

ПРИМЕР IX-6 [Фишер, 1958]. Шотландские дети, отдельно мальчики и девочки, были расклассифицированы по признаку цвета волос (табл. 71).

Таблица 71

Цвет волос у шотландских детей (к примеру IX-6)

Пол	Цвет волос					
	белокурый	рыжий	русый	каштановый	черный	всего
Мальчики	592	119	849	504	36	2 100
Девочки	644	97	677	451	14	1 783
Всего	1 136	216	1 526	955	50	3 883

В этой задаче речь идет о классификации каждого индивидуума по двум признакам: по полу и по цвету волос. К данным, представленным в таблице сопряженности признаков, необходимо применить критерий независимости, т. е. ответить на вопрос, есть ли связь между признаками «пол» и «цвет волос». Однако задачу можно сформулировать и по-другому: есть две выборки (мальчики и девочки) и сравниваются два выборочных распре-

деления признака «цвет волос», т. е. мы вернулись к критерию однородности из § 3 гл. VII.

Действительно, математически критерий однородности и критерий независимости идентичны. Различие заключается в биологической постановке вопроса. В следующем примере, с точки зрения биолога, есть смысл говорить лишь о критерии независимости.

Пример IX-7 [Рокицкий, 1973]. Оценивалась конституция каракульских овец при рождении и в полуторагодовалом возрасте (табл. 72). Есть ли зависимость между конституцией ягнят в разном возрасте?

Таблица 72

Конституция каракульских овец в разном возрасте (к примеру IX-7)

При рождении	В полуторагодовалом возрасте			Всего
	нежная	крепкая	грубая	
Нежная	16	52	21	89
Крепкая	15	485	325	826
Грубая	28	190	374	592
Всего	59	727	721	1507

§ 11. Об интерпретации статистических зависимостей

Пример IX-8 (Н. А. Агаджанян, А. Ю. Катков, 1981 г.). Западногерманский врач-онколог Э. ван Аакен наблюдал за 500 пожилыми бегунами и 500 небегающими людьми той же возрастной группы в течение шести лет. За это время 29 небегающих заболели раком, а среди бегунов заболело раком только четверо. Результаты наблюдений можно представить в виде таблицы сопряженности 2×2 (табл. 73).

Имеем $\chi^2_{\text{эксп}} = 18,05$; $P\{\widetilde{\chi}^2 \geqslant \chi^2_{\text{эксп}}\} < 0,001$. Налицо явная отрицательная корреляция: занятия бегом сопровождаются низкой заболеваемостью раком.

Можно ли на основании этих данных рекомендовать пожилым людям бег в качестве эффективного средства борьбы с раком? Очевидно, нет.

Здесь мы имеем наглядный пример того, как наличие статистической связи не выявляет и не указывает направления причинно-следственной связи. Действительно, наравне с выводом о том, что бег является причиной

Таблица 73

Таблица сопряженности 2×2 для данных примера IX-8

Пациенты	Заболевшие	Здоровые	Всего
Бегуны	4	496	500
Небегающие	29	471	500
Всего	33	967	1 000

низкой заболеваемости раком, правомерен и обратный вывод: бегом предпочитают заниматься здоровые люди.

Для того чтобы сделать выбор между этими альтернативами, нужны внестатистические соображения. Например, следовало бы изменить схему эксперимента: часть пожилых бегунов должна была бы прекратить занятия бегом, а часть небегающих должна была бы начать регулярные занятия бегом. Ясно, что такая насильственная ситуация неприемлема, и потому столь неоднозначными бывают выводы из подобных исследований.

«Статистическая зависимость, как бы ни была она сильна, никогда не может установить причинной связи: наши идеи о причине должны приходить извне статистики, в конечном счете из другой теории [. . .] Нам нет нужды углубляться в философское обсуждение этого вопроса; для наших целей необходимо только еще раз подчеркнуть, что статистическая зависимость любого сорта логически не влечет причинной [. . .].

Последователи Карла Пирсона и Юла в первом приступе энтузиазма, порожденного корреляционной техникой, легко делали опрометчивые выводы. Это продолжалось до тех пор, пока [. . .] Юл (1926) не напугал статистиков примерами высоких корреляций, которые, очевидно, не выражали причинных связей [. . .] Большинство этих «бессмысленных» корреляций действует через сопутствующие изменения во времени. Упомянутые примеры имели благотворный эффект, доводя до сознания статистиков, что причинная зависимость не может быть выведена ни из какого наблюдаемого совместного изменения, даже самого тесного» [Кендалл, Стьюарт, 1973, с. 374–375].

Задачи

IX-1 (В. Уэлдон, 1906 г.). Из 12 игральных костей половина окрашена в красный цвет, половина — в белый. Испытание I заключается в одновременном подбрасывании 12 костей и записи суммы выпавших на них очков.

Испытание II состоит в том, что кости, окрашенные в красный цвет, оставляют на столе в том положении, как они выпали в испытании I; кости, окрашенные в белый цвет, снова подбрасывают; записывают сумму очков, выпавших на 12 костях. Будут ли испытания I и II независимыми? Как бы вы смоделировали в такого рода эксперименте варьирование силы связи между испытаниями I и II?

IX-2. Измеряются два признака листа: 1) длина пластинки и 2) длина листа, т. е. сумма длины листа и длины черешка. Будут ли эти признаки независимыми или скоррелированными?

Таблица 74

Результаты измерения скорости крови обычным (x_i) и новым (y_i) методами (к задаче IX-7)

x_i	y_i	x_i	y_i
1 190	1 115	2 335	2 280
1 455	1 425	2 490	2 520
1 550	1 515	2 720	2 630
1 730	1 795	2 710	2 740
1 745	1 715	2 530	2 390
1 770	1 710	2 900	2 800
1 900	1 830	2 760	2 630
1 920	1 920	3 010	2 970
1 960	1 970		
2 295	2 300		

IX-3. [Терентьев, Ростова, 1977]. Оцените связь между массой тела и долей, какую масса мозга составляет от общей массы у обыкновенного тюленя:

Масса тела,	кг	7,5	12,5	17,5	22,5	37,5	27,5	32,5
Масса мозга,	%	4,10	2,24	1,12	0,85	0,55	0,68	0,55

Требуется ли здесь линеаризирующее преобразование?

IX-4. Проведите статистический анализ данных из примера IV-18 (см. табл. 14).

IX-5. Проведите статистический анализ данных из примера IV-19 (см. табл. 15).

IX-6. Покажите, что

$$\sum_{i=1}^{n}(x_i - m_x)(\widetilde{y}_i - \widetilde{m}_y) = \sum_{i=1}^{n} x_i\widetilde{y}_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} \widetilde{y}_i.$$

IX-7 [Браунли, 1977]. Скорость крови определяли двумя методами: обычным, путем непосредственного измерения, и новым, технически более простым (табл. 74). Если новый метод дает те же результаты, то коэффициент регрессии должен быть равен единице. Проверьте эту гипотезу. Можно ли для решения возникшей у физиолога задачи привлечь другую статистическую модель помимо линейной регрессии?

IX-8 [Урбах, 1975]. При облучении гамма-лучами наблюдается падение активности фермента (в % к контролю):

Доза, кР (D)	0	3	7,5	15	30	45	60
Активность фермента (A)	100	88,5	77,0	39,9	21,8	10,7	4,43

Предполагается, что активность фермента убывает по показательному закону $A = A_0 e^{-\gamma D}$. Оцените коэффициент γ.

IX-9 [Урбах, 1975]. Изучалось уменьшение темпов размножения двух штаммов бактерий при рентгеновском облучении (в % к контролю):

Доза, кР	12	3	4	5	6	7	
1-й штамм	96	87	83	77	71	63	02
2-й штамм	93	88	85	73	73	67	—

Сравните кривые доза-эффект.

IX-10. Давалась оценка хлебопекарного качества муки простого помола Q после прогревания ее при 170 °F в течение различных периодов времени (часы, T):

T	0,25	0,50	0,75	1,0	1,5	2,0	3,0	4,0	6,0	8,0
Q	93	71	63	54	43	38	29	26	22	22

Постройте график. Попробуйте использовать преобразования $\lg T$, $\lg Q$. Проведите регрессионный анализ.

IX-11 [Терентьев, Ростова, 1977]. Следует ли проводить оценку коэффициента корреляции r между числом боковых побегов и высотой растения у нивяника обыкновенного (табл. 75)?

Таблица 75

Распределение числа растений нивяника обыкновенного по количеству боковых побегов и по высоте растений (к задаче IX-11)

Высота	Количество боковых побегов							
	0	1	2	3	4	5	6	7
35	13	1						
45	62	6	3	1	1			
55	51	6	8	4	1			
65	17	0	8	5	3	1		
75	1	2	1	1	1	1	1	1

Таблица 76

Полярность (x_i) и гидрофобность (y_i) аминокислот (усл. ед.) (к задаче IX-14)

Аминокислота	x_i	y_i	Аминокислота	x_i	y_i
Алании	0,00	0,87	Лейцин	0,13	2,17
Аргинин	52,00	0,85	Лизин	49,50	1,64
Аспарагин	3,38	0,09	Метионин	1,43	1,67
Аспарагиновая кислота	49,70	0,66	Пролин	1,58	2,77
Валин	0,13	1,87	Серим	1,67	0,07
Гистидин	51,60	0,87	Тирозин	1,61	2,67
Глицин	0,00	0,10	Треонин	1,66	0,07
Глутамин	3,53	0,00	Триптофан	2,10	3,77
Глутаминовая кислота	49,90	0,67	Фенилаланин	0,35	2,87
Изолейцин	0,13	3,15	Цистеин	1,48	1,52

IX-12. Проведите статистический анализ данных из примера IX-6 (см. табл. 71).

IX-13. Проведите статистический анализ данных из примера IX-7 (см. табл. 72).

IX-14. Структурными элементами белков обычно служат 20 канонических аминокислот, важными характеристиками которых являются полярность (x_i) и гидрофобность (y_i) (табл. 76). Выясните, скоррелированы ли эти свойства. Какой критерий следует применить и почему?

IX-15 [Тьюки, 1981]. Проведите корреляционный анализ данных, представленных в табл. 77.

Таблица 77

Распределение ранних цветков *Ranunculus ficaria* по числу тычинок (x_i) и числу пестиков (y_2) (к задаче IX-15)

x_2 \ x_1	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	$\sum_{j=1}^{l} r_{2j}$
6																				1		1
7												1										1
8																						0
9												1										1
10		1						1														2
11				2				1														3
12			3	1	2	1	1	3	1	1												13
13		1	1	1	4	1	1	1			1	1										12
14		4	3		1	2	4	1	2	3	1	1										22
15			1	2	4	3	7	4	4	5	2	1		1	1							35
16	1			2			1	5	3	5	4	5	3	2								31
17					2	2	1	4	2	3	1	2	5	2	1							25
18				1		2	4	3	1	7	1	3	2	1		2						27
19						1	2		5	4	4	1	1	1	1	1						21
20					2		1	1	2	2		1	3	4	2	1						19
21					1				2	2		1	2	4	1	1						13
22								1	2	3		2	2	3	1	2						15
23									1	1		2	1	1	2			2				10
24									1	1		2										4
25														1	2						1	4
26													1		2							3
27								1		1								2				4
28																	1					1
29																						0
30																						0
31												1										1
$\sum_{i=1}^{k} n_{1i}$	1	6	8	9	16	12	22	26	26	38	14	23	20	20	13	7	1	4	0	1	1	268

Таблица 78

Распределение объема груди и роста (центральные значения интервалов в дюймах) для 4 995 женщин Великобритании (1951 г.) (к задаче IX-21).

Объем груди (x_1)	Рост (x_2)											Всего
	54	56	58	60	62	64	66	68	70	72	74	
56					1							1
54					1	2						3
52			1		3	4	1		1			10
50			1	3	5	4	1					14
48		1	3	9	7	6	3	1				30
46			4	11	17	17	7		1			57
44		2	11	26	50	45	17	10	1			162
42		2	11	42	85	73	31	12	3	2		261
40		2	20	76	132	131	71	31	9	4	3	479
38		2	36	98	158	203	126	65	17	3	1	709
36		6	48	188	317	410	263	89	15	1		1 337
34	1	9	67	210	376	427	196	59	8			1 353
32	3	5	39	131	163	122	31	8	1	1		504
30	1	4	11	18	25	10	2					71
28			2	1			1					4
Всего	5	33	254	813	1 340	1 454	750	275	56	11	4	4 995

IX-16. Докажите, что при $n = 2$ и $\varrho = 0$ $P\{\widetilde{r} = 1\} = P\{\widetilde{r} = -1\} = 1/2$.

IX-17. Используя тот факт, что $(\widetilde{b} - \beta)/\widetilde{s}_b \sim t(\nu)$, постройте доверительный интервал для коэффициента линейной регрессии β.

IX-18. Постройте доверительный интервал для коэффициента корреляции ϱ. Какое свойство:

$$\frac{\widetilde{r}\sqrt{n-2}}{\sqrt{1-\widetilde{r}^2}} \sim t(\nu), \quad \nu = n-2, \quad \text{или} \quad \frac{\widetilde{z} - E\widetilde{z}}{\sqrt{D\widetilde{z}}} \sim N(0;1)$$

— следует для этого использовать?

IX-19. Предложите критерий для сравнения нескольких (k) независимых коэффициентов корреляции.

IX-20. Проведите корреляционный анализ для парных наблюдений из примеров VI-4, VI-9 и задач VI-8, VI-9. Почему в этих случаях применяются парный t-критерий или парный критерий Вилкоксона?

IX-21 [Кендалл, Стьюарт, 1973]. Проведите статистический анализ данных, представленных в табл. 78.

IX-22. Проинтерпретируйте следующее высказывание Бернарда Шоу (1906): «Даже опытные статистики часто оказываются не в состоянии оценить, до какой степени смысл статистических данных искажается молчаливыми предположениями их интерпретаторов [...] Легко доказать, что ношение цилиндров и зонтиков расширяет грудную клетку, удлиняет жизнь и дает относительный иммунитет от болезней [...] Университетский диплом, ежедневная ванна, обладание тридцатью парами брюк, знание музыки Вагнера, скамья в церкви — короче всё, что подразумевает большие средства и хорошее воспитание, [...] может быть с помощью статистики представлено как магические чары, дарующие привилегии любого сорта».

Решения задач

I-1.

а) Урна содержит поровну белые и черные шары, символизирующие гаметы, несущие аллель A или аллель a соответственно. Шары извлекаются парами, каждая пара символизирует генотип зиготы.

б) Урна содержит поровну белые и черные шары, символизирующие пол новорожденного ребенка. Шары извлекаются по одному.

I-2.

а) мишень — отрезок $[0, \infty]$;

б) мишень — отрезок $[0, d]$;

в) мишень — отрезок $[0, 100]$.

I-3.

а) $A_1 \cap A_2 = \oslash$ означает, что отрезки $(a_1, b_1]$ и $(a_2, b_2]$ не пересекаются;

б) $A_1 \cap A_2 = A_1$ означает, что отрезок $(a_1, b_1]$ содержится в $(a_2, b_2]$, т. е. $a_2 \leqslant a_1 < b_1 \leqslant b_2$; $A_1 \cup A_2 = A_2$ означает то же самое; $A_1 \cap A_2 = \oslash$ означает, что либо $b_1 \leqslant a_2$, либо $b_2 \leqslant a_1$.

I-6. $P(\text{БЧК}) = 1/4$. Белая окраска появляется в двух случаях: когда выпадает белая грань и когда выпадает грань БЧК; следовательно, $P(\text{Б}) = 1/4 + 1/4 = 1/2$. Аналогично $P(\text{Ч}) = P(\text{К}) = 1/2$. Далее, два цвета одновременно появляются лишь в случае выпадения грани БЧК. Поэтому $P(\text{БЧ}) = P(\text{БК}) = P(\text{ЧК}) = 1/4$. Отсюда и следует, например, что $P(\text{БЧ}) = P(\text{Б}) \cdot P(\text{Ч})$, но $P(\text{БЧК}) > P(\text{Б}) \cdot P(\text{Ч}) \cdot P(\text{К})$.

I-8.

а) Вероятность не получить 6 при бросании одной кости равна 5/6, а при бросании четырех костей — $(5/6)^4$. Следовательно, искомая вероятность равна $6 - (5/6)^4 \approx 0{,}518$.

б) Вероятность того, что на обеих костях выпадет 6, равна 1/36; вероятность того, что этого не произойдет — 35/36. Вероятность того, что при двадцати четырех бросаниях не выпадет две 1, равна $(35/36)^4$. Следовательно, искомая вероятность равна $1 - (35/36)^{24} \approx 0{,}491$.

I-9. Обозначим медали первого ящика $З_1$, $З_2$, второго — $С_3$, $С_4$, третьего — $З_5$, $С_6$. Если первой извлечена золотая медаль, то ею равновероятно могут быть $З_1$, $З_2$ или $З_5$. Только один из этих трех равновероятных исходов — $З_5$, влечет событие: «вторая медаль — серебряная». Следовательно, искомая вероятность равна 1/3. Тот же результат имеем, если во вскрытом отделении обнаружена серебряная медаль.

I-10.

а) Вероятность того, что данный человек родится в данном месяце, равна 1/12. Вероятность того, что 12 человек родятся в заданные месяцы, равна $(1/12)^{12} = 12^{-12}$. Так как месяц рождения для каждого человека не оговорен, то число возможных размещений 12 человек по 12 месяцам равно 12. Следовательно, искомая вероятность равна $12! \cdot 12^{-12} = 11! \cdot 12^{-11} \approx 0{,}000054$.

б) Вероятность рождения одного человека в заданные 2 месяца равна 1/6, а шести человек, соответственно, $(1/6)^6 = 6^{-6}$. Число возможных размещений шести человек по двум месяцам равно 2^6. Число возможных сочетаний по 2 месяца из 12 в году равно $C_{12}^2 = \frac{12 \cdot 11}{2} = 66$. Искомая вероятность равна $6^6 \cdot 2^6 \cdot 66 = 3^{-6} \cdot 66 \approx 0{,}091$.

Эта задача иллюстрирует замечание французского математика Эмиля Бореля: «[. . .] как только задачи на вероятности становятся сколько-нибудь сложными, здравый смысл, даже руководимый светлым и глубоким умом, не может обойтись без помощи вычислений; он ведет самое большее к выводам [. . .] неполным и расплывчатым [. . .]».

II-3. Обозначим для краткости $E\widetilde{x} = m$. Тогда

$$D\widetilde{x} = E(\widetilde{x} - m)^2 = \sum_{i=0}^{\infty}(i - \boldsymbol{m})^2 p_i = \sum_{i=0}^{\infty}(i^2 - 2im + m^2)p_i =$$
$$= \sum_{i=0}^{\infty} i^2 p_i - 2m \sum_{i=0}^{\infty} i p_i + m_2 \sum_{i=0}^{\infty} p_i =$$
$$= E\widetilde{x}^2 - 2m \cdot m + m^2 = E\widetilde{x}^2 - (E\widetilde{x})^2.$$

II-6. Пусть $\{p_{ij}\}$ — совместное распределение случайных величин $\widetilde{x}_1$ и $\widetilde{x}_2$. Тогда

$$E(\widetilde{x}_1+\widetilde{x}_2)=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(i+j)p_{ij}=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}ip_{ij}+$$
$$+\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}jp_{ij}=\sum_{i=0}^{\infty}ip_{i\cdot}\sum_{j=0}^{\infty}jp_{\cdot j}=E\widetilde{x}_1+E\widetilde{x}_2.$$

II-7. Так как

$$D(\widetilde{x}_1+\widetilde{x}_2)=E(\widetilde{x}_1+\widetilde{x}_2)^2-[E(\widetilde{x}_1+\widetilde{x}_2)]^2=E(\widetilde{x}_1^2+2\widetilde{x}_1\widetilde{x}_2+\widetilde{x}_2^2)-$$
$$-(E\widetilde{x}_1+E\widetilde{x}_2)^2=E\widetilde{x}_1^2+2E(\widetilde{x}_1\widetilde{x}_2)+E\widetilde{x}_2^2-(E\widetilde{x}_1)^2+2E\widetilde{x}_1E\widetilde{x}_2+$$
$$+(E\widetilde{x}_2)^2=\left[E\widetilde{x}_1^2-(E\widetilde{x}_1)^2\right]+\left[E\widetilde{x}_2^2-(E\widetilde{x}_2)^2\right]+2[E(\widetilde{x}_1\widetilde{x}_2)-E\widetilde{x}_1E\widetilde{x}_2]=$$
$$=D\widetilde{x}_1+D\widetilde{x}_2+2[E(\widetilde{x}_1\widetilde{x}_2)-E\widetilde{x}_1E\widetilde{x}_2],$$

то следует убедиться в том, что для независимых случайных величин $E(\widetilde{x}_1,\widetilde{x}_2)=E\widetilde{x}_1\cdot E\widetilde{x}_2$. Имеем

$$E(\widetilde{x}_1\widetilde{x}_2)=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}ijp_{ij}=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}ijp_{i\cdot}p_{\cdot j}=$$
$$=\left(\sum_{i=0}^{\infty}ip_i\right)\left(\sum_{j=0}^{\infty}jp_j\right)=E\widetilde{x}_1E\widetilde{x}_2.$$

II-8. По теореме II-5 § 3 имеем

$$J(Y)=J_1(Y)\cdot J_2(Y)=e^{\lambda_1(Y-1)}e^{\lambda_2(Y-1)}=e^{(\lambda_1-\lambda_2)(Y-1)}=e^{Y-1},$$

где $\lambda=\lambda_1+\lambda_2$.

II-9. На первый взгляд согласуется. Однако у нас нет объективного критерия, который бы оценил это. В гл. VII такой критерий будет дан (см. задачу VII-16).

Для вычисления ожидаемых частот $p(i)$ и численностей $np(i)$ используйте рекуррентное отношение

$$p(i+1)=\frac{n-i}{i+1}\frac{p}{q}p(i),$$

а именно:

$$p(0) = q^n = (2/3)^{12} = 0{,}0077073; \quad np(0) = 202,75;$$

$$p(1) = \frac{n}{1} \cdot \frac{p}{q} p(0) = 12/1 \cdot \frac{1/3}{2/3} \cdot 0{,}0077073 = 0{,}046244; \quad np(1) = 1\,216,49;$$

$$p(2) = \frac{n-1}{2} \cdot \frac{p}{q} p(1) = 11/2 \cdot \frac{1/3}{2/3} \cdot p(1) = 0{,}12717;\ np(2) = 3\,345{,}33 \text{ и т. д.}$$

II-10. Вероятность того, что все 9 детей — девочки, равна $\frac{9!}{9!0!}\left(\frac{1}{2}\right)^9 = 2^{-9} = \frac{1}{512} \approx 0{,}002$. Мала или велика эта вероятность? Этот вопрос обсуждается в § 2 гл. IV.

II-11. Для вычисления ожидаемых частот $p(i)$ и численностей $np(i)$ используйте рекуррентное отношение

$$p(i+1) = \frac{m}{i+1} p(i),$$

т. е.

$$p(0) = \frac{1}{0!} m^0 e^{-m} = e^{-m} = 0{,}8147;$$

$$p(1) = \frac{m}{1} p(0) = 0{,}1670;$$

$$p(2) = \frac{m}{2} p(1) = 0{,}0171 \quad \text{и т. д.}$$

III-2.

$$E(\widetilde{x}_1 \widetilde{x}_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x,y)dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_1(x)f_2(y)dxdy =$$

$$= \int_{-\infty}^{\infty} xf_1(x) \left(\int_{-\infty}^{\infty} yf_2(y)dy \right) dx = \int_{-\infty}^{\infty} xf_1(x) E\widetilde{x}_2 dx =$$

$$= E\widetilde{x}_2 \int_{-\infty}^{\infty} xf_1(x)dx = E_1 \cdot E_2.$$

III-3.

$$E\widetilde{y} = E\left(\frac{\widetilde{x} - E\widetilde{x}}{\sqrt{D\widetilde{x}}}\right) = \frac{1}{\sqrt{D\widetilde{x}}}(E\widetilde{x} - E\widetilde{x}) = 0.$$

$$D\widetilde{y} = D\left(\frac{\widetilde{x} - E\widetilde{x}}{\sqrt{D\widetilde{x}}}\right) = \frac{1}{D\widetilde{x}}(D\widetilde{x} - 0) = 1.$$

III-4.

а) $\Phi(0) = 0{,}5$;

б) $\Phi(-1) = 0{,}159$;

в) $\Phi(2) = 0{,}977$.

III-5.

а) $P\{-3 < \widetilde{u} < 2\} = \Phi(2) - \Phi(-3) = 0{,}977 - 0{,}0014 \approx 0{,}976$;

б) $P\{-1 < \widetilde{u} < \infty\} = \Phi(\infty) - \Phi(-1) = 1 - 0{,}159 = 0{,}841$;

в) $P\{|\widetilde{u}| < 1\} = \Phi(1) - \Phi(-1) = 0{,}841 - 0{,}159 = 0{,}682$;

г) $P\{|\widetilde{u}| \leqslant 2\} = 0{,}951$;

д) $P\{|\widetilde{u}| \leqslant 3\} = 0{,}9973$;

е) $P\{|\widetilde{u}| \leqslant 3\} = 1 - P\{|\widetilde{u}| \leqslant 3\} = 1 - 0{,}9973 = 0{,}0027$.

III-6. $P\{|\widetilde{u}| \geqslant \gamma\} = 1 - 2P\{\widetilde{u} < -\gamma\} = 1 - 2\Phi(-\gamma) = \beta$, откуда $\Phi(\gamma) = (1 - \beta)/2$. Следовательно:

1) $(1 - \beta)/2 = 0{,}05$, $\gamma \approx 1{,}64$;

2) $(1 - \beta)/2 = 0{,}025$, $\gamma \approx 1{,}96$;

3) $(1 - \beta)/2 = 0{,}005$, $\gamma \approx 2{,}58$.

III-7. Так как для нормированного нормального распределения $\widetilde{u} = (\widetilde{x} - \mu)$, то $\widetilde{x} = \widetilde{u}\sigma + \mu = 2\widetilde{u} + 1$. Поэтому

$$P\{-3 < \widetilde{u} < 7\} = P\{-3 < (2\widetilde{u} + 1) < 7\} =$$
$$= P\{-2 < \widetilde{u} < 3\} = \Phi(3) - \Phi(-2) \approx 0{,}978.$$

III-8. Аналогично задаче III-7 $\widetilde{x} = \widetilde{u}\sigma + \mu = 2\widetilde{u} + 4$, откуда $2\widetilde{x} - 7 = 4\widetilde{u} + 8 - 7 = 4\widetilde{u} + 1$. Неравенство $|4\widetilde{u} + 1| < 2$ эквивалентно $-2 < 4\widetilde{u} + 1 < 2$, или $-0{,}75 < \widetilde{u} < 0{,}25$, откуда

$$P\{|2\widetilde{x} - 7| > 2\} = 1 - P\{|2\widetilde{x} - 7| \leqslant 2\} =$$
$$= 1 - P\{-0{,}75 \leqslant \widetilde{u} \leqslant 0{,}25\} = 1 - \Phi(0{,}25) + \Phi(-0{,}75) = 0{,}628.$$

III-9. Случайная величина $\widetilde{x} = \widetilde{x}_1 - \widetilde{x}_2$ имеет нормальное распределение: $\widetilde{x}_1 - \widetilde{x}_2 \sim N(-1; \sqrt{5})$. Следовательно, $\widetilde{x} = \widetilde{u}\sigma + \mu = \sqrt{5u} - 1$, откуда $-4 < \sqrt{5}\widetilde{u} - 1 < 8$; $-1{,}34 < \widetilde{u} < 4{,}02$. Значит, искомая вероятность равна $\Phi(4{,}02) - \Phi(-1{,}34) \approx 0{,}91$.

III-10. Случайная величина $\widetilde{x} = \widetilde{x}_1 + 2\widetilde{x}_2 - 3$ распределена по нормальному закону с $\mu = \mu_1 + 2\mu_2 - 3 = 3$; $\sigma^2 = \sigma_1^2 + 4\sigma_2^2 = 9 + 64 = 73$. Следовательно, $|\widetilde{x}_1 + 2\widetilde{x}_2 - 3| < 8$ эквивалентно неравенству $|\widetilde{x}| < 8$. Далее $\widetilde{x} = \widetilde{u} \cdot \sigma + \mu = \widetilde{u}\sqrt{73} + 3$, откуда искомая вероятность равна

$$1 - P\{-8 < \widetilde{u}\sqrt{73} + 3 < 8\} = 1 - P\{-1,29 < \widetilde{u} < 0,59\} \approx 0,375.$$

III-11. $\widetilde{x} = 5\widetilde{u} + 2$, откуда $-\gamma < 5\widetilde{u} + 2 < \gamma$, или $-\dfrac{\gamma + 2}{5} < \widetilde{u} < \dfrac{\gamma - 2}{5}$. Следовательно,

$$P\{|\widetilde{x}| < \gamma\} = 1 - \Phi\left(-\frac{\gamma + 2}{5}\right) - \Phi\left(-\frac{\gamma - 2}{5}\right) = 0,95,$$

т. е. по таблице надо найти такое значение $\delta = \dfrac{\gamma + 2}{5}$, что $\Phi[-\delta] + \Phi[-(\delta - 0{,}8)] = 0{,}05$. Из таблицы находим, что при $\delta \approx 2{,}51$ это равенство почти выполняется. Следовательно, $\gamma = 5\delta - 2 = 10{,}55$.

III-12. По определению $\widetilde{\chi}_1^2 = \widetilde{u}^2$, поэтому $E\widetilde{\chi}_1^2 = E\widetilde{u}^2 = D\widetilde{u} = 1$; $D\widetilde{\chi}_1^2 = E[\widetilde{\chi}_1^2 - E\widetilde{\chi}_1^2]^2 = E(\widetilde{\chi}_1^2)^2 - (E\widetilde{\chi}_1^2)^2 = E\widetilde{u}^4 - 1$. Имеем

$$E\widetilde{u}^4 = \int_{-\infty}^{\infty} x^4 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^3 e^{-\frac{1}{2}x^2} d\left(-\frac{x^2}{2}\right) =$$
$$= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^3 d(e^{-x^2/2}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} 3x^2 e^{-x^2/2} - \frac{1}{\sqrt{2\pi}} x^3 e^{-x^2/2} \Bigg|_{-\infty}^{\infty} =$$
$$= 3\int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 3D\widetilde{u} = 3, \quad \text{откуда } D\widetilde{\chi}_1^2 = 3 - 1 = 2.$$

III-13.

а) $P \leqslant 0{,}001$;

б) $P < 0{,}025$;

в) $P > 0{,}975$.

III-14.

а) $P\{\widetilde{t} \geqslant 2{,}2\} = 0{,}025$;

б) $P\{|\widetilde{t}| \geqslant 2{,}2\} = 0{,}05$;

в) $\gamma = 2{,}36$;

г) $\gamma = 3{,}50$.

III-15.

а) $0{,}005 < P < 0{,}01$;

б) $0{,}01 < P < 0{,}025$;

в) $P < 0{,}0005$.

IV-4. При построении обычной (линейной) гистограммы за точку отсчета на оси абсцисс возьмите $x_i = 140^\circ$ либо постройте круговую гистограмму [см.: Мардиа, 1978].

V-1. Функция правдоподобия $L(\Theta) = C_n^k \Theta^k (1-\Theta)^{n-k}$, ее логарифм $\ln L = \ln C_n^k + k \ln \Theta + (n-k)\ln(1-\Theta)$; уравнение правдоподобия $\dfrac{d(\ln L)}{d(\Theta)} = \dfrac{k}{\Theta} - \dfrac{n-k}{1-\Theta} = 0$, откуда $\Theta = \dfrac{k}{n}$.

V-2. Функция правдоподобия $L(\Theta) = \left(e^{-\Theta}\dfrac{\Theta^{x_1}}{x_1!}\right)\ldots\left(e^{-\Theta}\dfrac{\Theta^{x_n}}{x_n!}\right) = e^{-n\Theta}\dfrac{\Theta^{x_1+\ldots+x_n}}{x_1!\ldots x_n!}$, ее логарифм $\ln L = -n\Theta + (x_1 + \ldots + x_n)\ln\Theta - \ln(x_1!\ldots x_n!)$; уравнение правдоподобия $\dfrac{\partial(\ln L)}{\partial\Theta} = -n + \sum_{i=1}^n x_i \dfrac{1}{\Theta} = 0$, откуда $\Theta = \dfrac{1}{n}\sum x_i$.

V-3. Функция правдоподобия $L(\Theta_1,\ldots,\Theta_l)=\frac{n!}{k_1!k_2!\ldots k_l!}\Theta_1^{k_1}\ldots\Theta_l^{k_l}$. Так как $\Theta_1+\ldots+\theta_l=1$, то независимы только $(l-1)$ из них $-\Theta_1,\ldots,\Theta_{l-1}$, а $\Theta_l=1-\Theta_1-\ldots-\Theta_{l-1}$. Поэтому $\ln L=\ln\frac{n!}{k_1!\ldots k_l!}+k_1\ln\Theta_1+{}$ $+\ldots+k_{l-1}\ln\Theta_l+k_l\ln(1-\Theta_1-\ldots-\Theta_{l-1})$. Решаем $(l-1)$ уравнений правдоподобия:

$$\frac{\partial(\ln L)}{\partial\Theta_i}=\frac{k_i}{\Theta_i}-\frac{k_l}{1-\Theta_1-\ldots-\Theta_{l-1}}=0,$$

откуда

$$k_i\Theta_i=k_i(1-\Theta_1-\ldots-\Theta_{l-1}).$$

Суммируя обе части последних уравнений по $i=1,\ldots,l-1$, прибавляя затем к каждой из них $k_l\Theta_l$, получим $k_l=n\Theta_l$, поскольку $\sum_{i=1}^{l}\Theta_i=1$ и $\sum_{i=1}^{l}k_i=n$. Отсюда $k_l/\Theta_l=n$ и $\Theta_i=k_i\Theta_l/k_l=k_i/n$.

