Г. Ф. ЛАКИН

БИОМЕТРИЯ

Издание четвертое,
переработанное и дополненное

Допущено
Государственным комитетом СССР
по народному образованию
в качестве учебного пособия
для студентов биологических специальностей
высших учебных заведений

МОСКВА «ВЫСШАЯ ШКОЛА» 1990

ОГЛАВЛЕНИЕ

ББК 28
Л 19
УДК 57.087.1

Рецензенты:

Научно-исследовательский институт и музей антропологии им. Д. Н. Анучина (директор института, д-р биол. наук В. П. Чтецов); д-р биол. наук, проф. Л. А. Животовский (Институт общей генетики им. Н. И. Вавилова АН СССР)

Лакин Г. Ф.

Л 19 Биометрия: Учеб. пособие для биол. спец. вузов—4-е изд., перераб. и доп.—М.: Высш. шк., 1990.—352 с.: ил.

ISBN 5-06-000471-6

В книге рассмотрены основные понятия биометрии, числовые характеристики описания совокупности эмпирических данных, законы распределения, построение статистических оценок, параметрические и непараметрические методы проверки статистических гипотез, дисперсионный, корреляционный и регрессионный анализ и некоторые вопросы планирования экспериментов. 4-е издание (3-е — 1980 г.) содержит значительные поправки и дополнения.

Л $\frac{1901000000(4309000000)-177}{001(01)-90}$ 131—90

ББК 28
57

Учебное издание

Лакин Георгий Филиппович

БИОМЕТРИЯ

Научный редактор В. Е. Дерябин. Редактор А. С. Орлова. Младшие редакторы Е. В. Бурова, Е. И. Попова. Художественный редактор Т. А. Колесникова. Технический редактор Е. И. Герасимова. Корректор Г. И. Кострикова

ИБ № 7779

Изд. № Е—548. Сдано в набор 11.09.89. Подп. в печать 06.02.90. Формат 60×88¹/₁₆. Бум. офс. № 2. Гарнитура литературная. Печать офсетная. Объем 21,56 усл. печ. л. 21,56 усл. кр.-отт. 21,89 уч.-изд. л. Тираж 20 000 экз. Зак. № 1674. Цена 1 руб.

Издательство «Высшая школа», 101430, Москва, ГСП-4, Неглинная ул., д. 29/14.

Московская типография № 8 Государственного комитета СССР по печати, 101898, Москва, Центр, Хохловский пер., 7.

ISBN 5-06-000471-6

СПИСОК УСЛОВНЫХ ОБОЗНАЧЕНИЙ

$A, B, C, \ldots$ — постоянные величины; факторы и их градации в дисперсионном комплексе; случайные события

A — условная средняя, нулевой класс, где $A=0$

$a, b, c, \ldots$ — числеиность групп (градаций) факторов $A, B, C, \ldots$ в дисперсионных комплексах; параметры, определяющие соотношение между аргументом и функцией; варианты в клетках четырехпольной таблицы

a — отклонения вариант от условной средней, нулевого класса

As — коэффициент асимметрии ряда распределения

b_n — условный момент n-го поридка

b_{yx} и b_{xy} — коэффициенты регрессии Y/X и X/Y

Cs — показатель точности оценки средней величины

Cv — коэффициенты вариации

D — сумма квадратов отклонений вариант от их средних (девиата)

d — разность между сравниваемыми величинами

$\bar{d}$ — средняя разность между попарно сравниваемыми вариантами

Ex — показатель (коэффициент) эксцесса распределения

F — критерий соответствия (согласия) Р. Фишера

f — эмпирические частоты вариант в данной совокупности

f' — вычисленные или ожидаемые частоты вариант

f_{xy} — частоты вариант в клетках корреляционной таблицы

f_- — первая функция нормального распределения

H — величина, используемая в дисперсионном анализе; критерий Краскелла — Уоллиса

H_A — символ альтернативной гипотезы

H_0 — символ нулевой гипотезы

h^2 — показатель силы влияния факторов на результативный признак

i — порядковый номер варианты

J — критерий соответствия Ястремского

K — биномиальный коэффициент; коэффициент взаимной сопряженности между качественными признаками

k — число степеней свободы

$\lim$ — символ, обозначающий границы вариации признака

Me — медиана

Mo — мода

m — начальные моменты распределения; классы биномиального ряда; число случаев, благоприятствующих наступлению ожидаемого события; абсолютная численность альтернатив, обладающих данным (плюсовым) признаком; доза вещества, вызывающая эффект у 100% подопытных особей

N — объем генеральной совокупности; объем дисперсионного комплекса; число классов (групп) вариационного ряда; общая сумма сопоставляемых рядов

n — объем выборки; число членов ряда регрессии; численность вариант в отдельных градациях дисперсионного комплекса

n_i — варианты в градациях анализируемого комплекса
$\bar{n}$ — усредненная величина входящих в выборку числовых показателей
Q — коэффициент ассоциации Юла
P — вероятиость события; доверительная вероятность
p — доля вариант, обладающих данным признаком (+)
q — доля вариант, не обладающих данным признаком (—)
R — раиг, или порядковый номер варианты в ранжированном ряду; размах вариации; коэффициент корреляции знаков
r — коэффициент корреляцин
s_x — среднее квадратическое отклонение
s_x^2 — выборочная дисперсия (варианса)
$s_{\bar{x}}$ — ошибка выборочной средией
s_{yx} и s_{xy} — ошибки регрессии Y по X и X по Y
T — критерий Уилкоксона для независимых выборок
Tr — показатель трансгрессин рядов распределения
t — нормированное отклоненне; критерий Стьюдеита
U — критерий Уилкоксона (Манна — Уитни) для сопряженных выборок
X — критерий Ван-дер-Вардена
$\bar{x}$ — средняя арифметическая (общая)
$\bar{x}_A$ и $\bar{x}_B$ — средние арифметические из суммы членов градации A и B дисперсионного комплекса
$\bar{x}_g$ — средняя геометрическая
$\bar{x}_h$ — средняя гармоническая
$\bar{x}_i$ — групповая средняя в однофакторных дисперсионных комплексах
$\bar{x}_q$ — средияя квадратическая
$\bar{x}_Q$ — средняя кубическая
$\bar{x}_y$ и $\bar{y}_x$ — частные или групповые средние X по Y и Y по X
$\bar{\bar{x}}$ — средняя из суммы групповых средних
$X, Y, Z, \ldots$ — переменные величины; признаки
$x, y, z, \ldots$ — числовые значения признаков, варианты или даты (по терминологии Р. Фишера)
z — преобразованный коэффициент корреляции (по Фишеру); критерий зиаков
α (альфа) — уровень значимости оценок
β (бета) — критерий достоверности Блекмана
γ (гамма) — показатель линейиости связи
Δ (дельта) — предельная величина ошибки выборочных показателей; величина, определяющая границы доверительного интервала
η (эта) — центральные моменты; корреляционное отношение
λ (ламбда) — величина классового интервала
μ (ми, мю) — генеральная средняя; математнческое ожндание
ν (ню) — число огранниенй свободы вариацни
ρ (ро) — генеральный коэффициент корреляции
Σ (сигма прописная) — знак суммирования
σ (сигма строчная) — среднее квадратическое отклонение (генеральное)
σ^2 — генеральная дисперсия (варианса)
φ (фи) — преобразованная доля вариант (по Фишеру)
φ^2 — показатель взаимной сопряженности между группамн варьирующих признаков
χ^2 (хи-квадрат) — критерий соответствия Пирсона
ψ (пси) — величина, применяемая в формуле X-критерия Ван-дер-Вардена

ПРЕДИСЛОВИЕ К ЧЕТВЕРТОМУ ИЗДАНИЮ[1]

Настоящее (четвертое) издание «Биометрии» переработано и дополнено новыми данными. Пересмотру подвергнута каждая глава; особенно заметные изменения внесены в гл. I. Значительные поправки и дополнения сделаны в главах, посвященных дисперсионному и корреляционно-регрессионному анализу. Изменен и самый порядок расположения глав. Все это положительно сказалось на системе изложения материала, заимствованного из разных разделов биологии, медицины, антропологии, педагогики, основ сельскохозяйственных наук и других смежных отраслей знания.

При подготовке книги были учтены многочисленные отзывы, критические замечания и пожелания читателей, которые автор получил после выхода в свет третьего издания (1980).

Автор выражает всем этим лицам свою признательность. Автор также благодарен А. А. Кузьмину (г. Архангельск) за большую помощь при подготовке настоящего издания, профессору Н. Б. Вассоевичу (МГУ) за оказанную им помощь в работе, а также О. А. Лакиной (супруге автора), принимавшей участие в работе по совершенствованию рукописи.

Автор

ПРЕДИСЛОВИЕ К ПЕРВОМУ ИЗДАНИЮ

В творческой деятельности биолога, в какой бы области он ни работал, большое место занимают вопросы статистической обработки опытных данных, сравнительная оценка результатов наблюдений. Поэтому вполне понятен тот большой интерес, который проявляют биологи самых различных специальностей к справочным пособиям и руководствам по биометрии.

В нашей стране и за рубежом издан целый ряд таких пособий, и, несмотря на то что многие из них доступны лишь сравнительно узкому кругу специалистов, все издания быстро разошлись. Спрос на учебные пособия и справочную литературу по биометрии в нашей стране необычайно велик. Особенно сильно ощущается недостаток в общедоступных руководствах по этому пред-

[1] В связи с кончиной автора во время подготовки учебного пособия к переизданию доработка рукописи и научное редактирование осуществлены канд. биол. наук В. Е. Дерябиным, им же написано послесловие к книге.

мету; в них очень нуждаются широкие круги биологов и специалистов смежных с биологией дисциплин.

Настоящее руководство рассчитано главным образом на студентов и аспирантов, изучающих этот предмет, а также на преподавателей и научных работников, которые не имеют специальной математической подготовки и нуждаются в таком руководстве. Полезные сведения найдут в этой книге и учителя средних школ, занимающиеся проведением опытов на пришкольных участках, в лабораториях и в производственных условиях сельскохозяйственных предприятий.

От приступающих к изучению биометрии требуются знания математики хотя бы в объеме программы средней общеобразовательной школы.

Вычислительная работа значительно облегчается при пользовании математическими таблицами, особенно таблицами для извлечения квадратных корней и возведения в квадрат целых и дробных чисел; весьма удобны «Пятизначные таблицы» Е. Пржевальского, «Математические таблицы» П. П. Андреева (1958) и «Таблицы умножения» О. Рурк (1949, 1965). Можно воспользоваться и другими пособиями, в частности широко известными «Таблицами Барлоу» (1965). Специальные статистические таблицы, необходимые для практической работы, помещены в Приложениях.

При написании учебного пособия автор использовал различные литературные источники, а также собственный экспериментальный материал, собранный преимущественно в Сухумском питомнике обезьян.

Наряду с описанием методики вычисления суммарных биометрических показателей в книге рассматриваются и вопросы теоретического характера, даются обоснования некоторых наиболее важных формул. Примеры и графики иллюстрируют отдельные положения курса. Такое построение учебного пособия облегчает усвоение предмета и позволяет использовать биометрические методы не слепо, а осмысленно, с пониманием конструктивных особенностей различных математических формул и уравнений.

При подготовке рукописи автор учел полезные советы проф. М. Ф. Нестурха; большую помощь в работе оказали рецензенты проф. В. В. Алпатов и канд. биол. наук Ю. С. Куршакова и В. П. Чтецов. Автор благодарен также и коллективу кафедры антропологии МГУ (зав. кафедрой — проф. Я. Я. Рогинский), который рекомендовал рукопись к печати.

Специфика биометрии, ее место в системе биологических наук. С формальной точки зрения биометрия представляет совокупность математических методов, применяемых в биологии и заимствованных главным образом из области математической статистики и теории вероятностей. Наиболее тесно биометрия связана с математической статистикой, выводами которой она преимущественно пользуется, но и биометрия влияет на развитие математической статистики. Взаимодействуя между собой, они взаимно обогащают друг друга. Однако отождествлять биометрию с математической статистикой и теорией вероятностей нельзя.

Биометрия имеет свою специфику, свои отличительные черты и занимает определенное место в системе биологических наук. Современная биометрия — это раздел биологии, содержаиием которого является планирование наблюдений и статистическая обработка их результатов; математическая статистика и теория вероятностей — разделы математики, теоретические, фундаментальные науки, рассматривающие массовые явления безотносительно к специфике составляющих их элементов.

Теория вероятностей исследует законы, действующие в сфере массовых событий и случайных величин, а математическая статистика занимается разработкой выборочного метода, вопросами вероятностной оценки статистических гипотез. Биометрия — прикладная наука, исследующая конкретные биологические объекты с применением математических методов[1]. Биометрия возникла из потребностей биологии. В пограничных областях между биологией и математикой сложились и другие направления математической биологии. Каждое направление имеет свои задачи и применительно к ним использует соответствующие математические методы. Общим для всех направлеиий математической биологии является *дедуктивный* подход к решению конкретных задач, когда на первое место выдвигаются математические модели с последующей проверкой их опытом. Биометрия же опирается преимущественно на *индуктивный* метод,

---

мин «вариациоиная статистика» (от лат. variatio — изменеиие, колебаиие и status — состояиие, положение вещей) понимают как статистическую обработку результатов измерений. Оба термиа недостаточно точны. Отсюда попыткн заменить их термином «биологическая статистика» (А. В. Леонтович, 1909; П. Ф. Рокицкий, 1973), который также ие лишен недостатков.

Учитывая отсутствие общепринятого названия предмета, Ю. Л. Поморский (1935) пришел к выводу, что из всех предложенных терминов наиболее удачиым следует считать термин «биометрия», как наиболее четко отражающий содержание этого предмета. Следуя этому выводу, автор предпочитает пользоваться первоначальным термином «биометрия», как более кратким и удобным в обращении, применяя его в том смысле, который ои приобрел после классических работ Р. Фишера.

[1] Здесь и в дальнейшем слово «объект» обозиачает то, что изучается или может быть изучено, на что направлено внимание исследователя.

отправляясь от конкретных фактов, которые она анализирует с помощью математических методов.

Характерной особенностью биометрии является также то, что ее методы применяют при анализе не отдельных фактов, а их совокупностей, т. е. явлений массового характера, в сфере которых обнаруживаются закономерности, не свойственные единичным наблюдениям.

Значение биометрии в исследовательской работе и профессиональной подготовке специалистов биологического профиля. Связи современной биологии с математикой многосторонни, они все более расширяются и углубляются. В настоящее время трудно указать область знания, в которой не применялись бы математические методы. Даже в такой, казалось бы, очень далекой от математики области, как анатомия человека, не обходятся без применения биометрии. Примером тому может служить работа Е. М. Маргорина, изучавшего возрастную изменчивость органов у человека. Он писал: «В идеале для определения возрастных различий надо было бы изучать один и тот же орган в его индивидуальном развитии, т. е. у одного и того же человека... Но практически это ограничено пределами анатомии, изучаемой на живом организме, да и требует много времени для наблюдений. Поэтому к решению вопроса приходится подходить косвенным путем, сравнивая один и тот же орган в разные возрастные периоды у разных лиц. Но тогда на сцену выступает новая закономерность — индивидуальная изменчивость, накладывающая существенный отпечаток на весь ход изучения возрастных различий»[1]. Понятно, что в таких случаях достоверные выводы, как считает Е. М. Маргорин, можно получить не на 2—6 наблюдениях, а на гораздо большем их числе; тут без применения биометрии не обойтись.

Биометрия необходима и при изучении наследуемости и повторяемости хозяйственно важных признаков, измерении связей между ними и во многих других случаях. Применение биометрии оказалось полезным во многих областях прикладной биологии. Так, благодаря биометрическому анализу массовых антропологических измерений антропологам удалось подойти к довольно точному обоснованию принципов раскроя и стандартизации обуви и одежды, изготовляемой для массового потребления. Биометрические показатели легли в основу количественной оценки физического развития человека, его спортивных и трудовых достижений. Несомненно, что значение биометрии для наук, изучающих биологические объекты, будет возрастать тем более, чем успешнее применяются достижения счетно-вычислительной техники.

---

[1] См.: *Маргорин Е. М.* Изучение возрастных анатомических различий в свете индивидуальной изменчивости // Арх. анат., гист. и эмбр. 1960. Т. 39. № 10. С. 108, 109.

Конечно, не всякое исследование опирается на биометрию. В биологии с успехом применяют и чисто описательные методы, не требующие количественных оценок получаемых результатов. Но там, где исследования проводят с использованием счета или меры, применение биометрии становится совершенно необходимым. В таких случаях пренебрежение методами биометрии или неправильное их применение приводит к неоправданным затратам труда и времени, а главное — к мало убедительным, а нередко и ошибочным выводам.

В качестве примера можно привести одну из попыток опровергнуть закон расщепления, открытый Г. Менделем. В 1939 г. были опубликованы опыты Н. Е. Ермолаевой, из которых якобы следовало, что частота встречаемости доминантного признака во втором поколении гибридов не совпадает с ожидаемой величиной 3/4. Отсюда был сделан вывод о несостоятельности упомянутого закона Менделя. Заинтересовавшись работой Ермолаевой, акад. А. Н. Колмогоров подверг ее данные статистическому анализу и пришел к прямо противоположному выводу. В статье, опубликованной в одном из номеров журнала «Доклады Академии наук СССР» (1940), он писал: «Материал этот, вопреки сомнению самой Н. Е. Ермолаевой, оказывается блестящим подтверждением законов Менделя». Ошибка Ермолаевой явилась следствием пренебрежительного отношения к биометрии, недооценки ее роли в исследовательской работе. Приведенный пример показывает, во-первых, что пренебрежение биометрическими методами при изучении варьирующих объектов приводит к неубедительным и даже ошибочным выводам, а во-вторых, что неумелое, формальное применение биометрии создает лишь видимость строгой научности, а в действительности приносит не пользу, а вред.

Биометрия — формальная наука. Применение ее к анализу изучаемых явлений требует известной осторожности. Недаром, по образному выражению Гексли, биометрию сравнивают с жерновом, «который всякую засыпку смелет, но ценность помола определяется исключительно ценностью засыпанного»[1].

Биометрия призвана вооружать исследователей методами статистического анализа, воспитывать у них статистическое мышление, раскрывая перед ними диалектику связи между частью и целым, причиной и следствием, случайным и необходимым в явлениях живой природы. Поистине трудно переоценить значение биометрии в подготовке научно-педагогических кадров.

Этапы истории. Биометрия как относительно самостоятельная научная дисциплина сложилась во второй половине XIX в. Однако ее истоки восходят к более раннему периоду в истории естествознания: к тому времени, когда измерения биологических объ-

[1] *Константинов П. Н.* Основы сельскохозяйственного опытного дела. М., 1952. С. 206.

ектов стали рассматривать как *метод научного познания.* Пришедшее на смену феодализма буржуазное общество нуждалось в развитии точных знаний о природе; актуальным для этого времени стал афоризм Г. Галилея (1564—1642): «Измеряй все измеримое и сделай неизмеримое измеримым».

В 1614 г. появилась книга Сантарио (1561—1636) «О статической медицине». В 1680 г. вышла в свет книга Борелли (1608—1679) «О движении животных». В 1768 г. французский гипполог Буржеля издал свой труд «Экстерьер лошади». В этой книге приведен набор измерений, необходимых для определения пригодности лошадей к той или иной службе. Характерно, что в это же время, т. е. в XVIII столетии, развивается *военная антропология,* опирающаяся на результаты измерения тела мужчин призывного возраста в целях отбора пригодных к несению военной службы. Основанием для количественной оценки строения тела животных и человека служил, очевидно, тот факт, что внешние параметры тела животных, а также и строение тела человека находятся в определенной связи с их физическими и психическими свойствами. Чтобы точнее выразить эту связь, визуальную оценку свойств тела животных и людей по их внешнему виду (экстерьеру) стали дополнять его измерениями. А так как результаты измерений варьировали, нужно было исследовать эту изменчивость. В 1718 г. в Лондоне вышла в свет книга французского математика А. де Муавра (1667—1754) «Учение о случаях». Измерив рост у 1375 взрослых женщин и расположив результаты измерений в ряд, он обнаружил закономерность, соответствующую известному в теории вероятностей *закону нормального распределения* (см. гл. IV). Возникла необходимость интеграции методов биологии с методами теории вероятностей и математической статистики.

Теория вероятностей и математическая статистика возникли в середине XVII в. независимо друг от друга. Стимулом к обосованию теории вероятностей послужило развитие денежных отношений в буржуазном обществе. Известную роль при этом сыграли азартные игры — метание монет, игральных костей, картежные игры, — которые оказались простыми моделями, позволившими заметить закономерность в поведении случайных событий массового характера. У истоков теории вероятностей стояли французские ученые П. Ферма (1601—1665) и Б. Паскаль (1623—1662), а также голландский математик и естествоиспытатель Х. Гюйгенс (1629—1696).

Весомый вклад в становление теории вероятностей внесли Я. Бернулли (1654—1705) и А. де Муавр. Однако наиболее существенное развитие получила эта теория в трудах таких выдающихся математиков, как П. Лаплас (1749—1827), К. Гаусс (1777—1855), С. Пуассон (1781—1840), а также П. Л. Чебышев (1821—1894) и его петербургская школа.

Развитие математической статистики связаио с проблемами государствоведения. К середине XVII столетия в экономически развитых страиах Европы накопилось такое количество сведений о демографии, страховом деле, а также в области торговли, здравоохранения и других отраслях хозяйства, что разбираться в них при помощи способов описательной статистики стало почти невозможным. Назрела острая необходимость поиска новых методов анализа статистических данных, их теоретического обоснования. Задача сводилась к тому, чтобы по части судить о состоянии целого, т. е. по выборке делать заключение о всей совокупности общественных явлений в целом, полное описание которых становилось делом очень трудоемким и дорогим. Разработка теории выборочного метода сближала математическую статистику с выводами теории вероятностей, что явилось важной вехой на пути к возникновению биометрии.

Первым, кто удачно объединил методы антропологии и социальиой статистики с выводами теории вероятностей и математической статистикн, был бельгийский антрополог и статистик А. Кетле (1796—1874). В 1835 г. в Брюсселе вышла книга Кетле «О человеке и развитии его способностей или опыт социальной физики». Второе издание этой книги появилось в 1869 г. под заглавием «Социальиая физика или опыт исследования о развитии человеческих способностей»[1]. На большом фактическом материале Кетле впервые показал, что самые различные физические особенности человека и даже его поведение подчиняются в общем закону распределения вероятностей, описываемому формулой Гаусса-Лапласа. В другом труде, «О социальной системе и законах, управляющих ею» (1848), Кетле описал человеческое общество не как сумму индивидов или сообщество людей, проживающих на определенной территории, а как иекую систему, подчиняющуюся строгим законам природы, не зависящим от воли людей. Наконец, в труде «Антропология» (1871) Кетле показал, что открытые им статистические закономерности распространяются не только на человеческое общество, но и на все другие живые существа.

Из работ Кетле следовало, что задача статистики заключается не в одном лишь сборе и классификации статистических данных, а в их анализе, целью которого должно быть открытие закономерностей, действующих в сфере массовых явлений. Знание этих закономерностей и должно было превратить статистику в источник научного познания социальных и биологических явлений.

Исследования Кетле явились поворотным пунктом в истории статистической науки. Кетле одним из первых убедительно показал, что случайности, наблюдаемые в живой природе, вследствие

[1] Эта книга переведена на русский язык в 1911 и 1913 гг.

их повторяемости обнаруживают внутреннюю тенденцию, которую можно исследовать и описать точными математическими методами.

А. Кетле заложил основы биометрии. Создание же математического аппарата этой науки принадлежит английской школе биометриков XIX в., во главе которой стояли Ф. Гальтон (1822—1911) и К. Пирсон (1857—1936). Эта школа возникла под влиянием гениальных трудов Ч. Дарвина (1809—1882), совершившего переворот в биологической науке. Опровергнув господствующее тогда представление о неизменности биологических видов, Дарвин противопоставил ему эволюционное учение, положив в основу принцип естественного отбора. Этот принцип базируется на статистическом характере причинно-следственных отношений, складывающихся в живой природе; он подтверждает гегелевскую концепцию о внутренней связи между случайностью и необходимостью, между причиной и следствием, частью и целым.

Революция, совершенная Дарвином в биологической науке, поставила перед учеными целый ряд больших и неотложных задач, среди которых на первом плане оказалась проблема изменчивости и наследственности организмов. Решение этой проблемы явилось мощным стимулом к развитию экспериментальных методов и, как следствие, к развитию биометрии.

Одним из тех, кто испытал на себе влияние гениального труда Дарвина «Происхождение видов» (1859), был его двоюродный брат Ф. Гальтон. Сильное впечатление произвели на Гальтона и труды Кетле, особенно его «Социальная физика» и «Антропология». Поэтому неудивительно, что именно Гальтону принадлежит первая попытка применить статистические методы к решению проблемы наследственности и изменчивости организмов. Начиная с 1865 г. Гальтон опубликовал ряд оригинальных работ по антропологии и генетике. На большом фактическом материале он подтвердил вывод Кетле о том, что не только физические, но и умственные способности человека распределяются по закону вероятностей, описываемому формулой Гаусса — Лапласа.

Достойным продолжателем исследований Гальтона явился его ученик К. Пирсон — профессор Лондонского университета. Получив в 1884 г. кафедру прикладной математики и механики, Пирсон занялся изучением проблемы наследственности и изменчивости организмов. Он создал математический аппарат биометрии; развил учение о разных типах кривых распределения, разработал метод моментов (1894) и критерий согласия «хи-квадрат» (1990). Пирсон ввел в биометрию такие показатели, как *среднее квадратическое отклонение* (1894) и *коэффициент вариации* (1896). Ему принадлежит усовершенствование методов корреляции и регрессии Гальтона (1896, 1898). Вместе с Д. Гальтоном и Уэльдоном Пирсон организовал выпуск журнала «Биометрика» (1901), редактором которого он оставался до конца своей

жизни. Этот журнал сыграл важную роль в пропаганде биометрических методов, в создании английской школы биометриков.

Разработанные Гальтоном и Пирсоном биометрические методы вошли в золотой фонд математической статистики. Однако попытки Гальтона применить эти методы к решению проблемы наследственности организмов оказались неудачными. Гальтон и Пирсон полагали, что по внешнему сходству между родственниками можно судить о степени их родства. Это было ошибкой, на которую указал датский ученый В. Иогансен (1857—1927). В опытах с фасолью Иогансен пришел к важному выводу о том, что биологические проблемы должны решаться с помощью математики, но не как математические задачи. «Статистике,— писал Иогансен,— всегда должен предшествовать биологический анализ, иначе результаты могут быть «статистической ложью». Математика должна оказывать помощь, а не служить в качестве руководящей идеи»[1]. Это был новый, реалистический подход к оценке роли математических методов в биологических исследованиях.

Значение биометрии в исследовательской работе биологов стало очевидным уже тогда, когда были открыты статистические законы, действующие в сфере массовых явлений. Но биологи не сразу оценили всю важность этих открытий: во-первых, потому, что статистические методы базировались на больших количествах наблюдений, а во-вторых, они требовали большой вычислительной работы, к чему у биологов, привыкших к работе на малочисленных выборках, не было навыка.

Положение стало меняться после того, как была обоснована *теория малой выборки*. Пионером в этой области явился ученик К. Пирсона В. Госсет (1876—1937), опубликовавший в журнале «Биометрика» свой труд под псевдонимом «Стьюдент»[2]. Дальнейшее развитие теория малой выборки получила в трудах Пирсона и особенно Р. Фишера (1890—1962), внесшего огромный вклад в биометрию, обогатив ее новыми методами статистического анализа. На протяжении ряда лет Фишер работал в качестве научного сотрудника Ротамстедской сельскохозяйственной опытной станции, а с 1933 г. — в должности профессора кафедры прикладной математики Лондонского университета. Затем, с 1943 по 1957 г., Фишер заведовал кафедрой генетики в Кембридже.

Удачно соединяя в своем лице биолога-экспериментатора и математика-статистика, Фишер привнес в биометрию не только новые методы, но и новые идеи. Он заложил основы *планирования экспериментов* — теории, которая в настоящее время получила дальнейшее развитие и стала относительно самостоя-

[1] *Иогансен В.* Элементы точного учения об изменчивости и наследственности. М., 1933. С. 103.

[2] См.: *Student.* The probable Error of Mean // Biometrica. 1908. Vol. 6. P. 1—25.

тельным разделом биометрии. Фишер ввел в биометрию целый ряд новых терминов и понятий и убедительно показал, что планирование экспериментов и обработка их результатов — это две неразрывно связанные задачи статистического анализа. Классические труды Фишера явились новой вехой в истории биометрии. Они доказали, что биометрия — не просто наставление к использованию различных технических приемов, применяемых при обработке результатов наблюдений, а нечто большее — наука, занимающаяся статистическим анализом массовых явлений в биологии.

Рассматривая историю биометрии, нельзя не отметить тот огромный вклад в развитие теории вероятностей и математической статистики, который внесли такие ученые нашей страны, как С. Н. Бернштейн (1880—1968), А. Я. Хинчин (1894—1958), Е. Е. Слуцкий (1880—1948), А. И. Хотимский (1892—1939), Б. С. Ястремский (1877—1962), В. И. Романовский (1879—1954), В. С. Немчинов (1894—1964) и многие другие, особенно А. Н. Колмогоров и его школа, получившие мировое признание.

Первый учебник по теории вероятностей был издан в России в 1846 г. акад. В. Я. Буняковским (1804—1889). А первая полная сводка биометрических методов была составлена в 1909 г. акад. А. В. Леонтовичем (1869—1943). В 1910 г. появились «Очерки по теории статистики» А. И. Чупрова (1874—1926). В 1916 г. вышло в свет третье издание руководства по статистическим методам А. А. Кауфмана (1864—1919).

Поток биометрической литературы заметно возрос в нашей стране после Великой Октябрьской социалистической революции. Биометрические методы стали применять в самых различных отраслях биологии и смежных наук. В. В. Алпатов (1898—1979) и его последователи с успехом применили биометрию в пчеловодстве. Н. А. Плохинский (1899—1987), П. Ф. Рокицкий (1902—1977), А. С. Серебровский (1892—1948) и др. — в области генетики и селекции животных. М. В. Игнатьев (1894—1959) и Ю. П. Зыбин установили количественные параметры для раскроя и стандартизации обуви и одежды, изготовляемой для массового потребления. В. В. Бунак (1891—1979), И. И. Шмальгаузен (1884—1963) применили биометрию к изучению закономерностей роста и развития организмов животных и человека. П. В. Терентьев (1903—1970) обосновал метод корреляционных плеяд, зоогеографическое «правило оптимума». А. А. Ляпунов (1911—1973) был одним из основателей математической биологии и кибернетики в нашей стране. В кратком разделе невозможно перечислить имена всех ученых, внесших свою лепту в развитие биометрии.

Возрастающая роль биометрии в исследовательской работе естественно сказалась на подготовке специалистов биологического профиля. Первым, кто еще в 1919 г. начал читать студентам Московского университета курс биометрии с основами генетики,

был С. С. Четвериков (1880—1959). В 1924 г. он читал уже самостоятельный курс «Введение в биометрию». В дальнейшем курс биометрии в МГУ читали В. В. Алпатов, М. В. Игнатьев и др.

Основателем *Ленинградской школы биометриков* был Ю. А. Филипченко (1882—1930), организовавший при Ленинградском университете первую в нашей стране кафедру генетики. Филипченко не только умело применял биометрию в исследовательской работе, но и пропагандировал ее в нашей стране. Написанное им руководство по биометрии «Изменчивость и методы ее изучения» еще при жизни автора выдержало четыре издания (1923, 1925, 1927, 1929). После смерти Филипченко курс биометрии в ЛГУ читал его ученик А. И. Зуйтин, погибший в 1942 г. в блокадном Ленинграде. Затем биометрию в ЛГУ стал преподавать П. В. Терентьев, талантливый ученик С. С. Четверикова, внесший заметный вклад в развитие биометрии. Он первый ввел преподавание биометрии применительно к большому практикуму по зоологии позвоночных животных. П. В. Терентьев был одним из первых организаторов курсов повышения квалификации биологов по вопросам применения математических методов в биологии и четырех Всесоюзных совещаний (1958—1964), посвященных этому вопросу.

Большая работа по обучению специалистов лесного и сельского хозяйства биометрическим методам была проведена Ю. Л. Поморским (1893—1954). Много сделано в области пропаганды статистических методов А. К. Митропольским и другими учеными нашей страны. По учебным руководствам Ю. А. Филипченко, Ю. Л. Поморского, В. И. Романовского и других ученых воспиталось целое поколение отечественных биометриков.

Итак, биометрия в своем историческом развитии прошла долгий и сложный путь — от чисто словесного описания биологических объектов к их измерениям, от статистических сводок и таблиц к статистическому анализу массовых явлений. В истории биометрии можно отметить несколько периодов, или этапов. *Первый период,* описательный, берет свое начало в XVII столетии. В это время происходит переход от словесного описания и элементарного количественного учета биологических объектов к их числовым характеристикам. Измерения рассматриваются как метод научного познания живой природы.

Второй период, начавшийся в первой половине XIX в., ознаменован работами А. Кетле. В это время закладываются основы биометрии как науки, целью которой является не описание явлений, а их анализ, направленный на открытие статистических закономерностей, которые действуют в сфере массовых явлений. Биометрию рассматривают одновременно и как науку, и как метод научного познания.

Третий период, формалистический, характеризуется возникновением и развитием английской биометрической школы во главе

с Ф. Гальтоном и К. Пирсоном. В это время создают математический аппарат биометрии и предпринимают попытки применить его к изучению проблемы наследственности и изменчивости организмов.

Четвертый период, рационалистический, начинается с 1902 г. классическими исследованиями Иогансена, показавшего, что в области биологических исследований первое место должно принадлежать биологическому эксперименту, а не математике. Математические методы должны применяться как вспомогательный аппарат при обработке экспериментальных данных.

Пятый период в развитии биометрии открывают классические работы Стьюдента и Р. Фишера. В это время создаются основы теории малой выборки, теории планирования экспериментов, вводятся в содержание биометрии новые термины и понятия. Все эти новшества связаны с революцией в биологии, с ломкой устаревших принципов и понятий в области исследовательской работы, с усилением процесса математизации биологии. Происходит все более заметная специализация биометрии, применения ее методов в самых различных областях биологии, медицины, антропологии и других смежных науках.

ГЛАВА I

ОСНОВНЫЕ ПОНЯТИЯ БИОМЕТРИИ. ГРУППИРОВКА ПЕРВИЧНЫХ ДАННЫХ

I.1. ПРЕДМЕТ И ОСНОВНЫЕ ПОНЯТИЯ БИОМЕТРИИ

Предметом биометрии служит любой биологический объект, изучаемый с применением счета или меры, т. е. с количественной стороны в целях более или менее точной оценки его качественного состояния. При этом, как уже сообщалось, имеются в виду не единичные, а групповые объекты, т. е. явления массовые, в сфере которых проявляют свое действие статистические законы. Например, врач принял больного и назначил необходимое ему лекарство — это единичное явление, отдельный акт. Если же врач принял несколько больных или подверг неоднократному осмотру одного и того же больного, — это массовое явление независимо от того, каким был объект наблюдения — единичным или групповым.

Обычно наблюдения проводят на групповых объектах, например на особях одного и того же вида, пола и возраста, которые рассматривают как составные элементы, или члены группового объекта, и называют *единицами наблюдения.* Множество относительно однородных, но индивидуально различимых единиц, объединенных для совместного (группового) изучения, называют *статистической совокупностью.*

Понятие статистической совокупности — одно из фундаментальных биометрических понятий. Оно базируется на принципе качественной однородности ее состава. Нельзя объединять в одну совокупность особей разного пола и возраста, когда речь идет о нормах питания, стандартизации обуви и одежды, поскольку заведомо известно, что с возрастом и в зависимости от пола индивидов меняются их потребности в питании и закономерно изменяются размеры и пропорции тела. Недопустимо изучать закономерность модификационной изменчивости на генетически неоднородном материале, объединяя в одну совокупность чистопородных и гибридных особей и т. д.

Наряду с понятием статистической совокупности существует понятие *статистического комплекса.* Так, если статистическая совокупность состоит из относительно однородных единиц, то статистический комплекс слагается из разнородных групп, объединяемых для совместного (комплексного) изучения. При этом

каждая группа, входящая в состав комплекса, должна состоять из однородных элементов. Например, в массе подопытных животных наряду с контролем может быть образовано несколько групп, отличающихся друг от друга по возрасту, породной или видовой принадлежности и т. п., на которых испытывают действие изучаемого агента. При испытании различных доз удобрений каждую опытную делянку рассматривают как отдельную группу, входящую в состав статистического комплекса.

Вопрос о форме объединения биометрических данных экспериментатор решает сам в зависимости от объекта и цели исследования. Объединяемые в статистическую совокупность или статистический комплекс результаты наблюдений представляют некую систему, не сводимую к сумме составляющих ее единиц или компонентов. В статистических совокупностях и в статистических комплексах существует внутренняя связь между частью и целым, единичным и общим, которая находит свое выражение в статистических закономерностях, обнаруживаемых в сфере массовых явлений. Эти закономерности являются той теоретической платформой, на которой базируется биометрия.

1.2. ДИАЛЕКТИКА СВЯЗИ МЕЖДУ ЕДИНИЧНЫМ И ОБЩИМ

Между частью и целым, единичным и общим, на первый взгляд не существует разницы. Ведь «отдельное не существует иначе как в той связи, которая ведет к общему». Равно как и «общее существует лишь в отдельном, через отдельное»[1]. Нельзя представить поле пшеницы или ржи без множества произрастающих на нем растений данной культуры.

Однако связь между единичным и общим, частью и целым непростая. Диалектика этой связи заключается в том, что «всякое общее лишь приблизительно охватывает все отдельные предметы. Всякое отдельное неполно входит в общее...»[2]. Такова философская сторона рассматриваемых понятий.

1.3. ПРИЗНАКИ И ИХ СВОЙСТВА

В общем смысле под словом *«признак»* подразумевают свойство, проявлением которого один предмет отличается от другого. В области биологии признаками, по которым проводят наблюдения над объектами, служат такие характерные особенности в строении и функциях живого, которые позволяют отличать одну единицу наблюдения от другой, сравнивать их между собой. Например, исследователя интересует содержание зерен в колосьях пшеницы или ржи, возделываемой на специально подготовлен-

---

[1] *Ленин В. И.* Полн. собр. соч. Т. 29, С. 318.
[2] Там же.

ном участке. Массив данной культуры и будет объектом наблюдения, а признаком — количество зерен в колосьях отдельных растений, которые являются единицами наблюдения, составляя в общей массе, подвергаемой изучению, статистическую совокупность.

Характерным свойством биологических признаков является *варьирование* величины признаков в определенных пределах при переходе от одной единицы наблюдения к другой. Например, подсчитывая наличие зерен или колосков в колосьях, взвешивая детенышей животных одного и того же помета, определяя жирность молока у животных однородной группы и в других подобных случаях, нетрудно заметить, что величина каждого признака колеблется, образуя совокупность числовых значений признака, по которому проводят наблюдение. Эти колебания величины одного и того же признака, наблюдаемые в массе однородных членов статистической совокупности, называют *вариациями* (от лат. variatio — изменение, колебания), а отдельные числовые значения варьирующего признака принятого называть *вариантами* (от лат. varians, variantis — различимый, изменяющийся) [1].

1.4. КЛАССИФИКАЦИЯ ПРИЗНАКОВ

Все биологические признаки варьируют, но все они поддаются непосредственному измерению. Отсюда возникает деление признаков на *качественные,* или *атрибутивные,* и *количественные.*

Качественные признаки не поддаются непосредственному измерению и учитываются по наличию их свойств у отдельных членов изучаемой группы. Например, среди растений можно подсчитать количество экземпляров с разной окраской цветков — белой, розовой, красной, фиолетовой и т. д. В массе животных также нетрудно отличить и учесть особей разного пола и масти — серых, вороных, гнедых, пестрых и др.

Количественные признаки поддаются непосредственному измерению или счету. Их делят на *мерные,* или *метрические,* и *счетные,* или *меристические.* Длина колосьев, урожайность той или иной культуры, мясная и молочная продуктивность животных — все это мерные признаки, варьирующие непрерывно: их величина может принимать в определенных пределах (от—до) любые числовые значения. Счетные признаки, такие, например, как число зерен или колосков в колосьях, яйценоскость и другие подобные признаки, варьируют прерывисто или дискретно: их числовые значения выражаются только целыми числами.

Если результаты наблюдений группируются в противопоставляемые друг другу группы, их варьирование в отличие от рядо-

---

[1] По терминологии Р. Фишера, числовые значения признаков называются *датами.*

вой изменчивости называют *альтернативным* и признаки, по которым проводят наблюдение, — альтернативными. Примером могут служить случаи, когда противопоставляют особи женские мужским, больные — здоровым, высокорослые — низкорослым, успевающие — неуспевающим и т. д.

Деление признаков на качественные и количественные весьма условно. Например, в массе однородных индивидов, доступных измерению, можно выделить группы высоких, средних и низких, а также успевающих и неуспевающих и т. д. Вместе с тем в каждом качественном признаке, например в окраске листьев, цветков и плодов, можно обнаружить целую гамму количественных переходов, или градаций, и измерить их. И все же, несмотря на очевидную условность приведенной классификации, она необходима хотя бы потому, что количественные признаки распределяются в вариационный ряд, а качественные не распределяются (см. ниже). А при разных способах группировки исходных данных применяют разные способы их обработки.

На языке математики величина любого варьирующего признака является *переменной случайной величиной.* В отличие от постоянных величин, обозначаемых начальными буквами латинского алфавита, переменные величины принято обозначать последними в латинском алфавите прописными буквами X, Y, Z, а их числовые значения, т. е. варианты, — соответствующими строчными буквами того же алфавита: x_1, x_2, x_3, ..., x_n или y_1, y_2, y_3, ..., y_n и т. д. Общее обозначение любой варианты отмечают символом x_i, y_i и т. д., где индекс i символизирует общий характер варианты (даты) [1].

1.5. ПРИЧИНЫ ВАРЬИРОВАНИЯ РЕЗУЛЬТАТОВ НАБЛЮДЕНИЙ

Биологические признаки варьируют под влиянием самых различных, в том числе и случайных, причин. Наряду с естественным варьированием на величине признаков сказываются и ошибки, неизбежно возникающие при измерениях изучаемых объектов. Опыт показал, что как бы точно ни были проведены измерения, они всегда сопровождаются отклонениями от действительного значения измеряемой величины, т. е. не могут быть проведены абсолютно точно. Разница между результатами измерений и действительно существующими значениями измеряемой величины называется *погрешностью* или *ошибкой.*

Ошибки возникают из-за неисправности или неточности измерительных приборов и инструментов (технические ошибки), личных качеств исследователя, его навыков и мастерства в ра-

[1] Классики отечественной биометрии обозначали варнанты буквой V.

боте (личные ошибки) и от целого ряда других, не поддающихся регулированию и неустранимых причин (случайные ошибки).

Технические и личные ошибки, объединяемые в категорию *систематических,* т. е. неслучайных ошибок, можно в значительной степени преодолеть, совершенствуя технические средства, условия работы и личный опыт. Эти меры позволяют свести размеры таких ошибок до минимума, которым можно пренебречь. Случайные же ошибки, как независимые от воли человека, остаются и сказываются на результатах наблюдений.

Итак, варьирование результатов наблюдений вызывают причины двоякого рода: естественная изменчивость признаков и ошибки измерений. Однако по сравнению с естественным варьированием случайные ошибки измерений, как правило, невелики, поэтому варьирование результатов наблюдений рассматривают обычно как естественное варьирование признаков.

1.6. ФОРМЫ УЧЕТА РЕЗУЛЬТАТОВ НАБЛЮДЕНИЙ

Наблюдения над биологическими объектами проводят обычно по принятой исследователем программе. Результаты наблюдений фиксируют в дневниках, журналах, бланках, анкетах или других документах учета. Существует много различных форм и способов учета; выбор той или иной формы определяется задачей исследования и теми условиями, в которых оно проводится. Так, на маршрутных зоологических и ботанических экскурсиях, при проведении полевых опытов удобной формой учета служит дневник. В условиях лабораторного эксперимента результаты испытаний фиксируют в протоколах, журналах, учетных бланках и других формулярах.

1.7. ТОЧНОСТЬ ИЗМЕРЕНИЙ. ДЕЙСТВИЯ НАД ПРИБЛИЖЕННЫМИ ЧИСЛАМИ

Применяя биометрию к решению практических задач, исследователь имеет дело с измерениями биологических объектов. Обычно измерения проводят с точностью до десятых, сотых или тысячных долей единицы, более точные измерения производят реже. Практически каждый признак имеет свою меру. Едва ли необходимо измерять удой коровы за лактацию с точностью до одной сотой миллиграмма. Но было бы недостаточно точным выражать измерения жирномолочности не дробными, а целыми числами. Конечно, в особых случаях, таких, например, как дозирование или испытание ядов и других сильнодействующих веществ, измерения должны быть очень точными, выражаемыми не только тысячными, но и миллионными долями единицы.

Как показывает опыт, нет необходимости в точности измерений, когда эта точность практически не нужна. Данное положе-

ние относится и к измеряемым объектам, и к вычислениям обобщающих статистических характеристик. «Вычисления, — писал акад. А. Н. Крылов, — можно производить как угодно точно, но результат вычисления не может быть точнее тех данных... на которых оно основано»[1].

Разумеется, исследователь может иметь дело с точными числами, получаемыми в результате счета. Но гораздо чаще приходится оперировать приближенными числами, полученными в результате измерений. Такие математические операции, как нахождение логарифма чисел, деление, извлечение корня и другие действия, также в итоге дают приближенные числа.

Чтобы избежать грубых ошибок в работе и получать сопоставимые результаты, необходимо неукоснительно соблюдать признанные правила записи и округления приближенных чисел. Очень важно, чтобы числа, фиксируемые в документах учета, соответствовали *точности, принятой при измерении варьирующих объектов.* Так, если измерения проводят с точностью до одного десятичного знака, то результаты измерений нельзя записывать, например, в таком виде: 5,2; 4; 4,69; 4,083 и т. д. Правильная запись этих чисел будет такова: 5,2; 4,0; 4,7; 4,1.

Числа округляют следующим образом: если за последней сохраняемой цифрой следуют цифры 0, 1, 2, 3, 4, они отбрасываются (*округление с недостатком*); если же за последней сохраняемой цифрой следуют цифры 5, 6, 7, 8 и 9, то последняя сохраняемая цифра увеличивается на единицу (*округление с избытком*). Например, числа 45,346; 8,644; 9,425; 3,585 и 3,575 округляются до двух десятичных знаков так: 45,35; 8,64; 9,43; 3,59 и 3,58.

Многие исследователи считают более точным такое правило: если за последней сохраняемой цифрой следует цифра 5 (с нулями или без оных после нее), то округление осуществляется с недостатком при условии, что сохраняемая цифра *четная.* Если же сохраняемая цифра *нечетная,* то округление осуществляется с избытком. Например, числа 3,585 и 3,575 округляются до двух десятичных знаков таким образом: 3,58 и 3,58.

I.8. СПОСОБЫ ГРУППИРОВКИ ПЕРВИЧНЫХ ДАННЫХ

Зафиксированные в документах учета сведения об изучаемом объекте (или объектах) представляют тот первичный фактический материал, который нуждается в соответствующей обработке. Обработка начинается с упорядочения или систематизации собранных данных. Процесс систематизации результатов массовых наблюдений, объединения их в относительно однородные группы по некоторому признаку называется *группировкой.*

[1] *Крылов А. Н.* Лекции о приближенных вычислениях. М., 1933. С. 486.

Группировка — это не просто технический прием, позволяющий представить первичные данные в комплексном виде, но и глубоко осмысленное действие, направленное на выявление связей между явлениями. Ведь от того, как группируется исходный материал, во многих случаях зависят выводы о природе изучаемого явления. Один и тот же материал дает диаметрально противоположные выводы при разных приемах группировки. Нельзя группировать в одну и ту же совокупность неоднородные по составу данные, необдуманно выбирать способ группировки. Группировка должна отвечать требованию поставленной задачи и соответствовать содержанию изучаемого явления.

Таблицы. Наиболее распространенной формой группировки являются *статистические таблицы;* они бывают простыми и сложными. К *простым* относятся, например, четырехпольные таблицы, применяемые при альтернативной группировке, когда одна группа вариант противопоставляется другой; например, здоровые — больным, высокие — низким и т. д. В качестве примера такой группировки могут служить результаты обследования 265 учащихся младших классов на состояние нёбных миндалин (табл. 1).

Таблица 1

Школьные классы	Обнаружено детей		Всего
	здоровых	больных	
Третьи и четвертые	63	92	155
Пятые и шестые	71	39	110
Всего	134	131	265

Из табл. 1 видно, что заболевание нёбных миндалин, по-видимому, чаще встречается среди учащихся третьих и четвертых классов.

К *сложным* относятся многопольные таблицы, применяемые при изучении корреляционной зависимости и при выяснении причинно-следственных отношений между варьирующими признаками. Примером корреляционной таблицы служат классические данные Гальтона, показывающие наличие положительной зависимости между ростом родителей и ростом их детей (табл. 2).

В качестве примера группировки, применяемой при выяснении причинно-следственных отношений между признаками, приведены данные, полученные в Научно-исследовательском институте имени В. В. Докучаева при испытании гречихи сорта «Богатырь» на урожайность в зависимости от предшественников (табл. 3).

Таблица 2

Рост родителей, дюймы	Рост детей, дюймы								Всего
	60,7	62,7	64,7	66,7	68,7	70,7	72,7	74,7	
74							4		4
72			1	4	11	17	20	6	62
70	1	2	21	48	83	66	22	8	251
68	1	15	56	130	148	69	11		430
66	1	15	19	56	41	11	1		144
64	2	7	10	14	4				37
Всего	5	39	107	255	387	163	58	14	928

Из табл. 3 ясно, что в данных условиях лучшим предшественником для гречихи является, по-видимому, ячмень.

Таблица 3

Предшественники	Урожай гречихи по повторностям, ц/га			Средний урожай
	1	2	3	
Горох раннезеленый	23,7	20,1	20,5	21,4
Чечевица	23,6	25,1	21,1	23,2
Чина степная № 21	26,7	23,2	23,8	24,6
Ячмень	26,0	24,9	25,3	25,4

Приведенными таблицами не исчерпывается их многообразие. Здесь рассмотрены лишь типичные для курса биометрии примеры. Из этих примеров видно, что статистические таблицы имеют не только иллюстративное, но и аналитическое значение, позволяя обнаруживать связи между варьирующими признаками.

Статистические ряды. Особую форму группировки представляют так называемые *статистические ряды*. Статистическим называется ряд числовых значений признака, расположенных в определенном порядке. В зависимости от того, какие признаки изучаются, статистические ряды делят на атрибутивные, вариационные, ряды динамики и регрессии, а также ряды ранжированных значений признаков и ряды накопленных частот, являющихся производными вариационных рядов. Примером *атрибутивного ряда* могут служить данные, показывающие зависимость между содержанием гемоглобина Hb в крови и высотой организации позвоночных животных:

Класс животных	Рыбы	Амфибии	Рептилии	Птицы	Млекопитающие
Количество Hb, г/кг массы тела	1,6	2,9	3,8	11,2	11,7

Среди группировок видное место занимают вариационные ряды. На их описании следует остановиться более подробно. Ряды регрессии, динамики и другие будут рассмотрены в последующих главах.

Вариационным рядом или *рядом распределения* называют двойной ряд чисел, показывающий, каким образом числовые значения признака связаны с их повторяемостью в данной статистической совокупности. Например, из урожая картофеля, собранного на одной из опытных делянок, случайным способом, т. е. наугад, отобрано 25 клубней, в которых подсчитывали число глазков. Результаты подсчета оказались следующие: 6, 9, 5, 7, 10, 8, 9, 10, 8, 11, 9, 12, 9, 8, 10, 11, 9, 10, 8, 10, 7, 9, 11, 9, 10. Чтобы разобраться в этих данных, расположим их в ряд (в порядке регистрации результатов наблюдений) с учетом повторяемости вариант в этой совокупности:

Вариаиты x_i	6	9	5	7	10	8	11	12
Число иариант f_i	1	7	1	2	6	4	3	1

Это и есть вариационный ряд. Числа, показывающие, сколько раз отдельные варианты встречаются в данной совокупности, называются *частотами* или *весами* вариант и обозначаются строчной буквой латинского алфавита f. Общая сумма частот вариационного ряда равна объему данной совокупности, т. е. $\sum_{i=1}^{k} f_i = n$, где $\sum_{i=1}^{k}$ (греческая буква сигма прописная) обозначает действие суммирования, в даиом случае суммирование частот вариационного ряда от первого ($i=1$) до k-го класса, а n — общее число наблюдений, или объем совокупности.

Частоты (веса) выражают не только абсолютными, но и относительными числами — в долях единицы или в процентах от общей численности вариант, составляющих данную совокупность. В таких случаях весá называют *относительными частотами* или *частостями*. Общая сумма частостей равна единице, т. е. $\Sigma f_i/n = 1$, или $\Sigma(f_i/n)100 = 100\%$, если частоты выражены в процентах от общего числа наблюдений n. Замена частот частостями не обязательна, но иногда оказывается полезной и даже необходимой в тех случаях, когда приходится сопоставлять друг с другом вариационные ряды, сильно отличающиеся по их объемам.

Распределение исходных данных в вариационный ряд преследует определенные цели. Одна из них — ускорение работы при вычислении по вариационному ряду обобщающих числовых характеристик — средней величины и показателей вариации (см. гл. II). Другая сводится к выявлению закономерности варьирования учитываемого признака. Приведенный ряд удовлетворяет первой, но не удовлетворяет достижению второй цели. Чтобы ряд распределения полностью удовлетворял предъявляемым

к нему требованиям, его нужно строить по ранжированным значениям признака.

Под *ранжированием* (от франц. ranger — выстраивать в ряд по ранжиру, т. е. по росту) понимают расположение членов ряда в возрастающем (или убывающем) порядке. Так, в данном случае результаты наблюдений следует распределить так:

Вариаиты x_i	5	6	7	8	9	10	11	12
Частоты f_i	1	1	2	4	7	6	3	1

Этот упорядоченный ряд распределения в равной мере удовлетворяет достижению и первой, и второй целей. Он хорошо обозрим и наилучшим образом иллюстрирует закономерность варьирования признака.

В зависимости от того, как варьирует признак — дискретно или непрерывно, в широком или узком диапазоне, — статистическая совокупность распределяется в *безынтервальный* или *интервальный* вариационные ряды. В первом случае частоты относятся непосредственно к ранжированным значениям признака, которые приобретают положение отдельных групп или классов вариационного ряда, во втором — подсчитывают частоты, относящиеся к отдельным промежуткам или интервалам (от — до), на которые разбивается общая вариация признака в пределах от минимальной до максимальной варианты данной совокупности. Эти промежутки, или классовые интервалы, могут быть равными и не равными по ширине. Отсюда различают *равно- и неравноинтервальные вариационные ряды.* Примером неравноинтервального ряда распределения могут служить данные А. Ф. Ковшарь (1966), показывающие зависимость между числом стай сизых голубей и количеством особей в стае в гнездовой (с 15 марта по 15 августа) и послегнездовой (с 15 августа по 15 марта) периоды их жизни (табл. 4).

В неравноинтервальных рядах характер распределения частот меняется по мере изменения ширины классовых интервалов. Поэтому в качестве числовых характеристик таких рядов используют особые показатели (см. гл. II).

Неравноинтервальную группировку в биологии применяют сравнительно редко. Как правило, биометрические данные распределяются в равноинтервальные ряды, что позволяет не только выявлять закономерность варьирования, но и облегчает вычисление сводных числовых характеристик вариационного ряда, сопоставление рядов распределения друг с другом.

Приступая к построению равноинтервального вариационного ряда, важно правильно наметить ширину классового интервала. Дело в том, что грубая группировка (когда устанавливают очень широкие классовые интервалы) искажает типичные черты варьирования и ведет к снижению точности числовых характеристик ряда. При выборе чрезмерно узких интервалов точность обобща-

ющих числовых характеристик повышается, но ряд получается слишком растянутым и не дает четкой картины варьирования.

Таблица 4

Число особей в стае	Число встреч (частота)				Плотность распределения			
	в гнездовой период		в остальное время года		в гнездовой период		в остальное время года	
	абсолютные значения	значения в процентах	абсолютные значения	значения в процентах	абсолютная	относительная	абсолютная	относительная
Одиночки	6	18,20	1	1,70	6,00	18,20	1,00	1,70
2—5	19	57,60	9	15,25	6,33	19,20	3,00	5,08
5—10	4	12,10	4	6,78	0,80	2,42	0,80	1,36
10—20	2	6,10	12	20,34	0,20	0,61	1,20	2,03
20—30	1	3,00	13	22,03	0,10	0,30	1,30	2,20
30—50	1	3,00	11	18,65	0,05	0,15	0,55	0,93
50—100	0	0,00	9	15,25	0,00	0,00	0,18	0,31
Всего встреч	33	100	59	100	—	—	—	—

Примечание. Четыре последние графы понадобятся в дальнейшем (см. разд. II.1).

Для получения хорошо обозримого вариационного ряда и обеспечения достаточной точности вычисляемых по нему числовых характеристик следует разбить вариацию признака (в пределах от минимальной до максимальной варианты) на такое число групп или классов, которое удовлетворяло бы обоим требованиям. Эту задачу решают делением размаха варьирования признака на число групп или классов, намечаемых при построении вариационного ряда:

$$\lambda = \frac{x_{max} - x_{min}}{K}, \qquad (1$$

где λ — величина классового интервала; x_{max}, x_{min} — максимальная и минимальная варианты совокупности; K — число классов, на которые следует разбить вариацию признака.

Число классов (K) можно приблизительно наметить, пользуясь табл. 5.

Более точно величину K можно определить по формуле Стерджеса: $K = 1 + 3{,}32 \lg n$. При наличии в совокупности большого числа членов ($n > 100$) можно использовать формулу $K = 5 \lg n$ (К. Брукс, Н. Карузерс, 1963) [1].

[1] В тех случаях, когда по вариационному ряду вычисляют числовые характеристики (средние, дисперсии и др.) согласно рекомендации Д. Юла и

Таблица 5

Число наблюдений n (от — до)	Число классов K
25—40	5—6
40—60	6—8
60—100	7—10
100—200	8—12
>200	10—15

Вопрос о том, распределять ли собранные данные в интервальный или безынтервальный ряд, решают в зависимости от характера и размаха варьирования признака. Если признак варьирует дискретно и слабо, т. е. в узких границах (величиина λ оказывается равной единице или может быть приравнена к единице), данные распределяются в безынтервальный вариационный ряд. Если же признак варьирует в широких границах, то независимо от того, как он варьирует — дискретно или непрерывно, по данным строят интервальный вариационный ряд.

Техника построения вариационных рядов. Приступая к построению вариационного ряда, нужно в сводке исходных данных отыскать минимальную x_{min} и максимальную x_{max} варианты. Затем, используя формулу (1), определить величину классового интервала λ. Если окажется, что $\lambda=1$, собранный материал распределяется в безынтервальный вариационный ряд; если же $\lambda\neq 1$, исходные данные необходимо распределять в интервальный ряд. При этом точность величины классового интервала должна соответствовать точности, принятой при измерении признака. Например, жирномолочность коров ($n=60$), содержащихся на ферме, варьирует от 3,21 до 4,55%. В таком случае классовый интервал устанавливается следующим образом:

$$\lambda=\frac{4,55-3,21}{1+3,32\ \lg 60}=\frac{1,34}{1+5,90}=0,194\approx 0,19.$$

Если точность измерения данного признака ограничить десятыми долями единицы, величина классового интервала окажется следующей:

$$\lambda=\frac{4,6-3,2}{6,9}=0,2.$$

В обоих случаях результаты наблюдений должны распределяться в интервальный вариационный ряд.

При построении интервального вариационного ряда следует поступать так, чтобы минимальная варианта совокупности попадала примерно в середину первого классового интервала. Выполнение этого требования гарантирует построение вариационного ряда, наиболее полно отвечающего природе изучаемого явления,

М. Кендэла (1960), следует выделять 15—20 классовых интервалов независимо от числа наблюдений, что обеспечит достаточную компактность вычислений и практическую их точность. (*Прим. ред.*)

а следовательно, и наименьшие потери информации о точности вычисляемых статистических характеристик ряда. Этому требованию удовлетворяет формула

$$x_н = x_{min} - \lambda/2, \qquad (2)$$

где $x_н$ — нижняя граница первого классового интервала; x_{min} — минимальная варианта исследуемой совокупности; λ — величина классового интервала.

Так, при $x_{min}=3{,}21$ и $\lambda=0{,}19$ нижняя граница первого класса $x_н=3{,}21-0{,}19/2=3{,}115\approx3{,}12$. Прибавив к этой величине $\lambda=0{,}19$, определяем верхнюю границу первого класса: $3{,}12+0{,}19=3{,}31$. Затем находим верхнюю границу второго класса: $3{,}31+0{,}19=3{,}50$ и т. д., пока не получим интервал, в который попадает максимальная варианта совокупности ($x_{max}=4{,}55$).

Наметив классовые интервалы, остается распределить по ним все варианты совокупности, т. е. определить частоты каждого класса. Тут, однако, возникает вопрос: в какие классы относить варианты, которые по своей величине совпадают с верхней границей одного и нижней границей другого, соседнего класса? Например, в какой класс следует отнести варианту 3, 31 — в первый или во второй? Этот вопрос решается по-разному. Можно помещать в один и тот же класс варианты, которые больше нижней, но меньше или равны его верхней границе, т. е. по принципу «от и до включительно». Чаще, однако, поступают таким образом: верхние границы классов уменьшают на величину, равную точности, принятой при измерении признака, чем и достигается необходимое разграничение классов.

Следующий шаг ведет к замене классовых интервалов на их центральные или срединные значения. В результате *интервальный вариационный ряд превращается в безынтервальный ряд.* Необходимость такой замены вызывается тем, что обобщающие числовые характеристики (средняя, дисперсия и др.) вычисляются по безынтервальным рядам. Срединные значения классовых интервалов x_i, как это следует из формулы (2), отстоят от их нижних границ $x_н$ на величину, равную половине классового интервала.

Наиболее точно *центральную величину классового интервала* можно получить по формуле

$$x_i = \frac{1}{2}(x_н + x_к), \qquad (3)$$

где $x_к$ — конечная точка интервала, равная $x_{н+1}-\varepsilon$, т. е. началу следующего класса, уменьшенному на точность измерения признака.

Середины классов приобретают значения отдельных вариант и называются *классовыми вариантами* в отличие от конкретных вариант, составляющих данную совокупность. Описан-

ную методику можно продемонстрировать на конкретных примерах.

Пример 1. На свиноферме зарегистрировано 64 опороса. Количество поросят, полученных от каждой свиноматки, варьировало следующим образом:

8	10	6	10	8	5	11	7	10	6	9	7	8	7	9	11	8	9	10
8	7	8	11	8	7	10	8	8	5	11	8	10	12	7	5	7	9	7
10	5	8	9	7	12	8	9	6	7	8	7	11	8	6	7	9	10	

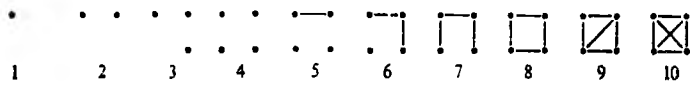
В этой совокупности $x_{\min}=5$ и $x_{\max}=12$. Отсюда

$$\lambda=\frac{12-5}{1+3{,}32\lg 64}=\frac{7}{1+6}=1.$$

Так как признак варьирует дискретно и $\lambda=1$, совокупность наблюдений следует распределить в безынтервальный вариационный ряд, т. е. непосредственно по ранжированным значениям признака, которые и будут классами данного ряда.

Чтобы при определении частот каждого класса не сбиться со счета, нужно построить вспомогательную (расчетную) таблицу, в которой первая графа заполняется классами (в данном случае ранжированными значениями признака), а вторая — служит для учета частот f_i, распределяемых по этим классам.

Разноску частот по классам производят следующим образом. Просматривая сводку результатов наблюдений, во второй графе расчетной таблицы с помощью условных знаков отмечают повторяемость вариант для каждого класса. При этом рекомендуется не выискивать одинаковые варианты в общей совокупности чисел, а *разносить* их по классам, что не одно и то же. Пренебрежение этой рекомендацией приводит к ошибкам, лишней затрате труда и времени на выполнение работы. Условными знаками при разноске наблюдений по классам могут быть точки, черточки и другие условные знаки. Удобным, особенно при обработке больших совокупностей, оказывается следующий шифр частот:

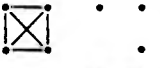
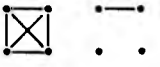


Наилучшим способом разноски вариант является раскладка бланков, каждый из которых соответствует отдельному наблюдению и содержит значения признаков. В этом случае значения признака, попадающие в один класс ряда, образуют отдельную стопку бланков. Всего таких стопок окажется столько, сколько образовано интервалов вариационного ряда. После окончания разноски ее результаты можно неоднократно проверить просмотром каждой из стопок бланков, т. е. просмотром

тех вариант, которые попали в каждый класс ряда. Это позв ляет полностью исключить ошибки, возможные при иных сп собах подсчета классовых частот.

Закончив разноску вариант по классам, переводят шиф частот в числа. В результате получается третья графа вспом гательной таблицы, содержащая частоты безынтервального ва риационного ряда (табл. 6).

Таблица 6

Классы x_i	Шифр частот	Частоты p_i
5		4
6		7
7		13
8		15
9		7
10		9
11		6
12		3
Сумма		64

Полученный вариационный ряд выражает зависимость меж ду отдельными вариантами и частотой их встречаемости в дан ной совокупности, т. е. закономерность варьирования учитыва мого признака.

Пример 2. На основании многолетних клинических набл дений, проводившихся в Сухумском питомнике обезьян, состав лена следующая выборка, включающая 100 анализов на соде жание кальция (мг%) в сыворотке крови низших обезьян (па вианов-гамадрилов):

3,6	12,9	12,3	9,9	12,7	11,7	10,8	10,4	10,9	10,2
4,7	10,4	11,6	11,7	12,1	10,9	12,1	9,2	10,7	11,5
3,1	10,9	12,0	11,1	13,5	11,2	13,5	10,1	14,0	10,0
11,6	12,4	11,9	11,4	12,8	11,4	10,9	12,7	13,8	13,2
11,9	10,8	11,0	12,6	10,0	10,3	12,7	11,7	12,1	13,8
12,2	11,9	11,6	10,6	11,1	10,7	12,3	11,5	11,2	11,5
12,7	10,5	11,2	11,9	9,7	13,0	9,6	12,5	11,6	9,0
11,5	12,3	12,8	12,6	12,8	12,5	12,8	11,4	12,5	12,3
14,5	12,3	12,6	11,7	12,2	12,3	11,6	12,0	13,5	12,5
11,6	11,9	12,0	11,4	14,7	11,3	13,2	14,3	13,2	14,2

Нужно сгруппировать эти данные в вариационный ряд. В данном случае признак варьирует непрерывно в пределах от 9,0 до 14,7 мг%. Устанавливаем величину классового интервала:

$$\lambda = \frac{14,7 - 9,0}{1 + 3,32 \lg 100} = \frac{5,7}{7,6} = 0,75 \approx 0,8.$$

Так как выборку приходится группировать в интервальный вариационный ряд, определяем нижнюю границу первого класса:

$$x_н = 9,0 - \frac{0,8}{2} = 8,6.$$

Затем намечаем следующие классовые интервалы: 8,6—9,4—10,2—11,0—11,8—12,6—13,4—14,2—15,0. Получилось восемь интервалов. Разграничиваем их на величину, равную точности измерения признака, т. е. уменьшаем верхние границы интервалов на 0,1 мг%. Строим вспомогательную расчетную таблицу и разносим все 100 вариант по намеченным классовым интервалам (табл. 7).

Переводим шифр частот в числа, сумма которых должна быть равна объему данной выборки, т. е. $\Sigma f_i = 100$. В результате получается интервальный вариационный ряд.

Срединные значения классов, приведенные в табл. 7, получены прибавлением к нижним границам классов 1/2 классового интервала — величины, равной 0,8/2=0,4, т. е., по формуле (3), 8,6+0,4=9,0; 9,4+0,4=9,8 и т. д. Таким образом, интервальный вариационный ряд превращен в ряд безынтервальный.

Пример 3. В результате учета яйценоскости 80 кур, содержащихся на птицеферме, было установлено, что признак варьирует от 208 до 250 яиц, полученных от несушки за 1 год. Определяем классовый интервал:

$$\lambda = \frac{250 - 208}{1 + 3,32 \lg 80} = \frac{42}{1 + 6,3} = 5,75 \approx 6.$$

Так как классовый интервал не равен единице, результаты наблюдений нужно распределять в интервальный вариационный ряд, несмотря на то что признак варьирует дискретно. Устанавливаем нижнюю границу первого класса: $x_н = 208 - 6/2 =$

Классы по уровню кальция в сыворотке крови, мг%	Срединные значения классов x_i	Шифр частот	Частоты p_i
8,6-9,3	9,0		2
9,4-10,1	9,8		6
10,2-10,9	10,6		15
11,0-11,7	11,4		23
11,8-12,5	12,2		25
12,6-13,3	13,0		17
13,4-14,1	13,8		7
14,2-14,9	14,6		5
Сумма			100

=205. Намечаем классовые интервалы: 205—211—217—223—229—235—241—247—253. Уменьшаем верхние границы классов на единицу: 205—210; 211—216; 217—222; 223—228; 229—234; 235—240; 241—246; 247—252. Дальнейшие действия, относящиеся к построению вариационного ряда, понятны из предыдущих примеров.

Графики вариационных рядов. Для того чтобы более наглядно представить закономерность варьирования количественных признаков, вариационные ряды принято изображать в виде графиков. Так, при построении графика безынтервального вариационного ряда по оси абсцисс откладывают срединные значения классов, по оси ординат — частоты. Высота перпендикуляров, восставляемых по оси абсцисс, соответствует частотам классов. Соединяя вершины перпендикуляров прямыми линиями, получают геометрическую фигуру в виде многоугольника, называемую *полигоном распределения частот*. Линия, соединяющая вершины перпендикуляров, называется *вариационной*

,ривой или кривой распределения частот вариационного ряда[1] рис. 1).