V-4. Пусть μ — математическое ожидание случайных величин $\widetilde{x}_1,\ldots,$ $\ldots,\widetilde{x}_n$. Тогда для каждого i

$$(\widetilde{x}_i-\widetilde{m})^2=[(\widetilde{x}_i-\mu)-(\widetilde{m}-\mu)]^2=(\widetilde{x}_i-\mu)^2-2(\widetilde{x}_i-\mu)(\widetilde{m}-\mu)+(\widetilde{x}-\mu)^2$$

и

$$n\widetilde{s}_0=\sum_{i=1}^{n}(\widetilde{x}_i-\mu)^2-2(\widetilde{m}-\mu)\sum_{i=1}^{n}(\widetilde{x}_i-\mu)+n(\widetilde{m}-\mu)^2=$$

$$=\sum(\widetilde{x}_i-\mu)^2-n(\widetilde{m}-\mu)^2,$$

откуда $\frac{\widetilde{s}_0^2}{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widetilde{x}_i-\mu}{\sigma}\right)^2-\left(\frac{\widetilde{m}-\mu}{\sigma}\right)^2$. Так как $\frac{\widetilde{x}_i-\mu}{\sigma}$ распределена со средним 0 и дисперсией 1, а $\frac{\widetilde{m}-\mu}{\sigma}$ со средним 0 и дисперсией $\frac{1}{n}$ (см. гл. V), то $M\left(\frac{\widetilde{s}^2}{\sigma^2}\right)=1-\frac{1}{n}=\frac{n-1}{n}$, т. е. $M\widetilde{s}^2=\frac{n-1}{n}\sigma^2$. Отсюда же следует, что $M\left(\frac{n}{n-1}\widetilde{s}_0^2\right)=\sigma^2$, т. е. оценка $\widetilde{s}^2=\frac{1}{n-1}\sum_{i=1}^{n}(\widetilde{x}_i-\widetilde{m})^2$ несмещенная.

V-5. Доверительным интервалом является область всех возможных значений, принимаемых случайной величиной.

V-6. $n = 245$; $m = 174{,}5$; $s^2 = 70{,}06$; $s = 8{,}37$; $s_m = 0{,}535$; $v = 4{,}8\,\%$; $\nu = 244$; $t_{0{,}025}(244) = 1{,}97$. С доверительной вероятностью 0,95: $173{,}4 < \mu < 175{,}6$; $57{,}11 < \sigma^2 < 83{,}02$. При вычислении доверительного интервала для σ^2 используем нормальную аппроксимацию, так как $n > 30$.

V-7. $n = 15$; $m = 1{,}47$; $s^2 = 0{,}0452$; $s = 0{,}213$; $s_m = 0{,}0549$; $v = 14{,}5\,\%$; $\nu = 14$. Из табл. III и V Приложения 1 $t_{0{,}025}(14) = 2{,}14$; $\chi^2_{0{,}025}(14) = 5{,}63$; $\chi^2_{0{,}974}(14) = 26{,}1$. С доверительной вероятностью 0,95: $1{,}35 < \mu < 1{,}59$; $0{,}024 < \sigma^2 < 0{,}112$.

V-8. $h = 0{,}232$. Поскольку $nh(1-h) > 25$, используем нормальную аппроксимацию: $\sqrt{h(1-h)}/n = 0{,}0086$ и с вероятностью 0,95: $0{,}215 < p < 0{,}249$.

V-9 (см. пример V-3). Ответ: $h = 0{,}04$. С вероятностью 0,95 процент больных диабетом в данном районе находится в интервале от 3,1 до 4,9.

V-10 (см. пример V-5). Ответ: $m = 42$. С доверительной вероятностью 0,95 среднее число колоний на чашке находится в пределах 29,3–54,7. Доверительный интервал очень велик, поскольку $n = 1$; однако в случае распределения Пуассона его все-таки можно оценить.

V-11. $(1-\alpha)\,100\,\%$-ный доверительный интервал для среднего квадратичного отклонения σ нормального распределения задается соотношением

$$s\sqrt{\frac{\nu}{\chi^2_{1-\alpha/2}}} < \sigma < s\sqrt{\frac{\nu}{\chi^2_{\alpha/2}}}.$$

V-12. Доверительный интервал для математического ожидания np биномиальной случайной величины $\widetilde{x}$ имеет вид $np_{\text{н}} < np < np_{\text{в}}$, где $p_{\text{н}}$ и $p_{\text{в}}$ — нижняя и верхняя границы доверительного интервала для параметра p (см. § 5).

V-13. $n\lambda_{\text{н}} < n\lambda < n\lambda_{\text{в}}$, где $\lambda_{\text{н}}$ и $\lambda_{\text{в}}$ — нижняя и верхняя границы доверительного интервала для параметра λ (см. § 6).

VI-1. $n_1 = 12$; $n_2 = 12$. Используя преобразование $\sqrt{x + 0{,}386}$, получаем $m_1 = 1{,}46$; $s_{m_1} = 0{,}174$; $m_2 = 1{,}92$; $s_{m_2} = 0{,}134$; $\nu = 22$; $|t_{\text{эксп}}| = 2{,}09 > t_{0{,}025}(22) = 2{,}07$. $U_{\text{эксп}} = 37{,}5$; $E\widetilde{U} = 72$; $k = 4$; $c_1 = 6$; $c_2 = 3$;

$c_3 = 6$; $c_4 = 2$; $D\widetilde{U} = 291$; $|u_{\text{эксп}}| = 2{,}02 > u_{0{,}025} = 1{,}96$. Способ II более эффективен на уровне значимости $\alpha = 0{,}05$ как по критерию Стьюдента, так и по критерию Вилкоксона–Манна–Уитни.

Если не использовать преобразование, то $m_1 = 2{,}1$; $m_2 = 3{,}5$; $s_{m_1} = 0{,}57$; $s_{m_2} = 0{,}50$; $|t_{\text{эксп}}| = 1{,}85 < t_{0{,}025}(22) = 2{,}07$ и мы приходим, по-видимому, к ложному выводу об отсутствии различия между двумя способами борьбы с насекомым.

VI-2. Представленные данные, несомненно, являются зависимыми, т. t,e. парными наблюдениями (см. задачу IX-20). Поэтому $n = 5$; $m_d = 2$; $s_d = 0{,}52$; $\nu = 4$; $|t_{\text{эксп}}| = 3{,}85 > t_{0{,}025}(4) = 2{,}77$. Увеличение чистого дохода от применения усиленного удобрения оказывается статистически существенным на уровне значимости $\alpha = 0{,}05$. По-видимому, имеет смысл применять усиленное удобрение, так как затрата на него каждого лишнего фунта дает прирост прибыли в среднем на два фунта ($m_d = 2$), т. е. окупается вдвое.

Если бы мы рассматривали эти данные как независимые выборки, то, имея $n_1 = n_2 = 5$; $m_1 = 18{,}5$; $m_2 = 20{,}5$; $s_{m_1} = 1{,}432$; $s_{m_2} = 1{,}688$; $\nu = 8$; $|t_{\text{эксп}}| = 0{,}90 < t_{0{,}025}(8) = 2{,}31$, мы пришли бы к неверному выводу о неэффективности усиленного удобрения.

VI-3. Лечение методом А и методом Б можно назначать попеременно по мере появления случаев заболевания.

VI-4. $h_1 = 0{,}056$; $h_2 = 0{,}361$; $h = 0{,}193$; $n = 457$. Так как $n_i h > 5$ и $n_i h(1 - h) > 5$, можно применить u-критерий: $|u_{\text{эксп}}| = 8{,}22 > u_{0{,}0005} = 3{,}29$.

Благотворность действия антикоагулянтов несомненна: $P\{\widetilde{u} \geqslant u_{\text{эксп}}\} < 0{,}001$.

VI-5. $m_1 = 124{,}9$; $m_2 = 101{,}0$; $s_1^2 = 618{,}81$; $s_2^2 = 425{,}33$; $s_1 = 24{,}83$; $s_2 = 20{,}62$; $s_{m_1} = 9{,}39$; $s_{m_2} = 7{,}9$; $\nu = 12$; $|t_{\text{эксп}}| = 1{,}96 < t_{0{,}025}(12) = 2{,}18$. Разница между средними прибавками массы у крыс незначима при $\alpha = 0{,}05$.

VI-6. $\nu_1 = 999$; $\nu_2 = 138$. Для длины черепов $F_{\text{эксп}} = 1{,}16 < F_{0{,}05} = 1{,}23$ и для ширины $F_{\text{эксп}} = 1{,}20 < F_{0{,}05} = 1{,}23$. Изменчивость обоих признаков у современных европейцев и древних египтян не различается при $\alpha = 0{,}1$.

VI-7. $\nu_1 = 11$; $\nu_2 = 4$; $m_1 = 0{,}296$; $m_2 = 0{,}41$; $s_1^2 = 0{,}00088$; $s_2^2 = 0{,}0040$; $s^2 = 0{,}0017$; $\nu = 15$; $|t_{\text{эксп}}| = 5{,}18 > t_{0{,}00005}(15) = 4{,}07$. Различие штаммов вируса гриппа по фоточувствительности высоко значимо (при $\alpha = 0{,}001$).

К подобному же выводу, но с меньшими вычислениями, мы приходим, используя критерий Вилкоксона–Манна–Уитни. Для этого достаточно заметить, что для штамма «Техас-77» все значения, кроме одного $(0{,}34)$, меньше всех значений для штамма «СССР-77». Значит, $U_{\text{эксп}} = 0{,}5$; $E\widetilde{U} = 30$; $D\widetilde{U} = 90$; $|u_{\text{эксп}}| = 3{,}11 > u_{0{,}001} - 3{,}09$, т. е. различие значимо при $\alpha = 0{,}002$.

VI-8. $\nu = 17$; $m_d = 2{,}1$; $s_d = 2{,}60$; $|t_{\text{эксп}}| = 0{,}81 < t_{0{,}05}(17) = 1{,}74$. По парному критерию Стьюдента наблюдаемые различия незначимы при $\alpha = 0{,}1$, т. е. можно считать, что артериальное давление восстанавливается после прерывания кровотока с последующим его возобновлением. Используя парный критерий Вилкоксона, приходим к тому же выводу: $N = 15$; $W_{\text{эксп}}44{,}5$; $E\widetilde{W} = 60$; $D\widetilde{W} = 307{,}4$; $k = 2$; $c_1 = 5$; $c_2 = 2$; $|u_{\text{эксп}}| = 0{,}88 < u_{0{,}05} = 1{,}64$.

VI-9. $\nu = 9$; $m_d = 1{,}58$; $s_d = 0{,}389$; $|t_{\text{эксп}}| = 4{,}06 > t_{0{,}005}(9) = 3{,}25$. Снотворное A существенно лучше снотворного B (при $\alpha = 0{,}01$). Если бы мы рассматривали эти данные как независимые, то пришли бы к неверному выводу об отсутствии различий: $m_1 = 2{,}33$; $m_2 = 0{,}75$; $\nu_1 = \nu_2 = 9$; $\nu = 18$; $s_1^2 = 4{,}009$; $s_2^2 = 3{,}201$; $|t_{\text{эксп}}| = 1{,}86 < t_{0{,}05}(18) = 1{,}73$.

VI-10. Случайная величина $\widetilde{t^2}$:

$$t^2 = F = \frac{(n_1+n_2+2)\left(n_2\sum\limits_{i=1}^{n_1}x_{1i} - n_1\sum\limits_{i=1}^{n_2}x_{2i}\right)^2}{(n_1+n_2)\left[n_1n_2\left(\sum\limits_{i=1}^{n_1}x_{1i}^2 - \sum\limits_{i=1}^{n_2}x_{2i}^2\right) - n_2\left(\sum\limits_{i=1}^{n_1}x_{1i}\right)^2 - n_1\left(\sum\limits_{i=1}^{n_2}x_{2i}\right)^2\right]}$$

— имеет F-распределение с параметрами $\nu_1 = 1$ и $\nu_2 = n - 2$. Считается, что по сравнению со статистикой t-критерия эта статистика экономичнее в вычислениях примерно на 30 %.

VI-11. Точный критерий Фишера — необходимо вычислять вероятности для следующих трех таблиц:

14	4	15	5	10	10
1	4	0	5	5	0

$p_1 = 0{,}059$; $p_2 = 0{,}005$; $p_3 = 0{,}057$; $\sum_{i=1}^{3} p_i = 0{,}12$. Действенность противогриппозной сыворотки неубедительна даже на уровне значимости $\alpha = 0{,}1$. Если же ограничиться уровнем $\alpha = 0{,}05$, то достаточно было вычислить значение $p_i = 0{,}059$, чтобы убедиться в отсутствии различий. При постановке эксперимента нарушен принцип рандомизации: выбор вакцинируемых не должен был зависеть от желания пациентов.

VII-1. Для анализа таблиц $r \times 2$ удобно использовать формулу Брандта–Снедекора:

$$\chi^2 = \frac{n^2}{n_{\cdot 1} n_{\cdot 2}} \left(\sum_{j=1}^{r} \frac{k_{j1}^2}{n_{j\cdot} - \frac{n_{\cdot 1}{}^2}{n}} = \right) \frac{n^2}{n_{\cdot 1} n_{\cdot 2}} \left(\sum_{j=1}^{r} \frac{k_{j2}^2}{n_{j\cdot} - \frac{n_{\cdot 2}{}^2}{n}} \right), \quad \nu = 14.$$

$\chi^2_{\text{эксп}} = 7{,}37 < \chi^2_{0{,}05}(14) = 23{,}7$. Данные различных авторов о расщеплении по окраске семядолей у гороха хорошо согласуются друг с другом. Для проверки согласия с ожидаемым расщеплением $k_1 : k_2 = 3 : 1$ удобно использовать формулу

$$u = \frac{|k_1 - 3k_2| - 1{,}5}{\sqrt{3(k_1 + k_2)}} \quad \text{или} \quad \chi^2 = \frac{(|k_1 - 3k_2| - 1{,}5)^2}{3(k_1 + k_2)}, \quad \nu = 1.$$

VII-3. $n = 400$; $m = 4{,}68$. Объединяем классы для $t = 0; 1$ и $i = 10, \ldots, 13$, чтобы ожидаемые численности превышали 5. Тогда число классов $l = 10$; $\nu = 9$; $\chi^2_{\text{эксп}} = 4{,}41 < \chi^2_{0{,}05}(9) = 16{,}9$. Распределение дрожжевых клеток по квадратам счетной камеры близко к пауссоновскому.

VII-4. $\nu = 1$; $\chi^2_{\text{эксп}} = 15{,}3 > \chi^2_{0{,}001}(1) = 10{,}8$. Две сравниваемые породы существенно различаются по устойчивости к американскому гнильцу.

VII-5. $\nu = 5$; $\chi^2_{\text{эксп}} = 61{,}4 > \chi^2_{0{,}001}(5) = 20{,}5$. Возрастной состав больных резко изменился, различия значимы при $\alpha = 0{,}001$.

VII-7. $\nu = l - 2$.

VII-8. $\nu = 11$; $\chi^2_{\text{эксп}} = 737{,}9l > \chi^2_{0{,}001} = 31{,}3$. Распределение числа рождений в разные месяцы высоко значимо отличается от равномерного. В этом случае вычисления удобно проводить по формуле

$$\chi^2 = \frac{1}{m} \sum_{i=1}^{r} (k_i - m)^2 = \frac{1}{m} \sum_{i=1}^{r} k_i^2 - rm.$$

VII-9. После объединения классов с малыми ожидаемыми численностями имеем $l = 15$ и $\nu = 14$; $\chi^2_{\text{эксп}} = 23{,}7 = \chi^2_{0{,}05}(14) = 23{,}7$. Отклонение от биномиального распределения можно считать значимым при $\alpha = 0{,}05$, но не при меньших значениях α.

VIII-1. $\nu_b = 3$; $\nu_w = 15$; $SS_b = 19{,}65$; $SS_w = 62{,}03$; $MS_b = 64{,}55$; $MS_w = 4{,}14$; $F_{\text{эксп}} = 15{,}61 > F_{0{,}001}(3;15) \approx 11{,}4$. Разные породы кур высоко значимо различаются по содержанию гемоглобина. Различие обусловлено разницей между породами «итальянские» и «бентамки»: $t = 2{,}87 > t_{0{,}025}(15) = 2{,}13$; и между породами «куропатчатые» и «бентамки»: $t = 2{,}48 > t_{0{,}025}(15) = 2{,}48$. Использование преобразования $\arcsin\sqrt{p}$ здесь не правомерно, поскольку в задаче речь идет о непрерывной, а не дискретной случайной величине (содержание гемоглобина).

VIII-2. $\nu_b = 7$; $\nu_w = 56$; $SS_b = 1{,}838$; $SS_w = 1{,}185$; $MS_b = 0{,}263$; $MS_w = 0{,}0212$; $F_{\text{эксп}} = 12{,}41 > F_{0{,}001}(7;56) = 4{,}1$. Четвертый, пятый и шестой индивиды, а также седьмой и восьмой образуют две группы, различия внутри которых не существенны. Все остальные попарные сравнения статистически значимы.

VIII-4. Применяя преобразование $\sqrt{x}$, получаем $\nu_b = 3$; $\nu_w = 36$; $SS_b = 0{,}7515$; $SS_w = 13{,}552$; $MS_b = 0{,}25055$; $MS_w = 0{,}3764$; $F_{\text{эксп}} = 0{,}665 < F_{0{,}05}(3;36) = 2{,}87$. Сравниваемые участки практически одинаковы по числу бактерий в почве.

VIII-5. Используем преобразование $\arcsin\sqrt{p}$: $\nu_b = 5$; $\nu_a = 3$; $\nu_w = 15$; $SS_b = 1{,}095$; $SS_a = 1{,}711$; $SS_w = 0{,}361$; $MS_b = 0{,}219$; $MS_a = 0{,}570$; MS_w, $= 0{,}021$; $F_{\text{эксп}} = MS_b/MS_w = 9{,}125 > F_{0{,}001}(5;15) = 7{,}6$; $F_{\text{эксп}} = MS_a/MS_w = 23{,}75 > F_{0{,}001}(3;15) = 9{,}3$.

Таким образом, значимыми оказываются различия как между сортами, так и между блоками.

VIII-6. Используем преобразование $\lg x$: $\nu_b = 3$; $\nu_w = 24$; $SS_b = 12{,}228$; $SS_w = 0{,}2287$; $MS_b = 4{,}0761$; $MS_w = 0{,}0095$; $F_{\text{эксп}} = 427{,}8 > F_{0{,}001}(3;24) = 7{,}55$. численности всех четырех видов планктона резко различаются.

VIII-7. В письме Ч. Дарвину Ф. Гальтон, по существу, сформулировал одну из первых непараметрических процедур сравнения двух независимых выборок равного объема, которая впоследствии получила название критерия Гальтона.

Более эффективным является применение критерия Крускала–Уоллиса. Для перекрестноопыленных растений имеем $H^*_{\text{эксп}}$=0,165<$\chi^2_{0,05}(3) =$ $= 7{,}82$, т. е. все четыре выборки перекрестноопыленных растений достаточно однородны. Для самоопыленных растений $\chi^2_{0,1}(3) = 6{,}25 < H_{\text{эксп}} =$ $= 6{,}2 < \chi^2_{0,05}(3) = 7{,}82$, т. е. выборки самоопыленных растений менее однородны, чем выборки перекрестноопыленных растений. Этот вывод может представлять самостоятельный интерес для биолога и быть предметом дальнейшего изучения.

Сравнение объединенных данных для перекрестноопыленных и самоопыленных растений с помощью критерия Вилкоксона–Манна–Уитни указывает на высоко значимое различие между ними: $U_{\text{эксп}} = 39,5 >$ $> U_{0,001}(15;15) = 40$, т. е. подтверждает вывод Ч. Дарвина и Ф. Гальтона.

К подобным выводам мы приходим, используя обычный дисперсионный анализ. Для перекрестноопыленных растений $\nu_b = 3$; $\nu_w = 11$; $SS_b = 13{,}036$; $SS_w = 170{,}116$; $MS_b = 4{,}345$; $MS_w = 15{,}465$; $F_{\text{эксп}} =$ $= 0{,}28 < F_{0,05}(3;11) = 3{,}59$. Для самоопыленных растений $\nu_b = 3$; $\nu_w = 11$; $SS_b = 25{,}772$; $SS_w = 33{,}159$; $MS_b = 8{,}591$; $MS_w = 3{,}014$; $F_{0,1}(3;11) = 2{,}66 < F_{\text{эксп}} = 2{,}85 < F_{0,05}(3;11) = 3{,}59$. Сравнение объединенных выборок дает $t_{0,025}(28) = 2{,}05 < t_{\text{эксп}} = 2{,}44 < t_{0,005}(28) = 2{,}76$.

IX-1. Испытания I и II не являются независимыми, силу связи можно изменять, подбрасывая повторно разное число костей.

IX-3. Графический анализ данных показывает, что зависимость между массой тела (x_1) и долей массы мозга от общей массы (x_2) у тюленя не является линейной. Действительно, $\nu = 5$ и $r_{\text{эксп}} = -0{,}83$; по абсолютной величине это значение меньше, чем $r_{0,025}(5) = 0{,}88$, т. е. корреляция незначима при $\alpha = 0{,}05$ или же она нелинейна. После логарифмического преобразования $\lg x_1$ и $\lg x_2$ зависимость линеаризуется, $r_{\text{эксп}} = -0{,}99$, и корреляция оказывается значимой при $\alpha = 0{,}001$, так как $r_{0,0005}(5) = 0{,}95$.

IX-4. $\widehat{y} = 1{,}02 + 2{,}06x$.

IX-5. $\nu = 4$; $\chi_{\text{эксп}} = 4{,}02 < \chi^2_{0,05}(4) = 9{,}49$. Зависимость между односторонностью в развитии рук и глазной односторонностью практически отсутствует.

IX-7. Если считать фиксированными результаты измерения обычным методом, то $\widehat{y} = 15{,}35 + 0{,}975x$; если фиксированными считать результаты измерения новым методом, то $\widehat{x} = 7{,}25 + 1{,}02y$ (регрессия x по y). В обоих

случаях гипотеза H_0: $\beta = 1$ принимается: $S_b = 0{,}025$; $\nu = 16$ и $|t_{\text{эксп}}| = |b - \beta|/s_b = (1{,}0 \text{ или } 0{,}8) < t_{0{,}025}(16) = 2{,}12$.

Строго говоря, это еще не означает, что оба метода дают одинаковые результаты. Необходимо, чтобы $\beta = 1$, и одновременно линия регрессии должна проходить через начало координат; процедуру проверки такой гипотезы см. у К. А. Браунли [1977, с. 333–336]. Проще, однако, использовать парный t-критерий: $m_d = 34{,}17$; $s_d = 14{,}06$; $|t_{\text{эксп}}| = 2{,}43 > t_{0{,}025}(16) = 2{,}12$, и гипотеза об одинаковости результатов, получаемых двумя методами, отвергается при $\alpha = 0{,}05$.

IX-8. $\widehat{y}_1 = 100{,}1 - 5{,}79x$; $\widehat{y}_2 = 98{,}5 - 5{,}34x$. Соответствующее уравнение регрессии есть $\ln A = 4{,}60 - 0{,}513D$, т. е. $\gamma = 0{,}51$.

IX-9. $\widehat{y}_1 = 100{,}1 - 5{,}79x$; $\widehat{y}_2 = 98{,}5 - 5{,}34x$.

IX-10. $\widehat{y} = 1{,}71 - 0{,}455x$, где $y = \lg Q$; $x = \lg T$.

IX-11. Нет. Уже на глаз видно, что выборочное двухмерное распределение не является нормальным.

IX-12. $\nu = 4$; $\chi^2_{\text{эксп}} = 10{,}47 > \chi^2_{0{,}05}(4) = 9{,}49$. Зависимость между полом ребенка и цветом волос у шотландских детей следует признать значимой при $\alpha = 0{,}05$.

IX-13. $\nu = 4$; $\chi^2_{\text{эксп}} = 160{,}53 > \chi^2_{0{,}001}(4) = 18{,}5$. Связь между конституцией ягнят при рождении и в полуторагодовалом возрасте несомненна.

IX-14. $\nu = 18$; $r_{\text{эксп}} = -0{,}256$; $|k_{\text{эксп}}| < r_{0{,}025} = 0{,}44$. Линейная корреляция между полярностью и гидрофобностью аминокислот практически отсутствует. Ранговая корреляция $r_{s_{\text{эксп}}} = -033$ также оказывается незначимой: $|r_{s_{\text{эксп}}}| < r_s(0{,}025) = 0{,}45$.

IX-15. $\nu = 266$; $|r_{\text{эксп}}| = 0{,}523 > r_{0{,}001} = 0{,}21$. Корреляция слабая, но статистически значимая. Для группированных данных рекомендуется использовать формулу

$$r = \frac{n\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_iy_i - \sum_{i=1}^{k} n_{i\cdot}x_i \sum_{j=1}^{l} n_{\cdot j}y_j}{\left[n\sum_{i=1}^{k} n_{i\cdot}x_i^2 - \left(\sum_{i=1}^{k} n_{i\cdot}x_i\right)^2\right]\left[n\sum_{j=1}^{l} n_{\cdot j}y_j^2 - \left(\sum_{j=1}^{l} n_{\cdot j}y_j\right)^2\right]}.$$

IX-17. Находим $b_{\text{н}}$ и $b_{\text{в}}$, такие, что $P\{b_{\text{н}} < \varrho < b_{\text{в}}\} = 1 - \alpha$, а именно: $b_{\text{н}} = b - t_{\alpha/2} \cdot s_{\text{в}}$ и $b_{\text{в}} = b + t_{\alpha/2} \cdot s_{\text{в}}$.

IX-18. Следует использовать z-преобразование Фишера или z^*-преобразование Хотеллинга (для малых выборок). Тогда можно найти $z_{\text{н}}$ и $z_{\text{в}}$, такие, что $P\left\{z_{\text{н}} < \frac{1}{2}\ln\frac{1+\varrho}{1-\varrho} < z_{\text{в}}\right\} = 1-\alpha$. Для этого надо найти $z_{\text{н}} = z - u_{\alpha/2} \cdot \frac{1}{\sqrt{n-3}}$, $z_{\text{в}} = z + u_{\alpha/2} \cdot \frac{1}{\sqrt{n-3}}$. Затем следует провести обратное преобразование (используя табл. XII*б* Приложения 1) и найти искомое $r_{\text{н}}$ и $r_{\text{в}}$. В случае малых выборок ($n > 10$) надо найти $z^*_{\text{н}} = z^* - u_{\alpha/2} \cdot \frac{1}{\sqrt{n-1}}$ и $z^*_{\text{в}} = z^* + u_{\alpha/2} \cdot \frac{1}{\sqrt{n-1}}$.

IX-19. Случайная величина $\widetilde{u}_i = (\widetilde{z}_i - E\widetilde{z}_i)/\sqrt{D}\widetilde{z}_i \sim N(0;1)$, следовательно, $\sum_{i=1}^k \widetilde{u}_i \sim \chi^2(\nu)$, $\nu = 1$. Гипотеза H_0: $\varrho_1 = \ldots = \varrho_k = \varrho$ равносильна гипотезе H_0: $E\widetilde{u}_1 = \ldots = E\widetilde{u}_k = 0$. Тогда если в качестве оценки нулевого математического ожидания использовать $\widetilde{m}_u = \frac{1}{k}\sum_{i=1}^k \widetilde{u}_i$, то $\sum_{i=1}^k \frac{\widetilde{u}_i - \widetilde{m}_u}{\widetilde{m}_u} \sim \chi^2(\nu)$.

IX-20. Для данных примера VI-4: $\nu = 14$; $r_{\text{эксп}} = 0{,}996$. Между результатами определения крахмала двумя методами имеется чрезвычайно высокая корреляционная зависимость.

Для данных примера VI-9: $r_{S_{\text{эксп}}} = 0{,}71 > r_{s(0{,}005)}(14) = 0{,}68$; корреляция значима при $\alpha = 0{,}01$.

Для данных задачи VI-2: $\nu = 3$; $r_{\text{эксп}} = 0{,}96 = r_{0{,}005}(3) = 0{,}96$; корреляция значима при $\alpha = 0{,}01$.

Для данных задачи VI-9: $\nu = 8$; $r_{\text{эксп}} = 0{,}80 > r_{0{,}005}(8) = 0{,}87$; корреляция значима при $\alpha = 0{,}001$.

Дисперсия разности двух случайных величин есть

$$D(\widetilde{x}_1 - \widetilde{x}_2) = D\widetilde{x}_1 + D\widetilde{x}_2 - 2\varrho\sqrt{D\widetilde{x}_1 D\widetilde{x}_2}.$$

Парные критерии автоматически учитывают компоненту, обусловленную корреляцией: $D(\widetilde{x}_1 - \widetilde{x}_2) = E(\widetilde{x}_1 - \widetilde{x}_2)^2 - [E(\widetilde{x}_1 - \widetilde{x}_2)]^2$.

IX-21. $r_{\text{эксп}} = 0{,}072$. Корреляция чрезвычайно слабая, хотя и статистически значима.

IX-22. «Действительно, по статистическим данным оказывается, что богатые люди, пользующиеся этими предметами, выше, здоровее и живут дольше, чем те люди, которые никогда не помышляют о приобретении таких вещей. Не требуется большой проницательности, чтобы видеть, что эта разница в действительности создается не цилиндрами и зонтиками, а тем богатством и питанием, о которых они свидетельствуют, и что золотые часы и членство в клубе на Пэл–Мэл имеют такие же превосходные свойства» (Б. Шоу, 1906 г.).

Приложение 1

Таблица I

Равномерно распределенные случайные числа (см. гл. I, § 4 и гл. IV, § 3)

85	017	85	532	13	618	23	157	86	952	02	438
16	719	82	789	69	041	05	545	44	109	05	403
65	842	27	672	82	186	14	871	22	115	86	529
76	875	20	684	39	187	38	976	94	324	43	204
93	640	39	160	41	453	98	319	41	548	93	137
99	478	70	086	71	265	11	742	18	226	29	004
65	119	26	486	47	353	43	361	99	436	42	753
70	322	21	592	48	233	93	806	32	581	21	828
58	113	41	278	11	679	49	540	61	777	67	954
44	655	81	225	31	133	36	768	60	452	38	697
02	295	13	487	98	662	07	092	44	673	61	303
85	035	54	881	36	587	43	310	48	897	48	493
01	197	80	935	28	021	61	570	23	350	65	710
97	907	19	078	40	646	31	352	48	625	44	369
63	268	96	905	28	797	57	048	46	359	74	294
53	841	59	684	67	411	09	243	56	092	84	369
53	712	71	399	10	916	07	959	21	225	13	018
11	434	51	908	62	171	93	732	26	958	02	400
62	375	99	292	21	177	72	621	66	995	07	289
28	337	20	923	87	929	61	020	62	841	31	374
38	631	79	430	62	421	97	959	67	422	69	992
49	172	16	332	44	670	35	089	17	691	89	246
89	232	57	327	34	679	62	235	79	655	81	336
02	844	15	026	32	439	58	587	48	274	81	330
40	387	65	406	37	929	08	709	60	623	22	237
80	240	44	177	51	171	08	723	39	323	05	798
44	910	99	321	72	173	56	239	04	595	10	835
33	663	86	347	00	926	44	916	34	823	51	770
86	430	19	102	37	420	41	876	76	569	24	358
31	379	68	588	81	675	15	694	43	438	36	879
03	474	37	386	36	964	73	661	46	986	37	162
97	742	46	762	42	811	45	720	42	533	23	732
16	766	22	766	56	502	67	107	32	907	97	852
12	568	59	926	96	966	82	731	05	037	29	315
55	595	63	564	38	548	24	622	31	624	30	990
16	227	79	439	49	544	85	482	17	379	32	378
84	421	75	331	57	245	50	688	77	047	44	767
63	016	37	859	16	955	56	719	98	105	07	175
33	211	23	429	78	645	60	782	52	420	74	438
57	608	63	244	09	472	79	654	49	174	60	962
16	227	79	439	49	544	85	482	17	379	32	378
84	421	75	331	57	245	50	688	77	047	44	767
63	016	37	859	16	955	56	719	98	105	07	175
33	211	23	429	78	645	60	782	52	420	74	438
57	608	63	244	09	472	79	654	49	174	60	962

Таблица II

Функция нормального распределения (см. § 4 гл. III)*. Даны значения $\Phi(u) = P\{\widetilde{u} < u\}$, $\widetilde{u} \sim N(0;1)$

u	0	1	2	3	4	5	6	7	8	9
−3	0,0014	$0,0^397$	$0,0^369$	$0,0^348$	$0,0^334$	$0,0^323$	$0,0^316$	$0,0^311$	$0,0^472$	$0,0^448$
−2	0,023	0,018	0,014	0,011	0,0082	0,0062	0,0047	0,0035	0,0026	0,0019
−1	0,159	0,136	0,115	0,097	0,081	0,067	0,055	0,045	0,036	0,029
−0,9	0,184	0,181	0,179	0,176	0,174	0,171	0,169	0,166	0,164	0,161
−0,8	0,212	0,209	0,206	0,203	0,200	0,198	0,195	0,192	0,189	0,187
−0,7	0,242	0,239	0,236	0,233	0,230	0 227	0,224	0,221	0,218	0,215
−0,6	0,274	0,271	0,268	0,264	0,261	0,258	0,255	0,251	0,248	0,245
−0,5	0,309	0,305	0,302	0,298	0,295	0,291	0,288	0,284	0,281	0,278
−0,4	0,345	0,341	0,337	0,334	0,330	0,326	0,323	0,319	0,316	0,312
−0,3	0,382	0,378	0,374	0,371	0,367	0,363	0,359	0,356	0,352	0,348
−0,2	0,421	0,417	0,413	0,409	0,405	0,401	0,397	0,394	0,390	0,386
−0,1	0,460	0,456	0,452	0,448	0,444	0,440	0,436	0,433	0,429	0,425
0,0	0,500	0,496	0,492	0,488	0,484	0,480	0,476	0,472	0,468	0,464
0,0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
1	0,841	0,864	0,885	0,903	0,919	0,933	0,945	0,955	0,964	0,971
2	0,977	0,982	0,986	0,989	0,9918	0,9938	0,9953	0,9965	0,9974	0,9981
3	0,9987	$0,9^303$	$0,9^331$	$0,9^352$	$0,9^366$	$0,9^377$	$0,9^384$	$0,9^389$	$0,9^428$	$0,9^452$

$*$ 0.0^3 означает 0.000: 0.9^3 означает 0,999 и т. д.

t-распределение Стьюдента (см. § 8 гл. III)*

ν	$P\{\lvert\tilde{t}\rvert \geqslant t\}$						ν	$P\{\lvert\tilde{t}\rvert \geqslant t\}$					
	0,10	0,05	0,02	0,01	0,002	0,001		0,10	0,05	0,02	0,01	0,002	0,001
	$P\{\tilde{t} \geqslant t\}$ или $P\{\tilde{t} < -t\}$							$P\{\tilde{t} \geqslant t\}$ или $P\{\tilde{t} < -t\}$					
	0,05	0,025	0,01	0,005	0,001	0,0005		0,05	0,025	0,01	0,005	0,001	0,0005
1	6,31	12,7	31,8	63,7	31,8	63,7	20	1,72	2,09	2,53	2,85	3,55	3,85
2	2,92	4,30	6,96	9,92	22,3	31,6	22	1,72	2,07	2,51	2,82	3,51	3,79
3	2,35	3,18	4,54	5,84	10,2	12,9	24	1,71	2,06	2,49	2,80	3,47	3,75
4	2,13	2,78	3,75	4,60	7,17	8,61	26	1,71	2,06	2,48	2,78	3,44	3,71
5	2,02	2,57	3,36	4,03	5,89	6,87	28	1,70	2,05	2,47	2,76	3,41	3,67
6	1,94	2,45	3,14	3,71	5,21	5,96	30	1,70	2,04	2,46	2,75	3,39	3,65
7	1,89	2,36	3,00	3,50	4,79	5,41	32	1,69	2,04	2,45	2,74	3,37	3,62
8	1,86	2,31	2,90	3,36	4,50	5,04	35	1,69	2,03	2,44	2,72	3,34	3,59
9	1,83	2,26	2,82	3,25	4,30	4,78	40	1,68	2,02	2,42	2,70	3,31	3,55
10	1,81	2,23	2,76	3,17	4,14	4,59	50	1,68	2,01	2,40	2,68	3,26	3,50
11	1,80	2,20	2,72	3,11	4,02	4,44	60	1,67	2,00	2,39	2,66	3,23	3,46
12	1,78	2,18	2,68	3,05	3,93	4,32	80	1,66	1,99	2,37	2,64	3,20	3,42
13	1,77	2,16	2,65	3,01	3,85	4,22	100	1,66	1,98	2,36	2,63	3,17	3,39
14	1,76	2,14	2,62	2,98	3,79	4,14	150	1,66	1,98	2,35	2,61	3,15	3,36
15	1,75	2,13	2,60	2,95	3,73	4,07	300	1,65	1,97	2,34	2,59	3,12	3,32
16	1,75	2,12	2,58	2,92	3,69	4,02	1000	1,65	1,96	2,33	2,58	3,10	3,30
17	1,74	2,11	2,57	2,90	3,65	3,97	∞	1,64	1,96	2,33	2,58	3,09	3,29
18	1,73	2,10	2,55	2,88	3,61	3,92							
19	1,73	2,09	2,54	2,86	3,58	3,88							

* В последней строке даны значения нормированной нормальной случайной величины $\tilde{t}(\infty) = \tilde{u} \sim N(0;1)$.