При построении графика интервального вариационного ряда по оси абсцисс откладывают границы классовых интервалов, по оси ординат — частоты интервалов. В результате получается

Рис. 1. Полигон распределения численности поросят в 64 опоросах свиноматок

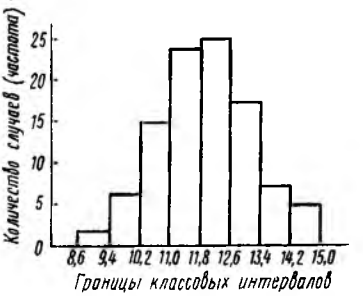


Рис. 2. Гистограмма распределения кальция (мг%) в сыворотке крови обезьян

Так называемая *гистограмма распределения частот.* На рис. 2 изображена гистограмма распределения кальция в сыворотке крови обезьян. Если из середин верхних сторон прямоугольников гистограммы опустить перпендикуляры на ось абсцисс, гистограмма превращается в полигон распределения, а линия, соединяющая середины верхних сторон прямоугольников гистограммы, будет представлять собой вариационную кривую.

Если по оси абсцисс откладывать значения классов, а по оси ординат — накопленные частоты с последующим соединением точек прямыми линиями, получается график, называемый *кумулятой.* На рис. 3 изображена кумулята распределения кальция в сыворотке крови обезьян. В отличие от вариационной кривой, имеющей куполообразную форму, кумулята имеет вид S-образной кривой. Накопленные частоты находят последовательным суммированием, или *кумуляцией* (от лат. cumulatio — увеличение, скопление) частот в направлении от первого класса до конца вариационного ряда. В данном случае частоты ряда распределения кальция в сыворотке крови обезьян кумулированы следующим образом:

Частоты f_i	2	6	15	23	25	17	7	5
Кумуляты частот Sf_i . .	2	8	23	46	71	88	95	100

Откладывая по оси абсцисс частоты, а по оси ординат — значения классов с последующим соединением геометрических

[1] Это понятие условно, так как такая линия является не кривой, а ломаной. (*Прим. ред.*)

точек прямыми линиями, как это показано на рис. 4, получают линейный график, называемый *огивой*.

По сравнению с эмпирическими вариационными кривыми, которые выглядят обычно в виде ломаных линий, кумулята и огива имеют более обтекаемую форму. Эта особенность позволяет в ряде случаев отдавать предпочтение этим графикам перед эмпирической вариационной кривой. Центральная точка ку-

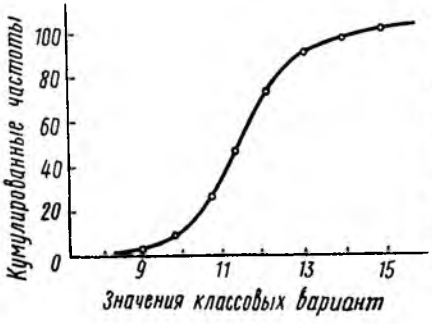


Рис. 3. Кумулята распределения кальция (мг%) в сыворотке крови обезьян

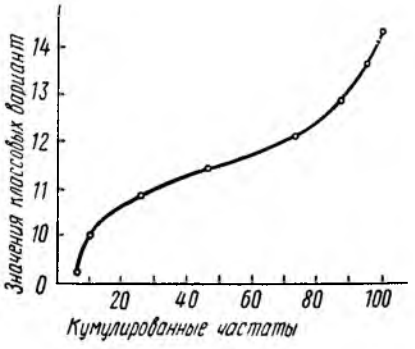


Рис. 4. Огива распределения кальция (мг %) в сыворотке крови обезьян

мулята совпадает с центром распределения совокупности, что дает возможность использовать ее при определении, например средних доз биологически активных веществ, вызывающих эффект у 50% подопытных индивидов. Огива позволяет сравнивать друг с другом одновременно несколько эмпирических распределений неравного объема.

Следует заметить, что неумелое построение графиков приводит к тому, что последние получаются либо в виде островершинных геометрических фигур с узким основанием, либо плосковершинными, чрезмерно растянутыми по оси абсцисс. В обоих случаях графики оказываются плохо обозримыми, нечетко отображающими закономерность варьирования.

Избежать этих недостатков позволяет *правило «золотого сечения»*, согласно которому основание геометрической фигуры должно относиться к ее высоте, как 1:0,62. Применительно к построению вариационной кривой масштабы на осях прямоугольных координат следует выбирать с таким расчетом, чтобы основание кривой было в 1,5—2,0 раза больше ее высоты (т. е. максимальной ординаты). Откладывая по оси абсцисс классы вариационного ряда, следует также доводить крайние из них до нулевых классов, в которых не содержится ни одной варианты. В результате вариационной кривой придается законченный, хорошо обозримый вид.

ГЛАВА II

ОСНОВНЫЕ ХАРАКТЕРИСТИКИ ВАРЬИРУЮЩИХ ОБЪЕКТОВ

II.1. СРЕДНИЕ ВЕЛИЧИНЫ

Вариационные ряды и их графики дают наглядное представление о варьировании признаков, но они недостаточны для полного описания варьирующих объектов. Для этой цели служат особые, логически и теоретически обоснованные числовые показатели, называемые *статистическими характеристиками.* К ним относятся прежде всего средние величины и показатели вариации. «Закономерность,— по словам В. И. Ленина,— не может проявляться иначе как в средней,... массовой... закономерности при взаимопогашении индивидуальных уклонений в ту или другую сторону» [1].

В отличие от индивидуальных числовых характеристик *средние величины* обладают большей устойчивостью, способностью характеризовать целую группу однородных единиц одним (средним) числом. И хотя средние величины абстрактны, они вполне понятны и ощутимы. Средний рост, средняя продуктивность, средний урожай, средняя успеваемость и другие средние — все это понятия абстрактные о конкретном. Значение средних заключается в их свойстве аккумулировать или уравновешивать все индивидуальные отклонения, в результате чего проявляется то наиболее устойчивое и типичное, что характеризует качественное своеобразие варьирующего объекта, позволяет отличать один групповой объект от другого.

В зависимости от того, как распределены первичные данные — в равно- или в неравноинтервальный вариационный ряд,— для их характеристики применяют разные средние величины. Именно при распределении собранных данных в неравноинтервальный вариационный ряд более подходящей обобщающей характеристикой изучаемого объекта служит так называемая *плотность распределения*, т. е. отношение частот или частостей к ширине классовых интервалов, как это показано в табл. 4. Кроме того, числовыми характеристиками таких рядов могут служить средние из абсолютных или относительных показателей плотности распределения. *Средняя плотность* показывает, сколько единиц данной совокупности приходится в среднем на интервал, равный единице измерения учитываемого признака. Так, по табл. 4 находим, что средние из относительных (процентных) показателей плотности распределения голубей в стае в гнездовой период $\bar{x}_1$ и в остальное время года $\bar{x}_2$

[1] *Ленин В. И.* Полн. собр. соч. Т. 26. С. 68.

оказываются следующими: $\bar{x}_1 = (1/7)(18{,}20+19{,}20+2{,}42+0{,}16+0{,}30+0{,}15+0{,}00) = 40{,}43/7 = 5{,}78 \approx 6$ и $\bar{x}_2 = (1/7)(1{,}70+5{,}08+1{,}36+\ldots+0{,}31) = 13{,}61/7 = 1{,}94 \approx 2$. Таким образом выясняется, что в среднем относительная плотность численности голубей в стае в гнездовой период в три раза выше, чем в остальное время года.

В качестве статистических характеристик равноинтервальных вариационных рядов применяют *степенные* и *структурные* (нестепенные) *средние величины*. Степенные средние вычисляют из общей формулы

$$M = \left[\frac{\sum x_i^k}{n}\right]^{1/k} \text{ или } M = \sqrt[k]{\frac{\sum x_i^k}{n}},$$

где M — средняя величина; x_i — варианта; n — число наблюдений, для которых вычисляют среднюю; k — величина, по которой определяют вид средней. Так, при $k=1$ получается средняя арифметическая, при $k=2$ — средняя квадратическая, при $k=-1$ образуется средняя гармоническая и т. д. Из структурных средних в биологии применяют медиану, моду и др.

Средние величины могут характеризовать только однородную совокупность вариант. Если средняя получена на качественно неоднородном материале или выбрана неправильно, без учета специфики характеризуемого явления или процесса, она окажется фиктивной. При наличии разнородных по составу данных их необходимо группировать в отдельные качественно однородные группы и вычислять групповые или частные средние.

Средние величины принято обозначать теми же строчными буквами латинского алфавита, что и варианты, с той лишь разницей, что над буквой, соответствующей средней величине, ставят черту. Так, если признак обозначен через X, то его числовые значения выражают буквой x_i, среднюю арифметическую — $\bar{x}$, среднюю гармоническую x_h и т. д.[1] При вычислении средних величин и других статистических характеристик не обязательно распределять исходные данные в вариационный ряд.

Средняя арифметическая $\bar{x}$. Из общего семейства степенных средних наиболее часто используют *среднюю арифметическую*. Этот показатель является центром распределения, вокруг которого группируются все варианты статистической совокупности. Средняя арифметическая может быть *простой* и *взвешенной*. Простую среднюю арифметическую определяют как сумму всех

[1] В старых руководствах по биометрии средние величины обозначали буквой M.

членов совокупности, деленную на их общее число:

$$\bar{x}=\frac{x_1+x_2+x_3+\ldots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i. \qquad (4)$$

В этой формуле x_i — значения вариант; $\sum_{i=1}^{n}$ — знак суммирования вариант в пределах от первой (x_1) до n-й варианты; n — общее число вариант, или объем данной совокупности.

Когда отдельные варианты повторяются, среднюю арифметическую вычисляют по формуле

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k}x_i f_i \qquad (5)$$

и называют *взвешенной средней,* причем весами, как это показывает формула (5), служат частоты вариант f_i.

При объединении групповых средних их весами будут объемы групп n_i, по которым эти средние вычислены. Общую (взвешенную) среднюю арифметическую нескольких однородных групп определяют по формуле

$$\bar{\bar{x}}=\frac{\bar{x}_1n_1+\bar{x}_2n_2+\bar{x}_3n_3+\ldots+\bar{x}_kn_k}{n_1+n_2+n_3+\ldots+n_k}=\frac{\sum(\bar{x}_in_i)}{\sum n_i}. \qquad (6)$$

Пример 1. Содержание гемоглобина в крови, взятой у взрослых мужчин ($n_1=30$), оказалось равным в среднем 69,8%. Тот же показатель для другой группы мужчин того же возраста ($n_2=20$) составил 64,9%. Нужно определить среднюю арифметическую из этих двух среднних. Если бы выборки были равновеликими, задача решалась бы просто: путем деления суммы частных средних на их число, т. е. $\bar{\bar{x}}=(69{,}8+64{,}9):2=67{,}35\approx\approx 67{,}4\%$. При разных объемах выборок такой расчет оказывается неточным, так как не учитываются веса частных средних. Взвешенная средняя будет равна

$$\bar{\bar{x}}=\frac{30\cdot 69{,}8+20\cdot 64{,}9}{30+20}=67{,}84\approx 67{,}8\ \%.$$

Средняя арифметическая — одна из основных характеристик варьирующих объектов. Она обладает рядом важных свойств.

1. Если каждую варианту статистической совокупности уменьшить или увеличить на некоторое произвольно взятое положительное число A, то и средняя уменьшится или увеличится на это число.

Доказательство: $\bar{x}^*=\frac{\sum(x_i-A)f_i}{\sum f_i}=\frac{\sum x_i f_i}{n}-\frac{A\sum f_i}{\sum f_i}=\bar{x}-A.$

Отсюда $\bar{x}=\bar{x}^*+A$. Это означает, что среднюю $\bar{x}$ можно вычис-

лять по уменьшенным на A членам выборки, прибавив к полученной величине вычтенное из вариант число A.

Пример 2. Имеются следующие шесть вариант: 7, 9, 15, 10, 11, 8. Средняя $\bar{x}=(1/6)(7+9+15+10+11+8)=60/6=10$. Вычтем из каждой варианты $A=7$ и вычислим среднюю арифметическую: $\bar{x}^*=\frac{0+2+8+3+4+1}{6}=\frac{18}{6}=3$. Прибавив к этой величине $A=7$, получим $\bar{x}=3+7=10$.

2. Если каждую варианту разделить или умножить на какое-либо одно и то же число A, то средняя арифметическая изменится во столько же раз.

Доказательство: $\bar{x}^*=\frac{\sum\left(\frac{x_i}{A}\right)f_i}{\sum f_i}=\frac{\sum x_i f_i}{A\sum f_i}=\frac{\bar{x}}{A}$. Это свойство позволяет вычислять среднюю $\bar{x}$ упрощенным способом, предварительно уменьшив каждую варианту в A раз, а затем умножив полученный результат на A, т. е. $\bar{x}=\bar{x}^*$.

Пример 3. Разделим каждую варианту данной совокупности на 2 и вычислим среднюю арифметическую:

$$\bar{x}^*=(1/6)(3{,}5+4{,}5+7{,}5+5{,}0+5{,}5+4{,}0)=30/6=5.$$

Умножив полученную величину на $A=2$, находим $\bar{x}=5\cdot 2=10$.

3. Сумма произведений отклонений вариант от их средней арифметической на соответствующие им частоты равна нулю[1].

Доказательство: $\sum[f_i(x_i-\bar{x})]=\sum f_i x_i-\bar{x}\sum f_i=\bar{x}n-\bar{x}n=0$.

Пример 4. Выше найдено, что средняя из совокупности шести вариант — 7, 9, 15, 10, 11 и 8 — равна 10. Определим сумму отклонений вариант от этой средней: $(7-10)+(9-10)+(15-10)+(10-10)+(11-10)+(8-10)=-6+6=0$.

4. Сумма квадратов отклонений вариант от их средней $\bar{x}$ меньше суммы квадратов отклонений тех же вариант от любой другой величины A, не равной $\bar{x}$, т. е.

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

Пример 5. Найдем сумму квадратов отклонений каждого члена данной совокупности от их средней $\bar{x}$, равной 10: $\Sigma(x_i-\bar{x})^2=(7-10)^2+(9-10)^2+(15-10)^2+(10-10)^2+(11-10)^2+(8-10)^2=9+1+25+1+4=40$. Теперь отыщем сумму квадратов отклонений тех же вариант от A, равного 8: $(7-8)^2+(9-8)^2+\ldots+(8-8)^2=64$.

Рассмотренные свойства средней арифметической позволяют преобразовывать многозначные и дробные числа, что облег-

[1] Под отклонением понимают разность между отдельными вариантами и их средней величиной, т. е. $(x_i-\bar{x})$.

чает работу по вычислению статистических характеристик (см. табл. 9).

Средняя гармоническая $\bar{x}_h$. Эту характеристику в отличие от средней арифметической, представляющей сумму вариант, отнесенную к их числу, определяют как *сумму обратных значений вариант, деленную на их число.* Для определения *простой* и *взвешенной* средней гармонической применяют формулы

$$\bar{x}_h = \frac{n}{\sum(1/x_i)} \quad (7а) \quad \text{и} \quad \bar{x}_h = \frac{n}{\sum[(1/x_i)\, f_i]}, \quad (7б)$$

в которых n — число произведенных наблюдений; x_i — значения вариант; f_i — частоты.

Чтобы уяснить суть средней гармонической, удобнее начать с рассмотрения соответствующих конкретных данных.

Пример 6. Пять доярок в течение 1 ч (60 мин) надоили следующее количество молока: первая — 10 л, вторая — 20, третья — 25, четвертая — 30 и пятая — 20 л; всего 105 л за 1 ч. Оценим эти итоги с помощью $\bar{x}$ и $\bar{x}_h$. Получим следующие результаты: $\bar{x} = (10+20+25+30+20):5 = 21$ л;

$$\bar{x}_h = \frac{5}{1/10 + 1/20 + 1/25 + 1/30 + 1/20} = \frac{5}{0{,}273} = 18{,}31 \text{ л}.$$

Разница между $\bar{x}$ и $\bar{x}_h$ весьма заметна. Какая же из этих средних верна? Возвращаясь к примеру, можно отметить, что, используя $\bar{x}$, можно легко определить общее количество надоенного пятью доярками молока: $21 \cdot 5 = 105$ л. Попробуем с помощью $\bar{x}$ вычислить время, затраченное в среднем одной дояркой на выдаивание 1 л молока. Получим результат: $60/21 = 2{,}86$ мин/л. Верно ли это? Проверим результат: первая доярка на выдаивание 1 л молока затратила $60/10 = 6$ мин, вторая — $60/20 = 3$, третья — 2,4, четвертая — 2, пятая — 3 мин. В среднем получается $(6+3+2{,}4+2+3)/5 = 16{,}4/5 = 3{,}38$ мин/л. Видно, что средняя арифметическая непригодна для определения среднего времени, затрачиваемого на выдаивание 1 л молока. Другой результат получается с применением средней гармонической: $60:18{,}31 = 3{,}28$ мин/л. Это точный результат.

Из приведенного примера видно, что средняя гармоническая применяется тогда, когда результаты наблюдений обнаруживают обратную зависимость, заданы обратными значениями вариант.

Средняя квадратическая $\bar{x}_q$. Для более точной числовой характеристики мер площади применяется средняя квадратическая. Этот показатель вычисляют по формулам

$$\bar{x}_q = \sqrt{\frac{\sum x_i^2}{n}} \quad (8а)$$

или при повторяемости отдельных вариант

$$\bar{x}_q = \sqrt{\frac{\sum f_i x_i^2}{n}}. \quad (8б)$$

Пример 7. Измеряли площадь корзинок у десяти наугад отобранных растений подсолнечника. Результаты измерений распределились следующим образом:

Площадь корзинок x_i, см²	50	95	130	175	200	220
Чнсло случаев f_i	1	1	2	3	2	1

Определим средний размер этого признака. Предварительно находим $\Sigma(f_i x_i^2) = 1 \cdot 50^2 + 1 \cdot 95^2 + 2 \cdot 130^2 + 3 \cdot 175^2 + 2 \cdot 200^2 + 1 \cdot 220^2 = 265\,600$. Отсюда $\bar{x}_q = 265\,600/10 = 163{,}0$ см². Средняя арифметическая в таких случаях оказывается меньше средней квадратической: $\bar{x} = (1 \cdot 50 + 1 \cdot 95 + 2 \cdot 130 + 3 \cdot 175 + 2 \cdot 200 + 1 \cdot 220)/10 = 1550/10 = 155$ см².

Средняя кубическая $\bar{x}_Q$. В качестве характеристики объемных признаков более точной является *средняя кубическая,* определяемая по формулам

$$\bar{x}_Q = \sqrt{\frac{\sum x_i^3}{n}} \quad (9а)$$

или при повторяемости отдельных вариант

$$\bar{x}_Q = \sqrt[3]{\frac{\sum (f_i x_i^3)}{n}}. \quad (9б)$$

Пример 8. Измеряли объем наугад отобранных 18 куриных яиц (учитывали полусумму большого и малого диаметра). Результаты измерений оказались следующими:

Днаметр янц x_i, см	4,7	4,8	5,0	5,4	5,6	6,0
Чнсло случаев f_i	2	4	6	3	2	1

Определим средний объем яиц по их диаметрам. Предварительно вычислим

$$\sum f_i x_i^3 = 2(4{,}7)^3 + 4(4{,}8)^3 + 6(5{,}0)^3 + 3(5{,}4)^3 + 2(5{,}6)^3 + 1(6{,}0)^3 = 2419{,}638.$$

Отсюда $\bar{x}_Q = \sqrt[3]{2419{,}638/18} = \sqrt[3]{134{,}42} = 5{,}12$.

Средняя геометрическая $\bar{x}_g$. Этот показатель представляет собой корень n-й степени из произведений членов ряда $\bar{x}_g = \sqrt[n]{x_1 x_2 x_3 \ldots x_n}$, где n — объем совокупности; при этом $x_i > 0$. Так, средняя геометрическая чисел 5, 8 и 25 равна

$$\bar{x}_q = \sqrt[3]{5 \cdot 8 \cdot 25} = \sqrt[3]{1000} = 10.$$

Обычно среднюю геометрическую вычисляют с помощью десятичных логарифмов по следующим рабочим формулам:

$$\lg \bar{x}_g = \frac{\sum \lg x_i}{n}; \quad (10)$$

$$\lg \bar{x}_g = \frac{\sum \lg (x_2/x_1)}{n}; \quad (11)$$

$$\lg \bar{x}_g = \frac{\lg x_к - \lg x_н}{n}. \quad (12)$$

формулу (10) применяют для вычисления средней геометрической из абсолютных прибавок величины признака; формула (11) служит для вычисления средней геометрической из относительных прибавок величины признака за равные промежутки времени, а формулу (12) используют для вычисления средней геометрической по разности между конечной $x_к$ и начальной $x_н$ прибавками величины признака.

Пример 9. По данным Дональдсона, масса тела лабораторных мышей изменялась с возрастом следующим образом (табл. 8).

Таблица 8

Возраст мышей, нед.	Масса тела, г	Прибавки массы тела за одну неделю, г		Логарифм прибавок массы тела	
		абсолютные	относительные	абсолютных	относительных
1	10	—	—	—	—
2	15	5	1,50	0,69897	0,17609
3	20	5	1,33	0,69897	0,12385
4	27	7	1,35	0,84510	0,13033
5	35	8	1,30	0,90309	0,11394
6	46	11	1,31	1,04139	0,11727
7	58	12	1,26	1,07918	0,10037
8	72	14	1,24	1,14613	0,09342
9	87	15	1,21	1,17609	0,08279
Сумма	—	77	10,50	7,58892	0,93806

Определим среднюю геометрическую из абсолютных недельных прибавок массы тела мышей за первые девять недель их жизни: $\lg \bar{x}_g = \frac{7,58892}{8} = 0,95861$, откуда $\bar{x}_g = 8,9$ г. Средняя геометрическая из абсолютных прибавок массы тела мышей оказывается меньше средней арифметической: $\bar{x} = 77/8 = 9,6$ г. Так как средняя геометрическая характеризует ряды динамики, обычно ее вычисляют не из абсолютных, а из относительных прибавок величины признака за определенные равные промежутки времени. Например, в данном случае $\lg \bar{x}_g = \frac{0,93806}{8} = 0,11726$, откуда $\bar{x}_g = 1,310$ г. Средняя арифметическая из относительных прибавок массы тела мышей $\bar{x} = \frac{10,50}{8} = 1,313$ г[1].

[1] Если собранные данные распределяются в интервальный вариационный ряд, среднюю геометрическую определяют по центральным значениям классовых интервалов, как и другие степенные средние.

Когда известны лишь два крайних значения изменяющегося во времени признака: начальная (базисная) и конечная величины, — среднюю геометрическую вычисляют по формуле (12). Так, применительно к рассматриваемому примеру начальная $x_н=10$ и конечная $x_к=87$, откуда среднюю геометрическую рассчитывают следующим образом:

$$\lg \bar{x}_g=(1/8)(\lg 87+\lg 10)=(1/8)(1{,}93952-1{,}000)=0{,}11744,$$

или $\bar{x}_g=1{,}310$ г — величина, которая была получена выше.

Пример 10. По данным М. А. Ольшанского (1950), в результате пятилетнего отбора (1936—1940) длина волокна у гибридного сорта хлопчатника увеличилась с 26,3 до 31,0 мм. Определим по формуле (12) среднегодовой эффект селекции этого признака:

$$\lg \bar{x}_g=(1/5)(\lg 31{,}0-\lg 26{,}3)=0{,}01428.$$

Отсюда $\bar{x}_g=1{,}0334$ мм, что составляет 3,9% от начальной (базисной) величины (26,3) этого признака.

Средняя геометрическая — более точная характеристика рядов динамики, чем средняя арифметическая. В этом можно убедиться, если последовательно перемножить средний прирост величины признака за учитываемый период времени начиная с базисной величины. Так, в рассмотренном примере 9 конечную величину признака ($x_к=87$ г) находят в результате следующего расчета: $10\cdot 1{,}310\cdot 1{,}310\cdot 1{,}310\cdot 1{,}310\cdot 1{,}310\cdot 1{,}310\cdot 1{,}310\cdot 1{,}310\times$ $\times 1{,}310=87$. Проверим этим способом точность средней арифметической ($\bar{x}=1{,}313$): $10\cdot 1{,}313\cdot 1{,}313\cdot 1{,}313\cdot 1{,}313\cdot 1{,}313\cdot 1{,}313\times$ $\times 1{,}313\cdot 1{,}313=88{,}3$. При сравнении первого результата со вторым видно, что средняя геометрическая более точно характеризует динамику явления, чем средняя арифметическая.

Однако средняя геометрическая, как правило, незначительно отличается по величине от средней арифметической. К тому же вычисление средней арифметической проще, чем средней геометрической. Поэтому вместо средней геометрической в качестве приближенной характеристики темпов динамики нередко используют среднюю арифметическую. При этом приходится учитывать и то, что средняя геометрическая дает хорошие (не искаженные) результаты лишь при наличии геометрической прогрессии, заложенной в самой динамике явления. Это обстоятельство несколько ограничивает область применения средней геометрической, которую вычисляют обычно в прогностических целях и при определении средних прибавок массы или размеров тела, возрастных изменений численного состава популяций за определенные (обычно равные) промежутки времени.

В заключение обзора степенных средних необходимо отметить, что между ними существуют определенные соотношения,

выражаемые следующим *рядом мажорантности* (неравенства): $\bar{x}_Q > \bar{x}_q > \bar{x} > \bar{x}_g > \bar{x}_h$.

II.2. ПОКАЗАТЕЛИ ВАРИАЦИИ

Средние величины не являются универсальными характеристиками варьирующих объектов. При одинаковых средних признаки могут отличаться по величине и характеру варьирования. Поэтому наряду со средними для характеристики варьирующих признаков используют и *показатели вариации*. Одним из таких показателей являются *лимиты* (от лат. limes — предел), обозначаемые символом lim. В биометрии под этим термином понимают значения минимальной x_{min} и максимальной x_{max} вариант совокупности.

Размах вариации R. Это показатель, представляющий собой разность между максимальной и минимальной вариантами совокупности, т. е. $R = x_{max} - x_{min}$. Чем сильнее варьирует признак, тем больше размах вариации, и, наоборот, чем слабее вариация признака, тем меньше будет размах вариации.

Лимиты и размах вариации — простые и наглядные характеристики варьирования, однако им присущи существенные недостатки: при повторных измерениях одного и того же группового объекта они могут значительно изменяться; кроме того, они не отражают существенные черты варьирования, что можно показать на следующем примере.

Возьмем два ряда распределения с одним и тем же весом входящих в их состав вариант, равным единице:

x_1	10	15	20	25	30	35	40	45	50	$\bar{x}_1 = 30$
x_2	10	28	28	30	30	30	32	32	50	$\bar{x}_2 = 30$

По числу вариант ($n = 9$), лимитам и размаху вариации эти ряды не отличаются друг от друга; их средние также равны между собой. Отличает их друг от друга характер варьирования, но эта особенность никак не отражается на лимитах и размахе вариации.

Более удобной характеристикой вариации мог бы служить показатель, который строится на основании отклонений вариант от их средней, т. е. $|x_i - \bar{x}| = d$. Сумма таких отклонений, взятая без учета знаков и отнесенная к числу наблюдений n, называется *средним линейным отклонением* $\bar{d} = \frac{\sum |x_i - \bar{x}|}{n}$. Так, если взять суммы отклонений вариант от их средней ($\bar{x} = 30$) для первого x_1 и второго x_2 приведенных здесь рядов, то получаются следующие результаты:

$d_1 = 20$	15	10	5	0	5	10	15	20	$\Sigma d_1 = 100$
$d_2 = 20$	2	2	0	0	0	2	2	20	$\Sigma d_2 = 48$

Отсюда $\bar{d}_1 = 100/9 = 11{,}1$ и $\bar{d}_2 = 48/9 = 5{,}3$. Таким образом, в первом случае варьирование сильнее, чем во втором.

Дисперсия и ее свойства s^2, σ^2. Несмотря на явное преимущество среднего линейного отклонения перед лимитами и размахом варнации, этот показатель не получил широкого применения в биометрии. Наиболее подходящим оказался показатель, построенный не на отклонениях вариант от их средних, а на квадратах этих отклонений, его называют *дисперсией* (от лат. dispersio — рассеяние [1]) и выражают формулами

$$s_x^2 = \frac{\sum_{l}^{k} (x_l - \bar{x})^2}{n} \quad \text{или} \quad s_x^2 = \frac{\sum_{l}^{k} f_l (x_l - \bar{x})^2}{n}, \qquad (13)$$

где $\sum_{l=1}^{k}$ — знак суммирования произведений отклонений варнаит x_i от их средней $\bar{x}$ на веса или частоты f_l этих отклонений в пределах от первого до k-го класса; n — общее число наблюденнй. Индекс x у символа дисперсин обозначает, что этот показатель характеризует варьирование числовых значений признака вокруг их средней величины.

Ценность дисперсии заключается в том, что, являясь мерой варьирования числовых значений признака вокруг их средней арифметической, она измеряет и внутреннюю изменчивость значений признака, зависящую от разностей между наблюдениями. Преимущество дисперсии перед другими показателями вариации состоит также и в том, что она разлагается на составные компоненты, позволяя тем самым оценивать влияние различных факторов на величину учитываемого признака.

Вместе с тем установлено, что рассчитываемая по формуле (13) дисперсия оказывается смещенной по отношению к своему генеральному параметру на величину, равную $n/(n-1)$. Чтобы получить несмещенную дисперсию, нужно в формулу (13) ввести в качестве множителя поправку на смещенность, называемую *поправкой Бесселя*. В результате формула (13) преобразу-

---

[1] Р. Фишер назвал эту характеристику «вариансой» (от англ. variance — изменение) вместо математического термина «дисперсия», который биометрики (Н. А. Плохинский, 1970; и др.) стали применять для обозначения суммы квадратов отклонений. Так возник разнобой в терминах математики и биометрии, что весьма нежелательно. Чтобы избежать указанного разнобоя в настоящем руководстве не делается различия между терминами «варианса» и «дисперсия», под которыми понимают средний квадрат отклонений, а для обозначения суммы квадратов отклонений вводится предложенный Дж. Юлом и М. Кендэлом (1960) термин «девиата» (от лат. deviatio — отклонение от чего-либо).

ется следующим образом:

$$s_x^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n} \cdot \frac{n}{n-1} = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n-1}. \qquad (14)$$

Разность n—1, обозначаемую в дальнейшем строчной буквой латинского алфавита k, называют *числом степеней свободы*, под которым понимают число свободно варьирующих единиц в составе численно ограниченной статистической совокупности.

Так, если совокупность состоит из n-го числа членов и характеризуется средней величиной $\bar{x}$, то любой член этой совокупности может иметь какое угодно значение, не изменяя при этом среднюю $\bar{x}$, кроме одной варианты, значение которой определяется разностью между суммой значений всех остальных вариант и величиной $n\bar{x}$. Следовательно, одна варианта численно ограниченной статистической совокупности не имеет свободы вариации. Отсюда число степеней свободы для такой совокупности будет равно ее объему n без единицы, т. е. $k=n-1$. А при наличии не одного, а нескольких ограничений свободы вариации число степеней свободы вариации будет равно $k=n-\nu$, где ν (греческая буква ню) обозначает число ограничений свободы вариации.

Дисперсия обладает рядом важных свойств, из которых необходимо отметить следующие.

1. Если каждую варианту совокупности уменьшить или увеличить на одно и то же постоянное число A, то дисперсия не изменится:

$$s_x^2 = \frac{1}{n-1} \sum [(x_i - A) - (\bar{x} - A)]^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2.$$

Пример 11. Рассмотрим вновь совокупность, состоящую из шести вариант: 7, 9, 15, 10, 11, 8. Характеристики этой совокупности $\bar{x}=10$ и $s_x^2=8$. Уменьшим каждую варианту на шесть единиц и вычислим дисперсию:

(x_i-6)	1	3	9	4	5	2	$\bar{x}^*=24/6=4$
$(x_i-\bar{x}^*)$	−3	−1	+5	0	+1	−2	
$(x_i-\bar{x}^*)^2$	9	1	25	0	1	4	$\frac{\Sigma (x_i-\bar{x}^*)^2}{n-1} = \frac{40}{5} = 8$

Отсюда следует, что дисперсию можно вычислить не только по значениям варьирующего признака, но и по их отклонениям от какой-либо постоянной величины A.

2. Если каждую варианту совокупности разделить или умножить на одно и то же постоянное число A, то дисперсия уменьшится или увеличится в A^2 раз.

Доказательство:

$$s_x^2 = \frac{1}{n-1}\sum\left(\frac{x_i}{A}-\frac{\bar{x}}{A}\right)^2 = \frac{1}{A^2(n-1)}\sum(x_i-\bar{x})^2.$$

А также $s_x^2 = \frac{1}{n-1}\sum(x_iA-\bar{x}A)^2 = \frac{1}{n-1}\sum(x_i-\bar{x})^2A^2.$

Пример 12. Проиллюстрируем это свойство на том же примере. Разделим каждую варианту на 2, т. е. 7/2=3,5; 9/2= = 4,5 и т. д., и вычислим дисперсию для совокупности уменьшенных вдвое вариант:

$x/2$	3,5	4,5	7,5	5,0	5,5	4,0	$\bar{x}^*=30/6=5$
$(x_i-\bar{x}^*)$	−1,5	−0,5	+2,5	0	+0,5	−1,0	
$(x_i-\bar{x}^*)^2$	2,25	0,25	6,25	0	0,25	1,0	$s_x^2=10/5=2$

Из этого свойства следует, что при наличии в совокупности многозначных вариант их можно сократить на какое-то постоянное число A и по результатам вычислить дисперсию. Затем полученную величину умножить на квадрат общего делителя A — что и даст искомую величину дисперсии. Так, в данном случае $s_x^2=2\cdot2^2=8$.

На основании математических свойств средней арифметической и дисперсии нетрудно составить сводку правил по преобразованию многозначных и дробных чисел, которую полезно использовать при обработке биометрических данных (табл. 9).

Таблица 9

Способы преобразования чисел	Какие поправки нужно внести в конечные результаты	
	при вычислении дисперсии	средней арифметической
$x-A$	Поправка не нужна	Прибавить число A
$(x-A)K$	Разделить на K^2	Разделить на K и прибавить число A
$(x-A)/K$	Умножить на K^2	Умножить на K и прибавить число A
x/K	Умножить на A^2	Умножить на A
xA	Разделить на A^2	Разделить на A

В этой таблице A — произвольно взятое число, обычно близкое к величине минимальной варианты x_{min}; K — произвольное число, позволяющее преобразовывать дробные числа; x — отдельные числовые значения признака, т. е. варианты.

Следует также иметь в виду, что вместо $\sum(x_i-\bar{x})^2$ можно использовать

$$\sum x_i^2-\frac{(\sum x_i)^2}{n};\quad n\left[\frac{\sum x_i^2}{n}-\frac{(\sum x_i)^2}{n}\right];\quad \sum x^2-n\bar{x}^2.$$

Отсюда можно вывести следующие рабочие формулы, удобные при вычислении дисперсии непосредственно по значениям варьирующего признака:

$$s_x^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]; \qquad (15)$$

$$s_x^2 = \frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right); \qquad (16)$$

$$s_x^2 = \frac{n}{n-1}\left[\frac{\sum x_i^2}{n} - \frac{(\sum x_i)^2}{n}\right], \qquad (17)$$

или при повторяемости отдельных вариант

$$s_x^2 = \frac{1}{n-1}\left[\sum f_i x_i^2 - \frac{(\sum f_i x_i)^2}{n}\right]; \qquad (15а)$$

$$s_x^2 = \frac{1}{n-1}\left(\sum f_i x_i^2 - n\bar{x}^2\right); \qquad (16а)$$

$$s_x^2 = \frac{1}{n-1}\left[\frac{\sum f_i x_i^2}{n} - \frac{(\sum f_i x_i)^2}{n}\right]. \qquad (17а)$$

Среднее квадратическое отклонение s_x. Наряду с дисперсией важнейшей характеристикой варьирования является среднее квадратическое отклонение — показатель, представляющий корень квадратный из дисперсии:

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}. \qquad (18)$$

Эта величина в ряде случаев оказывается более удобной характеристикой варьирования, чем дисперсия, так как выражается в тех же единицах, что и средняя арифметическая величина.

Дисперсия и среднее квадратическое отклонение наилучшим образом характеризуют не только величину, но и специфику варьирования признаков. Чтобы убедиться в этом, вернемся к рассмотренным выше рядам распределения, у которых одинаковый размах вариации и одинаковые средние показатели, но различный характер варьирования, и вычислим для них дисперсию и среднее квадратическое отклонение:

x_i	10	15	20	25	30	35	40	45	50	$\bar{x}_1 = 30$
$(x_i - \bar{x})$	−20	−15	−10	−5	0	+5	+10	+15	+20	
$(x_i - \bar{x})^2$	400	225	100	25	0	25	100	225	400	$\Sigma (x_i - \bar{x})^2 = 1500$

Отсюда $s_x^2 = 1500/(9-1) = 187{,}5$ и $s_x = \sqrt{187{,}5} = 13{,}7$.

x_2	10	28	28	30	30	30	32	32	50	$\bar{x}_2=30$
$(x_i-\bar{x})$	−20	−2	−2	0	0	0	+2	+2	+20	
$(x_i-\bar{x})^2$	400	4	4	0	0	0	4	4	400	$\Sigma(x_i-\bar{x})^2=816$

Отсюда $s_x^2=816/8=102$ и $s_x=\sqrt{102}=10{,}1$.

Из приведенных вычислений видно, что при одинаковых лимитах и размахе вариации дисперсия и среднее квадратическое отклонение оказались неодинаковыми: на величине этих показателей сказался различный характер варьирования признаков.

Поправка Шеппарда. При превращении интервального вариационного ряда в безынтервальный ряд частоты распределения относят к средним значениям классовых интервалов без учета внутриклассового разнообразия. Между тем варианты внутри классов распределяются неравномерно, накапливаясь больше у тех границ, которые ближе к средней арифметической ряда. Отсюда следует, что при вычислении обобщающих характеристик для непрерывно варьирующих признаков допускают систематическую погрешность, величина которой зависит от ширины классового интервала: чем шире интервал, тем больше и погрешность. На величине средней арифметической погрешность отражается слабо, тогда как на величине дисперсии она сказывается более сильно. Учитывая это обстоятельство, В. Шеппард (1898) установил, что разность между расчетной и фактической величиной дисперсии составляет $^1/_{12}$ квадрата классового интервала. Следовательно, при вычислении дисперсии по формуле (13) следует вносить поправку Шеппарда, т. е. вычитать эту величиию $\left(\frac{1}{12}\lambda^2\right)$ из s_x^2. Так, если взять распределение кальция (мг%) в сыворотке крови обезьян, для которого известны $s_x^2=1{,}60$ и $s_x=1{,}26$, то, учитывая ширину классового интервала $\lambda=0{,}8$ мг% и внося поправку Шеппарда, получим $s_x^2=1{,}60-0{,}8^2/12=1{,}55$ и $s_x=\sqrt{1{,}55}=1{,}24$.

Поправка Шеппарда вносится далеко не всегда. Ее обычно применяют или при высокой точности расчетов, или при наличии большого числа наблюдений ($n\geqslant 500$), распределяемых в интервальный вариационный ряд. Для получения обобщающих числовых характеристик небольших и средних по объему ($n<500$) совокупностей поправку Шеппарда не вносят.

Коэффициент вариации V, Cv. Дисперсия и среднее квадратическое отклонение применимы и для сравнительной оценки одноименных средних величин. В практике же довольно часто приходится сравнивать изменчивость признаков, выраженных разными единицами. В таких случаях используют не абсолютные, а относительные показатели вариации. Дисперсия и среднее квадратическое отклонение как величины, выражаемые теми же единицами, что и характеризуемый ими признак, для оценки

изменчивости разноименных величин непригодны. Одним из относительных показателей вариации является коэффициент вариации. Этот показатель представляет собой среднее квадратическое отклонение, выраженное в процентах от величины средней арифметической:

$$Cv = \frac{s_x}{\bar{x}} 100\%. \qquad (19)$$

Пример 13. Сравнивают два варьирующих признака. Один характеризуется средней $\bar{x}_1 = 2{,}4$ кг и средним квадратическим отклонением $s_1 = 0{,}58$ кг, другой — величинами $\bar{x}_2 = 8{,}3$ см и $s_2 = 1{,}57$ см. Следует ли отсюда, что второй признак варьирует сильнее, чем первый? Нет, не следует, так как среднее квадратическое отклонение определяют по отклонениям от средних, а они различны по величине. Кроме того, не вполне корректно сравнивать величины, выраженные разными единицами меры. Именно поэтому в подобных случаях уместно использовать безразмерные значения коэффициентов вариации. Сравнивая их в приводимом примере, находим, что сильнее варьирует не второй, а первый признак:

$$Cv_1 = 100\,(0{,}58/2{,}4) = 24{,}2\ \%\ \text{ и }\ Cv_2 = 100\,(1{,}57/8{,}3) = 18{,}9\ \%.$$

Различные признаки характеризуются различными коэффициентами вариации. Но в отношении одного и того же признака значение этого показателя Cv остается более или менее устойчивым и при симметричных распределениях обычно не превышает 50%. При сильно асимметричных рядах распределения коэффициент вариации может достигать 100% и даже выше. Варьирование считается слабым, если не превосходит 10%, средним, когда Cv составляет 11—25%, и значительным при $Cv > 25\%$.

Применяя коэффициент вариации в качестве характеристики варьирования, следует учитывать единицы размерности изучаемого признака: линейные или весовые (объемные). Акад. И. И. Шмальгаузен (1936) отмечал, что в таких случаях коэффициент вариации оказывается неодинаковым. Иллюстрацией тому могут служить данные Ю. Г. Артемьева (1939), исследовавшего варьирование величины внутренних органов у малых сусликов в зависимости от того, какими единицами меры выражены признаки (табл. 10).

Данные, приведенные в табл. 10, показывают, что при линейном выражении величины признака коэффициент вариации оказывается примерно в три раза меньше, чем при кубическом выражении того же признака. Причина такого явления — в математических свойствах Cv, которые надо учитывать, чтобы избежать возможных ошибок.

Таблица 1(

Органы	Коэффициент вариации при разном выражении признаков	
	линейном	кубическом
Сердце	3,4	10,2
Легкие	9,6	29,5
Селезенка	9,8	29,8
Почка	3,1	9,4
Печень	3,0	9,3

Нормированное отклонение *t*. Отклонение той или иной варианты от средней арифметической, отнесенное к величине среднего квадратического отклонения, называют *нормированным отклонением*:

$$t = \frac{(x_i - \bar{x})}{s_x}. \qquad (20$$

Этот показатель позволяет «измерять» отклонения отдельных вариант от среднего уровня и сравнивать их для разных признаков.

Пример 14. При обследовании группы подростков в возрасте от 15 до 16 лет установлено, что средний рост юношей характеризуется следующими показателями: $\bar{x} = 164{,}8$ см и $s_x = 5{,}8$ см. В группе оказался юноша, рост которого равен 172,4 см. Спрашивается: как велико отклонение роста этого юноши от средней величины данного признака в этой группе? Нормируя рост юноши ($\bar{x} = 172{,}4$), находим $t = \frac{172{,}4 - 164{,}8}{5{,}8} = +1{,}31$.

Получая значения нормированных отклонений для разных признаков, можно сравнить места, занимаемые особью, индивидом и т. п. по каждому из этих признаков в их распределениях. Пусть, например, нормированное отклонение у рассматриваемого юноши по ширине плеч равно — 0,41. Тогда можно утверждать что у него длина тела отклоняется от средней в сторону больших величин этого признака, а ширина плеч — в сторону малых, т. е. характерен относительно узкоплечий тип телосложения.

Нормированное отклонение используют также при работе с так называемым *нормальным распределением.*

II.3. СПОСОБЫ ВЫЧИСЛЕНИЯ СТЕПЕННЫХ СРЕДНИХ И ПОКАЗАТЕЛЕЙ ВАРИАЦИИ

Моменты распределения. Средние величины и показатели вариации вычисляются как на группированном, так и негруппированном в вариационные ряды исходном материале. Известны три основных способа вычисления обобщающих числовых характеристик: способ произведений, способ условной средней и способ сумм или кумулят. Каждый способ имеет свои конструктивные особенности, но любой из них приводит к одному и тому же конечному результату.

Чтобы раскрыть сущность каждого способа и облегчить понимание конструктивных особенностей других обобщающих показателей, с которыми придется встречаться в дальнейшем, необходимо познакомиться с понятием статистических моментов, или моментов распределения.

Моментами распределения называют суммы отклонений вариант x_i от какого-либо числа A, возведенные в k-ю степень и отнесенные к общему числу вариант n, составляющих данную совокупность. Иными словами, это величины, которые можно выразить в виде следующей общей формулы:

$$M = \frac{1}{n} \sum (x_i - A)^k.$$

Если отклонения вариант вычисляют по отношению к нулевой точке, то моменты называют *начальными* (m); если от средней арифметической, то моменты называют *центральными* и обозначают греческой буквой μ (ми). Если же отклонения вариант находят от произвольно выбранного числа A, моменты называют *условными* (b). В зависимости от степени отклонений статистические моменты подразделяют на моменты первого, второго и больших порядков. В области биометрии используют обычно моменты первого, второго, третьего и четвертого порядков. Формулы для их вычисления приведены в табл. 11.

Центральные моменты ряда распределения связаны с условными моментами следующим образом:

$$\mu_1 = b_1 - b_1 = 0;$$

$$\mu_2 = b_2 - b_1^2 = s^2;$$

$$\mu_3 = b_3 - 3b_1b_2 + 2b_1^3;$$

$$\mu_4 = b_4 - 4b_2b_3 + 6b_1^2b_2 - 3b_1^4.$$

Эти формулы используют при вычислении обобщающих характеристик вариационного ряда. При замене в этих формулах обозначений условных моментов b_i на начальные m_i можно

Таблица 1

Моменты распределении	Начальные	Центральные	Условные
Первого порядка	$m_1 = \frac{\sum f(x-0)}{n} = \frac{\sum fx}{n}$	$\mu_1 = \frac{\sum f(x-\bar{x})}{n}$	$b_1 = \frac{\sum f(x-A)}{n}$
Второго порядка	$m_2 = \frac{\sum f(x-0)^2}{n} = \frac{\sum fx^2}{n}$	$\mu_2 = \frac{\sum f(x-\bar{x})^2}{n}$	$b_2 = \frac{\sum f(x-A)^2}{n}$
Третьего порядка	$m_3 = \frac{\sum f(x-0)^3}{n} = \frac{\sum fx^3}{n}$	$\mu_3 = \frac{\sum f(x-\bar{x})^3}{n}$	$b_3 = \frac{\sum f(x-A)^3}{n}$
Четвертого порядка	$m_4 = \frac{\sum f(x-0)^4}{n} = \frac{\sum fx^4}{n}$	$\mu_4 = \frac{\sum f(x-\bar{x})^4}{n}$	$b_4 = \frac{\sum f(x-A)^4}{n}$

получить выражения для вычисления центральных моментов

Способ произведений. Основу этого способа составляет нахождение отклонений вариант от средней величины, характеризующей данную статистическую совокупность. Каждое отклонение возводится в степень, затем эти степени отклонений суммируются. В простейшем виде эта операция записывается так:

$$D = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum x)^2}{n}.$$

В совокупности варианты обычно повторяются, поэтому суммы отклонений вариант от средней должны умножаться на их веса или частоты (отсюда и название способа), т. е. рассчитываться по взвешенным суммам квадратов отклонений (девиат)

$$D = \sum_{i=1}^{k} (x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i x_i^2 - \frac{(\sum f_i x)^2}{n}.$$

Пример 15. По данным Г. Е. Бодренкова (1963), длина тела у личинок щелкуна, отобранных случайным способом в посеве озимой ржи и измеренных в миллиметрах, варьировала следующим образом: 7, 10, 14, 12, 15, 16, 12. Средняя длина личинок $\bar{x}=(1/7)(7+10+14+\ldots+12)=86/7=12{,}3$ мм. Чтобы определить показатели вариации, предварительно находим:

x_i	7	10	14	12	15	16	12	$\Sigma x_i=86$
x_i^2	49	100	196	144	225	256	144	$\Sigma x_i^2=1114$

Отсюда значение девиаты $D=1114-86^2/7=57{,}43$. Показатели вариации $s_x^2=\frac{D}{n-1}=\frac{57{,}43}{6}=9{,}57$; $s_x=9{,}57=3{,}094$ и коэффициент варнации $C_v=100\frac{3{,}094}{12{,}3}=24{,}2\,\%$.

Пример 16. Определить среднее число поросят в пометах 64 свиноматок (см. гл. I) и вычислить показатели вариации для этого распределения. Предварительно рассчитаем вспомогательные величины Σx_i^2, $\Sigma f_i x_i$ и $\Sigma f_i x_i^2$. Расчет проводится в табл. 12.

Таблица 12

Классы x_i	Частоты f_i	x_i^2	$f_i x_i$	$f_i x_i^2$
5	4	25	20	100
6	7	36	42	252
7	13	49	91	637
8	15	64	120	960
9	7	81	63	567
10	9	100	90	900
11	6	121	66	726
12	3	144	36	432
Сумма	64	—	528	4574

Подставляя найденные велнчины в формулы, имеем $\bar{x}=528/64=8{,}25$ поросят; $s_x^2=\frac{1}{63}\left(4574-\frac{528^2}{64}\right)=\frac{218}{63}=3{,}46$; $s_x=1{,}85$ и $Cv=100\frac{1{,}86}{8{,}25}=22{,}5\,\%$.

Пример 17. Годовой удой 80 коров, содержащихся на ферме, распределился следующим образом:

Удой, кг	2500—2600	2600—2700	2700—2800	2800—2900	2900—3000	3000—3100	3100—3200	3200—3300
Число коров	2	5	13	20	16	17	4	3

Вычислить характеристики для этого ряда. Предварительно превратим интервальный ряд в безынтервальный: 2550 2650 2750 2850 и т. д. Чтобы облегчить вычислительную работу, уменьшим каждую классовую варианту на $A=2500$ и разделим полученную величину на $K=10$. По преобразованным значениям классов: $(2550-2500):10=5$; $(2650-2500):10=15$; $(2750-2500):10=25$ и т. д. — рассчитываем вспомогательные величины (табл. 13).

Подставляя известные величины в формулы с учетом тех поправок, которые внесены в данном случае (см. табл. 9), находим: $\bar{x}=\frac{\Sigma f_i x_i}{n}K+A=\frac{3250}{80}\,10+2500=2906{,}25$ кг;

$$s_x^2=\frac{1}{n-1}\left[\sum f_i x_i^2-\frac{(\Sigma f_i x_i)^3}{n}\right]K^2=\frac{1}{79}\left(\frac{151\,500-3250^2}{80}\right)10^2=$$

$$=(151\,400-132\,031{,}25)/79=1\,936\,875/79=24517{,}40;\quad s_x=$$

$$=\sqrt{24517{,}40}=156{,}58\text{ и }C_v=100\,\frac{156{,}58}{2906{,}25}=5{,}4\ \%.$$

Таблица 13

Преобразованные значения классов x_i'	Частоты f_i	$f_i x_i'$	$x_i'^2$	$f_i x_i'^2$
5	2	10	25	50
15	5	75	225	1125
25	13	325	625	8125
35	20	700	1225	24500
45	16	720	2025	32400
55	17	935	3025	51425
65	4	260	4225	16900
75	3	225	5625	16875
Сумма	80	3250	—	151400

Способ условной средней. Нельзя не заметить, что вычисление статистических характеристик способом произведений, особенно при наличии многозначных чисел, представляет собой трудоемкий процесс. Гораздо легче рассчитать статистические характеристики упрощенным *способом условной средней,* называемым также *способом условного нуля.* Суть этого способа заключается в следующем.

Одну из вариант условно принимают за среднюю величину, обозначив ее через A. Обычно в качестве условной средней, или

нулевой точки отсчета, берут варианту или класс с наибольшей частотой, хотя это и не обязательно: в качестве условной средней можно принять любую варианту (при наличии негруппированных данных) или любой класс вариационного ряда. Наметив величину A, остается найти поправку, которую необходимо прибавить или вычесть (смотря по знаку) от условной средней, чтобы получить значение средней арифметической $\bar{x}$. Такой поправкой служит условный момент первого порядка $b_1 = \frac{\Sigma f_i (x_i - A)}{n}$ (см. табл. 11). Обозначив отклонения вариант от условной средней через a, получим $b_1 = \frac{\Sigma f_i a}{n}$. Отсюда формула для определения средней арифметической

$$\bar{x} = A + \frac{\Sigma f_i a}{n}. \tag{21}$$

Дисперсия, определяемая этим способом, равна разности между условным моментом второго и квадратом условного момента первого порядка, умноженной на величину $n/(n-1)$, которая называется *поправкой Бесселя*:

$$s_x^2 = \frac{n}{n-1}\left(b_2 - b_1^2\right). \tag{22}$$

В развернутом виде эта формула выглядит так:

$$s_x^2 = \frac{1}{n-1}\left[\Sigma f_i a^2 - \frac{(\Sigma f_i a)^2}{n}\right] \text{ или} \tag{23}$$

$$s_x^2 = \frac{n}{n-1}\left[\frac{\Sigma f_i a^2}{n} - \left(\frac{\Sigma f_i a}{n}\right)^2\right]. \tag{24}$$

Ниже приведены конкретные примеры применения этих формул.

Пример 18. Вычислить способом условной средней основные характеристики ряда распределения кальция (мг%) в сыворотке крови обезьян. Предварительно рассчитываем вспомогательные величины $\Sigma f_i a$ и $\Sigma f_i a^2$ (табл. 14).

В качестве условной средней A взята классовая варианта, равная 11,4. От этой величины находим отклонения классов: 9,0—11,4 = —2,4; 9,8—11,4 = —1,6 и т. д. (см. третью графу табл. 14). Подставляем найденные величины в формулы:

$$\bar{x} = 11{,}4 + 53{,}6/100 = 11{,}936 \approx 11{,}94 \text{ мг\%};$$

$$s_x^2 = \frac{100}{99}\left[\frac{187{,}52}{100} - \left(\frac{53{,}6}{100}\right)^2\right] = \frac{100}{99}(1{,}8752 - 0{,}2873) = 1{,}60.$$

Таблица 1

Классы x_i	Частоты f_i	$a=(x_i-A)$	f_ia	f_ia^2
1	2	3	4	5
9,0	2	—2,4	—4,8	11,52
9,8	6	—1,6	—9,6	15,36
10,6	15	—0,8	—12,0	9,60
11,4	23	0	0	0
12,2	25	+0,8	+20,0	16,00
13,0	17	+1,6	+27,2	43,52
13,8	7	+2,4	+16,8	40,32
14,6	5	+3,2	+16,0	51,20
Сумма	100	—	+53,6	187,52

Отсюда $s_x=\sqrt{1,60}=1,26$ и $C_v=100\,\frac{1,26}{11,94}\,10,6\,\%$.

Вычисление вспомогательных величин можно значительно упростить, если отклонения классовых вариант от условной средней A относить к величине классового интервала, т. е. вместо $a=(x_i-A)$ брать $a=(x_i-A)/\lambda$. Тогда во всех без исключения случаях (для равноинтервальных рядов) отклонения классовых вариант от условной средней A, где $a=0$, превращаются в ряд натуральных чисел 1, 2, 3, 4 и т. д., которые, как и в предыдущем случае, рассматриваются в сторону меньших, чем A значений вариант с отрицательным, а в сторону бо́льших, чем A, значений — с положительным знаком. При этом в формулы (21), (23) и (24) вносятся поправки на величину классового интервала:

$$\bar{x}=A+\lambda\,\frac{\sum f_ia}{n}; \qquad (25$$

$$s_x^2=\frac{\lambda^2}{n-1}\left[\sum f_ia^2-\frac{\sum f_ia}{n}\right] \text{ или} \qquad (26$$

$$s_x^2=\frac{\lambda^2 n}{n-1}\left[\frac{\sum f_ia^2}{n}-\left(\frac{\sum f_ia}{n}\right)^2\right]. \qquad (27$$

Применив эти формулы к рассматриваемому примеру, рассчитываем Σf_ia и Σf_ia^2 (табл. 15).

Подставляя известные величины в формулы (25) и (26 находим:

$$\bar{x}=11,4+0,8\,\frac{67}{100}=11,4+0,536=11,936\approx 11,94 \text{ мг\%};$$

$$s_x^2=\frac{0,8^2}{99}\left(293-\frac{67^2}{100}\right)=\frac{0,64}{99}\,(293-44,89)=\frac{158,79}{99}=1,60.$$

Таблица 15

Классы x_i	Частоты f_i	a	$f_i a$	$f_i a^2$	$f_i a^3$	$f_i a^4$
9,0	2	−3	−6	18	−54	162
9,8	6	−2	−12	24	−48	96
10,6	15	−1	−15	15	−15	15
11,4	23	0	0	0	0	0
12,2	25	+1	+25	25	+25	25
13,0	17	+2	+34	68	+136	272
13,8	7	+3	+21	63	+189	567
14,6	5	+4	+20	80	+320	1280
Сумма	100	—	+67	293	+553	2417

Две последние графы понадобятся в дальнейшем (см. разд. III.10).

К такому же результату приводит и формула (27):

$$s_x^2 = \frac{0{,}8^2 \cdot 100}{100-1}\left[\frac{293}{100} - \left(\frac{67}{100}\right)^2\right] = \frac{64}{99}(2{,}93 - 0{,}4489) = 1{,}60.$$

Пример 19. Обработаем этим способом данные об удоях 80 коров (см. табл. 13). Предварительно рассчитаем вспомогательные величины (табл. 16).

Таблица 16

Преобразованные значения классов x_i^*	Частоты f_i	Отклонения a	$f_i a$	$f_i a^2$
5	2	−3	−6	18
15	5	−2	−10	20
25	13	−1	−13	13
35	20	0	0	0
45	16	+1	+16	16
55	17	+2	+34	68
65	4	+3	+12	36
75	3	+4	+12	48
Сумма	80	—	+45	219

Подставляем найденные величины в формулы (с учетом необходимых поправок в связи с преобразованием многозначных чисел):

$$\bar{x} = \left(A - \lambda \frac{\sum f_i a}{n}\right) K + A = \left(35 + 10\,\frac{45}{80}\right) 10 + 2500 =$$
$$= 406{,}25 + 2500 = 2906{,}25 \text{ кг};$$

$$s_x^2 = \frac{\lambda^2}{n-1}\left[\sum f_i a^2 - \frac{(\sum f_i a)^2}{n}\right] K^2 = \frac{10^2}{79}\left(219 - \frac{45^2}{80}\right) 10^2 =$$

$$= \frac{100}{79}(219 - 25{,}3125)\,100 = \frac{1\,936\,875}{79} = 24517{,}4 \text{ и}$$

$$s_x = 24517{,}4 = 156{,}58.$$

Получился такой же результат, что и выше, а расчет средней $\bar{x}$ и дисперсии s_x^2 оказался проще.

II.4. СТРУКТУРНЫЕ СРЕДНИЕ И СПОСОБЫ ИХ ВЫЧИСЛЕНИЯ

Медиана (*Me*). Средняя арифметическа — одна из основных характеристик варьирующих объектов по тому или иному признаку. Однако она не лишена недостатков, так как очень чувствительна к увеличению числа наблюдений или к уменьшению за счет вариант, резко отличающихся по своей величине от основной массы. Поэтому на величину средней арифметической могут значительно влиять крайние члены ранжированного вариационного ряда, которые как раз и наименее характерны для данной совокупности. В связи с этим во многих случаях в качестве обобщающих характеристик совокупности более полезными могут оказаться так называемые *структурные средние*. Эти величины обычно представляют собой конкретные варианты имеющейся совокупности, которые занимают особое место в ряду распределения.

Одной из таких характеристик является *медиана* — средняя, относительно которой ряд распределения делится на две равные части: в обе стороны от медианы располагается одинаковое число вариант. При наличии небольшого числа вариант медиана определяется довольно просто. Для этого собранные данные ранжируют, и при нечетном числе членов ряда центральная варианта и будет его медианой. При четном числе членов ряда медиана определяется по полусумме двух соседних вариант, расположенных в центре ранжированного ряда. Например, для ранжированных значений признака — 12 14 16 18 20 22 24 26 28 — медианой будет центральная варианта, т. е. $Me = 20$, так как в обе стороны от нее отстоит по четыре варианты. Для ряда с четным числом членов — 6 8 10 12 14 16 18 20 22 24 — медианой будет полусумма его центральных членов, т. е. $Me = (14+16)/2 = 15$.

Для данных, сгруппированных в вариационный ряд, медиана определяется следующим образом. Сначала находят класс, в котором содержится медиана. Для этого частоты ряда кумулируют в направлении от меньших к большим значениям классов до величины, превосходящей половину всех членов данной

совокупности, т. е. $n/2$. Первая величина в ряду накопленных частот Σf_i, которая превышает $n/2$, соответствует медианному классу. Затем берут разность между $n/2$ и суммой накопленных частот Σf_i, предшествующей медианному классу, которая относится к частоте медианного класса f_{Me}; результат умножают на величину классового интервала λ. Найденную таким способом величину прибавляют к нижней границе $x_н$ медианного класса. Если же исходные данные распределены в безынтервальный вариационный ряд, названную величину прибавляют к полусумме соседних классовых вариант. В результате получается искомая величина медианы. Описанные действия выражаются в виде следующей формулы:

$$Me = x_н + \lambda \left(\frac{\frac{n}{2} - \Sigma f_l}{f_{Me}} \right), \qquad (28)$$

где $x_н$ — нижняя граница классового интервала, содержащего медиану, или полусумма соседних классов безынтервального ряда, в промежутке между которыми находится медиана; Σf_i — сумма накопленных частот, стоящая перед медианным классом; f_{Me} — частота медианного класса; λ — величина классового интервала; n — общее число наблюдений.

Пример 20. Вернемся к ряду распределения кальция (мг%) в сыворотке крови обезьян и вычислим для него медиану (табл. 17).

Таблица 17

Классы по содержанию кальция в сыворотке крови	Срединные значения классов	Частоты f_i	Накопленные частоты Σf_i
8,6—9,3	9,0	2	2
9,4—10,1	9,8	6	8
10,2—10,9	10,6	15	23
11,0—11,7	11,4	23	46
11,8—12,5	12,2	25	71
12,6—13,3	13,0	17	
13,4—14,1	13,8	7	
14,2—14,9	14,6	5	
Сумма	—	100	—

В данном случае $n/2 = 100/2 = 50$. Эта величина больше $\Sigma f_i = 46$, но меньше $\Sigma f_i = 71$. По $\Sigma f_i = 71$ определяем интервал, в котором находится медиана. Границы этого интервала: нижняя $x_н = 11{,}8$ и верхняя $x_в = 12{,}5$; его частота $f_{Me} = 25$. Классо-

вый интервал $\lambda=0{,}8$. Подставляя эти величины в формулу (28), находим

$$Me=11{,}8+0{,}8\left(\frac{50-46}{25}\right)=11{,}8+0{,}128\approx 11{,}93.$$

Теперь превратим интервальный вариационный ряд в безынтегральный и вычислим медиану по срединным значениям классовых интервалов. Медиана находится между значениями классов 11,4 и 12,2. Отсюда

$$Me=\frac{11{,}4+12{,}2}{2}+0{,}8\left(\frac{50-46}{25}\right)=11{,}8+0{,}128=11{,}93.$$

Пример 21. Определить медиану для ряда распределения количества поросят в пометах 64 свиноматок (табл. 18).

Таблица 1

Количество поросят в помете x_i	Число случаев f_i	Накопленные частоты Σf_i
5	4	4
6	7	11
7	13	24
8	15	39
9	7	
10	9	
11	6	
12	3	
Сумма	64	—

Здесь $n/2=64/2=32$. Эта величина превосходит $\Sigma f_i=24$ но меньше $\Sigma f_i=39$. Следовательно, медиана находится между 7-м и 8-м значениями классов и $x_н=(7+8)/2=7{,}5$. Частота медианного класса $f_{Me}=15$. Отсюда $Me=7{,}5+\frac{32-24}{15}=7{,}5+$ $+0{,}53=8{,}03$. Медиана этого распределения очень близка к средней арифметической $\bar{x}=8{,}25$ поросят.

Мода *(Mo)*. *Модой* называется величина, наиболее часто встречающаяся в данной совокупности. Класс с наибольшей частотой называется *модальным*. Он определяется довольно просто в безынтервальных рядах. Например, мода распределения численности поросят в пометах 64 свиноматок равна 8. Для определения моды интервальных рядов служит формула

$$Mo=x_н+\lambda\left(\frac{f_2-f_1}{2f_2-f_1+f_3}\right), \qquad (29$$

где $x_н$ — нижняя граница модального класса, т. е. класса с наибольшей частотой f_2; f_1 — частота класса, предшествующего

модальному; f_3 — частота класса, следующего за модальным; λ — ширина классового интервала.

Пример 22. Определить моду ряда распределения кальция (мг%) в сыворотке крови обезьян. Необходимые данные содержатся в табл. 17. Частота модального класса $f_2=25$, его нижняя граница $x_н=11{,}8$. Частота класса, предшествующего модальному, $f_1=23$, частота класса, следующего за модальным, $f_3=17$; $\lambda=0{,}8$. Подставляя эти данные в формулу (29), находим

$$Mo=11{,}8+0{,}8\left(\frac{25-23}{2\cdot 25-23-17}\right)=11{,}8+0{,}16=11{,}96.$$

Квантили. Наряду с медианой и модой к структурным характеристикам вариационного ряда относятся так называемые *квантили,* отсекающие в пределах ряда определеиную часть его членов. К ним относятся квартили, децили и перцентили (процентили). *Квартили* — это три значения признака (Q_1, Q_2, Q_3), делящие ранжированный вариационный ряд на четыре равные части. Аналогично, девять *децилей* делят ряд на 10 равных частей, а 99 *перцентилей* — на 100 равных частей.

В практике используют обычно перцентили P_3, P_{10}, P_{25}, P_{50}, P_{75}, P_{90} и P_{97}. Причем P_{25} и P_{75} соответствуют первому и третьему квартилям, между которыми находится 50% всех членов ряда, а P_{50} соответствует второму квартилю и равен медиане, т. е. $P_{50}=Me$. Любой перцентиль определяется рядом последовательных действий, которые можно выразить в виде следующей формулы:

$$P_j=x_н+\lambda\left(\frac{K-\Sigma f_l}{f_P}\right), \qquad (30)$$

где $x_н$ — нижняя граница класса, содержащего перцентиль P_j; она определяется по величине $K=L_jn/100$, превосходящей или равной Σf_i в ряду накопленных частот. Здесь P_j — выбранный перцентиль; n — общее число наблюдений; λ — ширина классового интервала; f_P — частота класса, содержащего искомый перцентиль; L_j — так называемый порядок перцентиля, показывающий, какой процент наблюдений имеет меньшую величину, чем P_j. Например, для P_{25} и P_{75} порядки окажутся соответственно равными 25 и 75%. Таким образом, как и при определении медианы, нахождение того или иного перцентиля связано с кумуляцией частот вариационного ряда в направлении от низшего (начального) класса к высшему.

Пример 23. Найти 50-й перцентиль ряда распределения годового удоя коров ($n=80$), для которого определены средняя арифметическая и показатели вариации (см. пример 18, табл. 13).

Удой, кг	2500—2600	2700—2800	2900—3000	3100—3200	330(
Частоты (f_i)	2	5	13	20	16	17	4	3
Σf_i	2	7	20	40	56	73	77	80

Величина $K=\frac{50\cdot 80}{100}=40$, она соответствует $\Sigma f_i=40$. Нижняя граница класса $x_н$, содержащего P_{50}, равна 2800; частота этого класса f_P равна 20; $\lambda=100$. Подставляем эти величины в формулу (30): $P_{50}=2800+100\frac{40-20}{20}=2900$. Найденная величина оказалась очень близкой к средней арифметической годового удоя коров этой группы $\bar{x}=2906{,}25$ кг.

Формула (30) применима и для нахождения перцентилей безынтервальных вариационных рядов.

Пример 24. Найдите 50-й перцентиль для ряда распределения численности поросят в пометах 64 свиноматок:

x_i	5	6	7	8	9	10	11	12
f_i	4	7	13	15	7	9	6	3
Σf_i	4	11	24	39	46	55	61	64

В данном случае $K=50\cdot 64/100=32$. Эта величина больше $\Sigma f_i=24$, но меньше $\Sigma f_i=39$. Следовательно, $x_н$ находится между 7-м и 8-м значениями классов, т. е. $x_н=(7+8)/2=7{,}5$; $f_P=15$. Отсюда $P_{50}=7{,}5+\frac{32-24}{15}=8{,}03$.

II.5. СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ ПРИ АЛЬТЕРНАТИВНОЙ ГРУППИРОВКЕ ВАРИАНТ

В данной главе уже упоминалось об альтернативной группировке исходных данных, когда одна группа вариант противопоставляется другой. Так, число мужских особей в популяции может противопоставляться числу женских особей, группа здоровых индивидов — группе больных индивидов и т. д. При альтернативной группировке данных их статистическими характеристиками будут как абсолютные, так и относительные численности противопоставляемых друг другу групп (альтернатив). Если абсолютную численность вариант, обладающих данным признаком, обозначить через m, то численность вариант противоположной группы, не имеющих данного признака, будет равна $n-m$, где n — общее число членов рассматриваемой совокупности. Например, среди $n=408$ новорожденных оказалось $m=208$ мальчиков. Тогда число новорожденных девочек $n-m=408-208=200$.

Численность альтернатив можно выразить в долях единицы, а также в процентах от их общего числа n. Обозначив долю вариант, обладающих учитываемым признаком, через p, получим $p=m/n$. Тогда доля вариант, не обладающих этим

признаком, обозначаемая буквой q, выразится отношением $q=(n-m)/n=1-(m/n)=1-p$. Для того чтобы численность противопоставляемых групп была выражена в процентах, достаточно каждую долю умножить на 100:

$$p=(m/n)\,100; \tag{31}$$

$$q=\frac{n-m}{n}\,100\,\%=100-p. \tag{32}$$

Очевидно, $p+q=1$ и $p+q=100\,\%$.

Относительные частоты или доли вариант при альтернативной группировке выполняют такую же роль, как средние величины для рядовой изменчивости признаков, когда исходные данные распределяются в вариационный ряд.