F-распределение Снедекора–Фишера (см. § 9 гл. III)

а) $P\{\widetilde{F} \geqslant F\} = 0{,}05$

ν_2	ν_1														
	1	2	3	4	5	6	8	10	15	20	30	60	150	500	∞
1	161	200	216	225	230	234	239	242	246	248	250	252	253	254	254
2	18,5	19,0	19,2	19,2	19,3	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,85	8,79	8,70	8,66	8,62	8,57	8,55	8,53	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96	5,86	5,80	5,75	5,69	5,65	5,64	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74	4,62	4,56	4,50	4,43	4,39	4,37	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06	3,94	3,87	3,81	3,74	3,70	3,68	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,64	3,51	3,44	3,38	3,30	3,26	3,24	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,35	3,22	3,15	3,08	3,01	2,96	2,94	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,14	3,01	2,94	2,86	2,79	2,74	2,72	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,98	2,85	2,77	2,70	2,62	2,57	2,55	2,54
12	4,75	3,89	3,49	3,26	3,11	3,00	2,85	2,75	2,62	2,54	2,47	2,38	2,33	2,31	2,30
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,60	2,46	2,39	2,31	2,22	2,17	2,14	2,13
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,49	2,35	2,28	2,19	2,11	2,05	2,02	2,01
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,41	2,27	2,19	2,11	2,02	1,96	1,93	1,92
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,35	2,20	2,12	2,04	1,95	1,89	1,86	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,30	2,15	2,07	1,98	1,89	1,83	1,80	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,25	2,11	2,03	1,94	1,84	1,78	1,74	1,73
26	4,23	3,37	2,98	2,74	2,59	2,47	2,32	2,22	2,07	1,99	1,90	1,80	1,74	1,70	1,69
28	4,20	3,34	2,95	2,71	2,56	2,45	2,29	2,19	2,04	1,96	1,87	1,77	1,70	1,67	1,65
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,16	2,01	1,93	1,84	1,74	1,67	1,64	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,08	1,92	1,84	1,74	1,64	1,56	1,53	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,10	1,99	1,84	1,75	1,65	1,53	1,45	1,41	1,39
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,91	1,75	1,65	1,55	1,42	1,33	1,27	1,25
300	3,87	3,03	2,63	2,40	2,24	2,13	1,97	1,86	1,70	1,61	1,50	1,36	1,26	1,19	1,15
1 000	3,85	3,00	2,61	2,38	2,22	2,11	1,95	1,84	1,68	1,58	1,47	1,33	1,22	1,13	1,08
∞	3,84	3,00	2,60	2,37	2,21	2,10	1,94	1,83	1,67	1,57	1,46	1,32	1,20	1,11	1,00

Продолжение табл. IV

б) $P\{\widetilde{F} \geqslant F\} = 0{,}025$

ν_2	ν_1														
	1	2	3	4	5	6	8	10	15	20	30	60	150	500	∞
1	648	800	864	900	922	937	957	969	985	993	1 001	1 010	1 015	1 017	1 018
2	38,5	39,0	32,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,5	14,4	14,3	14,2	14,1	14,0	13,9	13,9	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	8,98	8,84	8,66	8,56	8,46	8,36	8,30	8,27	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,76	6,62	6,42	6,33	6,23	6,12	6,06	6,03	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,60	5,46	5,27	5,17	5,07	4,96	4,89	4,86	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,90	4,76	4,57	4,47	4,36	4,25	4,19	4,16	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,43	4,30	4,10	4,00	3,89	3,78	3,72	3,68	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,10	3,96	3,77	3,67	3,56	3,45	3,38	3,35	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,85	3,72	3,52	3,42	3,31	3,20	3,13	3,09	3,08
12	6,55	5,10	4,47	4,12	3,89	3,73	3,51	3,37	3,18	3,07	2,96	2,85	2,78	2,74	2,72
14	6,30	4,86	4,24	3,89	3,66	3,50	3,29	3,15	2,95	2,84	2,73	2,61	2,54	2,50	2,49
16	6,12	4,69	4,08	3,73	3,50	3,34	3,12	2,99	2,79	2,68	2,57	2,45	2,37	2,33	2,32
18	5,98	4,56	3,95	3,61	3,38	3,22	3,01	2,87	2,67	2,56	2,44	2,32	2,24	2,20	2,19
20	5,87	4,46	3,86	3,51	3,29	3,13	2,91	2,77	2,57	2,46	2,35	2,22	2,14	2,10	2,09
22	5,79	4,38	3,78	3,44	3,22	3,05	2,84	2,70	2,50	2,39	2,27	2,14	2,06	2,02	2,00
24	5,72	4,32	3,72	3,38	3,15	2,99	2,78	2,64	2,44	2,33	2,21	2,08	2,00	1,95	1,94
26	5,66	4,27	3,67	3,33	3,10	2,94	2,73	2,59	2,39	2,28	2,16	2,03	1,94	1,90	1,88
28	5,61	4,22	3,63	3,29	3,06	2,90	2,69	2,55	2,34	2,23	2,11	1,98	1,89	1,85	1,83
30	5,57	4,18	3,59	3,25	3,03	2,87	2,65	2,51	2,33	2,20	2,07	1,94	1,85	1,81	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,53	2,39	2,18	2,07	1,94	1,80	1,71	1,66	1,64
60	5,29	3,93	3,34	3,01	2,79	2,63	2,41	2,27	2,06	1,94	1,82	1,67	1,56	1,51	1,48
125	5,15	3,80	3,22	2,89	2,67	2,51	2,30	2,15	1,94	1,82	1,68	1,52	1,40	1,34	1,30
300	5,08	3,74	3,16	2,83	2,61	2,45	2,23	2,09	1,88	1,75	1,62	1,45	1,31	1,23	1,18
1 000	5,04	3,70	3,13	2,80	2,58	2,42	2,20	2,06	1,85	1,72	1,58	1,41	1,26	1,16	1,09
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,19	2,05	1,83	1,71	1,57	1,39	1,24	1,13	1,00

в) $P\{\widetilde{F} \geqslant F\} = 0,01$

ν_2	ν_1														
	1	2	3	4	5	6	8	10	15	20	30	60	150	500	∞
1	4 052	5 000	5 403	5 625	5 764	5 859	5 981	6 056	6 157	6 209	6 261	6 313	6 345	6 361	6 366
2	98,5	99,0	99,2	99,2	99,3	99,3	99,4	99,4	99,4	99,4	99,5	99,5	99,5	99,5	99,5
3	34,1	30,8	29,5	28,7	28,2	27,9	27,5	27,2	26,9	26,7	26,5	26,3	26,2	26,1	26,1
4	21,2	18,0	16,7	16,0	15,5	15,2	14,8	14,5	14,2	14,0	13,8	13,7	13,5	13,5	13,5
5	16,3	13,3	12,1	11,4	11,0	10,7	10,3	10,1	9,72	9,55	9,38	9,20	9,09	9,04	9,02
6	13,7	10,9	9,78	9,15	8,75	8,47	8,10	7,87	7,56	7,40	7,23	7,06	6,95	6,90	6,88
7	12,2	9,55	8,45	7,85	7,46	7,19	6,84	6,62	6,31	6,16	5,99	5,82	5,72	5,67	5,65
8	11,3	8,65	7,59	7,01	6,63	6,37	6,03	5,81	5,52	5,36	5,20	5,03	4,92	4,88	4,86
9	10,6	8,02	6,99	6,42	6,06	5,80	5,47	5,26	4,96	4,81	4,65	4,48	4,38	4,33	4,31
10	10,0	7,56	6,55	5,99	5,64	5,39	5,06	4,85	4,56	4,41	4,25	4,08	3,97	3,93	3,91
12	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,30	4,01	3,86	3,70	3,54	3,43	3,38	3,36
14	8,86	6,51	5,56	5,04	4,70	4,46	4,14	3,94	3,66	3,51	3,35	3,18	3,07	3,03	3,00
16	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,69	3,41	3,26	3,10	2,93	2,82	2,78	2,75
18	8,29	6,01	5,09	4,58	4,25	4,01	3,71	3,51	3,23	3,08	2,92	2,75	2,64	2,59	2,57
20	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,37	3,09	2,94	2,78	2,61	2,50	2,44	2,42
22	7,95	5,72	4,85	4,31	3,99	3,76	3,45	3,26	2,98	2,83	2,67	2,50	2,38	2,33	2,31
24	7,85	5,61	4,75	4,25	3,90	3,67	3,36	3,17	2,89	2,74	2,58	2,40	2,29	2,23	2,21
26	7,72	5,23	4,64	4,14	3,82	3,59	3,29	3,09	2,82	2,66	2,50	2,33	2,21	2,15	2,13
28	7,64	5,45	4,57	4,07	3,75	3,53	3,23	3,03	2,75	2,60	2,44	2,26	2,14	2,09	2,06
30	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,98	2,70	2,55	2,39	2,21	2,09	2,03	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	2,99	2,80	2,52	2,37	2,20	2,02	1,90	1,84	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,82	2,63	2,35	2,20	2,03	1,84	1,70	1,63	1,60
125	6,84	4,78	3,94	3,47	3,17	2,95	2,66	2,47	2,19	2,03	1,85	1,65	1,49	1,41	1,37
300	6,72	4,68	8,85	3,38	3,08	2,86	2,57	2,38	2,10	1,94	1,76	1,55	1,36	1,28	1,22
1 000	6,66	4,63	3,80	3,34	3,04	2,82	2,53	2,34	2,06	1,90	1,72	1,50	1,35	1,19	1,11
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,51	2,32	2,04	1,88	1,70	1,47	1,29	1,15	1,00

г) $P\{\widetilde{F} \geqslant F\} = 0{,}005^*$

ν_2	ν_1														
	1	2	3	4	5	6	8	10	15	20	30	60	150	500	∞
1	162^2	200^2	216^2	225^2	231^2	234^2	239^2	242^2	246^2	248^2	250^2	253^2	254^2	254^2	255^2
2	199	199	199	199	199	199	199	199	199	199	199	199	199	200	200
3	55,6	49,8	47,5	46,2	45,4	44,8	44,1	43,7	43,1	42,8	42,5	42,1	42,0	41,9	41,8
4	31,3	26,3	24,3	23,2	22,5	22,0	21,4	21,0	20,4	20,2	19,9	19,6	19,4	19,4	19,3
5	22,8	18,3	16,5	15,6	14,9	14,5	14,0	13,6	13,1	12,9	12,7	12,4	12,3	12,2	12,1
6	18,6	14,5	12,9	12,0	11,5	11,1	10,6	10,3	9,81	9,59	9,36	9,12	8,98	8,91	8,88
7	16,2	12,4	10,9	10,1	9,52	9,16	8,68	8,38	7,97	7,75	7,53	7,31	7,17	7,10	7,08
8	14,7	11,0	9,60	8,81	8,30	7,95	7,50	7,21	6,81	6,61	6,40	6,18	6,08	5,98	5,95
9	13,6	10,1	8,72	7,96	7,47	7,13	6,69	6,42	6,03	5,83	5,62	5,41	5,28	5,21	5,19
10	12,8	9,43	8,08	7,34	6,87	6,54	6,1	5,85	5,47	5,27	5,07	4,86	4,73	4,67	4,64
12	11,8	8,51	7,23	6,52	6,07	5,76	5,3	5,09	4,72	4,53	4,33	4,12	3,99	3,93	3,90
14	11,1	7,92	6,68	6,00	5,56	5,26	4,86	4,60	4,25	4,06	3,86	3,66	3,53	3,46	3,44
16	10,6	7,51	6,30	5,64	5,21	4,91	4,52	4,27	3,92	3,73	3,54	3,33	3,20	3,14	3,11
18	10,2	7,21	6,03	5,37	4,96	4,66	4,28	4,03	3,68	3,50	3,30	3,10	2,96	2,90	2,87
20	9,94	6,99	5,82	5,17	4,76	4,47	4,09	3,85	3,50	3,32	3,12	2,92	2,78	2,72	2,69
22	9,73	6,81	5,65	5,02	4,61	4,32	3,94	3,70	3,36	3,18	2,98	2,77	2,64	2,57	2,55
24	9,55	6,66	5,52	4,89	4,49	4,20	3,83	3,59	3,25	3,06	2,87	2,66	2,52	2,46	2,43
26	9,41	6,54	5,41	4,79	4,38	4,10	3,73	3,49	3,15	2,97	2,77	2,56	2,43	2,36	2,33
28	9,28	6,44	5,32	4,70	4,30	4,02	3,65	3,41	3,07	2,89	2,69	2,48	2,35	2,28	2,25
30	9,18	6,35	5,24	4,62	4,23	3,95	3,58	3,34	3,01	2,82	2,63	2,42	2,28	2,21	2,18
40	8,83	6,07	4,98	4,37	3,99	3,71	3,35	3,12	2,78	2,60	2,40	2,18	2,04	1,96	1,93
60	8,49	5,80	4,73	4,14	3,76	3,49	3,13	2,90	2,57	2,39	2,19	1,96	1,81	1,73	1,69
125	8,17	5,53	4,49	3,91	3,54	3,28	2,93	2,70	2,37	2,18	1,98	1,74	1,56	1,47	1,42
300	8,00	5,39	4,37	3,80	3,43	3,17	2,81	2,59	2,26	2,07	1,87	1,61	1,43	1,31	1,25
1 000	7,92	5,33	4,31	3,74	3,37	3,11	2,77	2,54	2,21	2,02	1,81	1,56	1,36	1,22	1,13
∞	7,88	5,30	4,28	3,72	3,34	3,09	2,74	2,52	2,19	2,00	1,79	1,53	1,32	1,17	1,00

$*$ 162^2 означает $162 \cdot 10^2$.

$\partial)\ P\{\widetilde{F} \geqslant F\} = 0{,}001^*$

ν_2	ν_1														
	1	2	3	4	5	6	8	10	15	20	30	60	150	500	∞
1	405^3	500^3	540^3	563^3	576^3	586^3	598^3	606^3	616^3	621^3	626^3	630^3	633^3	636^3	637^3
2	999	999	999	999	999	999	999	999	999	999	1 000	1 000	1 000	1 000	1 000
3	167	149	141	137	135	133	131	129	127	126	125	125	124	124	124
4	74,1	61,3	56,2	53,4	53,7	50,5	49,0	48,1	46,8	46,1	45,4	44,9	44,5	44,1	44,1
5	47,2	37,1	33,2	31,1	29,8	28,8	27,6	26,9	25,9	25,4	24,9	24,4	24,1	23,8	23,8
6	35,5	27,0	23,7	21,9	20,8	20,0	19,0	18,4	17,6	17,1	16,7	16,3	16,0	15,8	15,8
7	29,3	21,7	18,8	17,2	16,2	15,5	14,6	14,1	13,3	12,9	12,5	12,2	11,9	11,7	11,7
8	25,4	18,5	15,8	14,4	13,5	12,9	12,0	11,5	10,8	10,5	10,1	9,80	9,57	9,39	9,33
9	22,9	16,4	13,9	12,6	11,7	11,1	10,4	0,89	9,24	8,90	8,55	8,26	8,04	7,86	7,81
10	21,0	14,9	12,6	11,3	10,5	9,92	9,2	8,75	8,13	7,80	7,47	7,19	6,98	6,81	6,76
12	18,6	13,0	10,8	9,63	8,89	8,38	7,7	7,29	6,71	6,40	6,09	5,83	5,63	5,46	5,42
14	17,1	11,8	9,73	8,62	7,92	7,43	6,8	6,40	5,85	5,56	5,25	5,00	4,80	4,64	4,60
16	16,1	11,0	9,00	7,94	7,27	6,81	6,19	5,81	5,27	4,99	4,70	4,45	4,26	4,10	4,06
18	15,4	10,4	8,49	7,46	6,81	6,35	5,76	5,39	4,87	4,59	4,30	4,06	3,87	3,71	3,67
20	14,8	9,95	8,10	7,30	6,46	6,02	5,44	5,08	4,56	4,29	4,00	3,77	3,58	3,42	3,38
22	14,4	9,61	7,80	6,81	6,19	5,76	5,19	4,83	4,33	4,06	3,78	3,53	3,34	3,19	3,15
24	14,0	9,34	7,55	6,59	5,98	5,55	4,99	4,64	4,14	3,87	3,59	3,35	3,16	3,01	2,97
26	13,7	9,12	7,36	6,41	5,80	5,38	4,83	4,48	3,99	3,72	3,44	3,20	3,01	2,86	2,82
28	13,5	8,93	7,19	6,25	5,66	5,24	4,69	4,35	3,86	3,60	3,32	3,08	2,89	2,73	2,69
30	13,3	8,77	7,05	6,12	5,53	5,12	4,58	4,24	3,75	3,49	3,22	2,98	2,79	2,63	2,59
40	12,6	8,25	6,60	5,70	5,13	4,73	4,21	3,87	3,40	3,15	2,87	2,64	2,44	2,28	2,23
60	12,0	7,76	6,17	5,31	4,76	4,37	3,87	3,54	3,08	2,83	2,55	2,31	2,11	1,93	1,89
100	11,5	7,41	5,85	5,01	4,48	4,11	3,61	3,30	2,84	2,59	2,32	2,07	1,87	1,68	1,62
200	11,2	7,15	5,64	4,81	4,29	3,92	3,43	3,12	2,67	2,42	2,15	1,90	1,68	1,46	1,39
500	11,0	7,01	5,51	4,69	4,18	3,82	3,33	3,02	2,58	2,33	2,05	1,80	1,57	1,32	1,23
∞	10,8	6,93	5,42	4,62	4,10	3,74	3,27	2,96	2,51	2,27	1,99	1,73	1,49	1,21	1,00

$*\ 405^3$ означает $405 \cdot 10^3$.

Окончание табл. IV

e) $P\{\widetilde{F} \geqslant F\} = 0{,}0005^*$

ν_2	ν_1														
	1	2	3	4	5	6	8	10	15	20	30	60	150	500	∞
1	162^4	200^4	216^4	225^4	231^4	234^4	239^4	242^4	246^4	248^4	250^4	252^4	253^4	254^4	254^4
2	200^1	200^1	200^1	200^1	200^1	200^1	200^1	200^1	200^1	200^1	200^1	200^1	200^1	200^1	200^1
3	266	237	225	218	214	211	208	206	203	201	199	198	197	196	196
4	106	87,4	80,1	76,1	73,6	71,9	69,7	68,3	66,5	65,5	64,6	63,8	63,2	62,7	62,6
5	63,6	49,8	44,4	41,5	39,7	38,5	36,9	35,9	34,6	33,9	33,1	32,5	32,1	31,7	31,6
6	46,1	34,8	30,4	28,1	26,6	25,6	24,3	23,5	22,4	21,9	21,4	20,9	20,5	20,2	20,1
7	37,0	27,2	23,5	21,4	20,2	19,3	18,2	17,5	16,5	16,0	15,5	15,1	14,7	14,5	14,4
8	31,6	22,8	19,4	17,6	16,4	15,7	14,6	14,0	13,1	12,7	12,2	11,8	11,6	11,3	11,3
9	28,0	19,9	16,8	15,1	14,1	13,3	12,4	11,8	11,0	10,6	10,2	9,80	9,53	9,32	9,26
10	25,5	17,9	15,0	13,4	12,4	11,8	10,9	10,3	9,56	9,16	8,75	8,42	8,16	7,96	7,90
12	22,2	15,3	12,7	11,2	10,4	9,74	8,94	8,43	7,74	7,37	7,00	6,68	6,45	6,25	6,20
14	20,2	13,7	11,3	9,95	9,11	8,53	7,78	7,31	6,65	6,31	5,95	5,66	5,43	5,24	5,19
16	18,9	12,7	10,3	9,08	8,29	7,74	7,02	6,57	5,94	5,61	5,27	4,98	4,76	4,57	4,52
18	17,9	11,9	9,69	8,47	7,71	7,18	6,48	6,05	5,44	5,12	4,78	4,50	4,28	4,10	4,06
20	17,2	11,4	9,20	8,02	7,28	6,76	6,08	5,66	5,07	4,75	4,42	4,15	3,93	3,75	3,70
22	16,6	11,0	8,82	7,67	6,94	6,44	5,78	5,36	4,79	4,47	4,15	3,88	3,66	3,48	3,44
24	16,2	10,6	8,52	7,39	6,68	6,18	5,54	5,13	4,55	4,25	3,93	3,66	3,44	3,27	3,22
26	15,8	10,3	8,27	7,16	6,46	5,98	5,34	4,94	4,37	4,07	3,75	3,48	3,27	3,09	3,04
28	15,5	10,1	8,07	6,98	6,28	5,80	5,18	4,78	4,22	3,92	3,61	3,34	3,13	2,95	2,90
30	15,2	9,90	7,90	6,82	6,14	5,66	5,04	4,65	4,10	3,80	3,48	3,22	3,00	2,82	2,78
40	14,4	9,25	7,33	6,30	5,64	5,19	4,59	4,21	3,68	3,39	3,08	2,82	2,60	2,41	2,37
60	13,6	8,65	6,81	5,82	5,20	4,76	4,18	3,82	3,30	3,02	2,71	2,45	2,23	2,03	1,98
100	13,0	8,21	6,42	5,47	4,87	4,44	3,89	3,54	3,03	2,75	2,44	2,18	1,95	1,74	1,67
200	12,5	7,90	6,16	5,23	4,64	4,23	3,68	3,34	2,83	2,56	2,25	1,98	1,74	1,50	1,42
500	12,3	7,72	6,01	5,09	4,51	4,10	3,56	3,21	2,72	2,45	2,14	1,87	1,61	1,34	1,24
∞	12,1	7,60	5,91	5,00	4,42	4,02	3,48	3,14	2,65	2,37	2,07	1,79	1,52	1,22	1,00

$*$ 162^4 означает $162 \cdot 10^4$; 200^1 означает $200^1 0^1$ и т. д.

Таблица V

Распределение χ^2 (см. § 7 гл. III)*

ν	$P\{\widetilde{\chi}^2 \geqslant \chi^2\}$							
	0,995	0,975	0,1	0,05	0,025	0,01	0,005	0,001
1	$0{,}0^4 4$	0,001	2,71	3,84	5,02	6,63	7,88	10,8
2	0,010	0,051	4,61	5,99	7,38	9,21	10,6	13,8
3	0,072	0,22	6,25	7,81	9,35	11,3	12,8	16,3
4	0,21	0,48	7,78	9,49	11,1	13,3	14,9	18,5
5	0,41	0,83	9,24	11,1	12,8	15,1	16,7	20,5
6	0,68	1,24	10,6	12,6	14,4	16,8	18,5	22,5
7	0,99	1,69	12,0	14,1	16,0	18,5	20,3	24,3
8	1,34	2,18	13,4	15,5	17,5	20,1	22,0	26,1
9	1,73	2,70	14,7	16,9	19,0	21,7	23,6	27,9
10	2,16	3,25	16,0	18,3	20,5	23,2	25,2	29,6
11	2,60	3,82	17,3	19,7	21,9	24,7	26,8	31,3
12	3,07	4,40	18,5	21,0	23,3	26,2	28,3	32,9
13	3,57	5,01	19,8	22,4	24,7	27,7	29,8	34,5
14	4,07	5,63	21,1	23,7	26,1	29,1	31,3	36,1
15	4,60	6,26	22,3	25,0	27,5	30,6	32,8	37,7
16	5,14	6,91	23,5	26,3	28,8	32,0	34,3	39,3
17	5,70	7,56	24,8	27,6	30,2	33,4	35,7	40,8
18	6,26	8,23	26,0	28,9	31,5	34,8	37,2	42,3
19	6,84	8,91	27,2	30,1	32,9	36,2	38,6	43,8
20	7,43	9,59	28,4	31,4	34,2	37,6	40,0	45,3
21	8,03	10,3	29,6	32,7	35,5	38,9	41,4	46,8
22	8,64	11,0	30,8	33,9	36,8	40,3	42,8	48,3
23	9,26	11,7	32,0	35,2	38,1	41,6	44,2	49,7
24	9,89	12,4	33,2	36,4	39,4	43,0	45,6	51,2
25	10,5	13,1	34,4	37,7	40,6	44,3	46,9	52,6
26	11,2	13,8	35,6	38,9	41,9	45,6	48,3	54,1
27	11,8	14,6	36,7	40,1	43,2	47,0	49,6	55,5
28	12,5	15,3	37,9	41,3	44,5	48,3	51,0	56,9
29	13,1	16,0	39,1	42,6	45,7	49,6	52,3	58,3
30	13,8	16,8	40,3	43,8	47,0	50,9	53,7	59,7
50	28,0	32,4	63,2	67,5	71,4	76,2	79,5	86,7
100	67,3	74,2	118,5	124,3	129,6	135,8	140,2	149,4

* $0{,}0^4 4$ означает 0,00004

Таблица VI

Точный критерий Фишера (см. § 4 гл. VI. Для случаев, когда $a_\alpha \neq a_{\alpha/2}$, они даны в виде дроби: $a_\alpha/a_{\alpha/2}$

Вид таблицы	$P\{\widetilde{a} \geqslant a\}$			Вид таблицы
	0,05	0,01	0,001	
a 0 0 1	19	99	999	a 0 0 1
a 1 0 1	38	198	1 998	a 0 1 1
a 2 0 1	57	297	2 997	a 0 2 1
a 1 1 1	77	397	3 997	a 1 1 1
a 2 1 1	115	595	5 995	a 1 2 1
a 0 0 2	5	13	44	a 0 0 2
a 1 0 2	9	22	75	a 0 1 2
a 2 0 2	12	32	107	a 0 2 2
a 1 1 2	16	39	131	a 1 1 2
a 2 1 2	22	55	185	a 1 2 2
a 0 0 3	3/4	7	17	a 0 0 3
a 1 0 3	5	11	26	a 0 1 3
a 2 0 3	7	15	36	a 0 2 3
a 1 1 3	9	17	42	a 1 1 3
a 2 1 3	12	24	57	a 1 2 3
a 0 0 4	3	5	10	a 0 0 4
a 1 0 4	4	8	16	a 0 1 4

Окончание табл. VI

Вид таблицы	$P\{\widetilde{a} \geqslant a\}$ 0,05	0,01	0,001	Вид таблицы	Вид таблицы	$P\{\widetilde{a} \geqslant a\}$ 0,05	0,01	0,001	Вид таблицы
a 2 0 4	5/7	10	21	a 0 2 4	a 0 0 8	2	3	5	a 0 0 8
a 1 1 4	6/7	11	24	a 1 1 4	a 1 0 8	3	4	7/8	a 0 1 8
a 2 1 4	8/9	15	31	a 1 2 4	a 2 0 8	3	5	9	a 0 2 8
a 0 0 5	2	4	8	a 0 0 5	a 1 1 8	4	6	9/10	a 1 1 8
a 1 0 5	3	6/7	11	a 0 1 5	a 2 1 8	4/5	7/8	12	a 1 2 8
a 2 0 5	4/5	8	15	a 0 2 5	a 0 0 9	2	3	5	a 0 0 9
a 1 1 5	5/6	9	17	a 1 1 5	a 1 0 9	2	4	6	a 0 1 9
a 2 1 5	7	11	21	a 1 2 5	a 2 0 9	3	5	8	a 0 9 0
a 0 0 6	2	4	7	a 0 0 6	a 1 1 9	3	5	8/10	a 1 1 9
a 1 0 6	3	5/6	9	a 0 1 6	a 2 1 9	4	6/7	10/11	a 1 2 9
a 2 0 6	4/5	6	12	a 0 2 6	a 1 0 7	3	5	8/9	a 0 1 7
a 1 1 6	4/5	7/8	13	a 1 1 6	a 2 0 7	3	6	10	a 0 2 7
a 2 1 6	6	9/10	17	a 1 2 6	a 1 1 7	4	6/8	11	a 1 1 7
a 0 0 7	2	3	6	a 0 0 7	a 2 1 7	5	8/9	14	a 1 2 7

Таблица VII

Непараметрические доверительные пределы для медианы (см. § 7 гл. V)

b	$P\{x_{(b)} < \zeta < x_{(n-b+1)}\}$					
	0,90	0,95	0,98	0,99	0,998	0,999
	$P\{-\infty < \zeta < x_{(n-b+1)}\}$ или $P\{x_{(b)} < \zeta < \infty\}$					
	0,95	0,975	0,99	0,995	0,999	0,9995
1	5–7	6–8	7–10	8–11	10–13	11–14
2	8–10	9–11	11–13	12–14	14–17	15–18
3	11–12	12–14	14–16	15–17	18–20	19–21
4	13–15	15–16	17–18	18–20	21–23	22–24
5	16–17	17–19	19–21	21–23	24–26	25–27
6	18–20	20–22	22–24	24–25	27–29	28–30
7	21–22	23–24	25–26	26–28	30–32	31–33
8	23–25	25–27	27–29	29–31	33–34	34–36
9	26–27	28–29	30–32	32–33	35–37	37–39
10	28–29	30–32	33–34	34–36	38–40	40–41
11	30–32	33–34	35–37	37–38	41–42	42–44
12	33–34	35–36	38–39	39–41	43–45	45–47
13	35–36	37–39	40–41	42–43	46–48	48–49
14	37–39	40–41	42–44	44–46	49–50	50–52
15	40–41	42–43	45–46	47–48	51–53	53–55
16	42–43	44–46	47–49	49–51	54–55	56–57
17	44–46	47–48	50–51	52–53	56–58	58–60
18	47–48	49–50	52–53	54–56	59–60	61–62
19	49–50	51–53	54–56	57–58	61–63	63–65
20	51–52	54–55	57–58	59–60	64–65	66–67
21	53–55	56–57	59–61	61–63	66–68	68–70
22	56–57	58–60	62–63	64–65	69–70	71–72
23	58–59	61–62	64–65	66–68	71–73	73–75
24	60–61	63–64	66–68	69–70	74–75	76–77
25	62–64	65–66	69–70	71–72	76–78	78–80
26	65–66	67–69	71–72	73–75	79–80	81–82
27	67–68	70–71	73–75	76–77	81–82	83–85
28	69–70	72–73	76–77	78–79	83–85	86–87
29	71–73	74–76	78–79	80–82	86–87	88–89
30	74–75	77–78	80–81	83–84	88–90	90–92