В качестве *характеристики альтернативного варьирования* служит среднее квадратическое отклонение s_p, которое определяют по формуле

$$s_p=\sqrt{p(1-p)}=\sqrt{pq}. \tag{33}$$

Этот показатель одинаково характеризует варьирование обеих альтернативных групп. Если же численность групп (альтернатив) выражена в процентах от их общего числа n, формула (33) принимает следующее выражение:

$$s_p=\sqrt{p(100-p)}. \tag{34}$$

Когда альтернативы выражены абсолютными числами, среднее квадратическое отклонение определяют по формуле

$$s_p=\sqrt{npq}. \tag{35}$$

Пример 25. Из общего числа $n=408$ новорожденных доля мальчиков составила $p=208/408=0{,}51$, а доля новорожденных девочек $q=200/408=0{,}49$. Среднее квадратическое отклонение долей выразится величиной: $s_p=\sqrt{0{,}51\cdot 0{,}49}=\sqrt{0{,}2499}=0{,}5$. Если выразить в процентах соотношение между количеством новорожденных мальчиков (51%) и девочек (49%), этот показатель будет равен

$$s_p=\sqrt{51(100-51)}=\sqrt{51\cdot 49}=\sqrt{2499}=50{,}0\,\%.$$

Для абсолютных значений альтернатив среднее квадратическое отклонение выразится величиной

$$s_p=\sqrt{408\cdot 0{,}51\cdot 0{,}49}=\sqrt{101{,}96}=10{,}1.$$

ГЛАВА III

ЗАКОНЫ РАСПРЕДЕЛЕНИЯ

III.1. ХАРАКТЕРНЫЕ ЧЕРТЫ ВАРЬИРОВАНИЯ

В любом более или менее симметричном вариационном ряду заметна одна характерная особенность — накапливание вариант в центральных классах и постепенное убывание их численности по мере удаления от центра ряда. Эта особенность варьирования количественных признаков встречается довольно

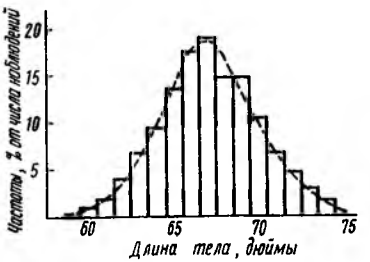


Рис. 5. Гистограмма изменчивости длины тела у 117 мужчин (по Н. Бейли, 1959)

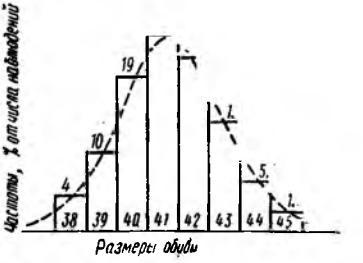


Рис. 6. Гистограмма распределения размеров мужской обуви среди населения центральных областей РСФСР

часто. Так, например, среди взрослого населения чаще встречаются люди среднего роста, а индивиды очень высокого или очень низкого роста — значительно реже. Однако в массе одновозрастных индивидов людей выше и ниже среднего роста оказывается примерно одинаковое количество.

Пусть большая масса людей одного пола и возраста разделена на отдельные группы так, чтобы в каждую группу вошли индивиды приблизительно одинакового роста. Пусть затем от каждой группы выделено по одному представителю, которые построены в одну шеренгу по ранжиру — от самого низкорослого до самого высокорослого. Если в затылок им поставить членов группы, получается «живая диаграмма» распределения, более или менее симметричная. На рис. 5 проиллюстрирована эта закономерность. Отмеченная черта варьирования обнаруживается не только в распределении людей по росту, но и по многим другим признакам, в частности по размерам обуви (рис. 6).

Впервые на эту закономерность варьирования обратил внимание А. Кетле (1835), исследовавший распределение нескольких тысяч американских солдат по росту (длине тела). «Человеческий рост,— писал он,— изменяющийся, по-видимому, самым случайным образом, тем не менее подчиняется самым точ

ным законам; и эта особенность свойственна не только росту; она проявляется также и в весе, силе, быстроте передвижений человека, во всех его физических... и нравственных способностях. Этот великий принцип... разнообразящий проявление человеческих способностей... кажется нам одним из самых удивительных законов мира»[1].

Описанная закономерность относится не только к человеку. Выше рассмотрены варьирование глазков в клубнях картофеля, распределение численности поросят в пометах свиноматок, кальция в сыворотке крови обезьян и удоев коров за лактацию. Подобных примеров можно привести много. Особенно примечательно, что не только распределение живых существ и продуктов их жизнедеятельности, но и случайные ошибки измерений подчиняются этой закономерности. «Не удивительно ли,— писал А. Кетле,— что случайные ошибки располагаются в таком совершенном порядке, и наши бессознательные промахи проявляются с такой симметрией, которая, кажется, могла бы быть результатом тщательно обдуманных расчетов»[2].

Таким образом, прослеживается широко распространенная в природе закономерность: в массе относительно однородных единиц, составляющих статистическую совокупность, большинство членов оказывается среднего или близкого к нему размера, и чем дальше они отстоят от среднего уровня варьирующего признака, тем реже встречаются в данной совокупности. И это независимо от формы распределения, что указывает на определенную связь между числовыми значениями варьирующих признаков и частотой их встречаемости в данной совокупности. Наглядным выражением этой связи и служат *вариационный ряд* и его линейный график — *вариационная кривая*. Эту закономерность можно воссоздать априори в виде математической модели, не опасаясь впасть в противоречие с фактами. Предварительно, однако, полезно напомнить некоторые фундаментальные понятия теории вероятностей.

III.2. СЛУЧАЙНЫЕ СОБЫТИЯ

Подброшенный камень падает вниз, брошенный в воду — тонет. По длине одной из сторон куба можно точно определить его объем. На языке теории вероятностей всякий результат, или исход, однократного испытания называется *событием*. Под испытанием, которое может повторяться бесконечно большое число раз, подразумевают комплекс условий, необходимых для того, чтобы тот или иной исход мог осуществиться. У каждого из

---

[1] *Кетле А.* Социальная физика или опыт о развитии способностей человека. Спб., 1911. Т. 1. С. 38—39.

[2] Там же. С. 330.

отмеченных здесь событий есть только один исход, заранее предсказуемый. Такие события называются *достоверными.* Если же при осуществлении комплекса условий события заведомо произойти не могут, они называются *невозможными.*

Существуют события, исход которых заранее непредсказуем. Если подбросить монету, то заранее нельзя сказать, как она упадет — вверх гербом илн решкой. Здесь исход испытания зависит от случая. События, исход которых при осуществлении комплекса условий точно предсказать нельзя, называют *случайными.*

Случайные события (обозначаются начальными прописными буквами латинского алфавита A, B, C,...) называются *несовместимыми* или *несовместными,* если в серии испытаний всякий раз возможно осуществление только одного из них. Например, при метании монеты она может упасть вверх гербом или решкой. Здесь два равновозможных и несовместных исхода. События, которые в данных условиях могут произойти одновременно, называются *совместными.*

III.3. ВЕРОЯТНОСТЬ СОБЫТИЯ И ЕЕ СВОЙСТВА

Случайное событие можно предсказать лишь с некоторой уверенностью или вероятностью, которую данное событие имеет. При этом *вероятность* рассматривают как числовую меру объективной возможности осуществления события A при единичном испытании и обозначают символом $P(A)$. Согласно классическому определению, вероятность события A выражается отношением числа благоприятствующих осуществлению этого события исходов m к числу всех равновозможных и несовместных исходов n, т. е.

$$P(A)=m/n. \tag{36}$$

Например, в урне находится 5 белых и 10 черных шаров. Наугад вынимают один шар. Какова вероятность, что вынутый шар окажется белым? Так как из общего числа 15 шаров в урне 5 белых, то из $n=15$ возможных исходов лишь $m=5$ «благоприятствуют» осуществлению ожидаемого события, т. е. появлению в однократном испытании белого шара. Отсюда вероятность этого события $P=5/15=1/3\approx 0{,}33$, т. е. чем больше шансов, благоприятствующих наступлению ожидаемого события, тем выше его вероятность.

Из «классического» определения вероятности следует, что она представляет *число, заключенное между нулем и единицей* ($0\leqslant P(A)\leqslant 1$), т. е. выражается в долях единицы (может быть выражена в процентах от общего числа испытаний). Очевидно, вероятность достоверного события A равна единице, а вероятность невозможного события $\bar{A}$ равна нулю. Из этих аксиома-

тических свойств вероятности следует, что вероятность события A и вероятность противоположного события $\bar{A}$ (не A) в сумме равны единице, т. е. $P(A)+P(\bar{A})=1$.

Считают, что события, имеющие очень малую вероятность, в единичных испытаниях не произойдут, т. е. такие события рассматривают как *практически невозможные*. Если же вероятность события достаточно велика, его принято считать *практически достоверным*. Этот принцип практической уверенности в прогнозировании исходов случайных событий позволяет использовать теорию вероятностей в практических целях.

Для упрощения символики принято значение вероятности ожидаемого события обозначать строчной латинской буквой p, т. е. тем же знаком, которым обозначается частость, а значение вероятности противоположного события — буквой q, т. е. $P(A)=p$ и $P(\bar{A})=q$, откуда $p+q=1$.

Вероятность, которую можно указать до опыта, называют *априорной*. Например, при подбрасывании монеты заранее известно, что она может упасть вверх гербом или решкой. Здесь только два возможных исхода, вероятность которых одна и та же: $p=1/2=0{,}5$. Иное дело, например, испытание действия на организм различных доз лекарственных или токсических веществ. В таких случаях результаты отдельных испытаний заранее, до опыта, указать невозможно; вероятность осуществления таких событий может быть установлена только на основании опыта, т. е. *апостериорно*.

Встречаются, однако, и такие события (и их немало), исход которых, как правило, отклоняется от их вероятности. Ярким примером такого рода событий служит соотношение полов в потомстве многих животных и человека. Известно, что пол потомства определяется в момент оплодотворения, когда в зиготу привносятся либо *XX*-, либо *XY*-хромосомы.. Вероятность появления в потомстве мужских и женских особей одна и та же: $p=1/2$. Это означает, что на 1000 новорожденных детей следует ожидать примерно равное число мальчиков и девочек. В действительности же равновеликого соотношения полов в потомстве не наблюдается. Так, по данным шведской статистики за 1935 г., на каждую тысячу новорожденных число родившихся девочек составляло на протяжении года от 473 до 491 при средней частоте, равной 482; относительная частота, или частость, новорожденных девочек $482/1000=0{,}482$, а частость рождения мальчиков $(1000-482)/1000=0{,}518$. Можно также привести данные австрийской статистики о новорожденных детях за период с 1866 по 1905 г., т. е. за 40 лет. Доля новорожденных мальчиков за этот период колебалась от 0,512 до 0,518 и составляла в среднем 0,515.

Из этих примеров видно, что при достаточно большом числе испытаний частости новорожденных девочек и мальчиков от-

клоняются, хоть и незначительно, от вероятности этих событий, равной 0,500. В таких случаях о вероятности событий принято судить по предельным значениям частости, обладающей устойчивостью. Отсюда в отличие от «классической» вероятности частости случайных событий, а точнее — их предельные значения, обладающие определенной устойчивостью, принято называть *статистической вероятностью* этих событий. Так что число 0,482 или округленно 0,48, возле которого колеблются значения частости рождения девочек, можно принять за вероятность этого события. В то же время статистическая вероятность появления в потомстве мальчиков будет равной 0,52.

III.4. ЗАКОН БОЛЬШИХ ЧИСЕЛ

Многочисленные опыты и наблюдения показали, что частости ожидаемых случайных событий приближаются к их вероятности по мере увеличения числа испытаний n. Так, если одну и ту же монету подбрасывать большое число раз, то невозможно ожидать, чтобы во всех без исключения случаях выпадал только герб или только решка. Ясно, что в каком-то числе случаев выпадет герб, а в других случаях — решка. Примечательно, что чем больше число испытаний, тем ближе к единице оказывается отношение выпавших гербов и решек, а частость каждого события становится ближе к его вероятности. Подтверждением тому служат результаты опытов с метанием монет, проведенные разными лицами (табл. 19).

Таблица 19

Кем проведен опыт	Число испытаний	Число случаев выпадения монеты гербом	Частость события	Отклонение частости от вероятности события
Бюффоном	4 040	2 048	0,5069	0,0069
Пирсоном:				0,0016
первый опыт	12 000	6 019	0,5016	
второй опыт	24 000	12 012	0,5005	0,0005

Из табл. 19 следует, что с увеличением числа испытаний отклонение частости ожидаемого результата от его вероятности ($p=0,5$) уменьшается. В этом факте проявляется действие *закона больших чисел*, теоретическое обоснование которому было дано Я. Бернулли (1713), а также П. Л. Чебышевым и другими математиками XIX столетия. Этот закон утверждает, что частость m/n события A будет сколь угодно близкой к его вероятности p, если число испытаний неограниченно возраста-

ет. Как было показано выше, частость события и его вероятность не совпадают. Разница между ними уменьшается при увеличении числа испытаний. Можно взять сколь угодно малое число ε и сравнивать его с разницей между частостью и вероятностью события. Вероятность того, что эта разница превысит число ε, будет стремиться к нулю при стремлении числа испытаний n к бесконечности, т. е.

$$P\left\{|m/n - p| > \varepsilon\right\} \underset{n \to \infty}{\longrightarrow} 0.$$

Этот вывод подтверждается и опытом Кетле: в урну помещали 20 белых и 20 черных шаров, затем извлекали из нее наугад один шар, регистрировали его и возвращали обратно. Каждое испытание повторяли многократно. Вероятность появления белого или черного шара оставалась при этом постоянной, равной 1/2. Результаты опыта Кетле приведены в табл. 20.

Таблица 20

Число вынутых шаров	Из них оказалось		Соотношение белых и черных шаров
	белых	черных	
4	1	3	0,33
16	8	8	1,00
64	28	36	0,78
256	125	131	0,95
1022	526	496	1,06
4096	2066	2030	1,02

Из табл. 20 видно, что с увеличением числа испытаний соотношение белых и черных шаров приближается к единице.

Закон больших чисел, как и другие статистические законы, о которых речь пойдет ниже, имеет объективный характер: их действие не зависит от сознания и воли людей. В качестве примера, иллюстрирующего действие закона больших чисел, можно рассмотреть русскую почтовую статистику за период с 1906 по 1910 г., приведенную в табл. 21 (по А. А. Кауфману, 1916).

Письмо без адреса или без указания места назначения, опущенное в почтовый ящик,— явление случайное. А между тем, как явствует из данных табл. 21, число таких случаев из года в год оставалось относительно постоянным. Такие факты были известны еще П. Лапласу. Затем А. Кетле и другие статистики собрали большой материал, убедительно свидетельствующий о наличии внутренней связи между случайностью и закономерностью, существующей в сфере массовых явлений. К. Маркс по этому поводу писал: «...внутренний закон, прокладывающий себе дорогу в этих случайностях и регулирующий

их, становится видимым лишь тогда, когда они охватываются в больших массах...»[1].

Таблица 2

Год	Всего поступило простых писем, млн.	Из них оказалось		На 1 млн. приходится писем	
		без адреса	без указания места назначения	без адреса	без указания места назначения
1906	983	26 112	28 749	27	29
1907	1076	26 977	26 523	25	25
1908	1214	33 515	26 112	27	21
1909	1357	33 643	28 445	25	21
1910	1507	40 101	36 513	27	24

III.5. БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Представим, что в отношении некоторого случайного события A производят n независимых испытаний при условии, что в каждом испытании вероятность p появления этого события постоянна. Будем учитывать только два исхода: появление события A либо противоположного ему события $\bar{A}$, тоже имеющего постоянную вероятность q, причем $p+q=1$. При этих условиях, если событие A в n испытаниях появится m раз, событие $\bar{A}$ будет встречаться $n-m$ раз. Вероятность любого исхода[2] ($P_n(m)$) независимо от того, в каком порядке эти события чередуются, выразится произведением $p^m q^{n-m}$ (по правилу умножения вероятностей), умноженным на биномиальный коэффициент $C_n^m = \frac{n!}{m!\,(n-m)!}$, т. е.

$$P_n(m) = C_n^m p^m q^{n-m}. \qquad (37$$

Эта формула (*формула Бернулли*) позволяет находить вероятность того, что из n взятых наугад элементов окажется m ожидаемых.

Пример 1. Какова вероятность появления 0, 1, 2, 3, 4, 5 особей мужского пола в числе пяти новорожденных? Можно показать, что возможна серия из пяти ($n=5$) повторных наблюдений с двумя исходами. При этом появление особей мужского пола (событие A) может иметь в серии различное числовое выражение m от 0 до 5, как и противоположное событие $\bar{A}$ (появление особей женского пола). В каждом испытании вероят

[1] *Маркс К., Энгельс Ф.* Соч. 2-е изд. Т. 25. Ч. II. С. 396.

[2] Читается: вероятность появления события A в n испытаниях m ра-

ность появления события $A(p)$ или $\bar{A}(q)$ будет одинаковой и равной 0,5. Вероятность того, что в «пятерке» не будет ни одной особи мужского пола, составит $P_5(0)=\frac{5!}{0!\cdot 5!}(0,5)^0\times$ $\times(0,5)^5=1\cdot 1\cdot 0{,}03125=0{,}03125$. Вероятность того, что в «пятерке» окажется одна особь мужского пола, составит $P_5(1)=$ $=\frac{5!}{1!\cdot 4!}(0,5)^1(0,5)^4=\frac{120}{1\cdot 24}(0,5)\,0{,}0625=0{,}15625$.

Аналогично определяем: $P_5(2)=10(0{,}25)(0{,}125)=0{,}3125$; $P_5(3)=0{,}3125$; $P_5(4)=0{,}15625$; $P_5(5)=0{,}03125$. Легко убедиться, что $\sum_{m=0}^{n} P_n(m)=1$. Точно так же можно вычислить вероятность осуществления m любых событий в n независимых испытаниях при условии постоянства вероятности появления события ($P_n(A)$). Совокупность этих вероятностей — $P_n(0)$, $P_n(1)$, ..., $P_n(m)$ — называется *биномиальным распределением.*

Можно показать, что

$$\sum_{m=0}^{n} P_n(m)=(p+q)^n. \tag{37a}$$

Так, при $n=2$ возможны следующие исходы:

Результаты испытаний AA $A\bar{A}$ $\bar{A}A$ $\bar{A}\bar{A}$

Вероятность исходов p^2 pq qp q^2

или $(p+q)^2=p^2+2pq+q^2=1$. При трех независимых испытаниях возможно $2^3=8$ исходов, вероятности которых распределятся следующим образом: $(p+q)^3=p^3+3p^2q+3pq^2+q^3$ и т. д. Следовательно, закон биномиального распределения выражается не только формулой Бернулли, но и формулой бинома Ньютона:

$$(p+q)^n=p^n+np^{n-1}q+\frac{n(n-1)}{1\cdot 2}p^{n-2}q^2+\ldots+q^n. \tag{38}$$

Так, например, при $n=10$ возможны $2^{10}=1024$ исхода, которые распределятся следующим образом: $P_{10}(m)=(0{,}5+0{,}5)^{10}=$ $=1/1024+10/1024+45/1024+120/1024+210/1024+252/1024+$ $+210/1024+120/1024+45/1024+10/1024+1/1024=1$. Если этот ряд представить в виде графика, как показано на рис. 7, получается полигон биномиального распределения, где ординаты соответствуют членам разложения бинома $(1/2+1/2)^{10}$. Из рис. 7 видно, что биномиальная кривая строго симметрична относительно максимальной ординаты, являющейся центром биномиального распределения.

Из приведенного примера также следует, что распределение вероятностей $\sum_{m=0}^{n} P_n(m) = (p+q)^n$ соответствует коэффициентам разложения бинома Ньютона, отнесенным к одному и тому же знаменателю, равному 2^r. Биномиальные коэффициенты легко вычислить при помощи арифметического треугольника Паскаля, в котором каждая цифра находится суммированием двух цифр, стоящих над ней (табл. 22).

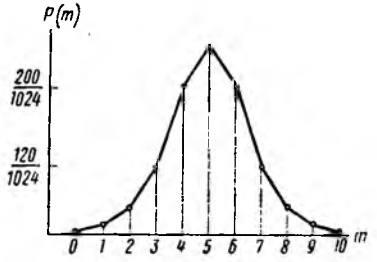


Рис. 7. Распределение вероятностей двучлена $\left(\frac{1}{2}+\frac{1}{2}\right)^{10}$

Сумма биномиальных коэффициентов для любой степени бинома, как это видно из табл. 22, равна 2^n.

Характер биномиального распределения не изменится от способа выражения исходов испытаний — в значениях вероятности или в абсолютных значениях частоты ожидаемого результата. В том и другом случае биномиальный закон выражает зависимость между частотой ожидаемого результата и числом независимых испытаний, проведен-

Таблица 22

n	*Биномиальные коэффициенты*	2^r
0	1	1
1	1 1	2
2	1 2 1	4
3	1 3 3 1	8
4	1 4 6 4 1	16
5	1 5 10 10 5 1	32
6	1 6 15 20 15 6 1	64
7	1 7 21 35 35 21 7 1	128
8	1 8 28 56 70 56 28 8 1	256
9	1 9 36 84 126 126 84 36 9 1	512
10	1 10 45 120 210 252 210 120 45 10 1	1024
	и т. д.	

ных в отношении случайного события A. Причем частота m появления ожидаемого события A в n независимых испытаний определяется его вероятностью p, которая остается постоянной в каждом отдельном испытании.

Закон биномиального распределения неоднократно подвергали экспериментальной проверке. Так, еще в 1912 г. В. И. Романовский 20 160 раз подбросил четыре одинаковые монеты, учитывая в каждом испытании комбинации «герб — решка». Результаты испытаний оказались следующие (табл. 23).

Из табл. 23 видно, что частоты встречаемости возможных комбинаций «герб — решка» распределились строго закономерно; их частости почти полностью совпали с вероятностью исходов для этих комбинаций.

Аналогичные результаты были получены и другими исследователями, проверявшими действие биномиального закона на самых различных моделях. Ф. Гальтон сконструировал специальный прибор, иллюстрирующий закономерность биномиального распределения. Этот прибор представляет собой небольшую дощечку с окаймленными краями, посреди которой на равном расстоянии в шахматном порядке вбиты мелкие гвозди (рис. 8). В нижней части доски помещены равные по размерам и открытые сверху отсеки, в верхней — расположено отделение с узкой щелью, направленной вниз, к середине доски. Через эту щель, находящуюся над гвоздями, насыпают дробь. При этом доску ставят наклонно под углом около 35° к поверхности стола, так что дробинки устремляются вниз, к отсекам. Ударяясь о гвозди, они отскакивают в различные стороны от линии, ведущей к центральному отсеку. Этот случайный процесс приводит к тому, что дробинки закономерно распределяются по отсекам, образуя столбиковую диаграмму, напоминающую гистограмму распределения частот по классам вариационного ряда.

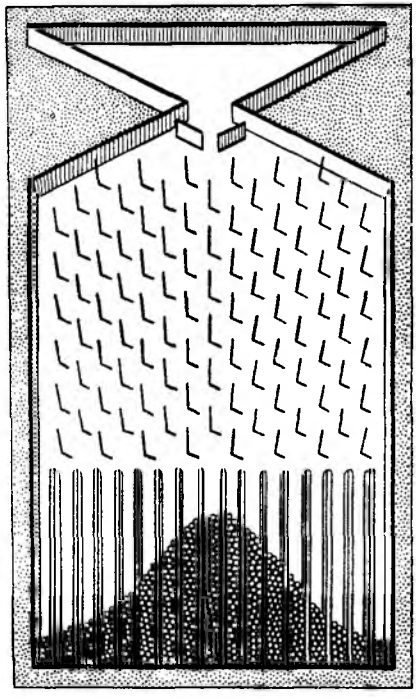


Рис. 8. Прибор Гальтона

Выпало одновременно		Абсолютные частоты комбинации	Частости, % (округленные данные)	Вероят исходо
гербов	решек			
4	0	1 181	6	6,2
3	1	4 909	24	25,0
2	2	7 583	38	37,5
1	3	5 085	25	25,0
0	4	1 402	7	6,2
Сумма		20 160	100	100,0

Для расчета теоретических (выравнивающих) частот рнационного ряда f' в формулу (37а) вводят множитель, равный сумме всех частот эмпирического вариационного р...

$$\sum f' = N(p+q)^n,$$

где $N=\Sigma f_i$; p — вероятность ожидаемого события; $q=1-p$; n — число членов ряда без единицы, т. е. $n=m-1$.

Пример 2. Существует следующее распределение числен... стн самок в 113 пометах лабораторных мышей:

Число самок в помете m	0	1	2	3	4	5	6	7	8
Число пометов с таким количеством самок	0	1	10	17	46	28	8	3	0

В данном случае ожидаемое и противоположное события имеют одну и ту же вероятность $p=q=0{,}5$. Число классов (без нулевых) равно семи. Сумма частот ряда $N=113$, а $n=7-1=6$. По треугольнику Паскаля (см. табл. 22) подбираем ряд биномиальных коэффициентов (K), численно равный 1 6 15 20 15 6 1 для случая $m=7$; сумма членов ряда K равна 64. Подставляем известные величины в формулу (39):

$$\sum f' = 113\left(\frac{1}{2}+\frac{1}{2}\right)^6 = 113\left(\frac{1}{64}+\frac{6}{64}+\frac{15}{64}+\frac{20}{64}+\frac{15}{64}+\frac{6}{64}+\right.$$

$$\left.+\frac{1}{64}\right) = 1{,}77+10{,}59+26{,}48+35{,}32+26{,}48+10{,}59+1{,}77 =$$

$=113{,}00$. Округляя числа, получаем теоретически ожидаемые частоты ряда:

m	1	2	3	4	5	6	7
f'	2	11	26	35	26	11	2

На моделях с известной вероятностью ($p=q=1/2$) ожидаемые частоты биномнального ряда легко рассчитываются по

следующей формуле:

$$f' = \frac{NK}{\Sigma K}. \qquad (40)$$

Так, ожидаемые частоты ряда распределения самок в 113 помётах лабораторных мышей рассчитывают по формуле (40) следующим образом (табл. 24):

Таблица 24

Классы m	Частоты f_i	Биномиальные коэффициенты K	$f' = \frac{NK}{\Sigma K}$	Округленные значения f'
1	1	1	1,77	2
2	10	6	10,59	11
3	17	15	26,48	26
4	46	20	35,32	35
5	28	15	26,48	26
6	8	6	10,59	11
7	3	1	1,77	2
Сумма	113	64	113,00	113

На моделях с неизвестной вероятностью значение p в формуле (39) приходится определять по средней величине полученных в опыте данных, т. е. исходить из статистической вероятности данного события. В таких случаях расчет теоретических (выравнивающих) частот биномиального ряда производят с помощью следующей формулы:

$$f' = N(p^m q^{n-m} K), \qquad (41)$$

где N и K имеют те же значения, что и в предыдущих формулах, а p — статистическая вероятность события, определяемая отношением m/n, т. е. по средней взвешенной $\bar{m} = \Sigma m f_i / N$ и $q = 1 - p$.

Пример 3. В табл. 23 приведены результаты опыта Романовского по проверке биномиального закона Бернулли. Выяснить, согласуются ли полученные в опыте данные с моделью биномиального распределения. Начнем с определения взвешенной средней ($\bar{m}$) ряда:

m	0	1	2	3	4	
f_i	6	24	38	25	7	$\Sigma f_i = N = 100$
mf_i	0	24	76	75	28	$\Sigma m f_i = 203$

Отсюда $\bar{m} = 203/100 = 2{,}03$ и $n = 4$. Тогда $p = \bar{m}/n = 2{,}03/4 = 0{,}5075 \approx 0{,}508$; $q = 1 - p = 0{,}492$. Получается следующая эмпирическая формула: $\Sigma f' = 100(0{,}508 + 0{,}492)^4$.

Для того чтобы рассчитать по этой формуле теоретически ожидаемые частоты ряда, необходимо, как и в предыдущем

случае, подобрать биномиальные коэффициенты, на которые будут умножены значения p и q. В данном случае ряд распределения состоит из пяти членов, ему соответствуют следующие биномиальные коэффициенты: 1 4 6 4 1.

Строим вспомогательную таблицу, первая графа которой заполняется «классами» вариационного ряда m. Во второй графе начиная со второго класса (сверху) помещен $p=0{,}508$, а $q=0{,}492$ вносим в третью графу таблицы, отступив на один шаг от последнего класса (табл. 25).

Таблица 25

m	p^m	q^{n-m}	K	$p^m q^{n-m} K$	$f'=100(p^m \times q^{n-m} K)$	Округленные значения
1	2	3	4	5	6	7
0	1,000	0,059	1	0,059	5,9	6
1	0,508	0,119	4	0,242	24,2	24
2	0,258	0,242	6	0,375	37,5	37
3	0,131	0,492	4	0,258	25,8	26
4	0,066	1,000	1	0,066	6,6	7
Сумма	—	—	16	1,000	100,0	100

Затем последовательно находим p^2, p^3, p^4 и т. д., а также q^2 q^3, q^4 и так до конца ряда. Именно: $p^2=(0{,}508)^2=0{,}258$; $p^3=(0{,}508)^3=0{,}131$ и т. д. Таким же образом рассчитаны значения q. Дальнейшие действия понятны из табл. 25. Указанием на то, что p и q рассчитаны правильно, служит равенство $\Sigma(p^m q^{n-m} K)=1$. Умножая величины $(p^m q^{n-m} K)$ на общее число наблюдений (N), которое в данном случае равно 100, получаем теоретические (выравнивающие) частоты f' ряда.

Если решать эту задачу на модели с известной вероятностью, приняв для каждого исхода комбинаций «герб—решка» $p=q=0{,}5$, то получается следующий результат:

$$\Sigma f' = 100(0{,}5+0{,}5)^4 = 100\left(\frac{1}{16}+\frac{4}{16}+\frac{6}{16}+\frac{4}{16}+\frac{1}{16}\right)=$$

$$=6{,}25+25{,}00+37{,}50+25{,}00+6{,}25=100.$$

Результат вычисления близок к тому, который был получен на модели с неизвестной вероятностью, хотя полного совпадения между этими данными нет.

III.6. РАСПРЕДЕЛЕНИЕ ПУАССОНА

Характер биномиальной кривой определяется двумя величинами: числом испытаний n и вероятностью p ожидаемого результа-

та. При $p=q$ биномиальная кривая строго симметрична и по мере увеличения числа испытаний приобретает все более плавный ход, приближаясь к своему пределу — *нормальной кривой* (см. ниже). Если $p \neq q$, биномиальная кривая становится асимметричной и тем сильнее, чем больше разница между p и q. Когда вероятность события очень мала и исчисляется сотыми и тысячными долями единицы, распределение частот таких редких событий в n независимых испытаний становится крайне асимметричным. Для описания такого рода распределений редких событий служит формула Пуассона

$$P_n(m)=\frac{a^m}{m!}e^{-a}=\frac{a^m}{m!\,e^a}, \qquad (42)$$

где $a \approx np$ — наивероятнейшая частота ожидаемого события; m — частота ожидаемого события в n независимых испытаний; $e=2{,}7183$ — основание натуральных логарифмов; $m!$ — факториал или произведение натуральных чисел $1 \cdot 2 \cdot 3 \cdot 4 \ldots m$.

Формула Пуассона позволяет определять вероятность для любых значений a от 0 до n. Например, для $a=2$ вероятность того, что событие A в данных условиях не осуществится, будет равна

$$P_0=\frac{2^0}{0!\,e^2}=\frac{1}{(2{,}7183)^2}=\frac{1}{7{,}389}=0{,}1353.$$

Вероятность единичного осуществления ожидаемого события при этих условиях выразится следующей величиной:

$$P_1=\frac{2}{1!\,e^2}=\frac{2}{(2{,}7183)^2}=\frac{2}{7{,}389}=0{,}2707.$$

Для трех случаев $P_3=\frac{2^3}{3!\,e^2}=\frac{8}{44{,}334}=0{,}1805$ и т. д.

(см. табл. III Приложений).

Чтобы формула Пуассона выражала не вероятности, а ожидаемые абсолютные частоты f' редкого события, ей придают следующий вид:

$$f'=n\frac{\bar{x}^m}{m!}e^{-\bar{x}}. \qquad (43)$$

Здесь f' — теоретические ординаты кривой распределения Пуассона, или ожидаемое число случаев редкого события в каждом отдельно взятом классе испытания — 0, 1, 2, 3, 4 и т. д.; n — число испытаний; $\bar{x}$ — среднее число фактически наблюдаемых случаев (взятое вместо a); объяснения остальных символов те же, что в формуле (42).

Распределение Пуассона — частный случай биномиального распределения. Оно, как и биномиальное распределение, прибли-

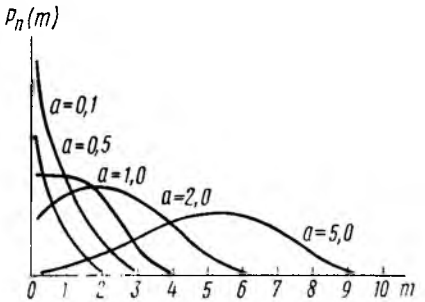


Рис. 9. График функции $P_n(m) = \frac{a^m}{m!} e^{-a}$ для разных значений a

жается к нормальной кривой (см. ниже) при возрастании числа $a \approx np$ (рис. 9).

По закону Пуассона распределяются редкие случайные события, встречающиеся в микробиологии, радиобиологии и других разделах современной биологии. Например, установлено, что численность перезимовавших клопов вредной черепашки на пробных площадках распределяется по закону Пуассона. По этому же закону распределяются частоты спонтанных мутаций у кишечной палочки. Подобных примеров можно привести много.

III.7. ПАРАМЕТРЫ ДИСКРЕТНЫХ РАСПРЕДЕЛЕНИЙ

Биномиальное распределение характеризуется двумя параметрами: *средним,* или *наивероятнейшим, числом* μ *ожидаемого результата* и *дисперсией частоты* σ_m^2 события A в n независимых испытаний. Первый параметр приближенно равен произведению числа испытаний n на вероятность p, которую событие A имеет в каждом классе испытаний, т. е. $\mu = np$. Второй параметр равен произведению числа испытаний n на вероятность p ожидаемого события A и вероятность q противоположного события $\bar{A}$, т. е. $\sigma_m^2 = npq$. Корень квадратный из дисперсии называется *стандартным отклонением.*

В отличие от биномиального распределения распределение редких событий, следующих закону Пуассона, характеризуется одним параметром — *средней величиной* ($np = \bar{m} = \bar{x}$), так как для этого распределения характерно равенство $\sigma_m^2 = \bar{m}$. Кроме того, распределение Пуассона, как и другие асимметричные распределения, характеризуется очень высоким коэффициентом вариации. Эти особенности распределения редких событий иллюстрирует опыт по облучению штамма бактерий α-частицами, результаты которого и их обработка приведены в табл. 26.

Характеристики этого распределения: $\bar{m} = 798/517 = 1{,}54$ и $s_m^2 = 790{,}3/516 = 1{,}53$. Отсюда $Cv = 100\sqrt{1{,}53}/1{,}54 = 80{,}2\%$. Совпадение по абсолютной величине дисперсии и средней арифметической указывает на то, что данное распределение следует закону Пуассона.

Расчет теоретических частот (f') по закону Пуассона производят по формуле (43). Это показано на примере 4.

Таблица 26

Поражаемость бактериальных клеток m	Число случаев f_i	mf_i	$m_i-\bar{m}$	$(m-\bar{m})^2$	$f_i(m_i-\bar{m})^2$
0	112	0	—1,54	2,3716	265,6192
1	168	168	—0,54	0,2916	48,9888
2	130	260	—0,46	0,2116	27,5080
3	68	204	+1,46	2,1316	144,9488
4	32	128	+2,46	6,0516	193,6512
5	5	25	+3,46	11,9716	59,8580
6	1	6	+4,46	19,8916	19,8916
7	1	7	+5,46	29,8116	29,8116
Сумма	517	798	—	—	790,2772

Пример 4. Воспользуемся данными опыта по облучению амма бактерий α-частицами (см. табл. 26) и рассчитаем для ..ого распределения теоретические частоты. В данном случае $i=517$; $\bar{x}\approx 1{,}5$; $e^{-1{,}5}=0{,}2231$[1]; классы испытаний m: 0 1 2 3 4 5 6 7, им соответствуют $m!=0!=1$; $1!=1$; $2!=1\cdot 2=2$; $3!=2\cdot 3=6$; ...; $7!=720\cdot 7=5040$, а также $1{,}5^2=2{,}25$; $1{,}5^3=3{,}375$; $1{,}5^4=5{,}063$; ...; $1{,}5^7=17{,}086$.

Подставляем известные величины в формулу (42):

$$f_0'=517\cdot 0{,}2231=115{,}34=115$$

$$f_1'=517\cdot 1{,}5\cdot 0{,}2231=173{,}04=173$$

$$f_2'=\frac{517\cdot 2{,}25\cdot 0{,}2231}{2}=129{,}77=130$$

$$f_3'=\frac{517\cdot 3{,}375\cdot 0{,}2231}{6}=64{,}88=65$$

$$f_4'=\frac{517\cdot 5{,}063\cdot 0{,}2231}{24}=24{,}35=24$$

$$f_5'=\frac{517\cdot 7{,}594\cdot 0{,}2231}{120}=7{,}30=7$$

$$f_6'=\frac{517\cdot 11{,}391\cdot 0{,}2231}{720}=1{,}82=2$$

$$f_7'=\frac{517\cdot 17{,}086\cdot 0{,}2231}{5040}=0{,}39=1$$

Сумма $516{,}89\approx 517$

[1] Значения показательной функции e^{-x} см. в справочниках по математике.

Расчет теоретических частот можно упростить, применяя табл. III Приложений, в которой содержатся значения вероятности $P(m)$ для каждого класса испытаний m и средней величины $a=\bar{x}$. Чтобы получить теоретические частоты f', достаточно значения вероятности $P(m)$, приведенные в табл. III Приложений для m и $\bar{x}$ (вместо $a=np$), умножить на общее число наблюдений n. Так, для $\bar{x}=1{,}5$ и $m=0$ в табл. III Приложений находим $P(m)=0{,}2231$. Умножая эту величину на n, равное 517, получаем $f_0'=115{,}34$. Затем для $m=1$ и $\bar{x}=1{,}5$ в той же таблице находим $P(m)=0{,}3347$ и $f_1'=517\cdot 0{,}3347=173{,}04$ и так поступаем до конца ряда, как это показано в табл. 27.

Таблица 27

Классы m	Частоты f_i	$P(m)$	Теоретические частоты f_i'	
			расчетные	с округлением
0	112	0,2231	115,34	115
1	168	0,3347	173,04	173
2	130	0,2510	129,77	130
3	68	0,1255	64,88	65
4	32	0,0471	24,35	24
5	5	0,0141	7,29	7
6	1	0,0035	1,81	2
7	1	0,0008	0,41	1
Сумма	517	—	516,90	517

При сравнении (визуальном) эмпирических частот с частотами, вычисленными по закону Пуассона, видно, что они согласуются между собой.

III.8. НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Случайные величины. Как было показано выше, варьирующие признаки в математике рассматривают как переменные случайные величины, способные в одних и тех же условиях испытания принимать различные числовые значения, которые заранее невозможно предсказать. Случайные величины делят на дискретные и непрерывные. Случайная величина называется *дискретной*, если она может принимать только определенные фиксированные значения, которые обычно выражаются целыми числами. Если же случайная величина способна принимать любые числовые значения, она называется *непрерывной*. Очевидно, что счетные признаки относятся к дискретным случайным величинам, тогда как признаки мерные, варьирующие непрерывно, являются величинами непрерывными.

Случайная величина X в серии независимых повторных испытаний может принимать самые различные значения, но в каждом отдельном испытании она принимает единственное из возможных значений x_i.

Закон распределения случайных величин. Функция $f(x)$, связывающая значения x_i переменной случайной величины x с их вероятностями p_i, называется *законом распределения этой величины.* Закон распределения случайной величины можно задать таблично, выразить графически в виде кривой вероятности и описать соответствующей формулой. Закон распределения дискретной случайной величины может, например, выражаться в виде биномиальной кривой и описываться формулой Бернулли, которая позволяет находить вероятные значения этой величины в серии независимых испытаний. В отношении же непрерывной случайной величины речь может идти лишь о тех значениях, которые она способна принять с той или иной вероятностью в интервале от и до. Этот интервал может быть каким угодно: и большим, и малым. Выдающиеся математики — А. Муавр (1733), И. Г. Ламберт (1765), П. Лаплас (1795) и К. Гаусс (1821) — установили, что очень часто вероятность P любого значения x_i непрерывно распределяющейся случайной величины x находится в интервале от x до $x+\mathrm{d}x$ и выражается формулой

$$P(X)=\frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}}\,\mathrm{d}x, \qquad (44)$$

где $\mathrm{d}x$ — малая величина, определяющая ширину интервала; π и e — математические константы (π — отношение длины окружности к ее диаметру, равное 3,1416...; $\mathrm{e}=2{,}7183$ — основание натуральных логарифмов); σ — стандартное отклонение, характеризующее степень рассеяния значений x_i случайной величины X вокруг средней μ, называемой *математическим ожиданием.* В показатель степени числа e входит *нормированное отклонение* $t=(x_i-\mu)/\sigma$ — величина, играющая важную роль в исследовании свойств нормального распределения, описываемого формулой (44).

Как видно из этой формулы, закон нормального распределения (нормальный закон) выражает функциональную зависимость между вероятностью $P(X)$ и нормированным отклонением t. Он утверждает, что вероятность отклонения любой варианты x_i от центра распределения μ, где $x_i-\mu=0$, определяется функцией нормированного отклонения t. Графически эта функция выражается в виде кривой вероятности, называемой *нормальной кривой.* Форма и положение этой кривой определяются только двумя параметрами: μ и σ. При изменении величины μ форма нормальной кривой не меняется, лишь график ее смещается вправо или влево. Изменение же величины σ влечет за собой

изменение только ширины кривой: при уменьшении σ кривая делается более узкой за счет меньшего рассеяния вариант вокруг средней, а при увеличении σ кривая расширяется. Во всех случаях, однако, нормальная кривая остается строго симметричной относительно центра распределения, сохраняя правильную колоколообразную форму (рис. 10).

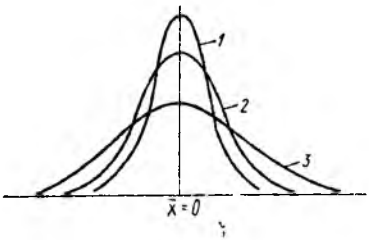


Рис. 10. Нормальные кривые (*1*, *2*, *3*) при разных значениях параметра σ ($\sigma_1<\sigma_2<\sigma_3$)

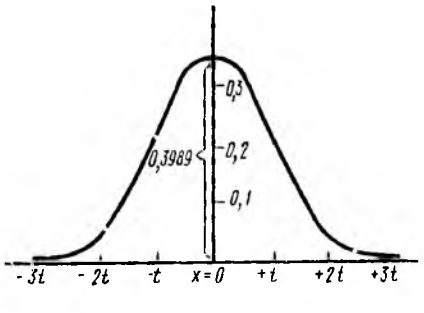


Рис. 11. Стандартизованная форма нормальной кривой (при $\sigma=1$)

Нормальная кривая с параметрами $\mu=0$ и $\sigma=1$ называется *нормальной* или *стандартизованной кривой*. Она описывается формулой

$$f(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}. \qquad (45)$$

Любую нормальную кривую можно привести к стандартной (вычитанием μ из x_i и делением на σ). Стандартная кривая (рис. 11) имеет площадь, равную единице. Ее вершина, т. е. максимальная ордината y_{max}, соответствует началу прямоугольных координат, перенесенному в центр распределения, где $x_i-\mu=0$. Вправо и влево от этого центра случайная величина X может принимать любые значения, и величина каждого отклонения ($x_i-\mu$) определяется функцией его нормированного отклонения $f(t)$. Вероятности P таких отклонений, соответствующие разным значениям t, приведены в табл. I Приложений.

Для того чтобы ордината выражала не вероятности, а абсолютные числовые значения случайной величины, т. е. выравнивающие частоты вариант эмпирического распределения, нужно в правую часть формулы (45) внести дополнительные множители: в числитель — общее число наблюдений n, умноженное на величину классового интервала λ, а в знаменатель — величину среднего квадратического отклонения эмпирического ряда распреде-

ления s_x. В результате можно записать формулу

$$f' = \frac{n\lambda}{s_x} f(t). \qquad (46)$$

Здесь f' — теоретические (выравнивающие) частоты вариационного ряда, а $f(t)$ — значения функции нормированного отклонения, рассчитанные по формуле (46). Эти значения содержатся в табл. II Приложений. Применяя табл. I и II Приложений, можно по двум показателям (средней арифметической $\bar{x}$ и среднему квадратическому отклонению s_x) вычислить теоретические частоты эмпирического вариационного ряда, рассчитать ординаты и построить график нормальной кривой. Сравнивая частоты эмпирического вариационного ряда с частотами, вычисленными по формуле (46), можно проверить, следует ли эмпирическое распределение нормальному закону.

Пример 5. По выборке, состоящей из 267 взрослых мужчин, для длины тела получен вариационный ряд (табл. 28).

Таблица 28

Центры интервалов x_i, см	Эмпирические частоты f_i	$t = \frac{x_i - \bar{x}}{s_x}$	Ординаты нормальной кривой $f(t)$	Теоретические частоты f'	
				расчетные	округленные
158	3	—2,77	0,0086	1,6	2
161	9	—2,03	0,0508	10,0	10
164	31	—1,29	0,1736	34,3	34
167	71	—0,55	0,3429	67,8	68
170	82	+0,19	0,3918	77,6	78
173	46	+0,93	0,2589	51,2	51
176	19	+1,67	0,0989	19,5	19
179	5	+2,41	0,0219	4,4	4
182	1	+3,15	0,0028	0,6	1
Сумма	267	—	—	267,0	267

Характеристики этого распределения: $\bar{x} = 169{,}22$ см и $s_x = 4{,}06$ см (эти показатели читатель может вычислить). Из табл. 28 видно, что расчет теоретических частот начинается с нормирования членов вариационного ряда, т. е. вычисления t. Затем по табл. II Приложений находят значение функции $f(t)$ для каждого нормированного отклонения t эмпирического ряда. Перемножая значения $f(t)$ на величину $n\lambda/s_x$, равную в данном случае $267 \cdot 3/4{,}06 \approx 198$, находят теоретические (выравнивающие) частоты данного распределения. Из рис. 12 видно, что представленные в виде линейного графика эмпирические и вычисленные

по нормальному закону частоты этого распределения согласуются между собой.

Параметры нормального распределения. Как было показано, нормальное распределение характеризуется двумя параметрами: средней величиной, или математическим ожиданием μ, и дисперсией σ_x^2 случайной величины X. Первый параметр равен сумме произведений отдельных значений x_i случайной величины X на их вероятности p_i, т. е.

$$\mu = x_1p_1 + x_2p_2 + x_3p_3 + \ldots + x_np_n = \sum_{i=1}^{n} x_ip_i.$$

Второй параметр равен сумме квадратов отклонений отдельных значений x_i случайной величины X от ее математического ожидания μ, т. е. $\sigma_x^2 = \sum [x_i - \mu(x)]^2$, или с учетом повторяемости f_i

$$\sigma_x^2 = \sum \{[x_i - \mu(x)]^2 f_i\}.$$

Формально математическое ожидание соответствует средней величине эмпирического распределения, однако, по существу, эти показатели отождествлять нельзя. Среднюю величину определяют как сумму всех членов ряда, отнесенную к их общему числу, а математическое ожидание представляет собой сумму произведений членов ряда на их вероятности. Эмпирическая средняя стремится к математическому ожиданию случайной величины по мере увеличения числа испытаний; при небольшом числе испытаний средняя может значительно отклоняться от своего математического ожидания.

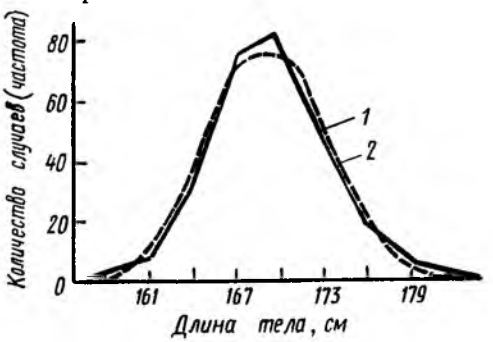


Рис. 12. Эмпирическая (*1*) и вычисленная по нормальному закону (*2*) кривые распределения длины тела у 267 мужчин

Основные свойства нормального распределения. Для нормального распределения характерно совпадение по абсолютной величине средней арифметической, медианы и моды. Равенство между этими показателями указывает на *нормальность данного распределения*. Вероятность отклонения любой варианты в ту или другую сторону от средней μ на t, $2t$ и $3t$, как это видно из табл. I Приложений, следующая:

$$P\{-t < |x - \mu| < +t\} = 0{,}6827;$$
$$P\{-2t < |x - \mu| < +2t\} = 0{,}9545;$$
$$P\{-3t < |x - \mu| < +3t\} = 0{,}9973.$$

Это означает, что при распределении совокупности наблюдений по нормальному закону из 10 000 вариант в интервале от $\mu-t$ до $\mu+t$ окажется 6827 вариант, или 68,3% от общего числа вариант, составляющих данную совокупность. В интервале от $\mu-2t$ до $\mu+2t$ будет находиться 9545 вариант, или 95,4% от числа всех вариант совокупности. И в интервале от $\mu-3t$ до $\mu+3t$ окажется 9973, или 99,7% от общего объема совокупности.

Следовательно, с вероятностью $P=0{,}6827$ можно утверждать, что наугад отобранная из нормально распределяющейся совокупности варианта не выйдет за пределы от $\mu-t$ до $\mu+t$, или в компактной форме $\mu\pm t$. Вероятность того, что случайно отобранная варианта не отклонится от средней μ более чем на $\mu\pm3t$, равна $P=0{,}9973$. Это означает, что 99,7% от всех вариант нормально распределяющейся совокупности находится в пределах $\mu\pm3\sigma$. Этот важный вывод известен в биометрии как правило *плюс—минус трех сигм.*

III.9. РАСПРЕДЕЛЕНИЕ МАКСВЕЛЛА

По нормальному закону распределяются многие биологические признаки, но не все: нередко встречаются и асимметричные распределения, которые, однако, не следуют закону Пуассона. Одним из трех распределений является *распределение, описываемое формулой Максвелла*

$$P(X)=\frac{2}{\sqrt{2\pi}}\,\frac{t^2}{a}\,e^{-\frac{t^2}{2}}\,dx. \qquad (47)$$

В этой формуле $a=0{,}6267\,\bar{x}$ — параметр распределения, определяемый через среднюю арифметическую $\bar{x}$ варьирующего признака; $t=x_i/a$, где x_i — числовые значения случайной величины X; dx — разность между двумя смежными значениями переменной величины X.

Указанием на то, что эмпирическое распределение следует закону Максвелла, служит равенство между средним квадратическим отклонением и величиной $0{,}674\,a$, т. е. $s_x=0{,}674\,a$, тогда как для распределения Пуассона характерно равенство $s_x{}^2=\bar{x}$.

Чтобы рассчитать по формуле (47) теоретические (выравнивающие) частоты, нужно проделать следующее. 1. Определить среднюю арифметическую эмпирического вариационного ряда и параметр a. 2. Разделить каждую классовую варианту x_i на величину a, что даст значения t. 3. Найти для каждого значения $t=x_i/a$ по табл. II Приложений значение функции $f(t)$. 4. Определить значения t^2/a. 5. Умножить значения t^2/a на удвоенную величину t и на величину классового промежутка $(\lambda=dx)$, т. е. определить $P=(t^2/a)2f(t)\lambda$. 6. Умножить значения P на общее

число наблюдений n, получить теоретические (выравнивающие частоты данного вариационного ряда, т. е. $f'=Pn$.

Пример 6. При скрещивании мелкоплодной линии томатов с крупноплодной линией того же сорта в первом поколении плоды получились не среднего, а несколько меньшего размера. Во втором поколении, т. е. при скрещивании представителей первого поколения между собой, масса отдельных плодов еще более приблизилась к массе плодов исходной мелкоплодной линии. Распределение массы плодов семенных гнезд, взятых с 928 растений расщепляющейся популяции томатов, показано в табл. 29.

Таблица 29

Классы x_i	Частоты $f(t)$	$t=\frac{x_i}{a}$	$f(t)$	t^2	$\frac{t^2}{a}$	$\frac{t^2}{a}2f(t)\lambda=p$	$pn=f'$
10	28	0,33	0,3778	0,109	0,0035	0,0264	25
20	93	0,66	0,3209	0,430	0,0141	0,0905	84
30	186	0,98	0,2468	0,966	0,0316	0,1571	146
40	148	1,31	0,1691	1,716	0,0562	0,1900	176
50	176	1,61	0,1040	2,686	0,0880	0,1830	170
60	102	1,97	0,0573	3,869	0,1268	0,1450	134
70	74	2,30	0,0283	5,267	0,1727	0,0977	91
80	46	2,62	0,0129	6,880	0,2256	0,0582	54
90	28	2,95	0,0051	8,708	0,2855	0,0291	28
100	19	3,28	0,0018	10,745	0,3523	0,0127	12
110	14	3,61	0,0005	13,032	0,4273	0,0043	4
120	6	3,93	0,0002	15,476	0,5074	0,0020	2
130	5	4,25	0,0001	18,164	0,5955	0,0011	1
140	2	4,59	0,0000	21,068	0,6907	0,0001	1
150	1	4,92	0,0000	24,186	0,7930	0,0000	0
Сумма	928	—	—	—	—	0,9972	928

Характеристики этого распределения следующие: $\bar{x}=48,7$ и $s_x=23,8$, откуда $a=0,6267\,\bar{x}=30,5$. Близость $s_x=23,8$ к величине $0,674\,a=20,6$ позволяет предположить, что данное распределение следует закону Максвелла. Расчет выравнивающих частот f' приведен в табл. 29. Значения $t=x_i/a$ получены так: $t_1=10/30,5=0,328=0,33$; $t_2=20/30,5=0,655=0,66$ и т. д. Значения $f(t)$ находят в табл. II Приложений: $f(t_1=0,33)=0,3778$; $f(t_2=0,66)=0,3209$ и т. д.

Величины предпоследней графы этой таблицы рассчитаны следующим образом: $t_1^2=(0,33)^2=0,109$; $t_1^2/a=0,109/30,5=0,0035$; $2f(t)=2\cdot0,3778=0,7556$; $(t_1^2/a)2f(t)\lambda=0,0035\cdot0,7556\times10=0,0264$. Умножая эту величину на $n=928$, получают искомое значение $f'=Pn=0,0264\cdot928=25$ и так до конца ряда.

В результате получается ряд теоретически вычисленных (выравнивающих) частот f'. Если эмпирические и вычисленные по закону Максвелла частоты этого ряда представить графически в

виде вариационных кривых, как это показано на рис. 13, можно убедиться в том, что они согласуются между собой. Следовательно, формула (47) для нахождения выравнивающих частот этого ряда выбрана правильно.

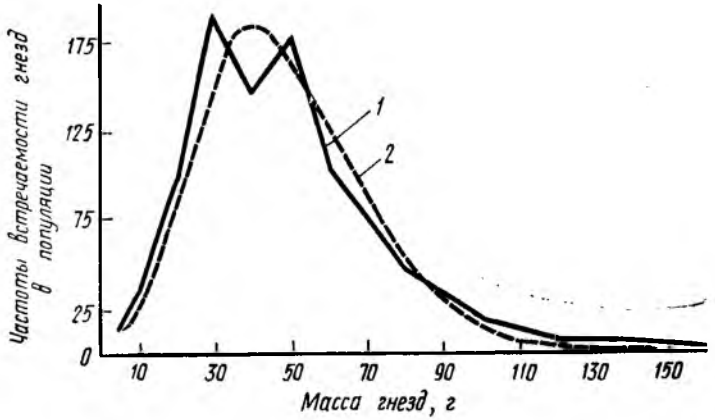


Рис. 13. Эмпирическая (*1*) и вычисленная по формуле Максвелла (*2*) кривые распределения массы гнезд томатов

III.10. ИЗМЕРЕНИЕ АСИММЕТРИИ И ЭКСЦЕССА

Среди эмпирических распределений *асимметрия* и *эксцесс* встречаются довольно часто. Заметить асимметрию и эксцесс можно по характеру распределения частот в классах вариационного ряда. Графически асимметрия выражается в виде скошенной вариационной кривой, вершина которой может находиться левее или правее центра распределения. В первом случае асимметрия называется *правосторонней* или *положительной,* а во втором — *левосторонней* или *отрицательной* (по знаку числовой характеристики). При правосторонней асимметрии ее пологая сторона находится правее (рис. 14), при левосторонней — левее центра распределения (рис. 15).

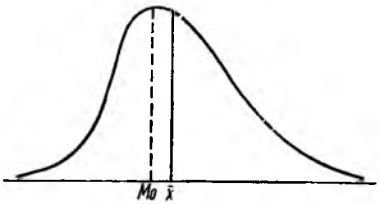


Рис. 14. Асимметричная кривая (положительная асимметрия)

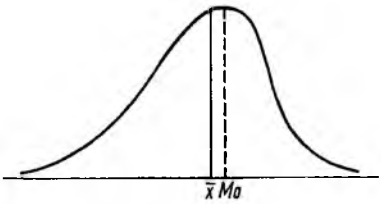


Рис. 15. Асимметричная кривая (отрицательная асимметрия)

Наряду с асимметричиыми встречаются *островершинные* и *плосковершинные распределения.* Островершнность кривой распределения вызывается чрезмерным накапливанием частот в центральных классах варнацнонного ряда, вследствие чего вершнна варнационной крнвой оказывается снльно поднятой вверх. В такнх случаях говорят о *положительном эксцессе распределения* (рис. 16). Кроме одновершинных встречаются н двух- и *многовершинные* кривые, а также *плосковершинные* н *двугорбые*

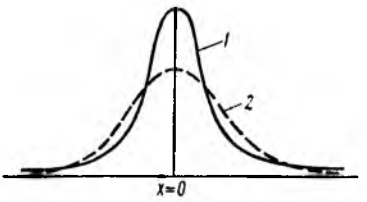


Рис. 16. Крутовершиинная кривая — положительиый эксцесс (*1*) в сравиеиии с иормальиой кривой (*2*)

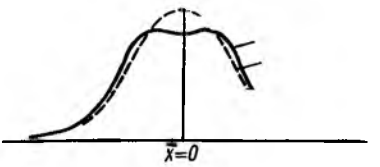


Рис. 17. Плосковершиинная кривая — отрицательиый эксцесс (*1* в сравиеиии с иормальной кривой (*2*)

крнвые, что свидетельствует о налнчнн у такого распределения *отрицательного эксцесса* (рис. 17).

Велнчнна асимметрни н эксцесса может быть разлнчной, поэтому важно ее не только обнаружнть, но н нзмернть. Для нзмерення аснмметрнн и эксцесса используют *центральные моменты распределения третьего и четвертого порядков.* В качестве показателя асимметрин As служнт центральный момент третьего порядка μ_3, отнесенный к кубу среднего квадратического отклонения s_x^3, т. е.

$$As=\frac{\mu_3}{s_x^3}=\frac{\sum_{i=1}^{k} f_i (x_i-\bar{x})^3}{n} \Big/ s_x^3. \qquad (48$$

Прн строго снмметрнчных распределеннях сумма третьих степеней отклонений вариант x_i от средней арифметической $\bar{x}$ равна нулю н $As=0$. Прн налнчни скошенностн распределения этот показатель будет нметь положнтельную (прн правосторонней асимметрнн) лнбо отрицательную велнчнну (прн левосторонией асимметрнн), которая н служит *мерой асимметрии.*

Показатель эксцесса, обозначаемый символом Ex, выражается формулой

$$Ex=\frac{\mu_4}{s_x^4}-3=\left[\frac{\sum_{i=1}^{k} f_i (x_i-\bar{x})^4}{n} \Big/ s_x^4\right]-3. \qquad (49$$

При отсутствии эксцесса $Ex=0$. В случае положительного эксцесса этот показатель приобретает положительный знак (+) и может иметь самую различную величину. При плосковершинности и двугорбости вариационной кривой коэффициеит Ex имеет отрицательный знак (—); предельная величина отрицательного эксцесса равна минус двум.

Вычисление показателей асимметрии и эксцесса по формулам (48) и (49), т. е. способом произведений непосредственно по центральным моментам распределения, оказывается довольно трудоемким, особенно при наличии в выборке многозначных чисел. Поэтому центральные моменты обычно вычисляют косвенным путем — через условные моменты распределения, которые, как было показано в гл. II, связаны определенным образом с центральными моментами. Вычисление условных моментов производят по-разному в зависимости от того, каким способом — условной средней или способом сумм — определяют коэффициенты асимметрии и эксцесса.

При вычислении показателей As и Ex способом условной средней A статистические моменты определяют по формулам

$$b_1=\sum f_i a/n;\quad b_2=\sum f_i a^2/n;\quad b_3=\sum f_i a^3/n \text{ и } b_4=\sum f_i a^4/n,$$

где $a=(x_i-A)/\lambda$; n — общее число наблюдений; f_i — частоты вариационного ряда; λ — классовый интервал[1]. Ниже приведены соответствующие примеры.

Пример 7. На практических занятиях студентам было предложено измерить в миллиметрах длину отобранных наугад 200 хвоинок сосны обыкновенной. В результате был получен вариационный ряд, по которому рассчитывали значения показателей асимметрии и эксцесса.

Определяем $\Sigma f_i a$, $\Sigma f_i a^2$, $\Sigma f_i a^3$ и $\Sigma f_i a^4$ (табл. 30).

Пользуясь итогами этой таблицы, определяем: $b_1=153/200=0{,}765$; $b_2=955/200=4{,}775$; $b_3=1059/200=5{,}295$ и $b_4=13\,687/200=68{,}435$, а также $b_1^2=0{,}5852$; $b_1^3=0{,}4477$; $b_1^4=0{,}3425$; $2b_1^3=0{,}8954$; $3b_1^4=1{,}0274$; $3b_1b_2=10{,}9586$; $6b_1^2b_2=16{,}7660$ и $4b_1b_3=16{,}2027$. Находим: $s_x^2=b_2-b_1^2=4{,}775-0{,}5852=4{,}1898$; $s_x=2{,}0469$; $s_x^3=8{,}5761$ и $s_x^4=17{,}5544$. Переходим к определению центральных моментов распределения: $\mu_3=5{,}295-10{,}9586+0{,}8954=-4{,}7682$; $\mu_4=68{,}4350-16{,}2027+16{,}7660-1{,}0274=67{,}9709$. Отсюда $As=-4{,}7682/8{,}5761=-0{,}5560$ и $Ex=67{,}9709/17{,}5544-3=3{,}8720-3=0{,}8720$. Полученные величины As и Ex показывают, что данное распределе-

[1] При вычислении коэффициентов As и Ex описанными способами среднее квадратическое отклонение определяют без внесения поправки Бесселя $n/(n-1)$, не умножая на величину классового интервала, поскольку условные моменты распределения b_1, b_2 и т. д. вычисляются без умножения на λ. Эти приемы облегчают вычисление коэффициентов асимметрии As и эксцесса Ex, не отражаясь на их величине.

ние имеет левостороннюю асимметрию и заметно выраженный эксцесс.

Таблица 3

Длина хвои, x_i, мм	Частоты f_i	a	$f_i a$	$f_i a^2$	$f_i a^3$	$f_i a^4$
125	2	−6	−12	72	−432	2592
175	2	−5	−10	50	−250	1250
225	4	−4	−16	64	−256	1024
275	5	−3	−15	45	−135	405
325	7	−2	−14	28	−56	112
375	25	−1	−25	25	−25	25
425	39	0	0	0	0	0
475	46	+1	+46	46	+46	46
525	31	+2	+62	124	+248	496
575	23	+3	+69	207	+621	1863
625	13	+4	+52	208	+832	3328
675	2	+5	+10	50	+250	1250
725	1	+6	+6	36	+216	1296
Сумма	200	—	+153	955	+1059	13687

Пример 8. Применим способ условной средней к расчету показателей *As* и *Ex* для ряда распределения кальция (мг%) в сыворотке крови обезьян. Необходимые данные содержатся в табл. 15: $\Sigma fa = +67$; $\Sigma fa^2 = 293$; $\Sigma fa^3 = 553$; $\Sigma fa^4 = 2417$. Отсюда $b_1 = +67/100 = 0{,}67$; $b_2 = 293/100 = 2{,}93$; $b_3 = 553/100 = 5{,}53$; $b_4 = 2417/100 = 24{,}17$. Находим: $b_1^2 = 0{,}449$; $b_1^3 = 0{,}301$; $b_1^4 = 0{,}202$; $2b_1^3 = 0{,}602$; $3b_1^4 = 0{,}606$; $3b_1b_2 = 5{,}889$; $4b_1b_3 = 14{,}820$; $6b_1^2b_2 = 7{,}983$. Отсюда $\mu_3 = 5{,}530 - 5{,}839 + 0{,}602 = 0{,}243$; $\mu_4 = 24{,}170 - 14{,}820 + 7{,}893 - 0{,}606 = 16{,}637$. Определяем показатели вариации: $s_x^2 = b_2 - b_1^2 = 2{,}93 - 0{,}449 = 2{,}481$; $s_x = 1{,}575$; $s_x^3 = 3{,}908$; $s_x^4 = 6{,}155$. В результате имеем $As = 0{,}243 : 3{,}908 = +0{,}062$; $Ex = 16{,}637/6{,}155 - 3 = 2{,}703 - 2 = -0{,}297$.

В данном случае показатели асимметрии и эксцесса оказались довольно низкими, что указывает на близость этого распределения к нормальной кривой. Как будет показано в гл. V, это предположение полностью подтверждается.

III.11. РАСПРЕДЕЛЕНИЕ ШАРЛЬЕ

Среди асимметричных распределений встречаются и такие, которые неплохо описывает *формула Шарлье*:

$$P(X) = \frac{n}{s_x} f(t)\left[1 + \frac{As}{6}(t^3 - 3t) + \frac{Ex}{24}(t^4 - 6t^2 + 3)\right]. \quad (50$$

Здесь $f(t)$ — функция нормированной разности $t = (a - b_1)/s_x$

где a — отклонения классовых вариант x_i от условной средней A, отнесенные к величине классового интервала λ, т. е. $a=(x_i-A)/\lambda$; $b_1=\Sigma fa/n$ — условный момент первого порядка; s_x — среднее квадратическое отклонение; n — общее число наблюдений.

Для нахождения теоретических (выравнивающих) частот по этой формуле необходимо: 1) вычислить s_x, As и Ex, а также b_1, n/s_x, $As/6$ и $Ex/24$; 2) определить для каждого класса вариационного ряда $t=(a-b_1)/s_x$ и величину $(n/s_x)f(t)$, предварительно выписав из табл. II Приложений (с положительными знаками) значения функции $f(t)$; 3) чтобы облегчить вычисление $(As/6)(t^3-3t)$ и $(Ex/24)(t^4-6t^2+3)$, предварительно рассчитать t^2, t^3, t^4, $3t$ и $6t^2$; 4) вычислить величины $[1+(As/6)(t^3-3t)]$ и $(Ex/24)(t^4-6t^2+3)$; 5) определить для каждого класса вариационного ряда $[1+(As/6)(t^3-3t)+(Ex/24)(t^4-6t^2+3)]$ и, умножив эту величину на $(n/s)f(t)$, найти выравнивающие частоты f' вариационного ряда.

Формулу (50) имеет смысл применять в тех случаях, когда эмпирическое распределение обнаруживает эксцесс. Если же распределение имеет только асимметрию, для нахождения выравнивающих частот f' достаточно использовать первое слагаемое формулы (50), т. е. исходить из

$$P(X)=\frac{n}{S_x}f(t)\left[1+\frac{As}{6}(t^3-3t)\right]. \qquad (51)$$

При этом, как показывает формула (51), теоретические частоты f' определяют умножением $[1+(As/6)(t^3-3t)]$ на величину $(n/s_x)(t)$.

Пример 9. Возвращаясь к материалам примера 7 о длине хвоинок сосны, в котором были получены значительные величины As и Ex, построим сглаживающую кривую типа распределения Шарлье с учетом только асимметричности распределения и затем с учетом одновременно As и Ex. Характеристики этого ряда: $s_x=2{,}047$; $As=-0{,}5560$, т. е. распределение имеет заметно выраженную левостороннюю асимметрию. Рассчитать теоретические (выравнивающие) частоты для этого распределения. Расчет приведен в табл. 31. Здесь $b_1=\Sigma fa/n=6{,}765$; $As/6=-0{,}093$; $n/s_x=97{,}7=98$. Остальные действия понятны из табл. 31.

Так как ряд распределения длины хвои у сосны обыкновенной имеет не только левостороннюю асимметрию, но и значительный положительный эксцесс ($Ex=+0{,}8720$), следует для нахождения выравнивающих частот этого ряда применить формулу (51). Расчет выравнивающих частот приведен в табл. 32.

В данном случае $Ex/24=+0{,}036$. Остальные величины объяснены выше.

Таблица 3

x_i	f_i	a	fa	$a-b_1$	$\frac{(a-b_1)}{S_x}=t$	t^3	$f(t)$	$\frac{n}{S_x}f(t)$	$1+\frac{A_s}{6}\times(t^3-3t)$	Произведение граф $9\cdot 10=f'$	
1	2	3	4	5	6	7	8	9	10	11	12
125	2	0	0	−6,765	−3,30	−35,94	0,0017	0,17	+3,42	0,58	1
175	2	1	2	−5,765	−2,82	−22,43	0,0075	0,74	+3,30	2,44	2
225	4	2	8	−4,765	−2,33	−12,65	0,026	2,55	+1,53	3,90	4
275	5	3	15	−3,765	−1,84	−6,23	0,073	7,15	+1,07	7,65	8
325	7	4	28	−2,765	−1,35	−2,46	0,160	15,68	+0,85	13,33	13
375	25	5	125	−1,765	−0,86	−0,64	0,275	26,95	+0,82	22,10	22
425	39	6	234	−0,765	−0,37	−0,051	0,372	36,46	+0,90	32,81	33
475	46	7	322	+0,235	+0,12	+0,17	0,396	38,81	+1,02	39,59	40
525	31	8	248	+1,235	+0,60	+0,22	0,333	32,63	+1,15	37,43	37
575	23	9	207	+2,235	+1,09	+1,30	0,220	21,56	+1,18	25,44	25
625	13	10	130	+3,235	+1,58	+3,94	0,114	11,17	+1,07	11,95	12
675	2	11	22	+4,235	+2,07	+8,87	0,047	4,61	+0,75	3,46	3
725	1	12	12	+5,235	+2,56	+16,78	0,015	1,47	+0,15	0,23	0
Сумма	200	—	1353	—	—	—	—	—	—	200,9	200

Таблица 3'

x_i	f_i	a	$\frac{a-b_1}{S_x}=t$	t^2	t^4	$\frac{n}{S_x}f(t)$	$1+\frac{A_s}{6}\times(t^3-3t)$	$\frac{Ex}{24}\times(t^4-6t^2+3)$	Произведение граф $7(8+9)=f'$	
1	2	3	4	5	6	7	8	9	10	11
125	2	0	−3,30	10,89	118,59	0,17	3,42	+2,020	0,93	1
175	2	1	−2,82	7,95	63,24	0,74	3,30	+0,667	2,94	3
225	4	2	−2,33	5,43	29,47	2,55	1,53	−0,004	3,89	4
275	5	3	−1,84	3,39	11,46	7,15	1,07	−0,211	6,14	6
325	7	4	−1,35	1,82	3,32	15,68	0,85	−0,166	10,73	11
375	25	5	−0,86	0,74	0,55	26,95	0,82	−0,032	21,24	21
425	39	6	−0,37	0,137	0,187	36,46	0,90	+0,085	35,91	36
475	46	7	+0,12	0,014	0,021	38,81	1,02	+0,106	43,70	44
525	31	8	+0,60	0,36	0,13	32,63	1,15	+0,035	38,57	38
575	23	9	+1,09	1,19	1,41	21,56	1,18	−0,098	23,33	23
625	13	10	+1,58	2,50	6,23	11,17	1,07	−0,208	9,63	10
675	2	11	+2,07	4,28	18,36	4,61	0,75	−0,156	2,74	3
725	1	12	+2,56	6,55	42,95	1,47	0,15	+0,239	0,58	0
Сумма	200	—	—	—	—	—	—	—	200,33	200

При сравнении эмпирических частот с частотами, вычисленными по формулам (50) и (51), видно, что они неплохо согласуются друг с другом. Более наглядное представление об этом дает рис. 18, на котором изображены эмпирическая (ломаная)

Рис. 18. Эмпирическая (*1*) и вычисленные по формуле Шарлье (*2, 3*) кривые распределения длины хвои сосны обыкновенной

и вычисленные (плавно идущие) вариационные кривые. Причем частоты, вычисленные по формуле (50), лучше согласуются с эмпирическими частотами, чем частоты, вычисленные по формуле (51), что, по-видимому, связано с более полным учетом информации о форме распределения при подборе теоретической кривой.