Таблица VIII

Критерий Вилкоксона–Манна–Уитни (см. § 7 гл. VI)

n_1	n_2	$\alpha = P\{\lvert\widetilde{U}\rvert \leqslant U\}$					
		0,1	0,05	0,02	0,01	0,002	0,001
		$\alpha/2 = P\{\lvert\widetilde{U}\rvert \leqslant U\}$					
		0,05	0,025	0,01	0,005	0,001	0,0005
2	5	0					
	6	0					
	7	0					
	8	1	0				
	9	1	0				
	10	1	0				
3	3	0	—				
	4	0	—				
	5	1	0				
	6	2	1				
	7	2	1	0			
	8	3	2	0			
	9	4	2	1	0		
	10	4	3	1	0		
4	4	1	0	—	—		
	5	2	1	0	—		
	6	3	2	1	0		
	7	4	3	1	0		
	8	5	4	2	1		
	9	6	4	3	1		
	10	7	5	3	2	0	
5	5	4	2	1	0	—	
	6	5	3	2	1	—	
	7	6	5	3	1	—	
	8	8	6	4	2	0	
	9	9	7	5	3	1	0
	10	11	8	6	4	1	0
6	6	7	5	3	2	—	—
	7	8	6	4	3	0	—
	8	10	8	6	4	1	0
	9	12	10	7	5	2	1
	10	14	11	8	6	3	2

Окончание табл. VIII

n_1	n_2	$\alpha = P\{\lvert\widetilde{U}\rvert \leqslant U\}$					
		0,1	0,05	0,02	0,01	0,002	0,001
		$\alpha/2 = P\{\lvert\widetilde{U}\rvert \leqslant U\}$					
		0,05	0,025	0,01	0,005	0,001	0,0005
7	7	11	8	6	4	1	0
	8	13	10	7	6	2	1
	9	15	12	9	7	3	2
	10	17	14	11	9	5	3
8	8	15	13	9	7	4	2
	9	18	15	11	9	5	4
	10	20	17	13	11	6	5
9	9	21	17	14	11	7	5
	10	24	20	16	13	8	7
10	10	27	23	19	16	10	8

Таблица IX

Парный критерий Вилкоксона (см. § 8 гл. VI)

N	$\alpha = P\{\widetilde{W} \leqslant W\}$					
	0,1	0,05	0,02	0,01	0,002	0,001
	$\alpha/2 = P\{\widetilde{W} \leqslant W\}$					
	0,05	0,025	0,01	0,005	0,001	0,0005
5	0					
6	2	0				
7	3	2	0			
8	5	3	1	0		
9	8	5	3	1		
10	10	8	5	3	0	
11	13	10	7	5	1	0
12	17	13	9	7	2	1
13	21	17	12	9	4	2
14	25	21	15	12	6	4
15	30	25	19	15	8	6

Окончание табл. IX

N	$\alpha = P\{\widetilde{W} \leqslant W\}$					
	0,1	0,05	0,02	0,01	0,002	0,001
	$\alpha/2 = P\{\widetilde{W} \leqslant W\}$					
	0,05	0,025	0,01	0,005	0,001	0,0005
16	35	29	23	19	11	8
17	41	34	27	23	14	11
18	47	40	32	27	18	14
19	53	46	37	32	21	18
20	60	52	43	37	26	21
21	67	58	49	42	30	25
22	75	65	55	48	35	30
23	83	73	62	54	40	35
24	91	81	69	61	45	40
25	100	89	76	68	51	45
26	110	98	84	75	58	51
27	119	107	92	83	64	57
28	130	116	101	91	71	64
29	140	126	110	100	79	71
30	151	137	320	109	86	78
31	163	147	130	118	94	86
32	175	159	140	128	103	94
33	187	170	151	138	112	102
34	200	182	162	148	121	111
35	213	195	173	159	131	120
36	227	208	185	171	141	130
37	241	221	198	182	151	140
38	256	235	211	194	162	150
39	271	249	224	207	178	161
40	286	264	238	220	185	172

Таблица X

Критерий Крускала–Уоллиса (см. § 7 гл. VIII)

n_1	n_2	n_3	$P\{\widetilde{H} \geqslant H\}$		n_1	n_2	n_3	$P\{\widetilde{H} \geqslant H\}$	
			0,05	0,01				0,05	0,01
1	2	5	5,00	—	2	6	6	5,41	7,47
1	2	6	5,14	—	3	3	3	5,60	7,20
1	3	3	5,14	—	3	3	4	5,73	6,75
1	3	4	5,20	—	3	3	5	5,65	7,08
1	3	5	4,96	—	3	3	6	5,62	7,41
1	4	4	4,97	6,67	3	4	4	5,60	7,14
1	4	5	4,99	6,95	3	4	5	5,63	7,45
1	5	5	5,13	7,31	3	4	6	5,61	7,50
2	2	3	4,71	—	3	5	5	5,71	7,54
2	2	4	5,33	—	3	5	6	5,60	7,59
2	2	5	5,16	6,53	3	6	6	5,63	7,73
2	2	6	5,35	6,65	4	4	4	5,69	7,65
2	3	3	5,36	—	4	4	5	5,62	7,76
2	3	4	5,44	6,44	4	4	6	5,68	7,80
2	3	5	5,25	6,82	4	5	5	5,64	7,77
2	3	6	5,35	6,97	4	5	6	5,66	7,94
2	4	4	5,45	7,04	4	6	6	5,72	8,00
2	4	5	5,27	7,12	5	5	5	5,78	8,00
2	4	6	5,34	7,34	5	5	6	5,73	8,03
2	5	5	5,34	7,27	5	6	6	5,76	8,12
2	5	6	5,34	7,38	6	6	6	5,80	8,22

Таблица XI

Распределение выборочного коэффициента корреляции Пирсона при $\varrho = 0$ (см. § 7 гл. IX)

ν	$P\{\lvert\widetilde{r}\rvert \leqslant r\}$					
	0,1	0,05	0,02	0,01	0,002	0,001
	$P\{\widetilde{r} \geqslant r\}$ или $P\{\widetilde{r} \leqslant -r\}$					
	0,05	0,025	0,01	0,005	0,001	0,0005
1	0,99	1,00	1,00	1,00	1,00	1,00
2	0,90	0,95	0,98	0,99	1,00	1,00
3	0,81	0,88	0,93	0,96	0,99	0,99
4	0,73	0,81	0,88	0,92	0,96	0,97
5	0,67	0,75	0,83	0,87	0,94	0,95
6	0,62	0,71	0,79	0,83	0,91	0,92
7	0,58	0,67	0,75	0,80	0,88	0,90
8	0,55	0,63	0,72	0,76	0,85	0,87
9	0,52	0,60	0,69	0,73	0,82	0,85
10	0,50	0,58	0,66	0,71	0,80	0,82
11	0,48	0,55	0,63	0,68	0,77	0,80
12	0,46	0,53	0,61	0,66	0,75	0,78
13	0,44	0,51	0,59	0,64	0,73	0,76
14	0,43	0,50	0,57	0,62	0,71	0,74
15	0,41	0,48	0,56	0,61	0,69	0,73
16	0,40	0,47	0,54	0,59	0,68	0,71
17	0,39	0,46	0,53	0,58	0,66	0,69
18	0,38	0,44	0,52	0,56	0,65	0,68
19	0,37	0,43	0,50	0,55	0,64	0,67
20	0,36	0,42	0,49	0,54	0,62	0,65
25	0,32	0,38	0,45	0,49	0,57	0,60
30	0,30	0,35	0,41	0,45	0,53	0,55
35	0,28	0,32	0,38	0,42	0,49	0,52
40	0,26	0,30	0,36	0,39	0,46	0,49
45	0,24	0,29	0,34	0,37	0,44	0,47
50	0,23	0,27	0,32	0,35	0,42	0,44
60	0,21	0,25	0,30	0,33	0,39	0,41
70	0,20	0,23	0,27	0,30	0,36	0,38
80	0,18	0,22	0,25	0,28	0,34	0,36
90	0,17	0,21	0,24	0,27	0,32	0,34
100	0,16	0,20	0,23	0,25	0,30	0,32
200	0,12	0,14	0,16	0,18	0,22	0,23
300	0,10	0,11	0,13	0,15	0,18	0,19
500	0,07	0,09	0,10	0,12	0,14	0,15
1000	0,05	0,06	0,07	0,08	0,10	0,10

Таблица XII

Преобразование коэффициента корреляция Пирсона (см. § 8 гл. IX)

а) $z = \frac{1}{2} \ln \frac{1+r}{1-r}$

r	0	1	2	3	4	5	6	7	8	9
0,0	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,1	0,10	0,11	0,12	0,13	0,14	0,15	0,16	0,17	0,18	0,19
0,2	0,20	0,21	0,22	0,23	0,24	0,26	0,27	0,28	0,29	0,30
0,3	0,31	0,32	0,33	0,34	0,35	0,37	0,38	0,39	0,40	0,41
0,4	0,42	0,44	0,45	0,46	0,47	0,48	0,50	0,51	0,52	0,54
0,5	0,55	0,56	0,58	0,59	0,60	0,62	0,63	0,65	0,66	0,68
0,6	0,69	0,71	0,73	0,74	0,76	0,78	0,79	0,81	0,83	0,85
0,7	0,87	0,89	0,91	0,93	0,95	0,97	1,00	1,02	1,05	1,07
0,8	1,10	1,13	1,16	1,19	1,22	1,26	1,29	1,33	1,38	1,42
0,9	1,47	1,53	1,59	1,66	1,74	1,83	1,95	2,09	2,30	2,65
0,99	2,65	2,70	2,76	2,83	2,90	2,99	3,11	3,25	3,45	3,80

Продолжение табл. XII

б) $r = (e^{2^z} - 1)/(e^{2^z} + 1)$

z	0	1	2	3	4	5	6	7	8	9
0,0	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,1	0,10	0,11	0,12	0,13	0,14	0,15	0,16	0,17	0,18	0,19
0,2	0,20	0,21	0,22	0,23	0,24	0,24	0,25	0,26	0,27	0,28
0,3	0,29	0,30	0,31	0,32	0,33	0,34	0,35	0,35	0,36	0,37
0,4	0,38	0,39	0,40	0,41	0,41	0,42	0,43	0,44	0,45	0,45
0,5	0,46	0,48	0,49	0,49	0,50	0,51	0,52	0,52	0,53	0,54
0,6	0,54	0,55	0,56	0,56	0,57	0,58	0,59	0,59	0,60	0,60
0,7	0,61	0,62	0,62	0,63	0,64	0,64	0,64	0,65	0,65	0,66
0,8	0,66	0,67	0,68	0,68	0,69	0,69	0,70	0,70	0,71	0,71
0,9	0,72	0,72	0,73	0,73	0,74	0,74	0,74	0,75	0,75	0,76
1	0,76	0,80	0,83	0,86	0,89	0,91	0,92	0,94	0,95	0,96
2	0,96	0,97	0,98	0,98	0,98	0,99	0,99	0,991	0,993	0,994

Таблица XIII

Распределение коэффициента корреляции Спирмена (см. § 9 гл. IX)

n	$P\{\lvert\widetilde{r}_S\rvert \geqslant r_S\}$					
	0,1	0,05	0,02	0,01	0,002	0,001
	$P\{\widetilde{r}_S \geqslant r_S\}$ или $P\{\widetilde{r_S} \leqslant -r_S\}$					
	0,05	0,025	0,01	0,005	0,001	0,0005
4	1,00					
5	0,90	1,00	1,00			
6	0,83	0,89	0,94	1,00		
7	0,71	0,79	0,89	0,93	1,00	1,00
8	0,64	0,74	0,83	0,88	0,95	0,98
9	0,60	0,70	0,78	0,83	0,92	0,93
10	0,56	0,65	0,75	0,79	0,88	0,90
11	0,54	0,62	0,71	0,76	0,85	0,87
12	0,50	0,59	0,68	0,73	0,82	0,85
13	0,48	0,56	0,65	0,70	0,79	0,82
14	0,46	0,54	0,62	0,68	0,77	0,80
15	0,44	0,52	0,60	0,65	0,75	0,78
16	0,43	0,50	0,58	0,64	0,73	0,76
17	0,41	0,48	0,57	0,62	0,71	0,75
18	0,40	0,47	0,55	0,60	0,70	0,73
19	0,39	0,46	0,54	0,58	0,68	0,71
20	0,38	0,45	0,52	0,57	0,66	0,70
21	0,37	0,44	0,51	0,56	0,65	0,68
22	0,36	0,43	0,50	0,54	0,63	0,67
23	0,35	0,42	0,49	0,53	0,62	0,65
24	0,34	0,41	0,48	0,52	0,61	0,64
25	0,34	0,40	0,47	0,51	0,60	0,63
26	0,33	0,39	0,46	0,50	0,59	0,62
27	0,32	0,38	0,45	0,49	0,58	0,61
28	0,32	0,38	0,44	0,48	0,57	0,60
29	0,31	0,37	0,43	0,48	0,56	0,59
30	0,31	0,36	0,43	0,47	0,55	0,58
40	0,26	0,31	0,37	0,41	0,48	0,51
50	0,24	0,28	0,33	0,36	0,43	0,46
60	0,21	0,26	0,30	0,33	0,39	0,42
80	0,19	0,22	0,26	0,29	0,34	0,36
100	0,17	6,20	0,23	0,26	0,31	0,33

Приложение 2

Биометрические аспекты популяционной генетики

Чацкий

К статистике давно в душе питаю страсть я,
И геология внушает мне участье. . . [1].

Е. П. Ростопчина. Возврат Чацкого в Москву. Продолжение комедии А. С. Грибоедова «Горе от ума», 1856[2]

Генетика вообще, а популяционная генетика в особенности, являются наиболее математизированными биологическими дисциплинами. Их математизация настолько глубока, что подобно тому, как в физике самостоятельно существует математическая физика, так и в биологии самостоятельно существуют *математическая теория естественного отбора* и ее фундаментальный раздел — *математическая популяционная (эволюционная) генетика*. Их творцами по праву признаны Роналд Айлмер Фишер (Ronald Aylmer Fisher, 1890–1962), Джон Бердон Сандерсон Холдэйн (John Burdon Sanderson Haldane, 1892–1964) и Сьюэл Райт (Sewall Wright, 1889–1988).

Обсуждение этих дисциплин выходит за рамки настоящей книги, однако в практической, экспериментальной популяционной генетике невозможно обойтись без прикладных аспектов биометрического (статистического) анализа, изложению которых и посвящено настоящее Приложение. Написание подобных приложений достаточно традиционно для генетических учебников и руководств. Образцами могут служить «Принципы генетики» Э. Синнота, Л. Данна и Ф. Добржанского (1958), «Генетика» И. Гершковича (1968), «Введение в популяционную и эволюционную генетику» Ф. Айалы, «Современная генетика» Ф. Айалы и Дж. Кайгера (1988), «Генетика человека» Ф. Фогеля и А. Мотульского (1989; 1990) и др. По этим книгам можно проследить эволюцию биометрических методов, применяемых в генетике.

П.1. Основные принципы биометрического анализа

Статистический анализ данных, именуемый *прикладной статистикой* (в биологии за ним укрепилось название *биометрия*), является неотъем-

[1] Современным «чацким» естественные науки «внушают» не меньшее «участье», но редко кто из них «питает страсть» к статистике.

[2] Цит. по: Ростопчина Е. П. Счастливая женщина. — М.: Правда, 1991. — С.3. 11.

лемой частью современной экспериментальной науки. В его применении можно выделить три основных этапа:

а) построение *математической (вероятностной) модели* изучаемого явления;

б) *статистическое планирование* экспериментов (наблюдений), призванных подтвердить или опровергнуть предложенную модель;

в) проверка адекватности модели, которая включает в себя *статистическое оценивание* параметров модели, *проверку статистических гипотез* о постулатах модели или вытекающих из них следствий, выявление *статистических связей*.

Для дискретных, целочисленных данных, рассматриваемых в данной книге, минимально необходимыми и достаточно универсальными являются: *метод максимального правдоподобия* — при оценке параметров, *критерий* χ^2 («хи-квадрат») — при проверке гипотез и выявления связей, оценка *необходимого объема выборки* — при планировании эксперимента.

В итоге предложенная модель либо принимается, либо отвергается, и тогда ищутся иные модели, более реалистичные, более совершенные, ибо совершенствование теоретических моделей и аналитических методов есть магистральный путь постижения научной истины. По образному выражению популяризатора науки В. Р. Полищука, математическая модель есть «формульный фантом» явления[3], а А. Н. Горбань и Р. Г. Хлебопрос выделили особое «царство» формальных и математических моделей, назвав его обитателей «матемазаврами»: «Разводите матемазавров! Это безобидные и даже полезные существа, заменяющие теоретикам кроликов и дрозофил»[4].

Приводимые ниже примеры из двух прикладных отраслей генетики популяций — *демографической* и *сельскохозяйственной* — призваны убедить читателя в необходимости и плодотворности использования методов биометрии при планировании и анализе результатов популяционных исследований.

[3] Памщук В. Р. Мастеровые науки. — М.: Наука, 1989. — С.186.

[4] Горбань А. Н., Хлебопрос Р. Г. Демон Дарвина: Идея оптимальности и естественный отбор. — М.: Наука, 1988. — С.11.

П.2. Пример из демографической генетики: Вторичное соотношение полов у человека

Популяционное исследование, при анализе которого впервые было использовано вероятностно-статистическое рассуждение, составляющее основу современной логики проверки статистических гипотез, принадлежит Джону Арбатноту (John Arbuthnot, 1667–1735, традиционная русская транскрипция — Арбетнот)[5]. Это был просвещеннейший и остроумнейший человек своего времени: математик, медик, педагог, публицист, музыкант, член Королевского общества и Коллегии врачей, учитель математики и личный врач королевы Анны; он увековечил себя созданием бессмертного образа Джона Булля, ставшего нарицательным прозвищем англичан и карикатурным символом Англии. По словам знаменитого историка Томаса Бабингтона Маколи (Th. B. Macaulay, 1800–1859), «История Джона Булля» — «самая остроумная и смешная политическая сатира, существующая на английском языке»[6].

Среди многочисленных друзей и корреспондентов Дж. Арбатнота были известные математики Джеймс Стирлинг (J. Stirling, 1692–1770) и Джон Килл (J. Keill, 1671–1721), он пользовался уважением самого Исаака Ньютона (I. Newton, 1642–1727). Его ближайшими друзьями были крупнейшие писатели и поэты Джонатан Свифт (J. Swift, 1667–1745), Томас Парнелл (Th. Parnell, 1679–1718), Джон Гей (J. Gay, 1685–1732) и Александер Поуп (A. Pope, 1688–1744). Свифт писал о нем: «Умом Доктор превосходил всех нас, а его гуманизм был под стать его уму. Я немедленно сжег бы своего „Гулливера“, когда б узнал, что на свете есть хотя бы дюжина Арбатнотов». Друзья основали знаменитый литературный кружок Scriblerus Club и под коллективным псевдонимом Martinus Scriblerus создали «Воспоминания Мартина Писаки», в которых талантливо и едко высмеяли такие пороки общества, как шарлатанство, педантизм, схоластика, псевдоученость.

Дж. Арбатнот был хорошо знаком с теорией вероятностей: он опубликовал перевод с латыни книги Христиана Гюйгенса (Ch. Huygens, 1629–1695) «О расчетах в азартной игре» с собственными дополнениями, где, в частности, предложил задачу, решить которую удалось много позже, когда сформировалось понятие геометрической вероятности. Тем самым задолго до Жоржа-Луи Леклерка де Бюффона (G.-L. de Buffon, 1707–1788) с его зна-

[5] Энциклопедический словарь Брокгауза и Эфрона. Биографии. — М.: Сов. энциклопедия, 1991. — С. 406.

[6] На русский язык лишь недавно переведены фрагменты этого памфлета, а также «Искусство политической лжи». См. сб.: Англия в памфлете: Английская публицистическая проза начала XVIII века. — М.: Прогресс, 1987. — С. 41–69, 271–282.

менитой задачей о бросании иглы на разграфленную плоскость Арбатнот предугадал ограниченность классического понятия вероятности[7]. В 1989 г. в библиотеке Эдинбургского университета была найдена рукопись, датированная 1694 г., в которой Дж. Арбатнот впервые обсуждал то, что теперь принято называть критериями значимости. Логику одного из них и пример вычислений он опубликовал в 1710 г. в небольшой статье с характерным для того времени названием «Аргумент в пользу Божественного Провидения, взятый из постоянной регулярности, наблюдаемой в рождении обоих полов»[8]. Обследовав записи в церковных (метрических) книгах о крещении новорожденных детей, он подсчитал количество мальчиков и девочек, родившихся в Лондоне за 82 года, с 1629-го по 1710-й, и обнаружил, что неизменно мальчиков рождалось больше, чем девочек.

Статистическое рассуждение, которое провел Арбатнот (его часто называют *аргументом Арбатнота*[9]), заключается в следующем. Если считать, что вероятности рождения мальчика или девочки одинаковы, т. е. равны $1/2$, то вероятность 82 раза подряд наблюдать повышенную частоту рождаемости мальчиков составляет $2^{-82} \approx 2 \cdot 10^{-25}$. Вычисленную вероятность следует, очевидно, признать невообразимо малой. На основании этого Дж. Арбатнот пришел к выводу, что обнаруженное явление явно не случайно. (Объяснил он его тем, что жизнь мужчин подвержена опасностям и Божественное Провидение предусмотрительно компенсирует их ожидаемые потери.)

Так было убедительно доказано явление, называемое *вторичным соотношением полов*, причины которого продолжают изучаться и в наше время. Впервые (но без вероятностного анализа) оно было отмечено в 1662 г. основателем демографии капитаном Джоном Граунтом (J. Graunt, 1620–1674), преуспевавшим лондонским торговцем тканями и галантереей, который благодаря своему увлечению наукой стал одним из членов-основателей Королевского общества. Данные за 1629–1664 гг. собрал и опубликовал Граунт, а Арбатнот дополнил их более поздними сведениями и предварительно изложил свои соображения в докладе Королевскому обществу при избрании его своим членом в 1704 г.

Эта работа Арбатнота не осталась незамеченной, она стимулировала дальнейшее развитие теории вероятностей и способствовала становлению математической статистики.

---

[7] См.: Гнеденко Б. В. Курс теории вероятностей. — М.: Наука, 1988. — С. 405.

[8] Arbuthnot J. An argument for Divine Providence, taken from the constant regularity observed in the birth of both sexe // Phil. Trans. Roy. Soc. — 1710 (1712). — V. 27. — № 328. — P. 186–190.

[9] Согласно Словарю иностранных слов, АРГУМЕНТ — логический довод, служащий основанием доказательства.

Вероятностно-статистические рассуждения, аналогичные аргументу Арбатнота (под его влиянием или независимо), одними из первых использовали Джон Мишелл (J. Michell, 1724–1793) для доказательства неравномерности распределения звезд на небосклоне, вычислив вероятности их скоплений типа созвездия Плеяд и двойных звезд, и Мари-Жан-Антуан-Никола де Каритат де Кондорсе (M.-J.-An.-N. de Caritat de Condorset, 1743–1794) для доказательства маловероятности того, что 7 королей подряд (с учетом средней продолжительности жизни) могли править Древним Римом на протяжении 275 лет, как это утверждали историки.

Процедура проверки случайности отклонений наблюдаемого явления в ту или иную (положительную или отрицательную) сторону от первоначально ожидаемого значения путем подсчета ожидаемой вероятности таких отклонений, как это сделал Дж. Арбатнот, впоследствии получила название *критерия знаков* и связывается с его именем[10].

Данные, собранные Дж. Граунтом и Дж. Арбатнотом, заслуживают того, чтобы на их примере продемонстрировать мощь современных методов статистического анализа.

Фрагмент этих данных — сведения о распределении новорожденных в Лондоне по полу и по годам за 20 лет, с 1660 по 1679 г., — представлен в табл. П.1. Полностью данные Арбатнота, помимо первоисточника, можно найти в статье историка науки А. П. Юшкевича[11], где впервые опубликован русский перевод письма Николая Бернулли Пьеру Ремону де Монмору от 23 января 1713 г. Здесь сделана, хотя и далеко не совершенная, но первая попытка более детального вероятностно-статистического анализа данных Арбатнота.

П.3. Простейшая вероятностная модель для соотношения полов

Простейшей вероятностной моделью, описывающей распределение новорожденных по полу, будет модель, удовлетворяющая следующим двум предположениям:

1. Вероятность p появления на свет мальчика при каждом акте рождения — величина постоянная.

[10] См.: Кендалл М. Дж., Стюарт А. Статистические выводы и связи. — М.: Наука, 1973. — С. 687; Холлендер М., Вулф А. Непараметрические методы статистики. — М.: Финансы и статистика, 1983. — С. 39.

[11] Юшкевич А. П. Николай Бернулли и издание «Искусства предположений» Якова Бернулли // Теория вероятностей и ее применения. — 1986. — Т. 31. — № 2. — С. 330–352.

Таблица П.1

Данные о соотношении полов у детей, родившихся в Лондоне в 1660–1679 гг. (по *J. Arbatnot, 1710*)

Год	Номер выборки, i	Наблюдаемые численности			Частоты рождения мальчиков, $\widehat{p}_i$
		мальчиков, a_i	девочек, b_i	всего, n_i	
1 660	1	3 724	3 247	6 971	0,5342
1 661	2	4 748	4 107	8 855	0,5362
1 662	3	5 216	4 803	1 0019	0,5206
1 663	4	5 411	4 881	1 0292	0,5257
1 664	5	6 041	5 681	1 1722	0,5154
1 665	6	5 114	4 858	9 972	0,5128
1 666	7	4 678	4 319	8 997	0,5200
1 667	8	5 616	5 322	10 938	0,5134
1 668	9	6 073	5 560	11 633	0,5220
1 669	10	6 506	5 829	12 335	0,5274
1 670	11	6 278	5 719	11 997	0,5233
1 671	12	6 449	6 061	12 510	0,5155
1 672	13	6 443	6 120	12 563	0,5129
1 673	14	6 073	5 822	11 895	0,5106
1 674	15	6 113	5 738	11 851	0,5158
1 675	16	6 058	5 717	11 775	0,5145
1 676	17	6 552	5 847	12 399	0,5284
1 677	18	6 423	6 203	12 626	0,5087
1 678	19	6 568	6 033	12 601	0,5212
1 679	20	6 247	6 041	12 288	0,5084
Всего	$r = 20$	$A = 116\,331$	$B = 107\,908$	$N = 224\,239$	$S_p = 10{,}3870$

Итоговые частоты рождения мальчиков $\widehat{P} = 0{,}51878$; $\overline{p} = 0{,}51935$.

Примечания. В статье Арбатнота приведены лишь наблюдаемые численности полов по годам.

Окончание табл. П.1

Обозначения и формулы для вычислений

$i = 1, 2, \ldots, r$ — порядковый номер выборки (года наблюдений); r — общее число выборок (лет);
a_i и b_i — наблюдаемые численности, соответственно, мальчиков и девочек, родившихся в i-м году;
$n_i = a_i + b_i$ — общее число новорожденных в i-м году (объем i-й выборки);
$A = \sum_{i=1}^{r} a_i$ и $B = \sum_{i=1}^{r} b_i$ — суммарные численности, соответственно, мальчиков и девочек, родившихся за r лет;
$N = \sum_{i=1}^{r} a_i + \sum_{i=1}^{r} b_i = A + B$ — общая численность новорожденных, зарегистрированных в течение r лет (общий объем выборки);
$\widehat{p}_i = n_i^{-1} a_i$ — частота рождений мальчиков, наблюдаемая в i-м году;
$S_p = \sum_{i=1}^{r} \widehat{p}_i$ — сумма частот за все годы;
$\widehat{P} = A/N$ — наблюдаемая частота рождений мальчиков за все r лет;
$\overline{p} = r^{-1} S_p$ — средняя частота рождений мальчиков.

2. Пол новорожденного не зависит от всех предыдущих или последующих актов рождения.

При таких условиях число новорожденных мальчиков в каждой i-й выборке объема n_i должно подчиняться *биномиальному распределению* с параметрами n_i и p, описание которого можно найти в любом руководстве по биометрии или статистике.

Когда вероятностная модель построена, современными статистическими методами можно проверить справедливость ее постулатов или вытекающих из них следствий, оценить параметры модели, обосновать объем и направление дальнейших исследований. Необходимые формулы статистик для планирования эксперимента (оценка необходимого объема выборок $\widehat{n}_i$), для проверки гипотез (статистики χ^2), для точечных ($\widehat{p}$ и $\overline{p}$) и интервальных оценок параметра p сведены в табл. П.2. Примеры вычислений по этим формулам показаны в табл. П.3. Результаты анализа данных Дж. Арбатнота представлены в табл. П.5, П.7 и на рис. П.2 и П.3.

Таблица П.2

Основные формулы для статистического анализа данных о вторичном соотношении полов

Планирование объема выборок n_i

H_0: $p = p_0$; $\quad H_1^*$: $p = p_1 > p_0$ или $p_1 - p_0 > 0$;

(I) $\quad \widehat{n}_i^* \geqslant [(p_0q_0)^{1/2} z\{\alpha\} + (p_1q_1)^{1/2} z\{\beta\}]^2 (p_1 - p_0)^{-2}$;

H_0: $p = p_0$; $\quad H_1^{**}$: $p = p_1 \neq p_0$ или $|p_1 - p_0| > 0$;

(II) $\quad \widehat{n}_i^{**} \geqslant [(p_0q_0)^{1/2} z\{\alpha/2\} + (p_1q_1)^{1/2} z\{\beta\}]^2 (p_1 - p_0)^{-2}$.

Проверка гипотез

H_0	χ^2	ν
a_i: $b_i = 1$: 1	$\chi_1^2\{1\text{: }1\} = (a_i - b_i)^2 n_i^{-1}$	$\nu_{[1:\ 1]} = 1$
$A : B = 1$: 1	$\chi_T^2 = (A - B)^2 N^{-1}$	$\nu_T = 1$
$\forall(a_i$: $b_i) = 1$: 1	$\chi_S^2 = \sum_{i=1}^r \chi_i^2\{1\text{: }1\}$	$\nu_S = r$
$(a_i$: $b_i) = A : B$	$\chi_i^2\{A : B\} = (Ba_i - Ab_i)^2 (ABn_i)^{-1}$	$\nu_{[A:B]} = 1$
$\forall(a_i$: $b_i) = A : B$		
(III)	$\chi_H^2 = \sum_{i=1}^r \chi_i^2\{A : B\}$	
(IV)	$\chi_H^2 = N^2 (AB)^{-1} p_0 q_0 (\chi_S^2 - \chi_T^2)$	$\nu_H = r - 1$
(V)	$\chi_H^2 = N^2 (AB)^{-1} (\sum_{i=1}^r a_i^2 n_i^{-1} - A^2 N^{-1})$	

Оценки параметра p

$P_H > \alpha \Rightarrow$	$\widehat{P} = A/N$
$P_H \leqslant \alpha \Rightarrow$	$\overline{p} = r^{-1} S_p$

Окончание табл. П.2

Стандартные ошибки оценок параметра p

$P_H > \alpha \Rightarrow$	$\delta\{\widehat{P}\} = (\widehat{P}\widehat{Q}/N)^{1/2}$
$P_H \leqslant \alpha \Rightarrow$	$\delta\{\overline{p}\} = [r^{-1}(r-1)^{-1}(\sum_{i=1}^{r} \widehat{p}_i^2 - r^{-1}S_p^2)]^{1/2}$

$(1-\alpha)\ 100\,\%$*-ные доверительные интервалы для параметра p*

$$P_H > \alpha \Rightarrow \widehat{P} - \delta\{\widehat{P}\}z\{\alpha/2\} \leqslant p \leqslant \widehat{P} + \delta\{\widehat{P}\}z\{\alpha/2\};$$
$$P_H \leqslant \alpha \Rightarrow \overline{p} - \delta\{\overline{p}\}t\{\nu, \alpha/2\} \leqslant p \leqslant \overline{p} + \delta\{\overline{p}\}t\{\nu, \alpha/2\}; \qquad \nu = r - 1$$

Обозначения и формулы для промежуточных вычислений

p — вероятность рождения мальчика; H_0 — проверяемая нулевая гипотеза; H_1 — альтернативная гипотеза; p_0 — значение p, постулируемое нулевой гипотезой; p_1 — значение p согласно гипотезе H_1; $q_0 = 1 - p_0$; $q_1 = 1 - p_1$; a_i и b_i — наблюдаемые численности полов; r — число выборок; $n_i = a_i + b_i$ — наблюдаемые объемы выборок; $\widehat{n}_i$ — планируемые объемы выборок; $\widehat{p}_i = n_i^{-1}a_i$ — наблюдаемые частоты рождения мальчиков; $\widehat{P}, \delta\{\widehat{P}\}$ и $\overline{p}, \delta\{\overline{p}\}$ — оценки параметра p и их стандартные ошибки соответственно в случаях однородности и неоднородности выборочных распределений; P — вероятность значимости критерия (см. разд. П.6). $S_p = \sum_{i=1}^{r} \widehat{p}_i$; $A = \sum_{i=1}^{r} a_i$; $B = \sum_{i=1}^{r} b_i$; $N = A + B$; $z\{\alpha\}$, $z\{\beta\}$, $z\{\alpha/2\}$ и $t\{\nu; \alpha/2\}$ — соответственно α-, β- и $\alpha/2$-квантили нормального (z) и t-распределения Стьюдента; символ $\forall$ означает «все», стрелка $\Rightarrow$ — «следовательно».

Вывод формул для $\widehat{n}_i$ см. в кн.: Бикел П., Доксам К. Математическая статистика. — М.: Финансы и статистика, 1983. — Вып.1. — С. 190–191; Браунли К. А. Статистическая теория и методология в науке и технике. — М.: Наука, 1977.

П.4. Понятие мощности критерия и планирование объема выборки

Современные представления о генетическом определении пола у человека позволяют выдвинуть проверяемую (нулевую) гипотезу H_0, что соотношение по полу равно 1: 1, или (что то же самое) H_0: $p = p_0 = 1/2$.

Таблица П.3

Примеры вычисления наблюдаемых значений статистик для данных Дж. Арбатнота

Планирование объема выборок n_i

$$H_0\colon p = p_0 = 0{,}500; \quad H_1^*\colon p \geqslant p_1 = 0{,}515;$$

$$\text{(I)} \qquad n_i^* \geqslant \left[\frac{(0{,}500 \cdot 0{,}500)^{1/2} 1{,}64 + (0{,}515 \cdot 0{,}485)^{1/2} 1{,}28}{0{,}515 - 0{,}500} \right]^2 \geqslant 9\,470;$$

$H_0\colon p = p_0 = 0{,}500; \quad H_1^{**}\colon |p_1 - p_2| > 0{,}015,$ т. е.

$$p \geqslant p_1 = 0{,}515 \quad \text{или} \quad p \leqslant p_1 = 0{,}485;$$

$$\text{(II)} \qquad n_i^{**} \geqslant \left[\frac{(0{,}500 \cdot 0{,}500)^{1/2} 1{,}96 + (0{,}515 \cdot 0{,}485)^{1/2} 1{,}28}{0{,}515 - 0{,}500} \right]^2 \geqslant 11\,660$$

Проверка гипотез

$$H_0\colon a_i\colon b_i = 1\colon 1;$$

$$\chi_1^2\{1\colon 1\} = (3\,724 - 3\,247)^2 / 6\,971 = 32{,}64;$$

$$\nu_{[1\colon\,1]} = 1;\ P_{[1\colon\,1]} = 1 \cdot 10^{-8} \text{ и т. д.}$$

$$H_0\colon A : B = 1\colon 1;$$

$$\chi_T^2 = (116\,331 - 107\,908)^2 / 224\,239 = 316{,}39;$$

$$\nu_T = 1;\ P_T = 8{,}9 \cdot 10^{-71}$$

$$H_0\colon \forall(a_i\colon b_i) = 1\colon 1;$$

$$\chi_S^2 = 32{,}64 + 46{,}40 + \ldots + 3{,}47 = 366{,}51;$$

$$\nu_S = 20;\ P_S = 1{,}7 \cdot 10^{-65}$$

$$H_0\colon (a_i\colon b_i) = A : B;$$

$$\chi_1^2\{A : B\} = \frac{(107\,908 \cdot 3\,724 - 116\,331 \cdot 3\,247)^2}{116\,331 \cdot 107\,908 \cdot 6\,971} = 6{,}71;$$

$$\nu_{[A:B]} = 1;\ P_{[A:B]} = 0{,}0096 \text{ и т. д.}$$

Окончание табл. П.3

$$H_0\colon \forall(a_i\colon b_i) = A : B;$$

(III) $\chi^2_H = 6{,}71 + 10{,}84 + \ldots + 5{,}25 = 50{,}19;$

(IV) $\chi^2_H = \dfrac{224\,239^2(336{,}51 - 316{,}39)}{4 \cdot 116\,331 \cdot 107\,908} = 50{,}19;$

$\nu_H = 20 - 1 = 19;\ P_H = 1{,}3 \cdot 10^{-4};$

(V) $\dfrac{224\,239^2}{116\,331 \cdot 107\,908}\left[\dfrac{3\,724^2}{6\,971} + \dfrac{4\,748^2}{8\,855} + \ldots + \dfrac{6\,247^2}{12\,288} - \dfrac{116\,331^2}{224\,239}\right] = 50{,}19$

Оценка параметра p

$$P_H \approx 10^{-4} \Rightarrow \overline{p} = 10{,}3870/20 = 0{,}51935 \approx 0{,}519$$

Стандартная ошибка оценки p

$$\delta\{\overline{p}\} = \left[\frac{0{,}5342^2 + 0{,}5382^2 + \ldots + 0{,}5084^2 - 10{,}3870^2/20}{20 \cdot 19}\right]^{1/2} = \\ = \pm 0{,}0018 \approx \pm 0{,}002$$

Конечный результат оценивания

$$\overline{p} \pm \delta\{\overline{p}\} = 0{,}519 \pm 0{,}002$$

---

95 %-ный доверительный интервал для параметра p

$$0{,}519 - 0{,}002 \cdot 2{,}09 < p < 0{,}519 + 0{,}002 \cdot 2{,}09$$

$$0{,}515 < p < 0{,}523$$

Примечания. Обозначения те же, что и в табл. П.2. P-значения и необходимые для вычислений значения квантилей $z\{\alpha = 0{,}05\} = 1{,}64$; $z\{\beta = = 0{,}10\} = 1{,}28$; $z\{\alpha/2 = 0{,}025\} = 1{,}96$ и $t\{\nu = 19; \alpha/2 = 0{,}025\} = 2{,}09$ находят в таблицах или с помощью пакетов прикладных программ.

При проверке статистической гипотезы возможны ошибочные решения двух типов:

а) можно ошибочно отвергнуть нулевую гипотезу, когда она верна, — *ошибка первого рода*;

б) можно ошибочно принять H_0, когда она неверна, — *ошибка второго рода*.

Вероятность ошибки первого рода, традиционно обозначаемая α, зависит от гипотезы H_0 и называется *уровнем значимости критерия*.

Вероятность ошибки второго рода традиционно обозначают β; она зависит от альтернативной гипотезы H_1. Ее дополнение $1-\beta$ называют *мощностью критерия*. Мощность можно интерпретировать как «качество» критерия, его чувствительность к отклонениям от H_0, как способность различать нулевую и альтернативную гипотезы.

Решение, какую вероятность ошибок первого и второго рода можно считать пренебрежимо малой, т. е. решение о *достаточности* эксперимента или наблюдения, является внестатистическим — оно есть акт интеллектуальной смелости[12]. В современной экспериментальной биологии общепринято использовать одно из трех значений уровня значимости α: чаще 0,05 или 0,01, реже 0,001.

При выборе значений вероятности β руководствуются соображением, что ошибки первого рода чреваты бóльшим риском, нежели ошибки второго рода. Считается, что β может быть в 2–4 раза больше, чем α. Так, если принимают $\alpha = 0{,}05$, то выбирают $\beta = 0{,}10$ или $0{,}20$, т. е. мощность $1-\beta$ равна $0{,}90$ или $0{,}80$.

На практике представление о мощности критерия позволяет оценить объем выборки $\widehat{n}_i$, который гарантировал бы различение гипотез H_0: $p = p_0$ и H_1: $p = p_1$ при заранее выбранных уровнях вероятностей α и β.

Наблюдения Дж. Граунта и Дж. Арбатнота, равно как и подавляющая масса данных всей последующей мировой демографической статистики, предоставляют существенную *априорную информацию* для конкретизации альтернативной гипотезы H_1, а именно, что истинная вероятность рождения мальчика p_1 превышает значение $p_0 = 0{,}500$. В таком случае альтернативная гипотеза становится *односторонней* H_1^*: $p = p_1 > p_0$ или $p_1 - p_0 > 0$, в противоположность менее конкретной *двусторонней* альтернативе H_1^{**}: $p = p_1 \neq p_0$ или $|p_1 - p_0| > 0$.

[12] Подобные действия в науке зиждятся на вненаучных интуитивных убеждениях и волевых решениях. См.: Фейнберг Л. Е. Две культуры: Интуиция и логика в искусстве и науке. — М.: Наука, 1992.

Допустим, что p_1 не меньше 0,515. Тогда объем выборки n_i, необходимый для проверки гипотезы H_0: $p_0 = 0{,}500$ против конкретной односторонней альтернативы H_1: $p_1 = 0{,}515$ при $\alpha = 0{,}05$ и $\beta = 0{,}10$ должен быть не меньше:

$$\widehat{n}_i^*\{p_0 = 0{,}500; p_1 \geqslant 0{,}515; \alpha = 0{,}05; \beta = 0{,}10\} \geqslant 9\,470$$

(см. формулу (I) в табл. П.2 и табл. П.3).

Объемы большинства выборок $n_i = a_i + b_i$, представленных в табл. П.1 (за исключением 1660, 1661 и 1666 гг.), удовлетворяют данному требованию. Это означает, что если истинная вероятность рождения мальчика больше или равна 0,515, то ее отличие от 0,500 при данных объемах выборок можно выявить статистическими методами.

Если же альтернативная гипотеза — *двусторонняя*, например, H_1^{**}: $|p_1 - p_0| > 0{,}015$, то в формуле для $\widehat{n}_i$ вместо $z\{\alpha\}$ следует использовать $z\{\alpha/2\}$ и при тех же значениях α и β потребовался бы существенно больший объем выборки:

$$\widehat{n}_i^{**}\{|p_1 - p_0| \geqslant 0{,}015; \alpha = 0{,}05; \beta = 0{,}10\} \geqslant 11\,660$$

(см. формулу (II) в табл. П.2 и табл. П.3).

В формулы (I) и (II) для вычисления $\widehat{n}_i$, помимо p_0 и p_1, т. е. конкретных различаемых значений параметра p, входят также $z\{\alpha\}$, $z\{\alpha/2\}$ и $z\{\beta\}$, которые суть α-, $\alpha/2$- и β-*квантили*, или *критические значения*, нормального распределения; их можно найти в любом руководстве, содержащем статистические таблицы, либо с помощью компьютера, используя распространенные пакеты статистических программ типа STATGRAPHICS, SYSTAT, SAS, SPSS, MICROSTAT, BMDP, STAT–ITCF, БИОСТАТ и др. Для наиболее распространных значений $\alpha = 0{,}05$ и $\beta = 0{,}10$ они равны, соответственно, $z\{\alpha/2 = 0{,}025\} = 1{,}96$, $z\{\alpha = 0{,}05\} = 1{,}64$ и $z\{\beta = 0{,}10\} = 1{,}28$.

П.5. Графический подход к планированию объема выборки

Способность критерия различать нулевую и альтернативную гипотезы повышается с ростом объема выборки. Наглядно эту тенденцию можно отобразить графически (рис. П.5). На обоих графиках абсцисса есть числовая ось для численностей новорожденных девочек b, а ордината — для численностей новорожденных мальчиков a. Сплошная линия, исходящая из начала координат под углом $45°$, есть геометрическое место точек, которое соответствует соотношению $A : B = 1{:}\ 1$, или, что то же самое,

значению параметра $p_0 = 1/2$, постулируемому нулевой гипотезой. Верхняя сплошная линия на обоих графиках соответствует значению $p_1 = 0{,}515$, ожидаемому согласно обеим альтернативным гипотезам, нижняя сплошная (рис. П.5, *Б*) — соответствует второму значению $p_1 = 0{,}485$, ожидаемому согласно двусторонней альтернативе.

Принимая, согласно выбранной вероятностной модели, что вторичное соотношение полов подчиняется биномиальному распределению с параметрами n и p, можно построить границы, разделяющие *выборочное пространство* (т. е. множество всех возможных выборочных значений a и b) на две области. Одна из них называется *критической*, или *областью отвержения* гипотезы. Вероятность попадания в нее равна выбранному уровню ошибки первого рода α при проверке H_0 или второго рода β для H_1. Вторая называется *областью принятия* гипотезы. Вероятность попадания в нее равна, соответственно, $1 - \alpha$ или $1 - \beta$. При больших объемах выборок эти границы можно считать прямолинейными и для их построения использовать известную аппроксимацию биномиального распределения нормальным распределением с параметрами: np — среднее значение и npq — дисперсия. Необходимые для расчетов формулы и пример вычислений представлены в табл. П.4. На рис. П.5 эти границы изображены тонкими пунктирными линиями.

На рис. П.5, *А* показана одна (верхняя) граница для 95 %-ной односторонней области принятия гипотезы H_0: $p_p = 1/2$ (она обозначена как $p_0^+\{\alpha = 0{,}05\}$); на рис. П.5, *Б* показаны две (верхняя — $p_0^+\{\alpha/2 = 0{,}025\}$ и нижняя — $p_0^-\{\alpha/2 = 0{,}025\}$) границы соответствующей 95%-ной двусторонней области. Для альтернативного значения $p_1 = 0{,}515$ на обоих графиках показана нижняя ($p_1^-\{\beta = 0{,}10\}$) граница 90 %-ной односторонней области принятия H_1, а на рис. П.5, *Б*, кроме того, — верхняя граница $p_0^+\{\beta = 0{,}10\}$ для второго альтернативного значения $p_1 = 0{,}485$.

Можно видеть, что вблизи от начала координат $(1 - \alpha)$ 100 %-ные и $(1 - \beta)$ 100 %-ные области принятия гипотез H_0 и H_1 перекрываются. Это означает, что при малых значениях a и b и заданных значениях вероятностей ошибок первого (α) и второго (β) родов невозможно различить гипотезы H_0 и H_1. С ростом же a и b границы этих областей постепенно расходятся и, начиная с какого-то момента, перестают пересекаться. Координаты точки их пересечения и соответствуют тем минимальным значениям a и b, которые необходимы для различения H_0 и H_1 (рис. П.5).

При односторонней альтернативе H_1^*: $p_1 \geqslant 0{,}515$ искомыми являются координаты точки пересечения границ $p_0^+\{\alpha = 0{,}05\}$ и $p_0^-\{\beta = 0{,}10\}$, а именно: $\widehat{a}^* = 4\,815$ и $\widehat{b}^* = 4\,655$ (рис. П.5, *А*). Соответственно, необхо-

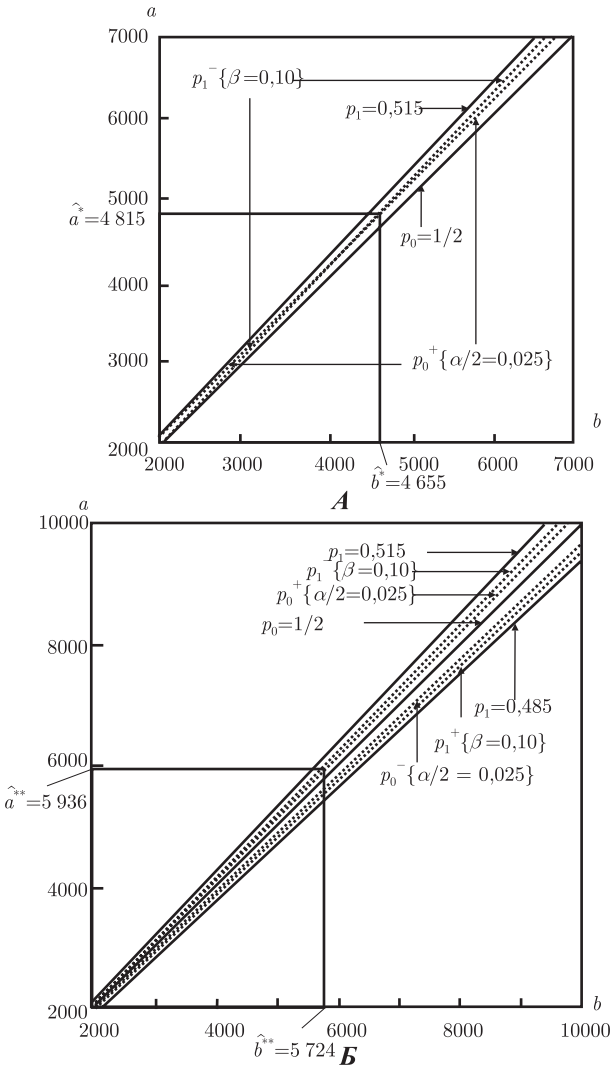


Рис. П.1. Графический подход к планированию объема выборки при односторонней (*А*) и двухсторонней (*Б*) альтернативах. Представлены фрагменты выборочного пространства для соотношения $A : B$, ограниченные значениями от 2 000 до 7 000 (А) и от 2 000 до 10 000 (Б); остальные пояснения см. в тексте

димый для различения гипотез объем выборки есть $\widehat{n}^* = 9\,470$, что совпадает со значением, найденным ранее алгебраическим путем по формуле I в табл. П.2 и П.3. При двусторонней альтернативе H_1^{**}: $|p_1 - p_0| > 0{,}015$ находят либо координаты точки пересечения границ $p_0^+\{\alpha/2 = 0{,}025\}$ и $p_1^-\{\beta = 0{,}10\}$, а именно: $\widehat{a}^{**} = 5\,936$ и $\widehat{b}^{**} = 5\,724$, либо симметрично: координаты $\widehat{a}^{**} = 5\,724$ и $\widehat{b}^{**} = 5\,936$ для точки пересечения границ $p_0^-\{\alpha/2 = 0{,}025\}$ и $p_1^+\{\beta = 0{,}10\}$ (рис.П.5, *Б*). Соответственно, $\widehat{n}^{**} = 11\,660$, что также совпадает со значением, найденным ранее.

Подобный графический прием, как будет показано далее в разд. П.9 и П.12, особенно плодотворен при статистическом анализе конкретных экспериментальных данных.

П.6. Проверка гипотез и оценка значимости критерия

До недавнего времени исследователи довольствовались выводами о справедливости или ложности нулевой гипотезы на основе сравнения вычисленного (наблюдаемого) значения статистики критерия $\chi^2_{\text{набл}}$ с табличным критическим значением $\chi^2\{\alpha\}$: если $\chi^2_{\text{набл}} < \chi^2\{\alpha\}$, то гипотеза H_0 не отвергается (принимается); если же $\chi^2_{\text{набл}} \geqslant \chi^2\{\alpha\}$, то H_0 отвергается (на уровне значимости α). Такая процедура именуется *проверкой гипотез*.

Современная доступность компьютерной техники позволяет вычислять конкретные, наблюдаемые значения уровня значимости, т. е. позволяет оценить, какова истинная *вероятность значимости* критерия, или так называемое *P-значение*. Она есть вероятность (Prob) того, что статистика χ^2 принимает значения, равные или большие полученного $\chi^2_{\text{набл}}$ при условии справедливости нулевой гипотезы:

$$P = \text{Prob}\{\chi^2 \geqslant \chi^2_{\text{набл}} | H_0\}.$$

Такая процедура более информативна, нежели обычная процедура проверки гипотез, ее называют проверкой или оценкой значимости (критерия). Искомые P-значения можно найти с помощью упомянутых ранее пакетов программ.

П.7. Проверка согласия с теоретически ожидаемым соотношением полов 1: 1

Когда объемы выборок n_i достаточно велики, проверить нулевую гипотезу о соотношении полов 1: 1, т. е. H_0: $p = p_0 = 1/2$, можно с помощью

Таблица П.4

Построение областей принятия и отвержения гипотез об ожидаемых соотношениях $A : B$

Координаты геометрического места точек для соотношений $A : B$*, ожидаемых при определенном значении* p

$$a = np; \qquad b = n - a$$

Координаты для границ $(1 - \&)\ 100\,\%$*-ных областей принятия гипотез*

верхние границы $p^+\{\&\}$

$$a^+ = a + z\{\&\}(npq)^{1/2}; \quad b^+ = n - a^+$$

нижние границы $p^-\{\&\}$

$$a^- = a - z\{\&\}(npq)^{1/2}; \quad b^- = n - a^-$$

Пример вычислений

$$p_1 = 0{,}515$$

$$\beta = 0{,}10; \quad z\{\beta = 0{,}10\} = 1{,}28$$

$$n_1 = 6\,000; \quad n_2 = 12\,000$$

Координаты геометрического места точек для соотношений $A : B$*, ожидаемых при* $p_1 = 0{,}515$

$$a_1 = 6\,000 \cdot 0{,}515 = 3\,090; \qquad b_1 = 6\,000 - 3\,090 = 2\,910;$$

$$a_2 = 12\,000 \cdot 0{,}515 = 6\,180; \quad b_2 = 12\,000 - 6\,180 = 5\,820;$$

Координаты для $p_1^-\{\beta = 0{,}10\}$ *— нижней границы 90 %-ной односторонней области принятия гипотезы* H_1*:* $p_1 = 0,515$

$$a_1^- = 3\,090 - 1{,}28(6\,000 \cdot 0{,}515 \cdot 0{,}485)^{1/2} = 3\,040; \quad b_1^- = 6\,000 - 3\,040 = 2\,960;$$

$$a_2^- = 6\,180 - 1{,}28(12\,000 \cdot 0{,}515 \cdot 0{,}485)^{1/2} = 6\,110; \; b_2^- = 12\,000 - 6\,110 = 5\,890.;$$

Окончание табл П.4

Примечания. При нахождении границ для односторонних $(1-\&)$ 100 %-ных областей принятия гипотез используют $\& = \alpha$ или $\& = \beta$, а для двухсторонних $\& = \alpha/2$. Соответствующие значения $\&$-квантилей нормального распределения, например, $z\{\beta = 0{,}10\} = 1{,}28$; $z\{\alpha = 0{,}05\} = 1{,}64$; $z\{\alpha/2 = 0{,}025\} = 1{,}96$; $z\{\alpha/2 = 0{,}005\} = 2{,}58$ и $z\{\alpha/2 = 0{,}0005\} = 3{,}29$ находят в статистических таблицах или с помощью статистических программ.

В качестве p используют конкретные значения, например, $p_0 = 1/2$ (см. рис. П.5 и П.2), $p_1 = 0{,}515$ (рис. П.5, *А*), $p_1 = 0{,}485$ (рис. П.5, *Б*) и $p_1 = 0{,}5188$ (см. рис. П.3).

Выбирают подходящие значения n_1 и n_2. Когда они большие, то при построении геометрического места точек для ожидаемых соотношений $A : B$ достаточно провести прямые через пары точек с координатами (a_1, b_1) и (a_2, b_2), соответствующие выбранным n_1 и n_2. При построении верхних и нижних границ для областей принятия гипотез достаточно провести прямые через пары точек с координатами (a_1^+, b_1^+) и (a_2^+, b_2^+) и с координатами (a_1^-, b_1^-) и (a_2^-, b_2^-).

Следует, однако, помнить, что при малых объемах выборок n эти границы становятся криволинейными, и для их построения указанные приближенные формулы неприменимы. В таких случаях можно использовать следующую итерационную процедуру вычислений, основанную на известном однозначном соответствии между биномиальным распределением и *F-распределением Снедекора–Фишера*. Сначала вычисляют приближенные значения a^+ и a^-. Затем в таблицах или с помощью пакетов статистических программ находят значение F, соответствующее данному $\&$ со степенями свободы $\nu_1 = 2(n - a^+ + 1)$ и $\nu_2 = 2a^+$. Вычисляют новое значение $(a^+)' = Fp(n+1)(q+Fp)^{-1}$. Его округляют до ближайшего целого числа, большего $(a^+)'$. Если это округленное значение отличается от предварительной оценки a^+ более, чем на единицу, то вычисления повторяют. Находят новое значение F' с $\nu_1 = 2[n-(a^+)'+1]$ и $\nu_2 = 2(a^+)'$ и вычисляют новое значение $(a^+)'' = F'p(n+1)(q+F'p)^{-1}$ и т. д. Аналогичные итерационные вычисления проводят и со значением a^-. Находят F с $\nu_1 = 2(a^- + 1)$ и $\nu_2 = 2(n - a^-)$, затем новое значение $(a^-)' = (np - Fq)(Fq + p)^{-1}$ и т. д. Соответствующие значения абсцисс находят вычитанием из n полученных значений ординат.

См.: *Поллард Дж.* Справочник по вычислительным методам статистики. — М.: Финансы и статистика, 1982. — С. 110–111.

критерия χ^2. В общем случае применимость этого критерия ограничена условием $n_i p_0 q_0 \geqslant 25$, где q_0 — традиционное обозначение для $(1 - p_0)$. В данном частном случае $p_0 = q_0 = 1/2$ и критерий χ^2 можно применять для выборок объема $n_i \geqslant 100$.

Удобной для проверки соотношения 1: 1 в каждой i-й выборке является формула $\chi_i^2\{1\colon 1\}$ с одной степенью свободы $\nu_{[1\colon 1]} = 1$ (см. табл. П.2). Примеры и результаты вычислений по этой формуле и соответствующие значения $P_{[1\colon 1]}$ приведены в табл. П.3 и П.5.

Можно видеть, что в большинстве случаев наблюдаются статистически высокозначимые отклонения от ожидаемого соотношения 1: 1. Во многих случаях найденные уровни значимости $P_{[1\colon 1]}$ значительно меньше 10^{-3}, и лишь в двух случаях (в 1677 и 1679 гг.) отклонения несущественны на уровне значимости $\alpha = 0{,}05$ (в табл. П.5 они отмечены знаком $\#$; двумя и тремя такими знаками отмечены, соответственно, случаи, когда $P_{[1\colon 1]} > > 0{,}01$ и $P_{[1\colon 1]} > 0{,}001$).

Проверку согласия с соотношением 1: 1 для суммарных численностей полов A и B можно провести по формуле χ_T^2 с одной степенью свободы $\nu_T = 1$, аналогичной формуле $\nu_{[1\colon 1]}$ (см. табл. П.2).

П.8. Вычисление малых P-значений

Найденное значение $\chi_T^2 = 316{,}39$ столь велико, что доступные компьютерные программы не позволяют удовлетворительно оценить соответствующее значение P_T. В таких случаях можно использовать аппроксимационные формулы, представленные в табл. П.6.

Когда используемая статистика χ^2 имеет одну степень свободы, то P-значения (если они меньше 10^{-4}) легко оценить по формуле (I). Если же $\nu \geqslant 2$, то можно использовать формулы для нормальной аппроксимации распределения χ^2 (т. е. аппроксимации его нормальным распределением). Аппроксимация Пайзера–Пратта, приведенная в табл. П.6, является одной из лучших. Полученные значения z^2 используют для оценки соответствующих P-значений по приведенной в этой же таблице формуле (II).

Таблица П.5

Основные результаты статистического анализа данных о соотношении полов у детей, родившихся в Лондоне в 1660–1679 гг.

Год	Наблюдаемые численности		Вычисленные значения частных статистик			
	мальчиков, a_i	девочек, b_i	$\chi^2_i\{1{:}\,1\}$	$P_{[1:\,1]}$	$\chi^2_i\{A:B\}$	$P_{[A:B]}$
1660	3 724	3 247	32,64	$1{\cdot}10^{-8}$	6,71	0,0096**
1661	4 748	4 107	46,40	$8{\cdot}10^{-11}$	10,84	0,0010***
1662	5 216	4 803	17,02	$4{\cdot}10^{-5}$	0,14	0,71
1663	5 411	4 881	27,29	$2{\cdot}10^{-7}$	2,04	0,15
1664	6 041	5 681	11,06	$9{\cdot}10^{-4}$	0,53	0,47
1665	5 114	4 858	6,57	$0{,}010^{\#\#}$	1,38	0,24
1666	4 678	4 319	14,32	$2{\cdot}10^{-4}$	0,06	0,81
1667	5 616	5 322	7,90	$5{\cdot}10^{-3\#\#\#}$	1,22	0,27
1668	6 073	5 560	22,62	$2{\cdot}10^{-6}$	0,52	0 47
1669	6 506	5 829	37,16	$1{\cdot}10^{-9}$	3,76	0,053
1670	6 278	5 719	26,05	$3{\cdot}10^{-7}$	1,01	0,32
1671	6 449	6 061	12,03	$5{\cdot}10^{-4}$	0,51	0,48
1672	6 443	6 120	8,30	$4{\cdot}10^{-3\#\#\#}$	1,73	0,19
1673	6 073	5 822	5,30	$0{,}021^{\#\#}$	3,17	0,075
1674	6 113	5 738	11,87	$6{\cdot}10^{-4}$	0,40	0,53
1675	6 058	5 717	9,88	$2{\cdot}10^{-3\#\#\#}$	0,84	0,36
1676	6 552	5 847	40,09	$3{\cdot}10^{-10}$	4,69	0,030*
1677	6 423	6 203	3,83	$0{,}050^{\#}$	5,06	0,025*
1678	6 568	6 033	22,71	$2{\cdot}10^{-6}$	0,32	0,57
1679	6 247	6 041	3,47	$0{,}062^{\#}$	5,25	0,022*
Значения сводных статистик	$\chi^2_T = 316{,}39$ $\nu_T = 1$ $P_T = 8{,}9\times$ $\times 10^{-71}$		$\chi^2_S = 366{,}51$ $\nu_S = 20$ $P_S = 1{,}7\times$ $\times 10^{-65}$		$\chi^2_H = 50{,}19$ $\nu_H = 19$ $P_H = 1{,}3\times$ $\times 10^{-4}$	

Примечания. Формулы вычисляемых статистик указаны в табл. П.2, а примеры вычислений — в табл. П.3.

P-значения, меньшие 10^{-9}, а именно: $P_{[1:\,1]}$ для 1661 и 1676 гг., P_T и P_S, оценены по аппроксимационным формулам (см. табл. П.6); остальные оценки P-значений получены с помощью программы MICROSTAT.

Знаком # обозначены случаи, статистически не отличающиеся от соотношения 1: 1 на уровнях значимости $\alpha = 0{,}05(\#)$, $\alpha = 0{,}01(\#\#)$ и $\alpha = 0{,}001(\#\#\#)$. Звездочками отмечены статистически существенные отклонения от соотношения $A : B$ на уровнях значимости $\alpha = 0{,}05(*)$, $\alpha = 0{,}01(**)$ и $\alpha = 0{,}001(***)$.

Окончание табл. П.5

Все вычисления проведены с числом знаков, максимально допустимым для используемого вычислительного средства (калькулятора или компьютера), и лишь конечный результат округлен до разумного предела значащих цифр.

Полученная оценка значимости критерия $P_T = 8,9 \cdot 10^{-71}$ свидетельствует о несомненности факта повышенной частоты рождаемости мальчиков.

П.9. Графический подход к проверке согласия с соотношением 1: 1

Никогда не следует пренебрегать возможностью визуализировать исходные данные и результаты их статистического анализа, т. е. представить их в наглядной и компактной графической форме. Вариант такого представления для данных Арбатнота показан на рис. П.2. Здесь каждая пара значений: a_i — число мальчиков, b_i — число девочек, родившихся в i-м г., — представлена одной точкой (a_i, b_i) в системе прямоугольных координат. Сплошная прямая, исходящая из начала координат под углом $45°$, есть геометрическое место точек, соответствующее соотношению по полу 1: 1.

Поразительно, сколь тонкие различия выявляет статистический анализ. Зрительно эмпирические точки не так уж далеко отстоят от этой прямой, тем не менее, поскольку все они располагаются по одну сторону от нее — над ней, наблюдаемое отклонение оказывается статистически высокозначимым: $P = 2^{-20} \approx 10^{-6}$ (критерий знаков).

Поведение эмпирических точек можно интерпретировать статистически еще более содержательно, если на графике вокруг теоретической прямой выделить $(1 - \alpha)\,100\,\%$-ные области принятия нулевой гипотезы, например, для общепринятых значений вероятностей $(1 - \alpha)$: 95 %, 99 и 99,9 % (пунктирные линии на рис. П.2 суть границы таких областей; их построение аналогично описанному в табл. П.4). Можно убедиться, что только две эмпирические точки (в 1677 и 1679 гг.) попадают в 95%-ную область принятия нулевой гипотезы. Это именно те года, в которые отклонения от соотношения 1: 1 по критерию χ^2 оказались несущественными на этом уровне значимости (отмечены знаком # в табл. П.5). Расположение всех остальных точек также находится в полном согласии с результатами анализа по критерию χ^2.

Таблица П.6

Оценка малых P-значений для статистики χ^2

$$\nu = 1 \quad \Rightarrow \quad P = 2(2\pi\chi^2)^{-1/2}e^{-\chi^2/2} \tag{I}$$

Нормальная аппроксимация Пайзера–Пратта

$$\begin{aligned} \nu \geqslant 2 \quad \Rightarrow \quad & P = (2\pi z^2)^{-1/2}e^{-z^2/2}; \\ & z^2 = (\chi^2 - \nu + 1)^{-2}(\chi^2 - \nu + 2/3 - \\ & - 0{,}08/\nu)^2\{(\nu-1)\ln[(\nu-1)/\chi^2] + \chi^2 - \nu + 1\} \end{aligned} \tag{II}$$

Примеры вычислений

$$\nu_T = 1, \qquad \chi^2_T = 316{,}39$$

(I) $$P_T = 2(2\pi 316{,}39)^{-1/2}e^{-316{,}39/2} = 8{,}9 \cdot 10^{-71};$$

$$\nu_S = 20, \qquad \chi^2_S = 366{,}51$$

(II) $$\begin{aligned} z^2 = (366{,}51 - 20 + 1)^{-2}(366{,}51 - 20 + 2/3 - 0{,}08/20)^2\{(20-1)\times \\ \times \ln[(20-1)/366{,}51] + 366{,}51 - 20 + 1\} = 290{,}71; \end{aligned}$$

$$P_S = (2\pi 290{,}71)^{-1/2}e^{-290{,}71/2} = 1{,}7 \cdot 10^{-65}$$

Примечание. Знак $\Rightarrow$ означает «следовательно».

Источники:

Peizer D. B., Pratt J. W. A normal approximation for binomial, F, beta, and other common, related tail probability, I // J. Amer. Stat. Assoc. — 1968. — Vol. 63. — № 324. — P. 1416–1456.

Ling R. F. A study of the accuracy of some approximations for t, χ^2 and F tail probability // J. Amer. Stat. Assoc. — 1978. — Vol. 73. — № 362. — P. 272–283.