* * *

Существуют иные способы работы с кривыми распределения, имеющими асимметрию и эксцесс. С одним из них, опирающимся на использование системы кривых Пирсона, можно ознакомиться в книге А. К. Митропольского (1971). Другой путь, часто пригодный в тех случаях, когда наблюдается значительная правосторонняя асимметрия, заключается в использовании нормализующих преобразований, когда исходные величины признака x трансформируются в виде $y=\psi(x)$. Конкретный вид функции преобразования должен быть таким, чтобы трансформированная величина y имела нормальное распределение. Среди различных возможных преобразующих функций наибольшее употребление нашла так называемая *формула лог-*

нормальной трансформации

$$y = \lg(x + x_0),$$

где x_0 — неизвестный параметр, значение которого подбирают так, чтобы преобразованные значения y имели распределение, максимально близкое к нормальному виду. Из всех существующих способов определения x_0 наиболее приемлемым является метод квантилей, описанный Е. И. Фортунатовой[1], с работой которой, написанной весьма доступно и снабженной числовыми примерами, желательно ознакомиться читателю.

ГЛАВА IV

ВЫБОРОЧНЫЙ МЕТОД И ОЦЕНКА ГЕНЕРАЛЬНЫХ ПАРАМЕТРОВ

IV.1. ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И ВЫБОРКА

Наблюдения над биологическими объектами могут охватывать все члены изучаемой совокупности без единого исключения или ограничиваться обследованием лишь некоторой части членов данной совокупности. В первом случае наблюдения называют *полными* или *сплошными,* во втором — *частичными* или *выборочными.* Полное обследование совокупности позволяет получать исчерпывающую информацию об изучаемом объекте, в чем и заключается преимущество этого способа перед способом выборочного наблюдения. Однако к сплошному наблюдению прибегают редко, так как эта работа сопряжена с большими затратами времени и труда, а также в силу практической невозможности или нецелесообразности проведения такой работы. Невозможно, например, учесть всех обитателей зоо- или фитопланктона даже небольшого водоема, потому что их численность практически необозрима. Нецелесообразно высевать всю партию семян для того, чтобы определить их всхожесть. В подавляющем большинстве случаев вместо сплошного наблюдения изучению подвергают некоторую часть обследуемой совокупности, по которой и судят о ее состоянии в целом.

Совокупность, из которой отбирают определенную часть ее членов для совместного изучения, называют *генеральной.* Отобранная тем или иным способом часть генеральной совокупности получила название *выборочной совокупности* или *выборки.* Общую сумму членов генеральной совокупности называют ее объемом и обозначают буквой N.

---

[1] См.: *Фортунатова Е. И.* О преобразовании ассиметричного распределения в нормальное // Вопр. антропологии. 1966. № 23. С. 49—65.

Теоретически объем генеральной совокупности ничем не ограничен ($N \to \infty$), т. е. генеральную совокупность представляют как бесконечно большое множество относительно однородных единиц или членов, составляющих ее содержание. Практически же объем генеральной совокупности всегда ограничен и может быть различным в зависимости от объекта наблюдения и той задачи, которую приходится решать. Например, при определении продуктивности животных той или иной породы или вида генеральную совокупность составят все особи данной породы или вида. Если же вопрос о продуктивности животных решают в зоне данной области или района, то генеральную совокупность составят все животные изучаемой породы, распространенной в данной области или районе.

Объем выборки, обозначаемый буквой *n,* может быть и большим, и малым, но он не может содержать менее двух единиц. Выборочный метод — основной при изучении статистических совокупностей. Его преимущество перед полным учетом всех членов генеральной совокупности заключается в том, что он сокращает время и затраты труда (за счет уменьшения числа наблюдений), а главное — позволяет получать информацию о таких групповых объектах, сплошное обследование которых практически невозможно или нецелесообразно.

Основное требование, предъявляемое к любой выборке, сводится к получению наиболее полной информации о состоянии генеральной совокупности, из которой выборка взята. Опыт показал, что правильно отобранная часть генеральной совокупности, т. е. выборка, довольно хорошо отображает структуру генеральной совокупности. Однако полного совпадения выборочных показателей с характеристиками генеральной совокупности, как правило, не бывает. Чтобы выборка наиболее полно отображала структуру генеральной совокупности, она должна быть достаточно представительной, или *репрезентативной* (от лат. represento — представляю). Репрезентативность выборки достигается способом *рандомизации* (от англ. random — случай) или случайным отбором вариант из генеральной совокупности, что обеспечивает равную возможность для всех членов генеральной совокупности попасть в состав выборки.

Существует два основных способа отбора вариант из генеральной совокупности: повторный и бесповторный. *Повторный отбор* производят по схеме «возвращения» учтенных единиц в генеральную совокупность, так что одна и та же единица может попасть в выборку повторно. При *бесповторном отборе* учтенные единицы не возвращаются в генеральную совокупность, каждая отобранная единица регистрируется только один раз. Повторный отбор не влияет на состав генеральной совокупности, и возможность каждой единицы попасть в выборку не меняется. При бесповторном отборе возможность единиц, состав-

ляющих генеральную совокупность, попасть в выборку меняется, так как каждый предшествующий отбор влияет на результаты последующего, а также и на состав генеральной совокупности, который тоже претерпевает изменения. В практике обычно применяют *бесповторный случайный отбор*. Так, если измеряют рост мужчин призывного возраста, то, измерив одного из них, вторично его уже не измеряют. Случайный повторный отбор служит теоретической моделью, с помощью которой изучают процессы, совершающиеся в статистических совокупностях, что имеет определенное познавательное значение.

Идеальный случайный отбор производится по методу жеребьевки или лотереи, а также с помощью таблицы случайных чисел, позволяющих полностью исключить субъективное влияние на состав выборки. Сущность этого метода заключается в следующем. На численно ограниченной, но довольно большой искусственной модели генеральной совокупности способом повторного случайного отбора образуется ряд чисел, которые заносят в таблицу таким образом, чтобы они имели одинаковое количество цифр. Этим облегчается использование такой таблицы в практических целях. Например, при трехзначности чисел цифру 8 заносят в таблицу в виде 008, а число 69 — в виде 069 и т. д. Числа записывают в таблицу в случайном порядке, поэтому ее и называют *таблицей случайных чисел*. Такая четырехзначная таблица помещена в Приложении (табл. IV).

Как пользоваться этой таблицей? Пусть из общего числа 120 животных, содержащихся в виварии, нужно отобрать для опыта 10 особей. Для того чтобы отбор был действительно случайным, исключающим субъективные влияния на состав выборки, необходимо поступить следующим образом. Всем животным вивария или только животным той группы, из которой намечено отобрать 10 особей, присваивают номера от 1 до n. Затем в таблице случайных чисел находят десять таких, которые не превышают n. Пусть $n=120$. Пользуясь табл. IV, условимся учитывать первые три цифры в каждом столбце этой таблицы (хотя можно исходить и из другого условия). В первом столбце находят числа 0905 и 0912. Согласно условию, это дает числа 90 и 91. Других нужных чисел в этом столбце нет. Во втором столбце таблицы находят числа 47 и 41. В третьем столбце обнаруживают числа 62, 84, 50 и 31. В четвертом столбце отыскивают остальные два числа: 39 и 87. Всего получилось десять чисел: 90, 91, 47, 41, 62, 84, 50, 31, 39 и 87. Особей с такими номерами включают в состав экспериментальной группы.

Наряду с простым случайным отбором в практике применяют и другие виды выборки из генеральной совокупности. К ним относится типический, серийный и механический отбор. *Типический отбор* используют в тех случаях, когда генеральная со-

вокупность расчленяется на отдельные (типические) группы. Например, в хозяйстве среди крупного рогатого скота находятся первотелки, группы коров по второму, третьему и другим отелам. В таких случаях из каждой группы случайным способом отбирают одинаковое, а чаще пропорциональное число единиц. Затем вычисляют групповые характеристики, объединяемые в общую характеристику генеральной совокупности.

При *серийном отборе,* как и при типическом, генеральную совокупность предварительно делят на группы (серии, гнезда), образуемые обычно по территориальному принципу. Затем по усмотрению исследователя из общего количества серий или гнезд отбирают некоторое их число для совместной обработки. При этом серии могут быть как равночисленными, так и состоять из разиого числа единиц. Например, из 30 групп подростков в возрасте от 14 до 15 лет намечено обследовать выборочно шесть групп. Членов этих групп и объединяют для совместного изучения. Таким образом, в отличие от типического отбора при серийной выборке из генеральной совокупности извлекают не отдельные единицы, а целые серии или гнезда относительно однородных единиц.

При *механическом отборе* генеральная совокупность разбивается на несколько равных частей или групп. Затем из каждой группы случайным способом отбирают по одной единице. Например, при обследовании посева ржи на урожайность намечено отобрать 100 растений (или колосьев). В таком случае поле ржи должно быть разбито на сто равных делянок. Следовательно, при механическом отборе число единиц равно численности групп, на которые разбита генеральная совокупность. Механический отбор может производиться и по другой схеме, когда в выборку попадает каждая десятая, сотая и т. п. единица генеральной совокупности. Например, при проведении ботанических или зоологических экскурсий можно регистрировать каждый пятый, десятый и т. п. экземпляр встреченных растений или животных данного вида.

Кроме типического, серийного и механического отбора в практике применяют и другие разновидности случайной выборки.

IV.2. ТОЧЕЧНЫЕ ОЦЕНКИ

Числовые показатели, характеризующие генеральную совокупность, называют *параметрами,* а числовые показатели, характеризующне выборку,— *выборочными характеристиками* или *статистиками.* Выборочные характеристики являются приближенными оценками генеральных параметров. Это величины случайные, варьирующие вокруг своих параметров. Оценки ге-

неральных параметров по выборочным характеристикам могут быть точечными и интервальными.

Генеральные характеристики, или параметры, принято обозначать буквами греческого алфавита, а выборочные характеристики — латинского. Выборочная средняя $\bar{x}$ является оценкой генеральной средней μ, выборочная дисперсия s_x^2 — оценкой генеральной дисперсии σ_x^2, а среднее квадратическое отклонение s_x — оценкой стандартного отклонения σ_x, характеризующего генеральную совокупность. Это *точечные оценки,* представляющие собой не интервалы, а числа («точки»), вычисляемые по случайной выборке.

Требования, предъявляемые к точечным оценкам. Выборочные характеристики как величины случайные, варьирующие вокруг своих генеральных параметров, в основном не совпадают с ними по абсолютной величине. Оценки должны удовлетворять по меньшей мере следующим требованиям: быть состоятельными, эффективными и несмещенными.

Для пояснения смысла этих свойств необходимо рассмотреть понятие выборочного распределения некоторой статистики. Пусть из бесконечно большой генеральной совокупности случайным образом извлекается большое число выборок, каждая из которых включает одно и то же количество наблюдений n. В каждой из этих выборок вычисляют значение статистики u. В силу случайных причин эти величины будут варьировать, образуя некоторое распределение, которое называют *выборочным распределением статистики.*

В тех случаях, когда распределение анализируемого признака не слишком сильно отличается от нормального вида, а объем выборок не слишком мал, очень часто выборочные распределения многих статистик оказываются нормальными. Поэтому их свойства можно описать только двумя параметрами: математическим ожиданием статистики μ_u и ее дисперсией σ_u^2.

Точечная оценка статистики называется *состоятельной,* если при увеличении объема выборки она стремится к величине генерального параметра. Так, для генеральной средней μ состоятельной оценкой является выборочная средняя $\bar{x}$, для генеральной дисперсии σ_x^2 состоятельной оценкой будет выборочная дисперсия s_x^2. Точечная оценка называется *эффективной,* если она имеет наименьшую дисперсию выборочного распределения по сравнению с другими аналогичными оценками, т. е. обнаруживает наименьшую случайную вариацию. Так, из трех показателей, описывающих положение центра нормального распределения некоторого признака X (средней арифметической, медианы и моды), наиболее эффективной оказывается первая $\bar{x}$, наименее эффективной — последняя Mo, так как для дисперсий этих оценок характерно $\sigma^2_{\bar{x}} < \sigma^2_{Me} < \sigma^2_{Mo}$. Оценка называется *несмещенной,* если математическое ожидание ее выборочных

ного распределения совпадает со значением генерального параметра.

Выборочная средняя является несмещенной оценкой генеральной средней, тогда как выборочная дисперсия представляет собой смещенную оценку относительно генерального параметра на величину $n/(n-1)$. Чтобы получить несмещенную оценку генеральной дисперсии, нужно при вычислении выборочной дисперсии, а следовательно, и среднего квадратического отклонения сумму квадратов отклонений (девиату) относить не к числу наблюдений n, а к числу степеней свободы ($k=n-1$).

Статистические ошибки. Выборочные характеристики, как правило, не совпадают по абсолютной величине с соответствующими генеральными параметрами. Величину отклонения выборочного показателя от его генерального параметра называют *статистической ошибкой* или *ошибкой репрезентативности*. Статистические ошибки присущи только выборочным характеристикам, они возникают в процессе отбора вариант из генеральной совокупности.

Для измерения ошибки репрезентативности некоторой статистики может служить *дисперсия выборочного распределения* σ_u^2 или найденное на ее основе значение среднего квадратического отклонения, которое называют также *квадратической ошибкой статистики* σ_u. Его величина показывает, насколько велика случайная вариация отдельных оценок по отношению к центру выборочного распределения, совпадающего со значением генерального параметра, если статистика несмещенная.

Из теории математической статистики известно, что в том случае, когда распределение исходного признака X не слишком сильно отличается от нормального вида, а объем выборки не слишком мал (на практике $n \geqslant 30$), квадратическая ошибка репрезентативности средней арифметической может быть найдена по формуле

$$\sigma_{\bar{x}} \approx s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \sqrt{\frac{s_x^2}{n}}^{1}. \qquad (52)$$

Эта формула трансформируется в несколько рабочих формул, из которых особенно удобны следующие:

$$s_x = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n(n-1)}} = \sqrt{\frac{\sum f_l (x_l - \bar{x})^2}{n(n-1)}} =$$

$$= \sqrt{\frac{1}{n-1}\left[\frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n}\right)^2\right]}. \qquad (53)$$

[1] Ошибку средней арифметической обозначают также буквой m.

Приведенные формулы применяют при вычислении ошибки средней арифметической способом произведений. Они показывают, что при простой случайной выборке величина ошибки зависит как от объема выборки, так и от размаха варьирования признака в генеральной совокупности.

Пример 1. Выше было найдено (см. пример 16 в гл. II), что среднее число поросят в опоросах 64 свиноматок равно 8,25. Определим ошибку для этой средней. Величины $\Sigma f_i x_i = 528$ и $\Sigma f_i x_i^2 = 4574$ находятся в табл. 12. Подставляя их в формулу (53), находим

$$s_{\bar{x}}^2 = \frac{1}{64 \cdot 63}\left(4574 - \frac{528^2}{64}\right) = \frac{1}{4032}(4574 - 4356) = \frac{218}{4032} = 0{,}054.$$

Отсюда $s_{\bar{x}} = \sqrt{0{,}054} = 0{,}233$. Среднюю арифметическую, сопровождаемую ошибкой, записываем так: $\bar{x} \pm s_{\bar{x}} = (8{,}25 \pm 0{,}233)$ поросят.

Если среднюю арифметическую вычисляют способом условной средней A, ее ошибку вычисляем по следующей формуле:

$$s_{\bar{x}} = \sqrt{\frac{\Sigma f a^2 - \frac{(\Sigma f a)^2}{n}}{n(n-1)}} = \sqrt{\frac{1}{n-1}\left[\frac{\Sigma f a^2}{n} - \left(\frac{\Sigma f a}{n}\right)^2\right]}. \tag{54}$$

Пример 2. Выше было найдено (см. табл. 15), что содержание кальция в сыворотке крови обезьян в среднем равно $\bar{x} = 11{,}94$ мг%. Определим ошибку для этой средней. В табл. 15 находим: $\Sigma fa = +67$; $\Sigma fa^2 = 293$; $n = 100$. Подставляем эти данные в формулу (53): $s_{\bar{x}}^2 = \frac{1}{99}\left[\frac{293}{100} - \left(\frac{67}{100}\right)^2\right] = \frac{1}{99}(2{,}93 - 0{,}45) = \frac{2{,}48}{99} = 0{,}025$. Отсюда $s_{\bar{x}} = \sqrt{0{,}025} = 0{,}158$.

При бесповторном отборе вариант из численно ограниченной генеральной совокупности ошибка выборочной средней, определяемая по формуле (53), оказывается несколько завышенной, особенно в тех случаях, когда объем выборки достаточно велик ($n > 25\%$ от N). Учитывая это обстоятельство, К. Пирсон (1898) предложил *поправку* $\sqrt{(N-n)/(N-1)}$, которую в этом случае необходимо вносить в качестве множителя в формулу (52). При этом вместо $(N-n)/(N-1)$ можно использовать приближенную величину $1-n/N$, где n/N — доля выборки, т. е. вычислять ошибку средней по формуле

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}. \tag{55}$$

Чем больше доля выборки, тем сильнее скажется поправка на величине ошибки средней. Если же доля выборки мала, что наиболее часто встречается в практике, поправка оказывается близкой к единице и величина ошибки средней практически не изменится. Поэтому в тех случаях, когда объем генеральной совокупности N достаточно велик по сравнению с объемом выборки n, величина поправки $1-n/N$ будет близка к единице и ею можно пренебречь.

Пример 3. Из общего числа 500 мужчин, подлежащих призыву на военную службу, выборочно изменен рост у 80 человек. Средний рост призывников оказался равен 170 см с дисперсией 66,3. Определим ошибку для этой средней:

$$s_{\bar{x}}=\sqrt{\frac{66,3}{80}\left(1-\frac{80}{500}\right)}=\sqrt{0,829\cdot 0,840}=\sqrt{0,696}=0,834.$$

Если же ошибку средней вычислить без поправки Пирсона, она оказывается следующей:

$$s_{\bar{x}}=\sqrt{66,3/80}=0,829=0,910.$$

Иногда возникает необходимость объединить несколько групповых средних с их ошибками. В таких случаях для определения ошибки общей (невзвешенной) средней $\bar{\bar{x}}$, если выборки равновелики, применяют формулу

$$s_{\bar{\bar{x}}}=\frac{1}{k}\sqrt{s_{\bar{x}}^2+s_{\bar{x}_2}^2+\ldots+s_{\bar{x}_k}^2}, \qquad (56)$$

где k — число слагаемых групповых средних.

Пример 4. На трех независимых равновеликих выборках получены следующие средние: $\bar{x}_1=10,2\pm 0,12$; $\bar{x}_2=11,5\pm 0,18$; $\bar{\bar{x}}=13,1\pm 0,09$. Общая средняя для этих выборок $\bar{\bar{x}}=(1/3)(10,2+11,5+13,1)=11,6$; ее ошибка

$$s_{\bar{\bar{x}}}=(1/3)\sqrt{0,12^2+0,18^2+0,09^2}=(1/3)\sqrt{0,055}=0,078.$$

Ошибку взвешенной средней из суммы частных или групповых средних определяют по формуле (52), в которой вместо s_x^2 берут взвешенную дисперсию $\bar{s}_x^2$; ее определяют по формуле

$$\bar{s}_x^2=\frac{(n_1-1)s_1^2-(n_2-1)s_2^2+\ldots+(n_k-1)s_k^2}{\sum_{l=1}^{k}n_l-\nu}=$$

$$=\frac{\sum_{l=1}^{k}(n_l-1)s_l^2}{\sum_{l=1}^{k}n_l-\nu}, \qquad (57)$$

где n_i — объемы независимых выборок из нормально распределяющейся генеральной совокупности; s_i^2 — дисперсии этих выборок; ν — число ограничений свободы вариации, равное числу независимых выборочных групп. При известном объеме генеральной совокупности N ошибка общей взвешенной средней $\bar{\bar{x}}$ определяется по формуле

$$s_{\bar{\bar{x}}} = \sqrt{\frac{\bar{s}_x^2}{n}\left(1 - \frac{n}{N}\right)}. \qquad (58)$$

Пример 5. Проводили учет урожая пшеницы на корню. Поле было разбито на пять типовых участков (групп). Каждый участок, в свою очередь, был разделен на более мелкие участки (метровки), из которых случайным бесповторным способом отбирали пропорциональное (по объему групп) количество колосьев с последующим взвешиванием массы содержавшихся в них зерен (в граммах). Результаты отбора и их обработка приведены в табл. 33.

Таблица 33

Типовые участки	Численность		Выборочные показатели		$(n_i-1)s_i^2$
	групп N	выборки n	средняя $\bar{x}$	дисперсия s_i^2	
1	100	20	9,8	4,6	87,4
2	200	40	8,5	3,2	124,8
3	400	80	7,7	3,6	284,4
4	150	30	12,1	4,1	118,9
5	150	30	10,2	3,8	110,2
Сумма	1 000	200	—	—	725,7

Взвешенная средняя

$$\bar{\bar{x}} = \frac{20 \cdot 9{,}8 + 40 \cdot 8{,}5 + 80 \cdot 7{,}7 + 30 \cdot 12{,}1 + 30 \cdot 10{,}2}{20 + 40 + 80 + 30 + 30} =$$

$$= \frac{1821{,}0}{200} = 9{,}105 = 9{,}11 \text{ г.}$$

Чтобы определить ошибку этой величины, нужно сперва рассчитать взвешенную дисперсию: $\bar{s}_x^2 = 725{,}7/(200-5) = 3{,}72$. Подставляя нужные величины в формуле (58), находим ошибку взвешенной средней:

$$s_{\bar{\bar{x}}} = \sqrt{\frac{3{,}72}{200}\left(1 - \frac{200}{1000}\right)} = \sqrt{0{,}0186 \cdot 0{,}80} = \sqrt{0{,}01488} = 0{,}122.$$

Статистические ошибки характеризуют варьирование выборочных показателей вокруг своих генеральных параметров. Они обладают теми же свойствами, что и среднее квадратическое отклонение. Чем сильнее варьирует признак, тем больше при

прочих равных условиях будет ошибка выборочных показателей, и, наоборот, при слабом варьировании признака ошибка выборочных показателей окажется меньше. Одно лишь свойство специфично для ошибок репрезентативности: *они уменьшаются при увеличении объема выборки,* т. е. при $n \to \infty$, $s_{\bar{x}} \to 0$. Это свойство статистических ошибок обусловлено действием закона больших чисел, по которому наиболее вероятный результат получается при наибольшем числе испытаний. Отсюда понятно значение ошибки: она указывает на точность, с какой выборочный показатель репрезентирует генеральный параметр. Чем меньше ошибка, тем ближе выборочная характеристика к величине генерального параметра, и, наоборот, чем больше ошибка, тем менее точно выборочная характеристика репрезентирует генеральный параметр.

Показатель точности оценок. Судить о точности, с какой определена та или иная выборочная средняя, позволяет отношение ошибки репрезентативности к своей средней. Этот показатель, обозначаемый символом *Cs* (обычно выражен в процентах), определяют по одной из следующих формул:

$$Cs = \frac{s_{\bar{x}}}{\bar{x}} 100; \tag{59}$$

$$Cs = \frac{Cv}{\sqrt{n}}. \tag{60}$$

Здесь Cv — коэффициент вариации, выраженный в процентах, n — объем выборки.

Пример 6. Сравнивают на точность определения средние: $\bar{x}_1 = (86{,}1 \pm 0{,}7)$ см и $\bar{x}_2 = (17{,}4 \pm 0{,}2)$ г. Так как средние выражены разными единицами, судить по абсолютной величине их ошибок о том, какая из них определена более точно, нельзя. Ответить на этот вопрос позволяет коэффициент *Cs*:

$$Cs = \frac{0{,}7}{86{,}1} 100 = 0{,}81\%; \quad Cs = \frac{0{,}2}{17{,}4} 100 = 1{,}15\%.$$

Из расчетов видно, что первая средняя определена более точно, чем вторая.

Показатель точности *Cs* нашел широкое применение особенно при сравнительной оценке результатов сельскохозяйственных опытов. Точность средних показателей, которыми оценивают результаты наблюдений, считают вполне удовлетворительной, если коэффициент *Cs* не превышает 3—5%.

Коэффициент *Cs* сопровождается ошибкой s_{Cs}, которую определяют по формуле

$$s_{Cs} = Cs \sqrt{\frac{1}{2n} + \left(\frac{Cs}{100}\right)^2}. \tag{61}$$

Ошибками репрезентативности сопровождаются и другие выборочные показатели, из которых необходимо отметить следующие.

Ошибка медианы $s_{Me}=s_{\bar{x}}\sqrt{\pi/2}=1{,}2533s_x/\sqrt{n}.$ **(62)**

Ошибка дисперсии $s_{s^2}=s_x^2/\sqrt{2n}.$ (63)

Ошибка среднего квадратического отклонения $s_s=\frac{s_x}{\sqrt{2n}}.$ (64)

Ошибка коэффициента вариации

$$s_{Cv}=\frac{Cv}{\sqrt{n-1}}\sqrt{\frac{1}{2}+\left(\frac{Cv}{100}\right)^2}\approx\sqrt{\frac{Cv^2}{2n}}. \quad (65)$$

Ошибка выборочной доли $s_p=\sqrt{\frac{p(1-p)}{n}}=\sqrt{pq/n}$ (66)

или с поправкой Пирсона $s_p=\sqrt{\frac{pq}{n}\left(1-\frac{n}{N}\right)}.$ (67)

Ошибка выборочной доли, выраженной в процентах,

$$s_{p\%}=\sqrt{\frac{p(100-p)}{n}}. \quad (68)$$

Ошибка абсолютной частоты $s_m=\sqrt{\frac{m(n-m)}{n}}=\sqrt{np(1-p)}.$ (69)

IV.3. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ

Доверительный интервал для генеральной средней. По известным выборочным характеристикам можно построить интервал, в котором с той или иной вероятностью находится генеральный параметр. Вероятности, признанные достаточными для уверенного суждения о генеральных параметрах на основании известных выборочных показателей, называют *доверительными*. Понятие о доверительных вероятностях предложено Р. Фишером. Оно вытекает из принципа, который положен в основу применения теории вероятностей к решению практических задач. Согласно этому принципу, маловероятные события считают практически невозможными, а события, вероятность которых близка к единице, принимают за почти достоверные. Обычно в качестве доверительных используют вероятности $P_1=0{,}95$; $P_2=0{,}99$ и $P_3=0{,}999$. Это означает, что при оценке генеральных параметров по известным выборочным показателям существует

риск ошибиться в первом случае один раз на 20 испытаний, во втором — один раз на 100 испытаний и в третьем — один раз на 1000 испытаний.

Доверительным вероятностям, как это видно из табл. I Приложений, соответствуют следующие величины нормированных отклонений:

вероятности $P_1 = 0{,}95$ соответствует $t_1 = 1{,}96$;

вероятности $P_2 = 0{,}99$ соответствует $t = 2{,}58$;

вероятности $P_3 = 0{,}999$ соответствует $t_3 = 3{,}29$.

Выбор того или иного порога доверительной вероятности исследователь осуществляет исходя из практических соображений той ответственности, с какой делаются выводы о генеральных параметрах.

С доверительной вероятностью тесно связан *уровень значимости* α, под которым понимают разность $\alpha = 1 - P$. Геометрически (рис. 19) эта величина представляет площадь под нормальной кривой выборочного распределения некоторой статистики, выходящую за пределы той его части, которая включает $P\%$ этой площади. Так, для $t = 1{,}96$ отклонения от центра нормального распределения включают 95% его площади. За пределами этих границ по обе стороны находится по 2,5% указанной площади, составляя тем самым 5%-ный уровень значимости.

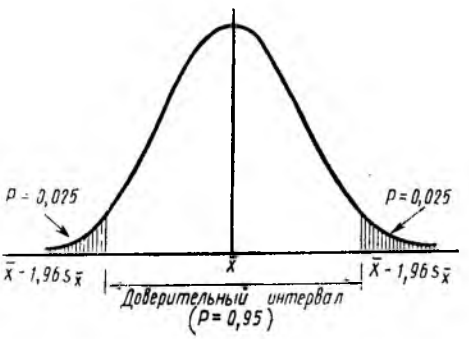


Рис. 19. 95%-ный доверительный интервал в границах от $\bar{x} - 1{,}96 s_{\bar{x}}$ до $\bar{x} + 1{,}96 s_{\bar{x}}$ нормальной кривой

Учитывая, что выборочное распределение некоторой статистики, например средней арифметической величины, при достаточно больших объемах выборок имеет нормальную форму, можно записать выражение

$$-t \leqslant \frac{\bar{x} - \mu}{s_{\bar{x}}} \leqslant t.$$

Это выражение означает, что вероятность того, что средняя $\bar{x}$, найденная по выборке, отклонится случайным образом от центра μ на какую-то долю квадратической ошибки $s_{\bar{x}}$, может быть оценена через нормированное значение по таблицам нормального распределения. Отсюда можно утверждать, что генеральная средняя μ находится с этой вероятностью в интервале

$$\bar{x} - ts_{\bar{x}} \leqslant \mu \leqslant \bar{x} + ts_{\bar{x}} \text{ или}$$

$$\bar{x} = \frac{ts_x}{\sqrt{n}} \leqslant \mu \leqslant \bar{x} + \frac{ts_x}{\sqrt{n}}. \qquad (70)$$

Величины $\bar{x}$ и s_x определяют по выборке, а t зависит только от одного из трех значений доверительной вероятности 0,95; 0,99; 0,999, принимая величины 1,96; 2,58; 3,29.

Пример 7. Распределение кальция в сыворотке крови обезьян, как было установлено выше, характеризуется следующими выборочными показателями: $\bar{x}=11{,}94$ мг%; $s_x=1{,}27$; $n=100$. Построим 95%-ный доверительный интервал для генеральной средней μ этого распределения:

$$11{,}94 - 1{,}96\frac{1{,}27}{\sqrt{100}} \leqslant \mu \leqslant 11{,}94 + 1{,}96\frac{1{,}27}{\sqrt{100}} \text{ или } 11{,}70 \leqslant \mu \leqslant 12{,}18.$$

Следовательно, с вероятностью $P=0{,}95$, или 95%, можно утверждать, что генеральная средняя данного нормального распределения находится между 11,70 и 12,18 мг%. Это довольно узкий доверительный интервал. Можно утверждать, что выборочная средняя $\bar{x}=11{,}94$ мг% является достаточно точной оценкой генерального параметра. На это указывает и показатель точности средней

$$Cs = 100\frac{1{,}27\sqrt{100}}{11{,}94} = 1{,}06\ \%.$$

Доверительный интервал для генеральной дисперсии и стандартного отклонения. Доверительный интервал для дисперсии нормально распределяющейся генеральной совокупности можно представить в таком виде: $P_н \leqslant \sigma^2_x \leqslant P_в$, где $P_н = s_x^2 - ts_x^2 \times \sqrt{2/n}$ — нижняя, а $P_в = s_x^2 + ts_x^2\sqrt{2/n}$ — верхняя границы доверительного интервала.

Пример 8. Определить границы 95%-ного доверительного интервала для генеральной дисперсии распределения кальция (мг%) в сыворотке крови обезьян. Для этого распределения $n=100$ и $s^2_x=1{,}60$. Ошибка выборочной дисперсии $s_{s^2}=1{,}60\times\sqrt{2/100}=1{,}60\cdot 0{,}1414=0{,}226$. Отсюда $P_н=1{,}60-1{,}96\cdot 0{,}226=1{,}60-0{,}44=1{,}16$ и $P_в=1{,}60+0{,}44=2{,}04$. Границы доверительного интервала для стандартного отклонения оказываются следующие: $P_и=\sqrt{1{,}16}=1{,}08$ и $P_в=\sqrt{2{,}04}=1{,}43$.

Более точно доверительный интервал для генеральной дисперсии строят с применением *χ^2(хи-квадрат)-критерия Пирсона* (см. гл. V). Критические (процентные) точки для этого критерия приведены в табл. VII Приложений. Они рассчитаны для

разных уровней значимости α и соответствующих порогов доверительной вероятности P.

При использовании критерия χ^2 для построения доверительного интервала применяют двусторонний уровень значимости, т. е. $\alpha=2{,}5\%$ (для $P_н$) и $P=100-2{,}5=97{,}5\%$ (для $P_в$) 95%-ного интервала. Границы 95%-ного доверительного интервала определяют по формулам

$$P_н=\frac{(n-1)s_x^2}{\chi^2_{2,5}}; \quad P_в=\frac{(n-1)s_x^2}{\chi^2_{97,5}},$$

где s^2_x — выборочная дисперсия, n — объем выборки.

Пример 9. В рассмотренном примере 8 применим описанный способ к определению границ 95%-ного доверительного интервала для генеральной дисперсии ряда распределения кальция (мг%) в сыворотке крови обезьян. Имеем $n=100$; $(n-1)s^2_x=991{,}60=158{,}40$. В табл. VII Приложений для $n-1=99$ и $\alpha=2{,}5\%$ находим $\chi^2=128{,}42$ и для $P=97{,}5\%$ — $\chi^2=73{,}36$. Отсюда $P_н=158{,}40/128{,}42=1{,}23$ и $P_в=158{,}40/73{,}36=2{,}16$. Границы доверительного интервала для стандартного отклонения оказываются следующими: $P_н=\sqrt{1{,}23}=1{,}11$; $P_в=\sqrt{2{,}16}=1{,}47$.

Доверительный интервал для коэффициента вариации. Границы доверительного интервала для генерального коэффициента вариации Cv определяют по следующим формулам:

$$P_н=\frac{Cv}{1+K\sqrt{1+2Cv^2}}; \quad P_в=\frac{Cv}{1-K\sqrt{1+2Cv^2}},$$

где $K=\frac{t}{\sqrt{2(n-1)}}$; $Cv=s_x/\bar{x}$.

Пример 10. Коэффициент вариацин, характеризующий варьирование кальция (мг%) в сыворотке крови обезьян, оказался равным 10,6%, илн $Cv=0{,}106$. Определим границы 95%-ного доверительного интервала для генерального параметра Cv. Предварительно вычисляем величину $K=\frac{1{,}96}{\sqrt{2(100-1)}}=0{,}139$.