Таким образом, указание области принятия гипотезы об ожидаемом соотношении есть не что иное, как наглядный способ проверки (частных) нулевых гипотез H_0: a_1: $b_1 = 1$: 1. Нахождение точки вне пределов заданной

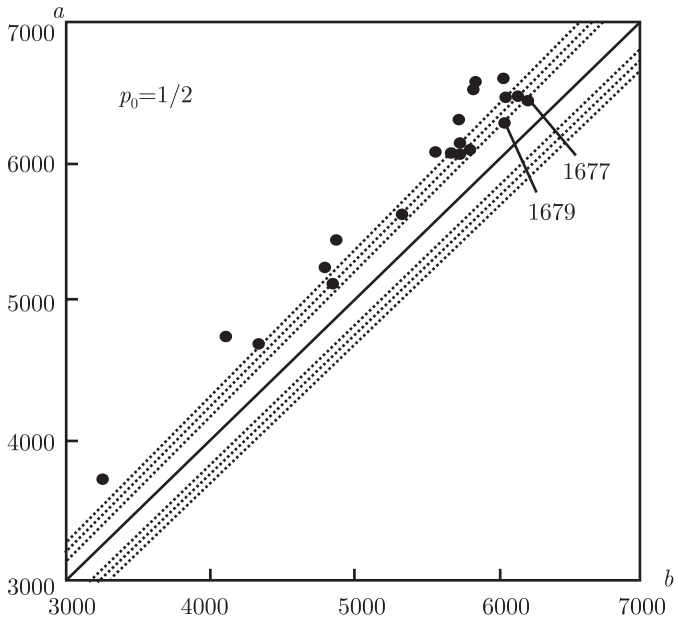


Рис. П.2. Графический подход к проверке согласия расщепления по полу у лондонских новорожденных в 1660–1677 гг. с соотношением 1: 1 (по Дж. Арбатноту, 1710). Ось абсцисс — количество новорожденных девочек (b); ось ординат — количество новорожденных мальчиков (a); точки с координатами (a_i, b_i) — наблюдаемое вторичное соотношение полов по годам. Сплошная прямая — геометрическое место точек для ожидаемого соотношения 1: 1 согласно нулевой гипотезе H_0: $p_0 = 1/2$. Все 20 эмпирических точек лежат над ней: мальчиков во все года появлялось на свет больше, чем девочек, и вероятность такого события мала: $2^{-20} \approx 10^{-6}$ (аргумент Арбатнота, или критерий знаков). Тонкие пунктирные линии — границы двухсторонних 95-, 99- и 99,9 %-ных областей принятия H_0. Только две точки (в 1677 и 1679 гг.) попадают в 95 %-ную область принятия H_0, т. е. только их отклонения от ожидаемого соотношения статистически несущественны на уровне значимости $\alpha = 0{,}05$ (ср. табл. П.5)

$(1 - \alpha)\,100\,\%$-ной области принятия, т. е. попадание в область отвержения, свидетельствует о ее существенном (на уровне значимости α) отклонении от ожидаемого соотношения.

Преимущества такого графического подхода очевидны: на одном рисунке можно уместить большие массивы исходных данных, совместив их с результатами статистического анализа.

П.10. Обобщенная проверка согласия с соотношением 1: 1

Проведенный анализ показал, что 90 % выборок (18 из 20) статистически значимо отклоняются от ожидаемого соотношения 1: 1 и только две выборки (10%) согласуются с ним. Много это или мало? Как оценить значимость наблюдаемых отклонений в целом для всей совокупности выборок? Ведь при полном согласии картина должна была быть диаметрально противоположной: в среднем только 5 % выборок могли бы значимо отличаться на уровне $\alpha = 0{,}05$ и всего лишь 1 % и 0,1 % — на уровнях $\alpha = 0{,}01$ и $\alpha = 0{,}001$, т. е. из 20 выборок в среднем лишь одной «дозволено» уклониться от проверяемого соотношения на уровне значимости $\alpha = 0,05$ и практически ни одной «не дозволено» — на более низких уровнях.

Чтобы оценить значимость наблюдаемых отклонений в целом для всей совокупности анализируемых выборок, можно использовать статистику χ^2_S, которая является простой суммой индивидуальных величин $\chi^2_i\{1\colon 1\}$ (см. табл. П.2). Поскольку выборки по годам можно считать независимыми, то независимыми будут и случайные величины $\chi^2_i\{1\colon 1\}$. Тогда их сумма χ^2_S также будет случайной величиной χ^2 с числом степеней свободы, равным числу независимых слагаемых $\nu_S = r$, и может служить статистикой критерия для проверки $H_0\colon \forall(a_i\colon b_i = 1\colon 1)$, т. е. гипотезы о том, что все выборки в совокупности удовлетворяют соотношению 1: 1.

Вычисленное значение $\chi^2_S = 366{,}51$ опять-таки столь велико, что можно оценить лишь порядок соответствующего значения P_s. Полученная оценка значимости критерия $P_S = 1{,}7 \cdot 10^{-65}$ на много порядков превосходит оценку Дж. Арбатнота и тем самым подтверждает и усиливает его вывод о явной неслучайности преобладания мальчиков среди новорожденных.

П.11. Проверка согласия с наблюдаемым соотношением суммарных численностей полов $A : B$

Дж. Арбатнот в своей статье выразил убеждение, что избыток мальчиков рождается с постоянной частотой, так как эта частота колеблется, на его взгляд, в узких фиксированных пределах. По данным табл. П.1, частоты рождения мальчиков по годам $\widehat{p}_i$ колеблются в пределах от 0,5084 в 1679 г. до 0,5362 в 1661 г.

Возникает вопрос: прав ли Арбатнот, т. е. насколько подобные колебания статистически допустимы и не противоречат основному постулату биномиальной модели о том, что вероятность рождения мальчика p есть величина постоянная.

Применительно к данным Арбатнота одними из первых пытались решить этот вопрос Н. Бернулли и В. Я. С'Гравесанде: они дали на него положительный ответ. Однако предложенный ими математический аппарат оказался несовершенным, и только в конце XIX — первой половине XX в. были изобретены мощные и эффективные (в математическом смысле этих слов) методы статистического анализа, в частности — критерий χ^2 (К. Пирсон) и метод максимального правдоподобия (Р. А. Фишер).

Итак, предшествующий анализ убедительно показал, что расщепление по полу у новорожденных не согласуется с простой гипотезой о соотношении 1: 1; каково же истинное соотношение — точно неизвестно. Если постулаты принятой модели соблюдены: вероятность рождения мальчика p постоянна и члены выборок и сами выборки независимы, то, согласно *закону больших чисел*, с ростом объема выборок ($n_i \to \infty$) итоговая частота $\widehat{P}$ стабилизируется ($\delta\{\widehat{P}\} \to 0$) и асимптотически приближается к своему истинному значению ($\widehat{P} \to p$). Тогда соотношение суммарных численностей мальчиков и девочек $A : B$ будет отражать реальное соотношение полов и служить его несмещенной и эффективной оценкой.

Поскольку объемы выборок n_i достаточно велики, то проверить, согласуются ли наблюдаемые расщепления по годам с этой оценкой, можно с помощью критерия χ^2. Удобной для вычислений является формула $\chi_i^2\{A : B\}$ с одной степенью свободы $\nu_{[A:B]} = 1$ (см. табл. П.2). Пример вычисления для первой выборки за 1660 г. приведен в табл. П.3. Результаты вычислений для всех выборок и соответствующие им значения $P_{[A:B]}$ представлены в последних двух графах табл. П.5.

Если бы материал был статистически однородным, то на уровне значимости $\alpha = 0{,}05$ только 5 % выборок должны были бы значимо отличаться от соотношения $A : B$ (практически одна выборка из 20 и ни одной — на более низких уровнях).

Действительно, во многих случаях найденные оценки значимости $P_{[A:B]}$ выше критического значения $\alpha = 0{,}05$, т. е. они удовлетворительно согласуются с оценочным соотношением $A : B$. Однако в пяти случаях (в 1660, 1661, 1676, 1677 и 1679 гг.) отклонения от проверяемого соотношения $A : B$ оказались статистически значимыми на уровне α=0,05, из них два (в 1660 и 1661 гг.) — на уровне α=0,01, один (в 1661 г.) — на уровне α=0,001

(в табл. П.5 они отмечены одной, двумя и тремя звездочками соответственно).

П.12. Графический подход к проверке согласия с суммарным соотношением $A : B$

В графической форме проверка согласия с суммарным соотношением $A : B$ представлена на рис. П.3, аналогичном рис. П.5 и П.2. Здесь наклонная сплошная прямая есть геометрическое место точек, соответствующее наблюдаемому соотношению суммарных численностей полов $A : B$, или (что то же самое) сводной выборочной оценке рождения мальчиков $\widehat{P} = 0{,}5188$. Окружающие ее тонкие пунктирные линии суть границы 95 %-, 99 %- и 99,9 %-ных областей принятия гипотез H_0: $A : B = A : B$ (их построение аналогично показанному в табл. П.4).

Как и должно было быть, половина эмпирических точек располагается над оценочной прямой, а другая половина — под ней.

На глаз все они группируются вокруг нее достаточно тесно (неспроста их разброс показался Дж. Арбатноту незначительным, а В. Я. С'Гравесанде и Н. Бернулли не смогли его опровергнуть). Тем не менее, за пределами 95 %-ной области принятия располагаются не 5 % данных, а 25 % — не одна, а пять точек: в 1660, 1661, 1676, 1677 и 1679 гг. Это именно те года, для которых по критерию χ^2 отклонения от суммарного соотношения полов $A : B$ оказались существенными на уровнях значимости $\alpha = 0{,}05$ и $\alpha = 0{,}01$ (см. табл. П.5).

П.13. Проверка однородности выборочных распределений

Итак, возникает подозрение, что параметр p не есть величина постоянная. Следовательно, нарушается, по меньшей мере, один из постулатов выдвинутой вероятностной модели, и вероятности рождения мальчиков в разные годы принципиально различны.

Проверяемая нулевая гипотеза H_0: $\forall(a_i\colon b_i = A : B)$ о согласии всей совокупности выборок с наблюдаемым суммарным соотношением $A : B$ называется гипотезой об *однородности* r биномиальных распределений. Соответствующей проверочной статистикой может служить χ^2_H для *таблиц сопряженности* $2 \times r$ (два пола и r сравниваемых выборок), которую можно вычислить как сумму r слагаемых $\chi^2_i\{A : B\}$ (см. формулу (III) в табл. П.2). В отличие от внешне похожей статистики χ^2_T число степеней свободы здесь на единицу меньше числа слагаемых $\nu_H = r - 1$, поскольку эти слагаемые

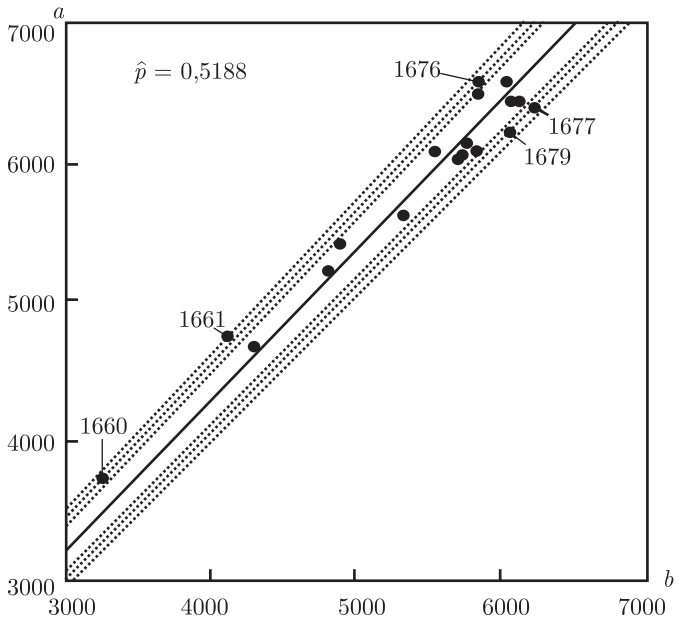


Рис. П.3. Графический подход к проверке согласия с наблюдаемым суммарным соотношением $A : B$. Ось абсцисс — количество новорожденных девочек (a); ось ординат — количество новорожденных мальчиков (b). Сплошная прямая соответствует выборочной оценке вероятности рождения мальчиков $\widehat{P} = 0{,}5188$, а окружающие ее пунктирные линии — суть границы двухсторонних 95-, 99- и 99,9 %-ных областей принятия проверяемой гипотезы H_0: $p = \widehat{P}$. Отмечены пять точек (в 1660, 1661, 1676, 1677 и 1679 гг.), для которых отклонения от выборочного значения $\widehat{P}$ существенны на уровне значимости $\alpha = 0{,}05$ (ср. табл. П.5)

не являются независимыми, так как в каждом из них в качестве оценки ожидаемого соотношения использованы суммарные численности A и B.

Формула (IV) в табл. П.2 принадлежит Маргарет Элиз Уоллас (M. E. Wallace) — ученице и сотруднице Р. А. Фишера, известной открытием *аффинности хромосом* и другими работами по генетике мыши. Формула Уоллас показывает, что статистики χ^2_H, χ^2_S и χ^2_T связаны между собой простым соотношением, потому ею удобно пользоваться, когда подсчитаны χ^2_S и χ^2_T и нет заинтересованности в значениях индивидуальных компонент $\chi^2_i\{A : B\}$. В данном частном случае, когда $p_o = q_0 = 1/2$, пер-

вый сомножитель в этой формуле равен $N^2(4AB)^{-1}$ (см. пример расчета в табл. П.3)[13].

Когда же нет заинтересованности в значениях ни $\chi_i^2\{A:B\}$, ни χ_S^2 или χ_T^2, тогда χ_H^2 признано удобным вычислять по известной формуле Брандта–Снедекора для таблиц сопряженности $2\times r$ (формула (V) в табл. П.2).

Полученное значение $\chi_H^2 = 50{,}19$ и соответствующая ему оценка $P_H = 1{,}3\cdot 10^{-4}$ свидетельствуют о статистически высокозначимой неоднородности (гетерогенности) расщепления по полу в анализируемой популяции.

П.14. Статистическая оценка вероятности рождения мальчиков

В последней строке табл. П.1 приведены две сводные величины: $\widehat{P}$ — доля (частота) рождений мальчиков за все $r = 20$ лет, $\overline{p}$ — среднее из $r = 20$ годовых частот $\widehat{p}_i$. Полученные их значения $\widehat{P} = 0{,}51878$ и $\overline{p} = 0{,}51935$ достаточно близки, однако только после проверки однородности выборочных распределений можно сказать, какую из них (исходя из статистических свойств) следует признать более реалистичной, более надежной оценкой вероятности рождения мальчиков.

Если бы данные были однородны (когда $P_H > \alpha$), то это означало бы, что все $r = 20$ выборок принадлежат одному биномиальному распределению с общим параметром p и величина $\widehat{P}$ была бы его эффективной оценкой, т. е. оценкой с минимальной стандартной ошибкой $\delta\{\widehat{P}\}$.

Поскольку же материал оказался существенно неоднородным ($P_H = 1{,}3\cdot 10^{-4}$), в качестве оценки вероятности рождения мальчика надежнее использовать не $\widehat{P}$, а *среднее* $\overline{p}$ из r выборочных частот $\widehat{p}_i$, представленных в последней графе табл. П.1.

Соответствующую *стандартную ошибку* $\delta\{\overline{p}\}$ можно оценить как ошибку выборочного среднего (см. табл. П.2 и П.3); в итоге для анализируемого фрагмента данных Дж. Арбатнота получаем оценку

$$0{,}519 \pm 0{,}002.$$

Теперь можно оценить *доверительный интервал*, в котором с *доверительной вероятностью* $1-\alpha$ находится искомый параметр p. Необходимое для этого значение $t\{\nu;\alpha/2\}$, т. е. соответствующее данному числу

[13] Эта формула опубликована в журнале Biometrika (1963. V. 50. № 3–4. P. 547–549); в рецензии М. Э. Уоллас на книгу: Bailey N. T. J. Introduction to the Mathematical Theory of Genetic Linkage. — Oxford: Oxford University Press, 1961.

степеней свободы $\nu = r - 1$ и уровню значимости α критическое значение *t-распределения Стьюдента*, можно найти в таблицах или с помощью указанных пакетов программ. Для данного случая $t\{\nu = 19, \alpha/2 = 0{,}025\} = 2{,}09$; следовательно, с вероятностью 95 % искомый параметр p находится в пределах

$$0{,}515 \leqslant p \leqslant 0{,}523,$$

что удовлетворительно согласуется с результатами мировой демографической статистики.

Указание доверительного интервала есть не что иное, как способ наглядной интерпретации результата проверки гипотезы H_0: $p = p_0$. Если бы полученный интервал накрывал постулированное значение $p_0 = 0,500$, то у нас не было бы оснований отвергнуть нулевую гипотезу. В данном случае доверительный интервал не накрывает это значение и вероятность того, что истинное значение параметра p находится за пределами этого интервала, достаточно мала — она не превышает выбранного уровня значимости $\alpha = 0,05$.

П.15. Основные итоги анализа вторичного соотношения полов и пути его дальнейшего изучения

Окончательные результаты проведенного анализа полезно представлять в компактном виде, наподобие таблицы дисперсионного анализа, выявляющего основные источники варьирования в изучаемом явлении (табл. П.7). Можно видеть, что значение статистики χ^2_S почти равно сумме значений статистик χ^2_T и χ^2_H (очевидно, с точностью до стремящегося к единице коэффициента $N^2(AB)^{-1}p_0q_0$ в формуле Уоллас). Это означает, что в совокупное отклонение от соотношения 1: 1 вносят вклад два практически аддитивных компонента изменчивости: отклонение от 1: 1 для суммарных численностей A и B и отклонение выборочных распределений от однородности.

В данном примере все три типа анализируемых отклонений оказались статистически высокозначимыми. В итоге выявлены два фундаментальных явления. Во-первых, вероятность рождения мальчика превышает 1/2 и, во-вторых, эта вероятность непостоянна.

Все последующие многочисленные исследования лишь подтверждали это. В частности, обнаружено, что вероятность рождения мальчика снижается с возрастом родителей и с порядком рождения. Эти факты имеют разумные биологические обоснования, поскольку процесс оплодотворения

Таблица П.7

Результаты анализа методом χ^2 вторичного соотношения полов среди детей, родившихся в Лондоне в 1660–1679 гг. (по *J. Arbuthnot, 1710*)

Анализируемый источник изменчивости	χ^2	ν	P
Отклонение суммарных численностей от 1: 1	$\chi^2_T = 316{,}39$	$\nu_T = 1$	$P_T = 8{,}9 \cdot 10^{-71}$
Отклонение от однородности	$\chi^2_H = 50{,}19$	$\nu_H = 19$	$P_H = 1{,}3 \cdot 10^{-4}$
Отклонение от 1: 1 для всей совокупности выборок	$\chi^2_S = 366{,}51$	$\nu_S = 20$	$P_S = 1{,}7 \cdot 10^{-65}$
$\chi^2_T + \chi^2_H = 316{,}39 + 50{,}19 = 366{,}58 \approx \chi^2_S$			

и эмбрионального пренатального развития зависит от гормонального и иммунологического статуса родителей. Различными гормональными воздействиями оказывается возможным снизить вероятность рождения мальчика до 0,3 или повысить до 0,8.

Современные исследования вторичного соотношения полов (и не только у человека) статистическими методами продвигаются в двух направлениях. Одно из них — поиск вероятностного распределения, адекватно описывающего это явление. Из изложенных результатов ясно, что дисперсия такого распределения должна быть большей, чем у биномиального. Такие модели обсуждаются в научной литературе — это, например, распределения Лексиса, отрицательное биномиальное, или модели, основанные на марковских процессах, когда параметр p непостоянен, и др. Другой путь — разложить явление на некие элементарные составляющие, для которых биномиальная модель выполняется.

Половая дифференциация — явление, широко распространенное в природе, и ее возникновение, назначение, механизмы и эффекты нуждаются в дальнейшем изучении. С точки зрения генетики, это, прежде всего, генетический полиморфизм, для плодотворного изучения которого необходимо сочетание идей и методов популяционной генетики с идеями и методами биометрии.

П.16. Пример из сельскохозяйственной генетики: Наследование окраски шерсти у крупного рогатого скота породы шортгорн

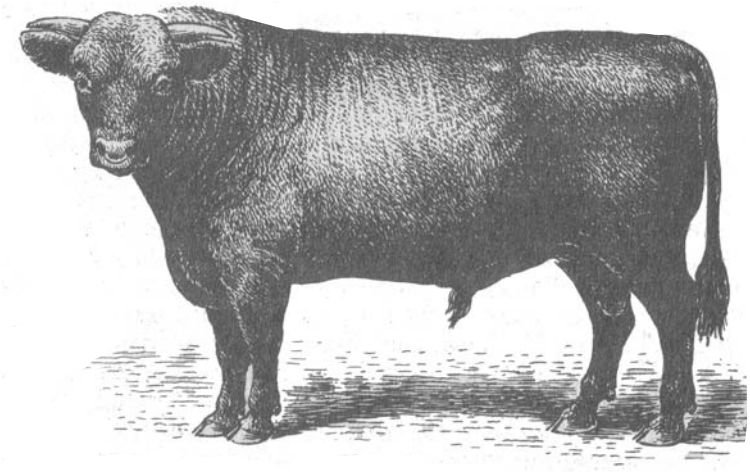
Первую попытку проверки справедливости менделевских закономерностей для домашних сельскохозяйственных животных предприняли в 1906 г. знаменитый биометрик Карл Пирсон (K. Pearson, 1857–1936) и его сотрудница Эми Баррингтон (A. Barrington). Для этого они обследовали многотомные *племенные книги* крупного рогатого скота, которые впервые в мире основал Джордж Коутс (G. Coates) в 1822 г. в Англии. В этих книгах по настоящий день регистрируются основные сведения о всех получаемых в племзаводах чистопородных и метисных (помесных) животных: масть (окраска шерсти), масса, экстерьер, происхождение, т. е. сведения о родителях и прародителях. Э. Баррингтон и К. Пирсон первыми осознали, какой богатейший для генетиков материал таят в себе такие племенные книги: на основе содержащейся в них информации можно проследить родословные зарегистрированных животных и выявить закономерности наследования фиксируемых в них признаков, в первую очередь качественных (например, масть).

П.17. Проблема выбора признака и модели его наследования

В качестве признака для анализа Э. Баррингтон и К. Пирсон выбрали окраску шерсти у крупного рогатого скота породы шортгорн (Shorthorn — короткорогая, по краниологическим признакам выделена в подвид *Bos taurus brachyceros*). В Англии это самая популярная и ценная мясо-молочная порода. Еще в XIX в. за отдельных ее особей предлагались баснословные суммы в десятки тысяч долларов. Как курьез можно отметить, что из двух быков-производителей этой породы, впервые импортированных в Америку в 1817 г., один носил кличку Moscow.

Обычно селекционеры-племеноводы предпочитают создавать одномастные породы, но при выведении породы шортгорн отбору по масти особого внимания, по-видимому, не уделялось. На одном и том же заводе масть животных варьирует от белой до темно-красной с различными их сочетаниями; темно- или светло-красные с белыми отметинами («звезды» во лбу, белые оторочки на ногах), пестрые (пятнистые) с крупными или мелкими пятнами, мраморные или совсем белые. Явно особый тип представляет *чалая* масть. В этом случае следует, очевидно, судить не о цветовом оттенке, промежуточном между красным и белым, а о чалости как явлении морфо-

генеза, когда шерсть у особей представляет собой смесь красных и белых волос. Здесь также наблюдаются вариации: красно- (или темно-) чалые, чалые, светло-чалые, чалые с отметинами.

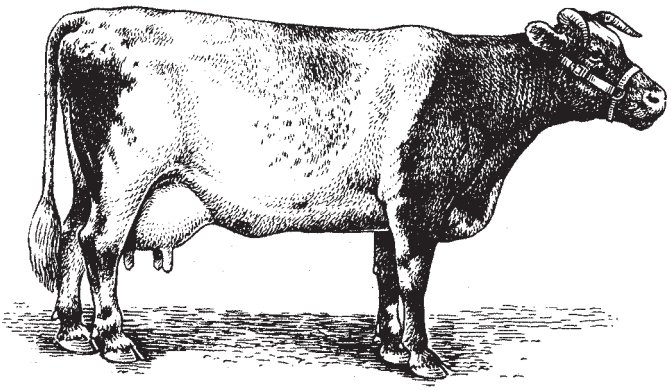


а

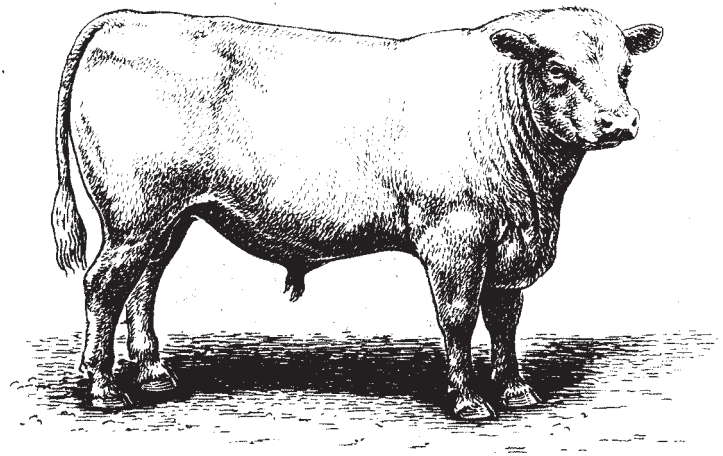
Рис. П.4. Три основных типа окраски шерсти у крупного рогатого скота породы шортгорн. *а* — красная; *б* — чалая, *в* — белая. Окраска чалых особей (гетерозиготы Rr) сильно варьирует, и иногда внешне их трудно отличить от особей со сплошной красной (гомозиготы RR) или сплошной белой (rr) окраской

В те времена сведения о биосинтезе пигментов и об образовании волосяного покрова были чрезвычайно скудны, и потому трудно было предугадать, какие из многочисленных вариантов окраски являются альтернативными, т. е. определяются аллелями одного локуса. Э. Баррингтон и К. Пирсон первыми предположили, что три масти — красная, белая и чалая — определяются одним геном с двумя кодоминантными аллелями (R — аллель красной окраски, r — аллель белой окраски) и что красные (К) и белые (Б) особи суть, соответственно, гомозиготы RR и rr, а чалые (Ч) — гетерозиготы Rr (рис. П.4). Ожидаемые согласно такой модели наследования соотношения фенотипов у потомства, получаемого от всех шести возможных типов спаривания между родителями, представлены в табл. П.17.

Данные, извлеченные Э. Баррингтон и К. Пирсоном из нескольких томов племенных книг Коутса, серьезно противоречили предложенной схе-



б



в

Таблица П.8

Частоты фенотипов, ожидаемые у потомства шортгорнов от каждого из возможных типов скрещивания между родителями согласно однолокусной с двумя кодоминантными аллелями (R и r) схеме наследования масти, предложенной Э. Баррингтон и К. Пирсоном (1906)

Типы скрещиваний		Потомки, %		
Фенотипы	Генотипы	К(RR)	Ч(Rr)	Б(rr)
К×К	$RR \times RR$	100		
К×Ч	$RR \times Rr$	50	50	
К×Б	$RR \times rr$		100	
Ч×Ч	$Rr \times Rr$	25	50	25
Ч×Б	$Rr \times rr$		50	50
Б×Б	$rr \times rr$			100

ме наследования: расщепления наблюдались во всех типах скрещиваний, а в семьях одномастных родителей выщеплялись пятнистые потомки. Обнаружив это, Баррингтон и Пирсон вообще усомнились в справедливости менделевских закономерностей. В этой связи Н. И. Вавилов (1887–1943) вспоминал, как еще в 1914 г. ему «пришлось быть в Лондоне на одном из вечеров в знаменитом Королевском институте, где Карл Пирсон демонстрировал пятнистых собачек и их потомство, выявлявших сложную картину расщепления, пытаясь при этом высмеять учение Менделя и его последователей»[14].

Ярое неприятие менделизма биометриками во главе с Карлом Пирсоном и Уолтером Фрэнком Рафейлом Уэлдоном (W. F. R. Weldon, 1860–1906) — драматическая (отчасти трагическая) страница в истории биологии. «Уэлдон бросил все свои силы на опровержение менделизма и в поисках исключений из его законов пересмотрел множество огромных томов заводских племенных книг породистых лошадей, подорвал свое здоровье и умер молодым»[15].

В 1908 г. Дж. Уилсон (J. Wilson), тщательно проанализировав историю шортгорнской породы крупного рогатого скота, пришел к выводу, что пятнистых, красных с белым особей следует объединять с красными в один фенотипический класс вне зависимости от степени их пятнистости, по-

[14] Вавилов Н. И. Предисловие к кн.: Мендель И. Г. Опыты над растительными гибридами. — М.-Л.: ОГИЗ Сельхозгиз, 1935.

[15] Кимура М. Молекулярная эволюция: теория нейтральности. — М.: Мир, 1985. — С.18.

скольку эта порода происходит от двух рас — белой романской и красной саксонской.

По его мнению, при такой классификации фенотипов данные, собранные Баррингтон и Пирсоном, удовлетворяют предложенной ими однолокусной с двумя кодоминантными аллелями модели наследования. Однако и в этом случае встречались не согласующиеся с данной моделью варианты выщеплений. Их Уилсон предложил объяснять неизбежными в племенном деле ошибками регистрации и ошибками установления родства.

Позднее предлагались более сложные модели — двух- и четырехлокусные, которые на качественном уровне, казалось, лучше описывают наблюдаемые расщепления, но к единому мнению исследователи не пришли. Один из них даже пошутил: единственное, что имеется общего в работах разных авторов, так это их противоречивость.

В 1917 г. изучением этой проблемы занялся Сьюалл Райт. Его работа в Journal of Heredity имеет непреходящее историческое значение как первое практическое применение закона Харди–Вайнберга, которое позволило получить веские количественные аргументы в пользу однолокусного с двумя кодоминантными аллелями способа наследования окраски шерсти у шортгорнов. Однако и ему для объяснения наблюдаемых отклонений от ожидаемых расщеплений пришлось постулировать наличие неких генов-модификаторов, искажающих проявление этого признака. Позднее, повторив исследование более тщательно на новом обширном материале, извлеченном из племенных книг трех стран (Англии, Америки и Канады за 1921 г.), он даже ввел новое понятие — *перекрывание фенотипов*. Проведенный Райтом анализ (проверка случайности скрещиваний, количественное сравнение трех различных моделей наследования, оценка степени «перекрывания фенотипов» и пр.) весьма поучителен, но, к сожалению, слишком громоздок для пересказа в данном учебнике. Поэтому интересующихся можно отослать к его изложению в 1-м (с. 193–195) и в 3-м (с. 526–533) томах капитального четырехтомного руководства «Эволюция и генетика популяций», написанного С. Райтом в 1968–1978 гг. в возрасте 79–89 лет.

Решающее слово в споре о наследовании масти у шортгорнов сказал в 1947 г. Иан Честер-Джоунс (I. Chester-Jones, 1916), впоследствии известный специалист по ветеринарной эндокринологии. Он тщательно обследовал небольшое стадо шортгорнов герцога Вестминстерского в Итоне, где непосредственному наблюдению были доступны три поколения особей, а регистрация находилась в руках исключительно компетентного человека. В результате оказалось, что основной причиной всех недоразумений является варьирование окраски гетерозигот от полностью красной до полностью

белой и отличить такие крайние варианты от гомозигот удается только после повторной экспертизы, изучая под микроскопом распределение пигментных зерен в образцах волос.

Наименее противоречивые данные о наследовании окраски у породы шортгорн опубликовал в 1937 г. Э. Робертс (E. Roberts, 1886–1969). Он исследовал относительно небольшое стадо шортгорнов Иллинойсского университета. Здесь идентификацию и регистрацию признаков проводили, по-видимому, наиболее тщательно: среди результатов нескольких сотен скрещиваний был лишь один исключительный, спорный случай, противоречащий обсуждаемой схеме наследования[16].

Данные о трех поколениях шортгорнов в той форме, как они опубликованы Робертсом, воспроизведены в табл. П.9 и П.10. Такая форма представления, предложенная еще Уилсоном, компактна, информативна и позволяет провести всесторонний статистический анализ данных, как то:

а) проверить предложенную однолокусную с двумя кодоминантными аллелями модель наследования масти, т. е. проверить согласие наблюдаемых расщеплений с менделевскими соотношениями;

б) проверить равновесность, т. е. согласие с законом Харди–Вайнберга в трех поколениях особей;

в) проверить случайность скрещиваний среди особей двух поколений: первого и второго — родителей и прародителей.

П.18. Проверка модели наследования

Согласно модели (см. табл. П.17), расщепление ожидается в потомстве только трех типов скрещиваний: К×Ч, Ч×Ч и Ч×Б. Проверку согласия с ожидаемыми для них менделевскими соотношениями 1К : 1Ч, 1К : 2Ч : 1Б и 1Ч : 1Б можно провести, как обычно, с помощью критерия χ^2. В табл. П.11 представлены результаты такой проверки для суммарных данных из табл. П.9 и П.10 и удобные для вычислений формулы $\chi^2\{1\colon 2\colon 1\}$ и $\chi^2\{1\colon 1\}$ (последняя уже использовалась ранее в табл. П.2).

Поскольку типы скрещиваний можно считать независимыми, все три статистики также будут независимыми и их сумма χ^2_M с суммарным числом степеней свободы $\nu = 4$ может служить статистикой критерия для проверки расщеплений в целом.

[16] Roberts. E. Color inheritancein Shorthorn cattle // J. Heredity. — 1937. — Vol. 18. — №4. — P. 167–168.

Таблица П.9

Результат скрещивания шортгорнов первого поколения (прародителей) в стаде Иллинойсского университета (по E. Roberts, 1937)

Типы скрещиваний	Наблюдаемые численности потомков			
	К	Ч	Б	∑
К×К	33			33
К×Ч	21	40		61
К×Б		43		43
Ч×Ч	9	30	12	51
Ч×Б		25	10	35
Б×Б			1	1
∑	63	138	23	224

Таблица П.10

Результат скрещивания шортгорнов второго поколению (родителей) в стаде Иллинойсского университета (по E. Roberts, 1937)

Типы скрещиваний	Наблюдаемые численности потомков			
	К	Ч	Б	∑
К×К	18			18
К×Ч	49	35		84
К×Б		12		12
Ч×Ч	38	66	41	145
Ч×Б		28	29	57
Б×Б			2	2
∑	105	141	72	318

Таблица П.11

Проверка однолокусной с двумя кодоминантными аллелями модели наследования масти у крупного рогатого скота породы шортгорн (по E. Roberts, 1937.)

Типы скрещиваний		Наблюдаемые численности потомков			Ожидаемые соотношения	Вычисленные значения статистик		
Фенотипы	Генотипы	a (К, RR)	b (Ч, Rr)	c (Б, rr)		χ^2	ν	P
К×Ч	$RR \times Rr$	70	75		1: 1	0,17	1	0,68
Ч×Ч	$Rr \times Rr$	47	96	53	1: 2: 1	0,45	2	0,80
Ч×Б	$Rr \times rr$		53	39	1: 1	2,13	1	0,14
$\sum$						$P_M = 0,60$ $\chi^2_M = 2,75$ $P_M = 0,60$	$\nu_M = 4$	

Статистики для проверки гипотез

H_0	χ^2	ν
К: Ч = 1: 1	$\chi^2\{1\colon 1\} = (a-b)^2(a+b)^{-1}$	$\nu_{[1\colon 1]} = 1$
К: Ч: Б = 1: 2: 1	$\chi^2\{1\colon 2\colon 1\} = (a+b+c)^{-1}4(a^2+b^2/2+c^2) - -(a+b+c)$	$\nu_{[1\colon 2\colon 1]} = 2$
Ч: Б = 1: 1	$\chi^2\{1\colon 1\} = (b-c)^2(b+c)^1$	$\nu_{[1\colon 1]} = 1$

Полученные значения всех четырех статистик и соответствующие им P-значения свидетельствуют о хорошем согласии анализируемых данных с предложенной моделью наследования окраски шерсти у крупного рогатого скота породы шортгорн.

П.19. Вероятностная модель популяции с двумя кодоминантными аллелями одного аутосомного гена и ее параметры

Построим вероятностную модель популяции с двумя кодоминантными аллелями одного аутосомного гена.

В общем случае для таких аллелей введем обозначения $A*1$ и $A*2$, при этом все три генотипа $A*1/A*1$, $A*1/A*2$ и $A*2/A*2$ различимы фенотипически: A1; A1,2 и A2, соответственно. Такая система обозначений рекомендуется международными комиссиями по генетической номенклатуре.

Модель, которая описывает такую популяцию, и схема статистического анализа соответствующих экспериментальных данных представлены в табл. П.12. Модель определяется двумя параметрами: p — *частота аллели* $A*1$ и F — *индекс фиксации*.

F как меру неравновесности популяции впервые ввел С. Райт в 1921 г. (фиксированной он называл популяцию, в которой отсутствуют гетерозиготы). Индекс фиксации задается любым из следующих трех тождественных выражений:

$$F = (P - p^2)(pq)^{-1} \equiv (2pq - Q)(2pa)^{-1} \equiv (R - q^2)(pq)^{-1},$$

т. е. является функцией частот аллелей (p и q) и генотипов (P, Q и R). Эту функцию можно интерпретировать как *коэффициент корреляции* между гаметами, т. е. как меру неслучайности их объединения. Область принимаемых им значений заключается в пределах от -1 до $+1$:

$$-1 \leqslant F \leqslant 1.$$

Когда гаметы объединяются случайно (свободно) и популяция находится в равновесии, тогда $F = 0$. Крайние значения F принимает, когда объединяются только одноименные гаметы: $A*1$ с $A*1$ и $A*2$ с $A*2$, тогда $F = 1$, или же когда объединяются лишь разноименные гаметы, тогда $F = -1$. В первом случае популяция будет состоять из одних гомозигот, во втором — из одних гетерозигот.

Индекс фиксации является обобщающим показателем неравновесности популяций. Когда популяция находится в стационарном состоянии, он связан с другими мерами неравновесности простыми соотношениями:

$$F = M(2-M)^{-1} = pq[w_1w_3 - w_2^2][w_1p + w_2q]^{-1}[w_2p + w_3q]^{-1} = \sigma^2(pq)^{-1}.$$

Здесь M — коэффициент *ассортативности* спаривания; w_1, w_2 и w_3 — относительные *приспособленности* генотипов $A*1/A*1$, $A*1/A*2$ и $A*2/A*2$ соответственно; σ^2 — дисперсия частоты аллели $A*1$ в *подразделенной* (расслоенной) популяции согласно *принципу Валунда*.

Соответственно, биологическая интерпретация индекса фиксации может быть самой разнообразной. Чаще всего его интерпретируют как *коэффициент инбридинга*: в случае инбридинга $F > 0$, а в случае аутбридинга $F < 0$. Когда инбридинг есть единственная и постоянно действующая на популяцию сила ($F = \mathrm{const}$), тогда соответствующие математические выражения для частот генотипов часто именуют *законом равновесия Райта*. Индекс фиксации может принимать положительные значения и по другим причинам: когда размер популяции мал (конечен) или когда популяция подразделена на субпопуляции. Отрицательные значения он может принимать, когда частоты аллелей у полов различны. Это возможно вследствие их дифференциальной жизнеспособности, или плодовитости (фертильности), либо вследствие различной интенсивности отбора среди мужских и женских особей. Последнее часто имеет место в популяциях сельскохозяйственных животных, где мужские особи (производители) подвергаются более жесткому отбору, нежели женские.

На рис. П.5 представлен один из возможных способов графического изображения того, как меняются частоты генотипов при различных значениях индекса фиксации. В частности, можно видеть, что индекс фиксации принимает отрицательные значения далеко не при любых значениях частот аллелей p и q.

Очевидно, что в модели Райта статистическая проверка согласия с законом Харди–Вайнберга сводима к проверке гипотезы о равенстве нулю индекса фиксации: H_0: $F = 0$.

П.20. Проверка согласия с законом Харди–Вайнберга и оценка параметров модели Райта

Статистическую проверку равновесности популяции часто проводят по результатам регистрации фенотипов особей в одной-единственной генерации. Однако такая процедура допустима далеко не всегда. Общим правилом здесь является следующее. Проверить, находится ли популяция в генетическом равновесии, можно в тех случаях, когда число аллелей меньше числа проявляемых фенотипов. Примерами могут служить:

а) две кодоминантные аллели, $A*1$ и $A*2$, и, соответственно, три различающихся фенотипа: A1, A1,2 и A2;

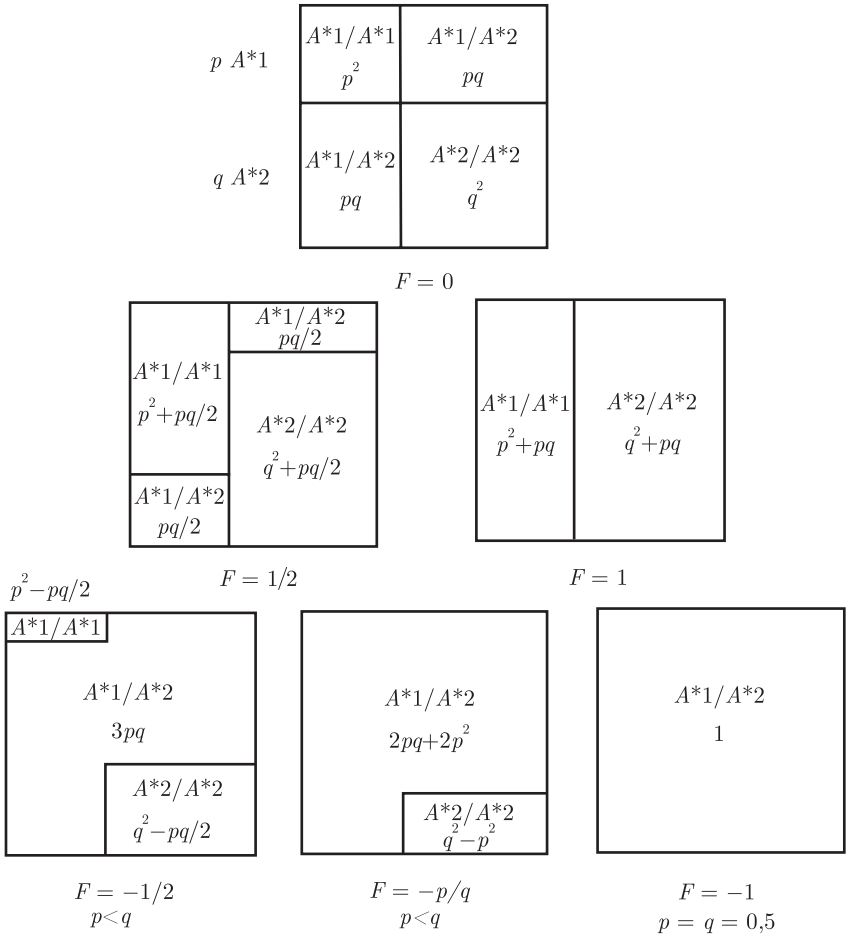


Рис. П.5.

б) системы, подобные системе АВ0 групп крови человека, — три аллели, две из которых кодоминантные, $AB0*A$, $AB0*B$, и одна рецессивная, $AB0*0$; соответственно различимы четыре фенотипа; A, B, 0 и AB;

в) две аллели — доминантная $*A$ и рецессивная $*a$, но ген сцеплен с полом; соответственно, различимы четыре фенотипа: самки A и a и самцы A и a.

Рис. П.5 (продолжение). Графическое представление модели Райта для популяции с двумя кодоминантными аллелями $A*1$ и $A*2$ одного аутосомного гена A.

Представлены соотношения генотипов в популяциях с различными значениями индекса фиксации F. Площадь каждого большого квадрата равна единице, а площади составляющих их прямоугольников пропорциональны частотам генотипов. Вверху — равновесная панмиктическая популяция, удовлетворяющая закону Харди–Вайнберга ($F = 0$). В центре — две популяции с положительными значениями индекса фиксации $F = 1/2$ и $F = 1$. Внизу — три популяции с отрицательными значениями индекса фиксации: $F = -1/2$ и $F = -1$ (при условии, что $p < q$).

С ростом положительных значений F происходит вытеснение гетерозигот из «популяционного пространства», вплоть до их полного отсутствия, когда индекс фиксации принимает свое крайнее положительное значение $F = 1$; при этом частоты аллелей p и q не зависят от F. Когда же из популяционного пространства вытесняются гомозиготы, то F принимает отрицательные значения. При $F = p/q$ (когда $p < q$) или $F = q/p$ (когда $q < p$) полностью исчезает лишь одна из гомозигот (более редкая). Полное отсутствие обеих гомозигот ($F = -1$) достигается, только когда частоты аллелей равны друг другу ($p = q = 1/2$)

Если, к примеру, ген не сцеплен с полом и аллель $*A$ полностью доминирует над $*a$, то различимы только два фенотипа A и a, и тогда для проверки равновесности необходима дополнительная информация о распределении фенотипов в двух последовательных генерациях (поколениях) родителей и потомков. В терминах статистики это означает, что статистика критерия для проверки равновесности должна обладать хотя бы одной степенью свободы.

Когда объемы выборок достаточно велики, проверить гипотезу о равновесности популяции в соответствии с законом Харди–Вайнберга, т. е. проверить нулевую гипотезу H_0: $F_0 = 0$, можно двумя путями: либо использовать критерий χ^2, либо найти интервальную оценку для индекса фиксации F.

В табл. П.12 указаны две формулы для соответствующей статистики χ^2_E с одной степенью свободы $\nu_E = 1$. Формула (III) не требует вычисления ожидаемых численностей генотипов. Но вычисление χ^2_i-компонент по формуле (IV) более информативно: они служат ориентиром для выявления тех классов, которые вносят наибольший вклад в наблюдаемые отклонения от закона Харди–Вайнберга, а знаки входящих в них разностей наблюдаемых и ожидаемых численностей $(n_i - \widehat{n})_i$ указывают на направление таких отклонений. Каждая такая компонента имеет одну степень свободы, но и их сумма χ^2_E также имеет всего одну степень свободы. В подобных случаях число степеней свободы для суммарной статистики χ^2_E легко опре-

Таблица П.12

Модель популяции с двумя кодоминантными аллелями $A*1$ и $A*2$ одного аутосомного гена A и схема статистического анализа экспериментальных данных

Аллели

$A*1$ $A*2$

Частоты аллелей

p q

Фенотипы

A1 *A1, 2* *A2*

Генотипы

$A*1/A*1$ $A*1/A*2$ $A*2/A*2$

Ожидаемые частоты генотипов (или фенотипов)

$P = p^2 + pqF$ $Q = 2pq - 2pqF$ $R = q^2 + pqF$

Наблюдаемые численности генотипов (или фенотипов)

n_1 n_2 n_3

Наблюдаемые численности аллелей

$m_1 = 2n_1 + n_2$ $m_2 = n_2 + 2n_3$

Планирование объема выборки N

H_0: $F = 0$; H_1^*: $F \geqslant F_1 > 0$ *(или $F \leqslant F_1 < 0$);*

(I) $$\widehat{N}^* \geqslant [(z\{\beta\}W + z\{\alpha\})F_1^{-1}]^2$$

H_0: $F = 0$; H_1^{**}: $|F| \geqslant |F_1|$;

(II) $$\widehat{N}^{**} \geqslant [(z\{\beta\}W + z\{\alpha/2\})F_1^{-1}]^2$$

Оценки параметров модели p и F

$$\widehat{p} = m_1(2N)^{-1};$$
$$\widehat{F} = (4n_1n_3 - n_2^2)(m_1m_2)^{-1}$$

Оценка ожидаемых численностей генотипов (фенотипов) при $F = 0$

$\widehat{n}_1 = N\widehat{p}^2$; $\widehat{n}_2 = 2N\widehat{p}\widehat{q}$; $\widehat{n}_3 = N\widehat{q}^2$;

Окончание табл. П.12

Проверка гипотез

H_0	χ^2	ν
$n_i - \widehat{n}_i = 0$	$\chi_i^2 = (n_i - \widehat{n}_i)^2(\widehat{n}_i)^{-1}$	$\nu_i = 1$
$F = 0$		
(III)	$\chi_E^2 = \widehat{F}^2 N$	
(IV)	$\chi_E^2 = \sum_{i=1}^3 \chi_i^3$	$\nu_E = 1$

Стандартные ошибки оценок p и F

$$\delta\{\widehat{p}\} = [\widehat{p}\widehat{q}(1+\widehat{F})(2N)^{-1}]^{1/2};$$
$$\delta\{\widehat{F}\} = \{[2\widehat{p}\widehat{q}(1-2\widehat{F})(1-\widehat{F})^2 + \widehat{F}(1-\widehat{F})(2-\widehat{F})](2\widehat{p}\widehat{q}N)^{-1}\}^{1/2}$$

$(1-\alpha)$ 100 %-ные доверительные интервалы для параметров p и F

$$\widehat{p} - z\{\alpha/2\}\delta\{\widehat{p}\} < p < \widehat{p} + z\{\alpha/2\}\delta\{\widehat{p}\}$$
$$\widehat{F} - z\{\alpha/2\}\delta\{\widehat{F}\} < F < \widehat{F} + z\{\alpha/2\}\delta\{\widehat{F}\}$$

Обозначения и формулы для промежуточных вычислений

$z\{\beta\}; z\{\alpha\}$ и $z\{\alpha/2\} - \beta$-, $\alpha-$ и $\alpha/2$-квантили нормального распределения;
$W = [(1-2F_1)(1-F_1)^2 + F_1(1-F_1)(2-F_1)(2pq)^{-1}]^{1/2}$, где p — заданное значение частоты аллели $A*1$, $q = 1-p$ и F_1 — заданное значение индекса фиксации;
$N = n_1 + n_2 + n_3$ и $\widehat{q} = 1 - \widehat{p}$.
Вывод формул для N и $\delta\{\widehat{F}\}$ см. в статьях: *Brown A. H. D.* The estimation of Wright's fixation index from genotypic frequences // Genetica. — 1970. — Vol. 41. — № 3. — P. 399–406; *Fyfe J. L., Bailey N. T. J.* Plant breeding studies in leguminous forage crops. I. Natural cross-breeding in winter bean // J. Argic. Sci. — 1951. — Vol. 4. — № 4. — P. 371–378.

делить как разность между числом различающихся фенотипов и числом определяющих их аллелей.

В табл. П.12 указаны также формулы для оценивания параметров p и F, для вычисления стандартных ошибок найденных оценок и для постро-

ения соответствующих доверительных интервалов. Формулы для оценок получены *методом максимального правдоподобия*. Такие оценки обладают хорошими статистическими свойствами: асимптотически, т. е. при больших объемах выборок, они являются *несмещенными, эффективными* и *нормально распределенными*. В частности, оценка $\widehat{F}$ является смещенной, но уже при умеренных объемах выборок $N > 25$ смещение становится пренебрежимо малым. Пример вычислений для популяции телят из табл. П.10 показан в табл. П.13.

Полученное значение $\chi^2_E = 3{,}41$ и соответствующее ему значение $P_E \approx 0{,}065$ свидетельствуют, что в данном случае нет оснований отвергать гипотезу о равновесии. Об этом же наглядно свидетельствует интервальная оценка индекса фиксации, поскольку полученный 95 %-ный доверительный интервал $-0{,}02 < F < 0{,}22$ накрывает проверяемое значение $F = 0$.

П.21. Проблема малой мощности критерия и планирование объема выборки

Анализ одного поколения не способен выявить динамические процессы, происходящие в популяции, выявить силы (факторы), стабилизирующие или нарушающие ее равновесие, ибо информация о состоянии популяции в данный момент времени не позволяет ничего сказать ни о ее прошлом, ни о ее будущем.

Тем не менее, использование только одного измерения, т. е. подсчет частот разных генотипов среди особей одной генерации и сравнение их с ожидаемыми по закону Харди–Вайнберга, остается распространенным приемом. Простота этой процедуры обманчива, поскольку для реалистичных значений интенсивности отбора или других факторов мощность статистического критерия оказывается крайне низкой, и, чтобы выявить статистически значимое отклонение от равновесия Харди–Вайнберга, зачастую нужны выборки гигантского объема. Так, например, если жизнеспособность генотипов различается на 10 %, то, чтобы выявить такое различие, необходима выборка объемом около 4 000 особей, а для выявления различия в 1 % понадобилась бы выборка порядка 400 000 особей.

В табл. П.12 приведены формулы для оценки объема выборки N, необходимого для различения гипотез H_0: $F_0 = 0$ и H_1: $F = F_1 \neq 0$ при односторонних альтернативах H_1^*: $F_1 \geqslant 0$ или $F_1 \leqslant 0$ (формула (I)) и при двусторонней альтернативе H_1^{**}: $|F_1| \geqslant 0$ (формула (II)). В эти формулы кроме конкретного, заданного, значения F_1 входят также конкретное значение параметра p (см. формулу для множителя W в конце таблицы)

Таблица П.13

Проверка равновесности популяции телят породы шортгорн в стаде Иллинойского университета, оценка ее параметров p (частота аллеля R) и F (индекс фиксации) и планирование объема выборки N

Аллели

R r

Частоты аллелей

p q

Фенотипы

К *Ч* *Б*

Генотипы

RR Rr rr

Наблюдаемые численности генотипов (или фенотипов)

$n_1 = 105$ $n_2 = 141$ $n_3 = 72$ $N = 318$

Наблюдаемые численности аллелей

$m_1 = 2 \cdot 105 + 141 = 351;$ $m_2 = 141 + 2 \cdot 72 = 285$

Оценки параметров p и F

$\widehat{p} = 351(2 \cdot 318)^{-1} = 0{,}5519 \approx 0{,}55;$ $\widehat{q} = 1 - 0{,}5519 = 0{,}4481 \approx 0{,}45;$

$\widehat{F} = (4 \cdot 105 \cdot 75 - 141^2)(351 \cdot 285)^{-1} = +0{,}10355 \approx +0{,}10$

Оценка ожидаемых численностей при $F = 0$

$\widehat{n}_1 = 318(0{,}5519)^2 = 96{,}86;$ $\widehat{n}_2 = 2 \cdot 318 \cdot 0{,}5519 \cdot 0{,}4481 = 157{,}29;$

$\widehat{n}_3 = 318(0{,}4481)^2 = 63{,}85;$

$\widehat{n}_1 + \widehat{n}_2 + \widehat{n}_3 = 96{,}86 + 157{,}29 + 63{,}85 = N = 318{,}00$

Проверка гипотез

H_0	χ^2	ν	P
$n_1 - \widehat{n}_1 = 0$	$\chi_1^2 = (105 - 96{,}86)^2(96{,}86)^{-1} = 0{,}68$	$\nu_1 = 1$	$P_1 = 0{,}41$
$n_2 - \widehat{n}_2 = 0$	$\chi_2^2 = (141 - 157{,}29)^2(157{,}29)^{-1} = 1{,}69$	$\nu_2 = 1$	$P_2 = 0{,}19$
$n_3 - \widehat{n}_3 = 0$	$\chi_3^2 = (72 - 63{,}85)^2(63{,}85)^{-1} = 1{,}04$	$\nu_3 = 1$	$P_3 = 0{,}31$
$F = 0$			
(III)	$\chi_E^2 = 0{,}10355^2 \cdot 318 = 3{,}41;$		
(IV)	$\chi_E^2 = 0{,}68 + 1{,}69 + 1{,}04 =$ $= 0{,}10355^2 \cdot 318 = 3{,}41;$	$\nu_E = 1$	$P_E = 0{,}065$

Окончание табл. П.13

Стандартные ошибки оценок p и F

$$\delta\{\widehat{p}\} = [0{,}5519 \cdot 0{,}4481(1 + 0{,}10)(2 \cdot 318)^{-1}]^{1/2} = \pm 0{,}0197 \approx \pm 0{,}02;$$
$$\delta\{\widehat{F}\} = \{[2 \cdot 0{,}55 \cdot 0{,}45(1 - 2 \cdot 0{,}10)(1 - 0{,}10)^2 + 0{,}10(1 - 0{,}10)(2 - 0{,}10)] \times$$
$$\times [2 \cdot 0{,}55 \cdot 0{,}45 \cdot 318]^{-1}\}^{1/2} = 0{,}06023 \approx 0{,}06$$

Вычисление 95 %-ных доверительных интервалов для параметров p и F

$$0{,}55 - 1{,}96 \cdot 0{,}02 < p < 0{,}55 \pm 1{,}96 \cdot 0{,}02;$$
$$0{,}10 - 1{,}96 \cdot 0{,}06 < F < 0{,}10 + 1{,}96 \cdot 0{,}06$$

Конечные результаты оценивания

$$\widehat{p} \pm \delta\{\widehat{p}\} = 0{,}55 \pm 0{,}02; \quad \widehat{F} \pm \delta\{\widehat{F}\} = 0{,}10 \pm 0{,}06$$

95 %-ные доверительные интервалы

$$0{,}51 < p < 0{,}59; \quad -0{,}02 < F < 0{,}22$$

Планирование объема выборки N при $p = 0{,}55$; $|F_1| = 0{,}10$; $\alpha = 0{,}05$ и $\beta = 0{,}10$

$$H_0\text{: } F = 0; \quad H_1^*\text{: } F \geqslant +0{,}10 \quad (\text{или } F \leqslant 0{,}10)$$
$$W = [(1 - 2 \cdot 0{,}10)(1 - 0{,}10)^2 +$$
$$+ 0{,}10(1 - 0{,}10)(2 - 0{,}10)(2 \cdot 0{,}55 \cdot 0{,}45)^{-1}]^{1/2} = 0{,}997$$
$$\widehat{N}^* \geqslant [(1{,}28 \cdot 0{,}997 + 1{,}64) \cdot 0{,}10^{-1}]^2 \geqslant 850; \qquad \text{(I)}$$
$$H_0\text{: } F = 0; \quad H_1^{**}\text{: } |F| \geqslant 0{,}10;$$
$$\widehat{N}^{**} \geqslant [(1{,}28 \cdot 0{,}997 + 1{,}96) \cdot 0{,}10^{-1}]^2 \geqslant 1047 \qquad \text{(II)}$$

Примечание:
P-значения и необходимые для вычислений значения квантилей: $z\{\alpha = 0{,}05\} = 1{,}64$; $z\{\alpha/2 = 0{,}025\} = 1{,}96$ и $z\{\beta = 0{,}10\} = 1{,}28$ находят в таблицах или с помощью пакетов прикладных программ.

и значения квантилей нормального распределения $z\{\alpha\}$, $z\{\alpha/2\}$ и $z\{\beta\}$ для выбранных значений α и β.

Допустим, требуется оценить, какой объем выборки нужен для выявления неравновесности популяции с параметрами, равными полученным выше оценкам $p = 0{,}55$ и $F = +0{,}10$ при $\alpha = 0{,}05$ и $\beta = 0{,}10$. При односторонней альтернативе H_1^*: $F_1 \geqslant 0{,}10$ для данных табл. П.10 оценка N по формуле (I) (табл. П.12 и П.13) равна

$$\widehat{N}^*\{p = 0{,}55; F_1 \geqslant 0{,}10; \alpha = 0{,}05; \beta = 0{,}10\} \geqslant 850.$$

Читатель может самостоятельно убедиться, что при односторонней альтернативе значения W и, соответственно, N всегда меньше для отрицательных значений F_1, чем для положительных, но незначительно, поэтому при вычислениях достаточно ограничиться положительными значениями F_1.

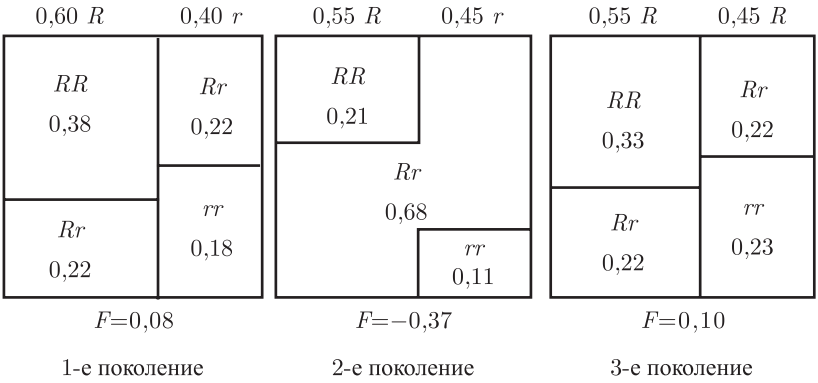


Рис. П.6. Изменение соотношения генотипов в трех последовательных поколениях шортгорнов в стаде Иллинойсского университета (по B. Roberts, 1937). Первое поколение особей подверглось жестокому отбору наподобие аутбридинга, о чем свидетельствует резкое возрастание доли гетерозигот во втором поколении и соответствующее статистически высокозначимое ($P_E = 2{,}1 \times 10^{-20}$) отрицательное значение индекса фиксации $F = -0{,}37 \pm 0{,}04$ (см. табл. П.15). В третьем поколении генетическое равновесие восстановилось, поскольку предыдущее поколение никакому вмешательству не подвергалось. Обращает на себя внимание тот факт, что частоты аллелей остались практически неизменными

При двусторонней альтернативе H_1^{**}: $|F_1| \geqslant 0{,}10$ оценка N равна

$$\widehat{N}^{**}\{p = 0{,}55; |F_1| \geqslant 0{,}10; \alpha = 0{,}05; \beta = 0{,}10\} \geqslant 1047.$$

Таким образом, для различения двух гипотез, H_0: $F = 0$ и H_1^{**}: $F \neq \neq 0$, при $p = 0{,}55$, уровне значимости $\alpha = 0{,}05$ и мощности критерия $1 - \beta = 0{,}90$ необходим объем выборки, примерно в три раза больший имеющегося. Посему в данном случае гипотеза о равновесии хотя и принимается, однако при этом нельзя исключить того, что при большем объеме выборки можно было бы выявить неравновесность с $|F| = 0{,}10$.

Исходные данные и основные результаты проведенного анализа удобно представить в компактном виде (табл. П.14).

Пытливому читателю предоставляется возможность самостоятельно провести подобный анализ для родительской и прародительской популяций шортгорнов в стаде Иллинойсского университета по данным табл. П.9 и П.10. Соответствующие данные и результаты их анализа представлены в табл. П.15 и П.16.

Оказывается, что в отличие от популяции телят популяция их родителей явно неравновесна: $F \pm \delta\{\widehat{F}\} = -0{,}37 \pm 0{,}04$; $\chi^2_E = 85{,}72$ и $P_E = = 2{,}1 \cdot 10^{-20}$ (см. табл. П.14). Столь малое P-значение можно оценить, используя аппроксимацию Пайзера–Пратта (см. табл. П.6).

Если же проанализировать популяцию прародителей, то вновь наблюдается равновесие: $F \pm \delta\{\widehat{F}\} = 0{,}08 \pm 0{,}05$; $\chi^2_E = 2{,}80$ и $P_E \approx 0{,}094$ (табл. П.16).

В графическом виде полученные результаты представлены на рис. П.6.

Одной из причин наблюдаемой нестабильности равновесия (нестационарности популяции) может быть неслучайность скрещиваний. Э. Робертс отметил, что первое поколение в обследованном стаде подверглось жесткому искусственному отбору с целью улучшения его производительных и продуктивных свойств.

П.22. Проверка случайности скрещиваний

При анализе причин отклонений от равновесия Харди–Вайнберга является проверка нулевой гипотезы о случайности (независимости) скрещиваний, в данном случае — о независимости от фенотипа скрещиваемых особей, т. е. об отсутствии ассортативности. Простейшая статистическая модель для скрещивания особей трех различающихся фенотипов A, B и C и схема анализа экспериментальных данных представлены в табл. П.17.

Таблица П.14

Проверка равновесности популяции телят шортгорнов в стаде Иллинойского университета (по E. Roberts, 1937)

Фенотипы	Генотипы	Наблюдаемые численности n_i	Ожидаемые численности $\widehat{n}_i$	Знаки разностей $n_i - \widehat{n}_i$	Вычисленные значения статистик		
					χ_i^2	ν_i	P_i
К	RR	105	96,86	$+$	0,68	1	0,41
Ч	Rr	141	157,29	$-$	1,69	1	0,19
Б	rr	72	63,85	$+$	1,04	1	0,31
$\sum$		$N = 318$	318,00		$\chi_E^2 = 3{,}41$ $\nu_E = 1$ $P_E = 0{,}065$		

Оценки параметров и объема выборки

$$\widehat{p} \pm \delta\{\widehat{p}\} = 0{,}55 \pm 0{,}02; \qquad \widehat{F} \pm \delta\{\widehat{F}\} = 0{,}10 \pm 0{,}06;$$

$$\widehat{N}^{**}\{p = \widehat{p}; F \geqslant |\widehat{F}| = 0{,}10; \alpha = 0{,}05; \beta = 0{,}10\} \geqslant 1\,047$$

Примечание:
P-значения найдены с помощью программы MICROSTAT.

Модель характеризуется тремя параметрами: P и Q — *частоты двух из трех фенотипов*, M — *коэффициент ассортативности скрещиваний*[17].

Коэффициент ассортативности M можно интерпретировать как *коэффициент корреляции* между скрещивающимися особями, как меру неслучайности скрещиваний. Он может принимать значения в пределах от -1 до $+1$:

$$-1 \leqslant M \leqslant 1.$$

Когда особи скрещиваются свободно, независимо от их фенотипов, то $M = 0$; когда скрещиваются особи предпочтительно одинакового фенотипа, то $M > 0$; когда же предпочтительным является скрещивание между разными фенотипами, тогда $M < 0$. Один из способов графического отображения этой модели для случая $M = 0$ показан на рис. П.7.

Критерием для проверки H_0: $M = 0$, т. е. гипотезы о случайности скрещиваний самок и самцов трех фенотипов, может служить статистика χ_A^2 с числом степеней свободы $\nu_A = 3$ (см. табл. П.17). Число степеней свободы в подобных ситуациях можно определить как разность между числом возможных типов спаривания и числом различимых фенотипов (в данном

[17] См.: Workman P. L. The analysis of simple genetic polymorphisms. // Human Biology. — 1969. — Vol. 41. — № 1. — P. 97–114.

Таблица П.15

Проверка равновесности популяции родителей шортгорнов в стаде Иллинойского университета (по E. Roberts, 1937)

Фенотипы	Генотипы	Наблюдаемые численности n_i	Ожидаемые численности $\widehat{n}_i$	Знаки разностей $n_i - \widehat{n}_i$	Вычисленные значения статистик		
					χ_i^2	ν_i	P_i
К	RR	132	189,87	$-$	17,64	1	$3{\cdot}10^{-5}$
Ч	Rr	431	315,26	$+$	42,49	1	10^{-10}
Б	rr	73	130,87	$-$	25,59	1	$4{\cdot}10^{-7}$
$\sum$		$N = 636$	636,00		$\chi_E^2 = 85{,}73$ $\nu_E = 1$ $P_E = 2{,}1 \cdot 10^{-20}$		

Оценки параметров и объема выборки

$$\widehat{p} \pm \delta\{\widehat{p}\} = 0{,}55 \pm 0{,}01; \qquad \widehat{F} \pm \delta\{\widehat{F}\} = -0{,}37 \pm 0{,}04;$$

$$\widehat{N}^{**}\{p = \widehat{p} = 0{,}55; F \geqslant |\widehat{F}| = -0{,}37; \alpha = 0{,}05; \beta = 0{,}10\} \geqslant 70$$

Примечание:
Малые P-значения найдены по аппроксимационной формуле из табл. П.6.

Таблица П.16

Проверка равновесности популяции прародителей шортгорнов в стаде Иллинойского университета (по E. Roberts, 1937)

Фенотипы	Генотипы	Наблюдаемые численности n_i	Ожидаемые численности $\widehat{n}_i$	Знаки разностей $n_i - \widehat{n}_i$	Вычисленные значения статистик		
					χ_i^2	ν_i	P_i
К	RR	170	161,52	$+$	0,45	1	0,50
Ч	Rr	198	214,96	$-$	1,34	1	0,25
Б	rr	80	71,52	$+$	1,01	1	0,31
$\sum$		$N = 448$	448,00		$\chi_E^2 = 2{,}80$ $\nu_E = 1$ $P_E = 0{,}094$		

$$\widehat{p} \pm \delta\{\widehat{p}\} = 0{,}60 \pm 0{,}16; \qquad \widehat{F} \pm \delta\{\widehat{F}\} = 0{,}08 \pm 0{,}05;$$

$$\widehat{N}^{**}\{p = \widehat{p} = 0{,}60; F \geqslant |\widehat{F}| = 0{,}08; \alpha = 0{,}05; \beta = 0{,}10\} \geqslant 1\,689$$

Примечание:
P-значения найдены с помощью программы MICROSTAT.

случае $\nu_A = 6-3 = 3$). Пример вычислений для данных Э. Робертса о скрещивании родительских особей породы шортгорн в стаде Иллинойсского университета приведен в табл. П.18.

	PA	QB	RC
PA	A×A P^2	A×A PQ	A×C PR
QB	A×B PQ	B×B Q^2	B×C QR
RC	A×C PR	B×C QR	C×C Q^2

Рис. П.7. Графическое представление модели скрещивания между особями трех фенотипов A, B и C при отсутствии ассортативности ($M = 0$)

Условием применимости статистики χ^2_a считается требование $\widehat{n}_i \geqslant 5$. Для анализируемых данных оно выполняется не совсем строго: наименьшая ожидаемая численность для скрещивания типа Б×Б (белых коров с белыми быками) $\widehat{n}_6 = 4{,}19$ почти на единицу меньше пяти (см. табл. П.18), но таким несоответствием можно пренебречь.

Исходные данные и основные результаты их анализа удобно представить в компактном виде (табл. П.19). Полученная оценка значимости критерия $P_A = 0{,}24$ не дает оснований отвергнуть гипотезу о случайности скрещиваний между родительскими особями шортгорнов. Наглядно это можно видеть на рис. П.8, который мало отличается от идеальной картины для $M = 0$ (см. рис. П.7).

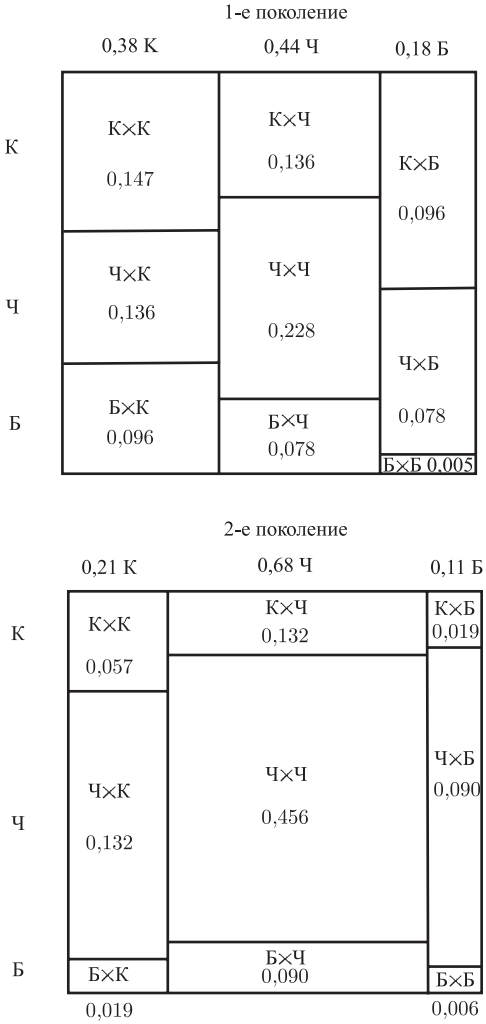


Рис. П.8. Изменение соотношения между различными типами скрещиваний в двух последовательных поколениях шортгорнов в стаде Иллинойсского университета (по E. Roberts, 1937). Скрещивание особей первого поколения явно неслучайно в отличие от скрещивания особей второго поколения. Этим, очевидно, можно объяснить, почему генетическое равновесие во втором поколении нарушилось, а в третьем — восстановилось (см. П.6)

Таблица П.17

Модель скрещиваний между особями трех фенотипов и схема магнетического анализа экспериментальных данных.

Фенотипы скрещивающихся особей

A B C

Ожидаемые частоты фенотипов

P Q R

Типы скрещиваний

$A\times A$ $A\times B$ $A\times C$ $B\times B$ $B\times C$ $C\times C$

Ожидаемые частоты типов скрещиваний

$$P^2(1-M)+MP \quad 2PR(1-M) \quad 2QR(1-M)$$
$$2PQ(1-M) \quad Q^2(1-M)+MQ \quad R^2(1-M)+MR$$

Наблюдаемые численности типов скрещиваний

n_1 n_2 n_3 n_4 n_5 n_6

Оценки частот фенотипов

$$\widehat{P}=(2n_1+n_2+n_3)(2N)^{-1} \quad \widehat{Q}=(n_2+2n_4+n_5)(2N)^{-1}$$
$$\widehat{R}=(n_3+n_5+2n_6)(2N)^{-1}$$

Оценки ожидаемых численностей типов скрещиваний при $M=0$

$$\widehat{n}_1=\widehat{N}P^2 \quad \widehat{n}_2=2N\widehat{P}\widehat{Q} \quad \widehat{n}_3=2N\widehat{P}\widehat{R}$$
$$\widehat{n}_4=N\widehat{Q}^2 \quad \widehat{n}_5=2N\widehat{Q}\widehat{R} \quad \widehat{n}_6=N\widehat{R}^2$$

Проверка правильности вычислений

$$N=\sum_{i=1}^{6}n_i=\sum_{i=1}^{6}\widehat{n}_i \qquad P+Q+R=\widehat{P}+\widehat{Q}+\widehat{R}=1$$

Проверка гипотез

H_0	χ^2	ν
$n_i-\widehat{n}_i=0$	$\chi_i^2=(n_i-\widehat{n}_i)^2(\widehat{n}_i)^{-1}$	$\nu_i=1$
$M=0$	$\chi_A^2=\sum_{i=1}^{6}\chi_i^2$	$\nu_A=3$

Данный вывод согласуется с тем, что полученное от этих родителей поколение телят находится в равновесии Харди–Вайнберга. Несмотря на то что популяция родительских особей была неравновесной, случайность их скрещиваний обусловила равновесность популяции телят — поколение F_3 (см. табл. П.14, П.15 и рис. П.6, П.8).

Таблица П.18

Пример вычислений для проверки случайности скрещиваний между родительскими особями шортгорнов в стаде Иллинойсского университета (по E. Roberts, 1937)

Фенотипы скрещивающихся особей

К	*Ч*	*Б*

Типы скрещиваний

К×*К*	*К*×*Ч*	*К*×*Б*	*Ч*×*Ч*	*Ч*×*Б*	*Б*×*Б*

Наблюдаемые численности типов скрещиваний

$n_1 = 18 \quad n_2 = 84 \quad n_3 = 12 \quad n_4 = 145 \quad n_5 = 57 \quad n_6 = 2 \quad N = 318$

Оценки частот фенотипов

$$\widehat{P} = (2 \cdot 18 + 84 + 12)(2 \cdot 318)^{-1} = 0{,}2075;$$
$$\widehat{Q} = (84 + 2 \cdot 145 + 57)(2 \cdot 318)^{-1} = 0{,}6777;$$
$$\widehat{R} = (12 + 57 + 2 \cdot 2)(2 \cdot 318)^{-1} = 0{,}1148$$

Оценки ожидаемых численностей типов скрещиваний

$$\widehat{n}_1 = 318(0{,}2075)^2 = 13{,}69; \qquad \widehat{n}_4 = 318(0{,}6777)^2 = 146{,}05;$$
$$\widehat{n}_2 = 318 \cdot 2 \cdot 0{,}2075 \cdot 0{,}6777 = 89{,}44; \qquad \widehat{n}_5 = 318 \cdot 2 \cdot 0{,}6777 \cdot 0{,}1148 = 49{,}48;$$
$$\widehat{n}_3 = 318 \cdot 2 \cdot 0{,}2075 \cdot 0{,}1148 = 15{,}15; \qquad \widehat{n}_6 = 318(0{,}1148)^2 = 4{,}19$$

Проверка правильности вычислений

$$\widehat{P} + \widehat{Q} + \widehat{R} = 0{,}6777 + 0{,}1148 = 1{,}0000;$$
$$\sum_{i=1}^{6} n_i = 13{,}69 + 89{,}44 + 15{,}5 + 146{,}05 + 49{,}48 + 4{,}19 = 318{,}00 = N$$

Противоположная картина наблюдается при анализе скрещивания прародительских особей F_1 (табл. П.20). В результирующих таблицах также стоит указывать значения частных компонент χ_1^2 и знаки входящих в них разностей $(n_i - \widehat{n}_i)$, поскольку, как уже говорилось, они несут по-

Окончание табл. П.18

Проверка гипотез χ^2

H_0	χ^2	ν	P
$n_1 - \widehat{n}_1 = 0$	$\chi^2_1 = (18 - 13{,}70)^2(13{,}70)^{-1} = 1{,}36$	$\nu_1 = 1$	$P_1 = 0{,}24$
$n_2 - \widehat{n}_2 = 0$	$\chi^2_2 = (84 - 89{,}45)^2(89{,}45)^{-1} = 0{,}33$	$\nu_2 = 1$	$P_2 = 0{,}57$
$n_3 - \widehat{n}_3 = 0$	$\chi^2_3 = (12 - 15{,}15)^2(15{,}15)^{-1} = 0{,}65$	$\nu_3 = 1$	$P_3 = 0{,}42$
$n_4 - \widehat{n}_4 = 0$	$\chi^2_4 = (145 - 146{,}04)^2(146{,}04)^{-1} = 0{,}01$	$\nu_4 = 1$	$P_4 = 0{,}92$
$n_5 - \widehat{n}_5 = 0$	$\chi^2_5 = (57 - 49{,}47)^2(49{,}47)^{-1} = 1{,}14$	$\nu_5 = 1$	$P_5 = 0{,}29$
$n_6 - \widehat{n}_6 = 0$	$\chi^2_6 = (2 - 4{,}19)^2(4{,}19)^{-1} = 1{,}14$	$\nu_6 = 1$	$P_6 = 0{,}29$
$M = 0$	$\chi^2_A = 1{,}36 + 0{,}33 + \ldots + 1{,}14 = 4{,}63$	$nu_A = 3$	$P_A = 0{,}24$

Примечание:
P-Значения найдены с помощью программы MICROSTAT.

лезную, содержательную информацию, позволяя выявить, какие конкретно типы скрещиваний вносят основной вклад в отклонение от случайности. В данном случае наблюдается явный избыток скрещиваний типа К×Б и дефицит скрещиваний типа Б×Б: обе соответствующие оценки значимости $P_3 \approx P_6 \approx 0{,}02$ ниже критического уровня $\alpha = 0{,}05$ (в табл. П.20 они отмечены звездочкой $*$).

По-видимому, выявленная неслучайность скрещиваний между особями первого поколения и обусловила неравновесность второго поколения (см. табл. П.15 и рис. П.6).

В табл. П.17 не даны формулы для оценки параметра M и его стандартной ошибки. Решение соответствующего уравнения правдоподобия здесь не имеет явного вида и требует применения специальных вычислительных приемов (в случае надобности необходимо обращаться за помощью к профессионалу-математику).

Предложенная модель предполагает, что либо преимущество имеют скрещивания всех подобных фенотипов с подобными же (*положительная ассортативность*), либо преобладают скрещивания между всеми разными фенотипами (*отрицательная ассортативность*). Такая модель удовлетворительно описывает, например, некоторые популяции растений, в которых

Таблица П.19

Проверка случайности скрещиваний между родительскими особями шортгонов в стаде Иллинойсского университета (по Е. Робертс, 1937).

Типы скрещиваний	Наблюдаемые численности n_i	Ожидаемые численности $\widehat{n}_i$	Знаки разностей $n_i - \widehat{n}_i$	Вычисленные значения статистик		
				χ_i^2	ν_i	P_i
К×К	18	13,69	+	1,36	1	0,24
К×Ч	84	89,44	−	0,33	1	0,57
К×Б	12	15,15	−	0,65	1	0,42
Ч×Ч	145	146,05	−	0,01	1	0,92
Ч×Б	57	49,18	+	1,14	1	0,29
Б×Б	2	4,19	−	1,14	1	0,29
$\sum$	$N = 318$	318,00		$\chi_A^2 = 4{,}63$ $\nu_A = 3$ $P_A = 0{,}24$		

Оценки частот фенотипов

$$\widehat{P} = 0{,}2075 \qquad \widehat{Q} = 0{,}6777 \qquad \widehat{R} = 0{,}1148$$

Таблица П.20

Проверка случайности скрещиваний между прародительскими особями шортгорнов в стаде Иллинойсского университета (по Е. Робертс, 1937).

Типы скрещиваний	Наблюдаемые численности n_i	Ожидаемые численности $\widehat{n}_i$	Знаки разностей $n_i - \widehat{n}_i$	Вычисленные значения статистик		
				χ_i^2	ν_i	P_i
К×К	33	31,25	+	0,01	1	0,92
К×Ч	61	75,13	−	2,66	1	0,10
К×Б	43	30,36	+	2,26	1	0,022*
Ч×Ч	51	43,75	+	1,20	1	0,27
Ч×Б	35	35,36	−	0,004	1	0,95
Б×Б	1	7,14	−	5,20	1	0,023*
$\sum$	$N = 224$	223,99		$\chi_A^2 = 14{,}42$ $\nu_A = 3$ $P_A = 0{,}002$		

$$\widehat{P} = 0,3794 \qquad \widehat{Q} = 0,4420 \qquad \widehat{R} = 0,1786$$

Примечание:
Звездочкой $*$ отмечены отклонения, значимые на уровне $\alpha = 0{,}05$.
P-значения оценены с помощью программы MICROSTAT.

определенная доля, равная M, размножается самоопылением, а остальные $(1-M)$ — путем свободного скрещивания.

В данном случае при скрещивании особей первого поколения (прародителей) только для одного из трех типов разнородных скрещиваний К×Б наблюдается статистически значимая положительная ассортативность и только для одного из трех типов однородных скрещиваний Б×Б наблюдается статистически значимая отрицательная ассортативность. По-видимому, модель нуждается в усложнении: вместо одного интегрального коэффициента ассортативности скрещивания M в нее следует ввести два (или более) дифференциальных коэффициента.

Что касается скрещиваний с белыми коровами, то в 1952 г. Дж. М. Рендел (J. M. Rendel) обнаружил: у некоторых белых коров снижена фертильность вследствие дисфункции половых органов (матки). Это явление в ветеринарии получило название «болезнь белых телок».

Итак, биометрический анализ убедительно показывает, что на протяжении всего трех последовательных поколений в стаде шортгорнов Иллинойсского университета происходили (вызванные вмешательством человека) серьезные генетико-популяционные изменения. Исходная генетически равновесная популяция прародителей подверглась жесткому отбору и явно неслучайному скрещиванию — аутбридингу: статистически значимо преобладали скрещивания гомозиготных красных особей RR с гомозиготными белыми rr и избегались скрещивания между белыми особями. Соответственно, популяция особей следующего поколения оказалась генетически явно неравновесной с высоким отрицательным значением индекса фиксации $F=-0{,}37$. Особям этого поколения была предоставлена свобода скрещиваний, в результате популяция особей третьего поколения вновь стала равновесной в полном согласии с законом Харди–Вайнберга. Примечательно, что при этом частоты аллелей R и r сохранились практически неизменными.

Вряд ли плодотворность применения биометрического анализа в популяционной генетике можно было бы продемонстрировать столь убедительно на примере природных популяций. Воистину прав был С. Райт, когда писал, что выведение пород домашних животных больше напоминает природный эволюционный процесс, нежели селекционные эксперименты в лаборатории, и что надо следовать примеру Ч. Дарвина, который, прежде чем осознать созидающую роль естественного отбора, много внимания уделил вопросам искусственного отбора у домашних животных.

Новые генетические механизмы и их роль в генетико-популяционных процессах

Достижения клеточной и молекулярной биологии последних лет обогатили генетику открытием новых явлений и механизмов, которые существенно расширили и углубили наши представления о генетико-популяционных и эволюционных процессах.

Их открытие связано с бурно развивающимися исследованиями первичной структуры ДНК и, в особенности, с изобретением полимеразной цепной реакции. В результате открылась возможность анализа ДНК из таких источников, как мумии, насекомые, законсервированные в янтаре на миллионы лет, останки (кости и зубы) древнейших организмов. Появился новый раздел —- молекулярная палеогенетика популяций. Другая, прикладная отрасль — молекулярная судебная медицина, основанная на ДНК-типировании; ее разрешающая способность превышает разрешающую способность дактилоскопии, а возможности много шире: достаточно иметь следовые количества крови, спермы, несколько волос или одну луковицу волоса, чтобы идентифицировать их принадлежность.

Два относительно недавно открытых явления — импринтинг и прогрессивная амплификация — играют важную роль в генетико-популяционных процессах.

Импринтинг. Разработанные и широко используемые в биологии развития методы переноса ядер, а также генетический анализ мейотического нерасхождения хромосом у мыши показали, что некоторые гены экспрессируются по-разному в зависимости от того, передаются они от матери или от отца. Это явление было названо *генетическим* (или молекулярным) *импринтингом*.

Причиной импринтинга являются эпигенетические, зависимые от происхождения гамет, модификации содержащегося в них генома. Явление заключается в том, что «импринтабельная» (т. е. способная к импринтингу) аллель, пройдя через гаметогенез у одного пола, способна экспрессироваться у потомка, но, пройдя через гаметогенез у противоположного пола, инактивируется и не экспрессируется у потомка. Такую инактивированную аллель называют *импринтной*. Очевидно, что должны существовать гены, которые контролируют этот процесс, их называют *генами-импринторами*.

Эволюционно процесс импринтинга достаточно консервативен, поскольку он оказался широко распространенным как у растений, так и у животных.

Импринтинг проявляется на разных уровнях организации генетического материала. Различают импринтинг целого генома, импринтинг индивидуальных хромосом, локусов и генов. Основным молекулярным механизмом признано метилирование цитозиновых остатков в ДНК. Однако, по-видимому, возможны и другие механизмы не только на уровне первичной структуры ДНК, но и на более высоких уровнях хромосомной организации, когда активный хроматин переходит в неактивное «гетерохроматизированное» состояние и сохраняется в ряду последовательных клеточных делений. Ответственными за такие переходы могут быть *гены молчания* (сайленсеры — от англ. silencer), делающие невозможными транскрипцию и дальнейшую экспрессию инактивируемых генов.

На популяционном уровне импринтинг в одном локусе вызывает кажущийся недостаток гетерозигот. Соответствующее генетико-популяционное уравнение напоминает классическую формулу закона равновесия Райта:

$$P(AA) = p^2 + \Theta pq; P(Aa) = 2pq(1 - \Theta); P(aa) = q^2 + \Theta pq,$$

но здесь Θ — вероятность того, что аллель, полученная от матери, не экспрессируется у потомка, и в отличие от райтовского индекса фиксации F, который имеет смысл коэффициента корреляции и находится в пределах $-1 \leqslant F \leqslant 1$, параметр в находится в пределах $0 \leqslant \Theta \leqslant 1$. Это означает, что дифференциальный импринтинг (т. е. ситуация, когда $\Theta > 0$) всегда вызывает в популяции дефицит гетерозигот.

Популяционные последствия импринтинга кажутся аналогичными эффектам нуль-аллели, хромосомной делеции или неполной пенетрантности и на первый взгляд их невозможно отличить друг от друга. Однако, если изучать два и более последовательных поколения, то эффект импринтинга можно вычленить, поскольку он является временным и специфичным в отношении родительского возникновения аллелей. Другим последствием импринтинга может быть различие в частотах рекомбинации у самок и самцов: он будет приводить к завышению частоты рекомбинации у особей с импринтингом по сравнению с особями без него. Кроме того, импринтинг имитирует явление гибридной силы (гетерозиса, сверхдоминирования), т. е. преимущества гетерозигот по способному к нему локусу.

Прогрессивная амплификация. Еще до недавнего времени интроны считали инертной, индифферентной, «эгоистичной» частью ДНК. Теперь становится ясным, что они и другие нетранскрибируемые последователь-

ности могут влиять на активность и мутабильность транскрибируемых последовательностей. Наиболее ярко это проявляется в недавно открытом явлении прогрессивной (точнее, прогрессирующей) амплификации, которое может сопровождаться тяжелыми наследственными нарушениями. В настоящее время известно не менее шести наследственных заболеваний человека, причиной которых является массовая амплификация коротких GC-богатых тринуклеотидных последовательностей. Среди них синдром ломкой Х-хромосомы, болезнь Хантингтона, миотоническая дистрофия и др. Если в норме число таких повторов исчисляется единицами и десятками, то у больных их сотни и тысячи. Например, миотоническая дистрофия (МД)[18] обусловлена прогрессирующей амплификацией тринуклеотидной последовательности CTG в нетранскрибируемой области гена миотониновой протеинкиназы, превышающей 2 000 повторов (в норме они исчисляются от 5 до 24).

Совсем недавно было показано, что поли-(CTG)-последовательность сверхпрочно связывается с гистоновыми компонентами нуклеосом, чем блокируется активность прилежащих районов.

Это принципиально новый тип мутаций. Когда экспансия триплетных повторов превышает некоторый порог (52 повтора), последовательность становится нестабильной, особенно в процессе мейоза у матерей. Таким образом, получено молекулярное обоснование феномену «генетического предвидения» (anticipation), которое заключается в усилении тяжести болезни в последующих поколениях. Загадочным представляется двухвершинное распределение числа повторов в популяции здоровых людей. Среди обследованных людей (около 1 000 человек) различных национальностей повторы числом менее 5 не обнаружены; наиболее часты пятикратные и 10–13-кратные повторы. Другая загадка — тот факт, что в популяции негроидов Южной Африки (Нигерия и прилежащие государства), насчитывающей около 30 млн человек, синдром МД ни разу не был обнаружен. В то же время средняя мировая частота встречаемости этого заболевания составляет 1: 8 000. Эти загадки ждут своего решения.

Если до недавнего времени мы рассматривали естественный отбор как силу, действующую на уровне разнообразия (полиморфизма) продуктов генов, то теперь возникает принципиально новая мишень для отбора — полиморфизм генетического материала, под которым следует понимать разнообразие первичной, вторичной, третичной и прочих надструктур ДНК.

[18] МД — аутосомная доминантная миопатия с плейотропным эффектом, включающая продолжительные мышечные судороги, катаракту, аритмию сердца.

Рекомендуемая литература

(звездочкой ∗ отмечены книги, в которых обсуждаются биологические задачи)

Монографии, учебники, популярные книги по теории вероятностей

1. Борель Э. Вероятность и достоверность. — М.: Наука, 1969. — 110 с.
2. Гнеденко Б. В. Курс теории вероятностей. — 5-е изд. — М.: Наука, 1969. — 400 с.
3. Гнеденко Б. В., Хинчин А. Я. Элементарное введение в теорию вероятностей. — М.: Наука, 1967. — 167 с.

4*. Кендалл М., Моран П. Геометрические вероятности. — М.: Наука, 1972. — 192 с.

5. Колмогоров А. Н. Основные понятия теории вероятностей. — 2-е изд. — М.: Наука, 1974. — 119 с.

6*. Нейман Ю. Вводный курс теории вероятностей и математической статистики. — М.: Наука, 1968. — 448 с.

7*. Савельев Л. Я. Комбинаторика и вероятность. — Новосибирск: Наука, 1975. — 423 с.

8. Тутубалин В. Н. Теория вероятностей: Краткий курс и научно-методические замечания. — М.: Изд-во Моск. ун-та, 1972. — 230 с.

9*. Феллер В. Введение в теорию вероятностей и ее приложения. — 2-е изд. — М.: Мир, 1967. — Т.I. — 498 с.

Руководства и учебники по математической статистике и ее приложениям в науке и технике

10. Брандт З. Статистические методы анализа наблюдений. — М.: Мир, 1975. — 312 с.

11. Браун ли К.А̇. Статистическая теория и методология в науке и технике. — М.: Наука, 1977. — 407 с.

12*. Ван дер Варден Б. Л. Математическая статистика. — М.: ИЛ, 1960. — 434 с.

13. Гмурман В. Е. Теория вероятностей и математическая статистика. — 5-е изд. — М.: Высшая школа, 1977. — 479 с.

14. Головач А. В., Ерина А. М., Трофимов В. П. Критерии математической статистики в экономических исследованиях. — М.: Статистика, 1973. — 135 с.

15. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке: Методы обработки данных. — М.: Мир, 1980. — 610 с.

16. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке: Методы планирования эксперимента. — М.: Мир, 1981. — 516 с.

17. Дунин-Барковский И. В., Смирнов Н. В. Теория вероятностей и математическая статистика в технике: Общая часть. — М.: ГИТТЛ, 1955. — 556 с.

18. Кендалл М., Стьюарт А. Теория распределений. — М.: Наука, 1966. — 587 с.

19. Кендалл М., Стьюарт А. Статистические выводы и связи. — М.: Наука, 1973. — 900 с.

20. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. — М.: Наука, 1976. — 736 с.

21. Крамер Г. Математические методы статистики. — 2-е изд. — М.: Мир, 1975. — 648 с.

22*. Мардна К. Статистический анализ угловых наблюдений. — М.: Наука, 1978. — 239 с.

23*. Рао С. Р. Линейные статистические методы и их применения. — М.: Наука, 1968. — 548 с.

24. Смирнов Н. В., Дунин–Барковский И. В. Курс теории вероятностей и математической статистики для технических приложений. — 3-е изд. — М.: Наука, 1969. — 511 с.

25*. Тьюки Дж. Анализ результатов наблюдений: Разведочный анализ. — М.: Мир, 1981. — 693 с.

26*. Хальд А. Математическая статистика с техническими приложениями. — М.: ИЛ, 1956. — 664 с.

27. Хан Г., Шапиро С. Статистические модели в инженерных задачах. — М.: Мир, 1969.

Сборники задач по теории вероятностей и математической статистике

28. Большев Л. Н., Иванова Г. П. Сборник задач по математической статистике. — М.: Изд-во Моск. ун-та, 1980. — 71 с.

29*. Джермен М. Количественная биология в задачах и примерах. — М.: Мир, 1972. — 151 с.

30. Емельянов Г. В., Скитович В. П. Задачник по теории вероятностей и математической статистике. — Л.: Изд-во Ленингр. ун-та, 1967. — 331 с.

31. * Мешалкин Л. Д. Сборник задач по теории вероятностей. — М.: Изд-во Моск. ун-та, 1963. — 150 с.

32. Мостеллер Ф. Пятьдесят занимательных вероятностных задач. — М.: Наука, 1968. — 103 с.

Монография и руководства по непараметрическим методам математической статистики

33*. Ашмарин И. П., Васильев Н. Н., Амбросов В. А. Быстрые методы статистической обработки и планирование экспериментов. — 2-е изд. — Л.: Изд-во Ленингр. ун-та, 1975. — 78 с.

34. Кендэл М. Ранговые корреляции. — М.: Статистика, 1975. — 214 с.

35. Тюрин Ю. Н. Непараметрические методы статистики. — М.: Знание, 1978. — 64 с.

36. Bülling H., Trenkler G. Nichtparametrische statistische Methoden. — Berlin: New York, 1978. — 435 S.

37. Hollander M., Wolfe D. A. Nonparametric statistical methods. — N.Y.: Wiley, 1973. — 503 p.

38. Lehmann E. L. Nonparametrics: statistical methods based on ranks. — San Francisco: Holden-Day, 1975. — 457 p.

Руководства и учебники по биометрии и медико-биологическим применениям статистического анализа

39*. Ашмарин И. П., Воробьев А. А. Статистические методы в микробиологических исследованиях. — Л.: Медгиз, 1962. — 180 с.

40*. Бейли И. Статистические методы в биологии. — М.: ИЛ, 1962. — 260 с.

41. Гласе Дж., Стэнли Дж. Статистические методы в педагогике и психологии. — М.: Прогресс, 1976. — 495 с.

42*. Доспехов Б. А. Методика полевого опыта (с основами статистической обработки результатов исследований). — М.: Колос, 1979. — 416 с.

43*. Литтл Т. М., Хиллз Ф. Дж. Сельскохозяйственное опытное дело: Планирование и анализ. — М.: Колос, 1981. — 319 с.

44*. Максимов В. Н. Многофакторный эксперимент в биологии. — М.: Изд-во Моск. ун-та, 1980. — 279 с.

45*. Рокицкий П. Ф. Биологическая статистика. — 3-е изд. — Минск: Вышэйшая школа, 1973. — 320 с.

46*. Снедекор Дж. У. Статистические методы в приложении к исследованиям в сельском хозяйстве и биологии. — М.: Сельхозиздат, 1961. — 503 с.

47*. Терентьев П. В., Ростова Н. С. Практикум по биометрии. — Л.: Изд-во Ленингр. ун-та, 1977. — 152 с.

48*. Урбах В. Ю. Биометрические методы: статистическая обработка опытных данных в биологии, сельском хозяйстве и медицине. — М.: Наука, 1964. — 415 с.

49*. Урбах В. Ю. Статистический анализ в биологических и медицинских исследованиях. — М.: Медицина, 1975. — 295 с.

50*. Филипченко Ю. А. Изменчивость и методы ее изучения. — 5-е изд. — М.: Наука, 1978. — 238 с.

51*. Фишер Р. А. Статистические методы для исследователей. — М.: Госстатиздат, 1958. — 268 с.

52*. Bliss C. L. Statistics in biology: Statistical methods for research in the natural sciences. — N.Y., 1967. — Vol. 1. — 558 p.; — vol. 2. 1970. 639 p.

53*. Rasch D., Enderlein G., Herrendorfer G. Biometrie: Verfahren, Tabellen, angewandte Statistik. — Berlin, 1973. — 390 S.

54*. Sokal R. R., Rohlf F. J. Biometry: The principles and practice of statistics in biological research. — San Francisco, 1969. — 776 p.

55*. Weber E. Grundriss der biologischen Statistik: Anwendungen der mathematischen Statistik in Forschung. — 8 Aufl. — Jena, 1980. — 652 S.

Справочники, словари, энциклопедии, сборники таблиц по математической статистике

56. Бернстейн А. Справочник статистических решений. — М.: Статистика, 1968. — 162 с.

57. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. — 3-е изд. — М.: Наука, 1982. — 474 с.

58. Бронштейн И. Н., Семендяев К. А. Справочник по математике для инженеров и учащихся втузов. — Лейпциг: Тойбнер, 1979. — М.: Наука, 1980. — 975 с.

59. Закс Л. Статистическое оценивание. — М.: Статистика, 1976. — 598 с.

60. Корн Г., Корн Т. Справочник по математике для научных работников и инженеров: Определения, теоремы, формулы. — 4-е изд. — М.: Наука, 1978. — 831 с.

61. Математическая энциклопедия / Гл. ред. И. М. Виноградов. — М.: Советская энциклопедия. 1977. Т. 1, А–Г. — 1151 стлб.; Т. 2, Д–Коо, 1979, 1103 стлб., — издание продолжается.

62. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. — М.: Финансы и статистика, 1982. — 276 с.

63. Оуэн Д. Б. Сборник статистических таблиц. — 2-е изд. — М.: Наука, 1972. — 586 с.

64. Справочник по специальным функциям с формулами, графиками и таблицами / Под ред. М. Абрамовича и И. Стиган. — М.: Наука, 1979. — 830 с.

65*. Францевич Л. И. Обработка результатов биологических экспериментов на микро-ЭВМ «Электроника Б3-21»: Программы и программирование. — Киев: Наукова думка, 1979. — 89 с.

66. Янко Я. Математико-статистические таблицы. — М.: Госстатиздат, 1961. — 243 с.

67. Biometrika tables for statisticians / Ed. by E. S. Pearson, H. O. Hartley. — 4-th ed. — Cambridge, 1976. — Vol. 1, 2.

68. Biometrisches Wörterbuch. Bd.1, 2. — Berlin, 1968. — 1047 S.

69. CRC Handbook of tables for probability and statistics / Ed. by W. H. Beyer. — Cleveland, Ohio, 1968. — 2nd ed. — 642 p.

70. Greenwood J. A., Hartley H. O. Guide to tables in mathematical statistics. — Princeton: New Jersey, 1962. — 1014 p.

71. Hald A. Statistical tables and formulas. — N.Y.; L., 1952. — 97 p.

72. International encyclopedia of statistics / Ed. by W. H. Kruskal, J. M. Tanur. — N.Y.; L., 1978. Vol. 1, A–N. 666 p.; Vol. 2, O–Z. p 667–1350.

73. Koller S. Neue graphische Tafeln zur Beurteilung statistischen Zahlen. — 4 Aufl. — Darmstadt, 1969. — 167 S.

74. Rasch D., Herrendorfer G., Bock J. e. a. Verfahrensbibliothek: Versuchsplanung und auswertung. — Berlin, 1978. —Bd.1, 2.

75. Statistical tables and formulas with computer applications / Ed. by Z. Yamauti. — Tokyo, 1972.

76. Walsh J. E. Handbook on nonparametric statistics: in 3 vol. — Princeton: New Jersey, 1962–1968. — 1982 p.

Предметный указатель

Оглавление

Н. В. Глотов, Л. А. Животовский,
Н. В. Хованов, Н. Н. Хромов-Борисов

Биометрия

Дизайнер
Технический редактор
Компьютерный набор и верстка
Корректор

Подписано в печать Формат $60 \times 84^{1}/_{16}$.
Печать офсетная. Усл. печ. л. Уч. изд. л.
Гарнитура . Бумага офсетная №1.
Тираж экз. Заказ №