Подставляем известные значения в формулы:

$$P_н=\frac{0{,}106}{1+0{,}139\sqrt{1+2(0{,}106)^2}}=\frac{0{,}106}{1+0{,}139\cdot 0{,}11}=$$
$$=\frac{0{,}106}{1{,}141}=0{,}093, \text{ или } 9{,}3\%;$$
$$P_в=\frac{0{,}106}{1-0{,}139\sqrt{1+2(0{,}106)^2}}=\frac{0{,}106}{1-0{,}141}=$$
$$=\frac{0{,}106}{0{,}859}=0{,}123, \text{ или } 12{,}3\%.$$

Это означает, что при повторных выборках в данных условиях коэффициент вариации не превысит 12,3% и не окажется ниже 9,3%. Довольно узкий доверительный интервал (9,3—12,3%) указывает на то, что выборочный коэффициент вариации $Cv=10{,}6\%$ достаточно точно репрезентирует генеральный параметр Cv.

Доверительный интервал для доли. *Доля* — это средняя, которая характеризует количество единиц в выборке, имеющих учитываемый признак. Общее число таких единиц в генеральной совокупности составляет *генеральную долю* ($\bar{p}=m/N$). Границы доверительного интервала для генеральной доли — $P_н \leqslant \bar{P} \leqslant P_в$ — определяют так же, как и для генеральной средней рядовой изменчивости, т. е. $P_н=p-ts_m$ и $P_в=p+ts_m$. Эти формулы применяют тогда, когда выборочные доли p и q равны между собой или незначительно отклоняются от 50%-ной численности групп. Если же это условие не выполняется (при $75\% < p \leqslant 25\%$), доверительные границы для генеральной доли следует определять по формуле

$$P=\frac{1}{n+t^2}\left[\left(m+\frac{t^2}{2}\right)\pm t^2\sqrt{\frac{m(n-m)}{n}+\frac{t^2}{4}}\right], \qquad (71)$$

где n — число наблюдений; m — абсолютная численность одной из групп; t — нормированное отклонение, определяемое по значению вероятности (P).

Пример 11. В рабочем поселке с N числом жителей способом бесповторного случайного отбора было обследовано 150 человек, из которых 20 оказались больными. Определить вероятные границы генеральной доли больных в данном населенном пункте. Выборочная доля больных $p=m/n=20/150=0{,}13$, или 13%. Исходим из $P=0{,}95$ и соответственно $t=1{,}96\approx 2$. Подставляя известные данные в формулу (71), находим

$$P=\frac{1}{150+4}\left[\left(20+\frac{4}{2}\right)\pm 2\sqrt{\frac{20(150-20)}{150}+\frac{4}{4}}\right]=$$

$$=\frac{1}{154}\left(22\pm 2\sqrt{18{,}3}\right)=\frac{1}{154}(22\pm 8{,}56)=0{,}143\pm 0{,}056.$$

Отсюда границы доверительного интервала оказываются следующими: $P_н=0{,}143-0{,}056=0{,}087$, или 8,7%; $P_в=0{,}143+0{,}056=0{,}199$, или 19,9%.

Таким образом, с вероятностью $P=95\%$ можно утверждать, что генеральная доля больных находится между границами от 8,7 до 19,9% от общего числа лиц N, проживающих в данном населенном пункте.

ГЛАВА V

КРИТЕРИИ ДОСТОВЕРНОСТИ ОЦЕНОК

V.1. СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ И ИХ ПРОВЕРКА

В гл. IV было показано, что выборочные характеристики являются оценками генеральных параметров, которые, как правило, остаются неизвестными. Там же описаны точечные и интервальные способы оценки неизвестных параметров по значениям выборочных характеристик[1].

Ниже будут обсуждаться *сравнительные оценки* генеральных параметров по разности, наблюдаемой между сравниваемыми выборками. Это важно, так как ни одно исследование не обходится без сравнений. Сравнивать приходится данные опыта с контролем, урожайность одной культуры с урожайностью другой, продуктивность одной группы животных с продуктивностью другой и т. д.

О преимуществе той или иной из сравниваемых групп судят обычно по разности между средними долями и другими выборочными показателями — величинами случайными, сопровождаемыми ошибками репрезентативности.

Вопрос о достоверности выборочной разности с ее ошибкой приходится решать исходя из той или иной гипотезы, т. е. предположения или допущения относительно параметров сравниваемых групп, которое выражено в терминах вероятности и может быть проверено по выборочным характеристикам.

В области биометрии широкое применение получила так называемая *нулевая гипотеза* (H_0). Сущность ее сводится к предположению, что разница между генеральными параметрами сравниваемых групп равна нулю и что различия, наблюдаемые между выборочными характеристиками, носят не систематический, а исключительно случайный характер. Так, если одна выборка извлечена из нормально распределяющейся совокупности с параметрами μ_x и σ_x, а другая — из совокупности с параметрами μ_y и σ_y, то нулевая гипотеза исходит из того, что $\mu_x = \mu_y$ и $\sigma_x = \sigma_y$, т. е. $\mu_x - \mu_y = 0$ и $\sigma_x - \sigma_y = 0$ (отсюда и название гипотезы — нулевая).

Противоположная нулевой — *альтернативная гипотеза* (H_a) — исходит из предположения, что $\mu_x - \mu_y \neq 0$ и $\sigma_x - \sigma_y \neq 0$.

Для проверки принятой гипотезы, а следовательно, и достоверности оценки генеральных параметров по выборочным данным используют величины, функции распределения которых

---

[1] В настоящем пособии термин «оценка» применяется в двояком смысле: и как собственно оценка, выражаемая *числом*, и как самый процесс оценивания генеральных параметров по выборочным показателям.

известны. Эти величины, называемые *критериями достоверности,* позволяют в каждом конкретном случае выявить, удовлетворяют ли выборочные показатели принятой гипотезе. Функции распределения указанных величин табулированы, т. е. сведены в специальные таблицы, где содержатся значения функции для разных чисел степеней свободы k или объема выборки n и уровней значимости α.

Уровень значимости, или вероятность ошибки, допускаемой при оценке принятой гипотезы, может различаться. Обычно при проверке статистических гипотез принимают три уровня значимости: 5%-ный (вероятность ошибочной оценки $P=0{,}05$), 1%-ный ($P=0{,}01$) и 0,1%-ный ($P=0{,}001$). В биологических исследованиях часто считают достаточным 5%-ный уровень значимости. При этом нулевую гипотезу не отвергают, если в результате исследования окажется, что вероятность ошибочности оценки относительно правильности принятой гипотезы превышает 5%, т. е. $P>0{,}05$. Если же $P<0{,}05$, то принятую гипотезу следует отвергнуть на взятом уровне (α). Ошибка при этом возможна не более чем в 5% случаев, т. е. она маловероятна.

При более ответственных исследованиях уровень значимости может быть уменьшен до 1 или даже до 0,1%. Трем упомянутым уровням значимости (α) отвечают (при нормальности распределения используемого критерия) нормированные отклонения (t): при α_1 ($P=0{,}05$) нормированное отклонение $t_1=1{,}96$; при α_2 ($P=0{,}01$) — $t_2=2{,}58$; при α_3 ($P=0{,}001$) — $t_3=3{,}29$; и соответственно пороги доверительной вероятности $(1-\alpha)$ равны $P_1=0{,}95$, $P_2=0{,}99$ и $P_3=0{,}999$.

В области биометрии применяют два вида статистических критериев: *параметрические,* построенные на основании параметров данной совокупности (например, $\bar{x}$ и s^2_x) и представляющие функции этих параметров, и *непараметрические,* представляющие собой функции, зависящие непосредственно от вариант данной совокупности с их частотами. Первые служат для проверки гипотез о параметрах совокупностей, распределяемых по нормальному закону, вторые — для проверки рабочих гипотез независимо от формы распределения совокупностей, из которых взяты сравниваемые выборки. Применение параметрических критериев связано с необходимостью вычисления выборочных характеристик — средней величины и показателей вариации, тогда как при использовании непараметрических критериев такая необходимость отпадает.

При нормальном распределении признака параметрические критерии обладают большей мощностью, чем непараметрические критерии. Они способны более безошибочно отвергать нулевую гипотезу, если она не верна. Поэтому во всех случаях, когда сравниваемые выборки взяты из нормально распределя-

ющихся совокупностей, следует отдавать предпочтение параметрическим критериям.

В случае очень больших отличий распределений признака от нормального вида следует применять непараметрические критерии, которые в этой ситуации оказываются часто более мощными. В ситуациях, когда варьирующие признаки выражаются не числами, а условными знаками, применение непараметрических критериев оказывается единственно возможным.

Из параметрических критериев в биометрии применяют *t-критерий Стьюдента* и *F*-критерий Фишера. Первый используют для сравнительной оценки средних величин, второй — для оценки дисперсий. Ниже рассмотрен отдельно каждый из этих критериев.

V.2. ПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ

***t*-критерий Стьюдента (*t*-распределение).** Использование формулы Гаусса—Лапласа (44) для сравнительной оценки средних величин затруднено тем, что в качестве аргументов в эту формулу входят генеральные параметры μ и σ (которые, как правило, остаются неизвестными), тогда как при обработке и сравнении выборочных групп приходится пользоваться не генеральными, а выборочными характеристиками $\bar{x}$ и s_x. Учитывая это обстоятельство, английский математик В. Госсет (печатавшийся под псевдонимом Стьюдент), в 1908 г. нашел закон распределения величины $t=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, в которой генеральный параметр σ заменен на его выборочную характеристику s_x, т. е. нашел *закон распределения значений*

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}. \qquad (72)$$

Оказалось, что отношение разности между выборочной и генеральной средними к ошибке выборочной средней непрерывно распределяется согласно следующей формуле:

$$f(t)=C\left(1+\frac{t^2}{n-1}\right)^{-\frac{n-1}{2}} \quad \text{для} \quad -\infty<t<+\infty,$$

где C — константа, зависящая только от числа степеней свободы $k=n-1$.

Открытый Стьюдентом и теоретически обоснованный Р. Фишером *закон t-распределения* служит основой так называемой теории малой выборки, которая характеризует распределение выборочных средних в нормально распределяющейся совокупности в зависимости от объема выборки. *t*-распределение зави-

сит только от числа степеней свободы $k=n-1$, причем с увеличением объема выборки n t-распределение быстро приближается к нормальному с параметрами $\mu=0$ и $\sigma=1$ и уже при $n\geqslant 30$ не отличается от него. Это видно из табл. 34, в которой наряду с табулированными значениями функции нормального распределения приведены табулированные значения t-распределения для разных значений t.

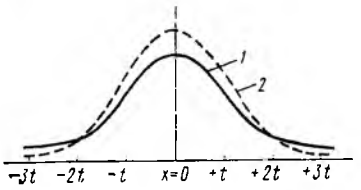


Рис. 20. Кривая t-распределения (*1*) при $n=3$ на фоне нормальной кривой (*2*)

Более наглядное представление о характере t-распределения дает рис. 20, на котором на фоне нормальной кривой изображена (более пологая) кривая t-распределения при $n=3$. t-распределение симметрично и отражает специфику распределения средней арифметической в случае малой выборки в зависимости от ее объема (n). Для выборок, объем которых превышает 30 единиц, величина t распределяется нормально и не зависит от числа наблюдений. Если же $n<30$, характер t-распределения находится в зависимости от числа наблюдений n.

Таблица 3

Распределение	Нормированное отклонение t						
	0,5	1,0	1,5	2,0	2,5	3,0	3,5
Нормальное	383	683	866	955	988	997	9995
Стьюдента при							
$n=3$	333	577	728	816	870	905	927
$n=20$	377	670	850	940	978	993	998
$n=30$	383	683	866	955	988	997	9995

Примечание. Значения фуикции даны числами после запятой.

Для практического использования t-распределения составлена специальная таблица (см. табл. V Приложений), в которой содержатся критические точки (t_{st}) (от англ. standard — норма, образец) для разных уровней значимости α и чисел степеней свободы k. Как пользоваться этой таблицей в разных случаях применения t-критерия, будет показано ниже.

Оценка разности средних. Сравнивая друг с другом две независимые выборки, взятые из нормально распределяющихся совокупностей с параметрами μ_1 и μ_2, можно предположить, что $\mu_1-\mu_2=D$, а дисперсия этой разности σ^2_D. Значения генеральных параметров неизвестны, однако несложно найти величины выборочных средних и разность между ними

$(\bar{x}_1 - \bar{x}_2) = d$. Нулевая гипотеза сводится к предположению, что $\mu_1 = \mu_2$. Критерием для проверки H_0-гипотезы служит отношение

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 + \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}},$$

где t — переменная величина, следующая t-распределению Стьюдента с числом степеней свободы $k = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$, а $s_{\bar{x}_1 - \bar{x}_2}$ — ошибка указанной разности, обозначаемая в дальнейшем символом s_d.

Так как, согласно H_0-гипотезе, $\mu_1 - \mu_2 = 0$, то t-критерий выражается в виде отношения разности выборочных средних к своей ошибке, т. е.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_d} = \frac{d}{s_d}.$$

H_0-гипотезу отвергают, если фактически установленная величина t-критерия (обозначаемая символом t_Φ) превзойдет или окажется равной критическому (стандартному) значению t_{st} этой величины для принятого уровня значимости α и числа степеней свободы $k = n_1 + n_2 - 2$, т. е. при условии $t_\Phi \geqslant t_{st}$.

Ошибку разности средних s_d определяют по следующим формулам:

а) для равночисленных выборок, т. е. при $n_1 = n_2$,

$$s_d = \sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}} = \sqrt{\frac{\sum (x_i - \bar{x}_1)^2}{n(n-1)} + \frac{\sum (x_i - \bar{x}_2)^2}{n(n-1)}} =$$

$$= \sqrt{\frac{\sum (x_i - \bar{x}_1)^2 + \sum (x_i - \bar{x}_2)^2}{(n-1)\,n}}; \qquad (73)$$

б) для неравночисленных выборок, т. е. при $n_1 \neq n_2$,

$$s_d = \sqrt{\frac{(n_1 - 1)\,s_1^2 + (n_2 - 1)\,s_2^2}{n_1 + n_2 - 2}\left(\frac{n_1 + n_2}{n_1 n_2}\right)} =$$

$$= \sqrt{\frac{\sum (x_i - \bar{x}_1)^2 + \sum (x_i - \bar{x}_2)^2}{n_1 + n_2 - 2}\left(\frac{n_1 + n_2}{n_1 n_2}\right)}. \qquad (74)$$

В этой формуле вместо $\left(\frac{n_1 + n_2}{n_1 n_2}\right)$ можно использовать $\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.

Пример 1. Изучали влияние кобальта на массу тела кроликов. Опыт проводнли на двух группах животных: опытной и контрольной. Были исследованы кролики в возрасте от полу-

тора до двух месяцев, массой тела 500—600 г. Опыт продолжался полтора месяца. Животных обеих групп содержали на одном и том же кормовом рационе. Однако опытные кролики в отличие от контрольных ежедневно получали добавку к рациону в виде водного раствора по 0,06 г хлористого кобальта на 1 кг живой массы тела. За время опыта животные дали следующие прибавки живой массы тела (табл. 35).

Таблица 3

Привесы, г		Отклонения от средней арифметической		Квадраты отклонений	
опыт	контроль	опыт	контроль	опыт	контроль
580	504	58	22	3364	484
692	560	54	34	2916	1 156
700	420	62	106	3844	11 236
621	600	17	74	289	5 476
640	580	2	54	4	2 916
561	530	77	4	5929	16
680	490	42	36	1764	1 296
630	580	8	54	64	2 916
	470		56		3 136
$\Sigma=5104$	$\Sigma=4734$	—	—	$\Sigma=18\,174$	$\Sigma=28\,632$
$\bar{x}_1=638$	$\bar{x}_2=526$	—	—	$\Sigma=46\,806$	

Средние арифметические привесов: в опыте $\bar{x}_1=5104/8=638$ г, в контроле $\bar{x}_2=4734/9=526$ г. Разница $|\bar{x}_1-\bar{x}_2|=d=112$ г. Чтобы установить, достоверна или случайна эта разница, нужно определить ошибку разности средних по формуле (74):

$$s_d=\sqrt{\frac{46\cdot806}{8+7}\,\frac{9+8}{9\cdot8}}=\sqrt{736{,}8}=27{,}14.$$

Отсюда $t_\Phi=112/27{,}13=4{,}1$. По табл. V Приложений для 1%-ного уровня значимости и числа степеней свободы $k=9+8-2=15$ находим $t_{st}=2{,}95$. Так как $t_\Phi>t_{st}$, нулевая гипотеза опровергается на высоком уровне значимости ($P<0{,}01$). Разница между средними величинами опыта и контроля оказалась в высшей степени достоверной.

Пример 2. На двух группах лабораторных мышей — опытной ($n_1=9$) и контрольной ($n_2=11$) — изучали воздействие на организм нового препарата. После месячных испытаний масса

тела животных, выраженная в граммах, варьировала следующим образом:

В опытной группе 80, 76, 75, 64, 70, 68, 72, 79, 83	$\bar{x}_1 = 74{,}1$
В контрольной группе 70, 78, 60, 80, 62, 68, 73, 60, 71, 66, 69	$\bar{x}_2 = 68{,}8$

Разница между средними $|\bar{x}_1 - \bar{x}_2| = 5{,}3$ г. Для определения ошибки этой разности предварительно рассчитаем девиаты: $D_1 = \Sigma(x_i - \bar{x})^2 = \Sigma x^2 - (\Sigma x)^2/n = (80^2 + 76^2 + \ldots + 83^2) - 667^2/9 = 302{,}89$ и $D_2 = (70^2 + 78^2 + 60^2 + \ldots + 69^2) - 757^2/11 = 443{,}64$. Отсюда ошибка разности средних выразится величиной $s_d^2 = \frac{302{,}80 + 443{,}64}{9 + 11 - 2}\left(\frac{9 + 11}{9 \cdot 12}\right) = \frac{14930{,}6}{1782{,}0} = 8{,}38$ и $s_d = \sqrt{8{,}38} = 2{,}89$. Критерий $t_\Phi = 5{,}30/2{,}89 = 1{,}83$. Для $k = 9 + 11 - 2 = 18$ и 5%-ного уровня значимости в табл. V Приложений находим $t_{st} = 2{,}10$. Так как $t_\Phi < t_{st}$, нулевая гипотеза остается в силе.

Неопровержение H_0-гипотезы нельзя рассматривать как доказательство равенства между неизвестными параметрами совокупностей, из которых извлечены сравниваемые выборки. В таких случаях вопрос о преимуществе одной статистической совокупности перед другой остается открытым. Ведь не исключено, что при повторных испытаниях H_0-гипотеза может оказаться несостоятельной. Более того, и в тех случаях, когда H_0-гипотеза опровергается, не следует спешить с окончательным выводом.

Следует заметить, что вышеизложенное применение t-критерия предполагает, что дисперсии сравниваемых групп одинаковы: $\sigma^2_1 = \sigma^2_2$. Если это не так, то величину критерия находят по формуле

$$t = \frac{|\bar{x}_1 - \bar{x}|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

а число степеней свободы — по следующим формулам:

а) при $n_1 = n_2$ $$k = n - 1 + \frac{2n - 2}{s_1^2/s_2^2 + s_2^2/s_1^2}. \qquad (75)$$

б) при $n_1 \neq n_2$ $$k = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \Big/ \left[\frac{(s_1^2/n_1)^2}{n_1 + 1} + \frac{(s_2^2/n_2)^2}{n_2 + 1}\right] - 2. \qquad (76)$$

Так, при изучении влияния кобальта на массу тела кроликов (см. пример 1) дисперсии равны $s_1^2 = \frac{1}{8 - 1} 18\,174 = 2596{,}3$ и $s_2^2 = \frac{1}{9 - 1} 28\,632 = 3579{,}0$ (см. табл. 35). Видно, что $s^2_2 > s^2_1$.

Следовательно, величину критерия необходимо определять с учетом неравенства дисперсий. Предварительно найдем $s^2_1/n_1 = 2596{,}3/8 = 324{,}54$ и $s^2_2/n_2 = 3579{,}0/9 = 397{,}67$. Величина t-критерия равна $t_\Phi = 638 - 526/\sqrt{324{,}54 + 397{,}67} = 4{,}17$. Затем определяем $(s^2_1/n_1)^2/(n_1+1) = 324{,}54^2/9 = 11702{,}6$ и $(s^2_2/n_2)^2/(n_2+1) = 397{,}67^2/10 = 15813{,}9$. В результате $k = 722{,}2^2/27516{,}6 - 2 \approx 17$. Для $k = 17$ и $\alpha = 1\%$ в табл. V Приложений находим $t_{st} = 2{,}90$. Так как $t_\Phi = 4{,}17 > t_{st} = 2{,}90$, то H_0-гипотеза отвергается.

Правильное применение t-критерия предполагает нормальное распределение совокупностей, из которых извлечены сравниваемые выборки, и равенство генеральных дисперсий. Если эти условия не выполняются, то t-критерий применять не следует. В таких случаях более эффективными будут непараметрические критерии.

Оценка средней разности между выборками с попарно связанными вариантами. Сравниваемые выборки нередко представляют собой ряды попарно связанных вариант, т. е. являются *зависимыми выборками*. В таких случаях оценкой разности между генеральными средствами $\mu_1 - \mu_2 = D$ будет *средняя разность*, определяемая из суммы разностей между попарно связанными вариантами сравниваемых групп, т. е.

$$d = \frac{\sum d_l}{n}. \tag{77}$$

Оценкой генеральной дисперсии σ^2 разности средних $\mu_1 - \mu_2 = D$ будет *выборочная дисперсия*

$$s^2 = \frac{\sum (d_l - \bar{d})^2}{n-1}. \tag{78}$$

В формулах (77) и (78) n — число парных наблюдений; $d_i = x_i - y_i$; величина $\bar{d}$ идентична разности средних, т. е.

$$\bar{d} = \frac{\sum d_l}{n} = (\bar{x}_1 - \bar{x}_2).$$

Ошибку средней разности $\bar{d}$, обозначаемую символом s_d, определяют по формулам

$$s_d = \sqrt{\frac{\sum (d_l - \bar{d})^2}{n(n-1)}} = \sqrt{\frac{\sum d_l^2 - \frac{(\sum d_l)^2}{n}}{n(n-1)}} \tag{79}$$

или

$$s_d = \sqrt{\frac{1}{n-1}\left(\frac{\sum d_l^2}{n} - \bar{d}^2\right)}. \tag{80}$$

Если члены генеральной совокупности распределяются нормально, то и разности между ними будут распределяться нормально и случайная величина $t=(\bar{d}-D)/s_{\bar{d}}$ будет иметь распределение Стьюдента с $k=n-1$ степенями свободы. H_0-гипотеза сводится к предположению, что $\mu_1-\mu_2=D=0$. Отсюда t-критерий выразится в виде отношения средней разности к своей ошибке, т. е. $t=\bar{d}/s_{\bar{d}}$. Если $t_\Phi \geqslant t_{st}$ для принятого уровня значимости и числа степеней свободы $k=n-1$, то нулевая гипотеза должна быть отвергнута.

Пример 3. На протяжении ряда лет в условиях Одесской опытной станции изучали влияние черного и апрельского пара на урожай ржи. Результаты опыта учитывали по массе 1000 зерен (табл. 36).

Таблица 36

Посев на пару	Масса 1000 зерен по годам опыта						Среднее
	1898	1899	1901	1902	1903	1904	
Апрельскому	31,6	24,2,	24,8	29,1	29,9	31,0	28,43
Черному	31,1	24,0	24,6	28,6	29,1	30,1	27,91
Разность d_i	0,5	0,2	0,2	0,5	0,8	0,9	$\bar{d}=0,52$
Квадрат разности d_i^2	0,25	0,04	0,04	0,25	0,64	0,81	$\Sigma d_i^2=2,03$

В табл. 36 приведены выборки с попарно связанными вариантами: несомненно, что каждый год имел свои специфические условия, которые одинаково влияли на урожай ржи, посеянной как по черному, так и по апрельскому пару. Поэтому обрабатывать полученные данные нужно с учетом тех условий, в которых проводили эксперимент. Из табл. 36 видно, что урожай ржи по апрельскому пару несколько выше, чем по черному. Средняя разность $\bar{d}=\Sigma d_i/n=3,1/6=0,52$ г. Определяем ошибку этой разности:

$$s_{\bar{d}}=\sqrt{\frac{1}{5}\left[\frac{2,03}{6}-(0,52)^2\right]}=\sqrt{\frac{0,34-0,27}{5}}=\sqrt{0,014}=0,12\text{ г.}$$

Критерий $t_\Phi=\frac{0,52}{0.12}=4,33$. Для $k=6-1=5$ и $\alpha=1\%$ $t_{st}=4,03$ (см. табл. V Приложений). Так как $t_\Phi>t_{st}$, то H_0-гипотезу отвергают на высоком уровне значимости $(0,001<P<0,01)$. Следовательно, с вероятностью $P>0,99$ можно утверждать, что разница между сравниваемыми выборками статистически достоверна.

Пример 4. В результате семилетних исследований урожайности ячменя и овса в условиях нечерноземной зоны РСФСР были получены следующие данные (табл. 37).

Таблица 37

Годы	Урожай, ц/га		Разница d_i	$d_i-\bar{d}$	$(d_i-\bar{d})^2$
	ячменя	овса			
1928	7,7	8,26	—0,56	—1,54	2,37
1929	9,0	7,22	1,78	0,80	0,64
1930	9,4	8,43	0,97	—0,01	0,00
1931	7,4	5,57	1,83	0,85	0,72
1932	7,4	6,35	1,05	0,07	0,00
1933	10,9	8,00	2,90	1,92	3,69
1934	8,0	9,13	—1,13	—2,11	4,45
Сумма	59,8	52,96	+6,84	—	11,87
Среднее	8,54	7,56	0,98	—	—

Разница между средним урожаем ячменя и овса составила 8,54—7,56=0,98 ц/га. Ошибка этой разницы $s_{\bar{d}}=\sqrt{\frac{11,87}{7\cdot6}}=$ $=\sqrt{0,283}=0,53$. Отсюда $t_{\Phi}=0,98/0,53=1,85$. Эта величина не превышает критический уровень $t_{st}=2,45$ для 5%-ного уровня значимости и числа степеней свободы $k=(7-1)=6$. Следовательно, нулевую гипотезу здесь отбросить нельзя.

Оценку средней разности можно произвести по доверительному интервалу, построенному на основании полученной разности $\bar{d}$ и ее ошибки $s_{\bar{d}}$. Если нижняя граница доверительного интервала окажется с положительным знаком, это будет свидетельствовать о достоверности разницы. Если же нижняя граница доверительного интервала будет с отрицательным знаком, это будет служить указанием на случайный характер наблюдаемой средней разности.

Так, в примере 3 $\bar{d}\pm ts_{\bar{d}}=0,52\pm1,96\cdot0,12=0,52\pm0,24$. Нижняя граница 95%-ного доверительного интервала (0,52—0,24=0,28) оказалась с положительным знаком, тогда как в примере 4 $\bar{d}\pm ts_{\bar{d}}=0,98\pm1,96\cdot0,53=0,98\pm1,04$ и нижняя граница доверительного интервала (0,98—1,04=—0,06) оказалась с отрицательным знаком, что не дает основания для отклонения нулевой гипотезы.

Оценка разности между долями. Выборочная доля зависит от числа единиц в выборке, имеющих учитываемый признак, а общее число таких единиц в генеральной совокупности определяет *генеральную долю* P. Оценкой разности между генеральными долями $P_1-P_2=D$ служит разность между выборочными долями $p_1-p_2=d$. Отношение этой разности к своей ошибке дает случайную величину $t=d/s_{d_p}$, которая

следует t-распределению Стьюдента. H_0-гипотезу, или предположение о том, что $\bar{P}_1=\bar{P}_2$, отвергают, если $t_\Phi \geqslant t_{st}$ для $k=n_1+n_2-2$ и принятого уровня значимости α. Ошибка разности между долями, взятыми из приблизительно равновеликих выборок (когда численности групп различаются не более чем на 25%), вычисляют по формуле

$$s_{d_p}=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}=\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}, \qquad (81)$$

где $q=1-p$. Если доли выражены в процентах от общего числа наблюдений, ошибку разности между ними определяют по формуле

$$s_{d_p}=\sqrt{\frac{p_1(100-p_1)}{n_1}+\frac{p_2(100-p_2)}{n_2}}. \qquad (82)$$

Сопоставляемые группы n_1 и n_2 могут быть выражены абсолютными числами m_1 и m_2. Ошибка наблюдаемой между ними разности определяется по следующей формуле:

$$s_{d_p}=\sqrt{\frac{m_1(n_1-m_1)}{n_1}+\frac{m_2(n_2-m_2)}{n_2}}, \qquad (83)$$

но так как $m_1/n_1=p_1$; $m_2/n_2=p_2$; $(n_1-m_1)/n_1=q_1$; $(n_2-m_2)/n_2=q_2$, то формулу (81) можно представить и в таком виде:

$$s_{d_p}=\sqrt{n_1p_1(1-p_1)+n_2p_2(1-p_2)}=\sqrt{n_1p_1q_1+n_2p_2q_2}.$$

Когда сравнивают доли из неравновеликих выборок и при $75\%<p<25\%$, ошибку разности между ними определяют по формуле

$$s_{d_p}=\sqrt{p(1-p)\frac{n_1+n_2}{n_1n_2}}=\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}; \qquad (84)$$

p определяют как средневзвешенную из p_1 и p_2 долей, или же из абсолютных численностей групп:

$$p=\frac{p_1n_1+p_2n_2}{n_1+n_2}=\frac{m_1+m_2}{n_1+n_2}. \qquad (85)$$

В этих формулах n_1 и n_2 — численности групп, на которых определяют доли $p_1=m_1/n_1$ и $p_2=m_2/n_2$. Если доли выражают в процентах от n, то вместо $q=1-p$ нужно брать $q=100-p$. Если же неравновеликие группы выражены абсолютными числами m_1 и m_2, ошибку разности между ними определяют по формуле

$$s_{d_p}=\sqrt{\frac{m_1+m_2}{n_1+n_2}\left(1-\frac{m_1+m_2}{n_1+n_2}\right)\frac{n_1+n_2}{n_1n_2}}. \qquad (86)$$

Пример 5. В потомстве от скрещивания двух золотистых хомячков с самками-альбиносами того же вида было получено в первой группе — 14 золотистых и 9 особей-альбиносов, а во второй группе — 12 золотистых и 8 особей-альбиносов. Разница между полученными в потомстве золотистыми особями составила 14—12=2 единицы. Определяем ошибку этой разницы:

$$s_{d_p}=\sqrt{\frac{14\cdot 9}{23}+\frac{12\cdot 8}{20}}=\sqrt{10{,}28}=3{,}2.$$

Критерий $t_ф=2/3{,}2=0{,}62$ (эта величина не превосходит точку $t_{st}=2{,}02$ для $k=23+20-2=41$ и 5%-ного уровня значимости; см. табл. V Приложений). Отсюда ясен вывод: H_0-гипотезу отвергнуть нельзя; разница между численностью золотистых хомячков, полученных в потомстве разных производителей, оказалась статистически недостоверной.

Пример 6. Изучали влияние эндотоксина на выживаемость облученных животных. Результаты опыта приведены в табл. 38.

Таблица 38

Группы животных	Выжило	Погибло	Всего
Контрольная	3 (21,4%)	11 (78,6%)	14
Опытная	23 (63,9%)	13 (36,1%)	36
Всего	26	24	50

Доля выживших в контроле $p_1=3/14=0{,}214$; в опыте $p_2=23/36=0{,}639$. Разница $d_p=0{,}639-0{,}214=0{,}425$. Нужно найти ошибку этой разницы. В данном случае объемы выборок ($n_1=14$ и $n_2=36$), из которых взяты для сравнения доли животных, различаются более чем на 25%. Определяем взвешенную долю: $p=\frac{0{,}214\cdot 14+0{,}639\cdot 36}{14+36}=\frac{3+23}{50}=0{,}52$; $q=1-0{,}52=0{,}48$.

Подставляем найденные величины p и q в формулу (84):

$$s_{d_p}=\sqrt{0{,}52\cdot 0{,}48\left(\frac{1}{14}+\frac{1}{36}\right)}=\sqrt{0{,}025}=0{,}157.$$

Критерий $t=0{,}425/0{,}157=2{,}71$ превосходит критическую точку $t_{st}=2{,}70$ для $k=50-2=48$ и 1%-ного уровня значимости. Нулевая гипотеза опровергается на высоком уровне значимости ($P<0{,}01$). Следовательно, с вероятностью более 99% можно судить о положительном действии эндотоксина на выживаемость подопытных животных.

Описанные выше критерии проверки равенства долей в двух выборках оказываются пригодными при не слишком больших и не слишком малых значениях p ($25\% < p < 75\%$). Особенно это относится к случаю небольших выборок. Свободным от подобного рода ограничений и поэтому более универсальным оказывается способ проверки равенства долей, основанный на использовании *угловой трансформации (φ-преобразования Фишера)*. При этом методе сравниваемые доли выражают в процентах с введением *поправки Йейтса* на непрерывность, равной $^1/_2 n$, которую вычитают из большей и прибавляют к меньшей доле. Затем по таблице значений $\varphi = 2 \arcsin \sqrt{p}$ (см. табл. VIII Приложений) находят величины для *исправленных долей:* $p_1\% + 100/(2n)$ и $p_2\% - 100/(2n)$, берут их разность и относят ее к ошибке, определяемой по формуле

$$s_{d_\varphi} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \qquad (87)$$

Условием для непринятия нулевой гипотезы служит следующее выражение:

$$t_\Phi = \frac{\varphi_1 - \varphi_2}{\sqrt{1/n_1 + 1/n_2}} = (\varphi_1 - \varphi_2) \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \geqslant t_{st}$$

для $k = n_1 + n_2 - 2$ и принятого уровня значимости.

Так, относительно оценки влияния эндотоксина на выживаемость облученных животных (см. табл. 38) имеем: контроль ($p_1\%$) $21{,}4 + 100/(2 \cdot 14) = 24{,}97\%$; опыт ($p_2\%$) $63{,}9 - 100/(2 \cdot 36) = 62{,}51\%$. По табл. VIII Приложений для $p_1 = 24{,}97$ находим $\varphi_1 = 1{,}047$ и для $p_2 = 62{,}51$ $\varphi_2 = 1{,}824$. Отсюда критерий достоверности

$$t_\Phi = \frac{1{,}824 - 1{,}047}{\sqrt{\frac{1}{14} + \frac{1}{36}}} = \frac{0{,}777}{0{,}315} = 2{,}47.$$

Эта величина превосходит критическую точку $t_{st} = 1{,}96$ для $k = 14 + 36 - 2 = 48$ и $\alpha = 5\%$ (см. табл. V Приложений). Нулевая гипотеза опровергается на уровне значимости ($0{,}01 < P < 0{,}05$).

Оценка разности между выборочной и генеральной долями. При оценке разности между известной генеральной долей $\bar{P}$ и долей выборки p нулевая гипотеза сводится к предположению, что разница между ними возникла случайно. Критерий Стьюдента в таких случаях выражается в виде отношения разности $(\bar{P} - p) = d_p$ к своей ошибке, которую определяют по формуле

$$s_{d_p} = \sqrt{\frac{\bar{p}(1 - \bar{P})}{n}}, \qquad (88)$$

где n — объем выборки. Как и в предыдущих случаях, условием для непринятия нулевой гипотезы служит критерий $t_ф = \left| \frac{\bar{P} - p}{s_{d_p}} \right| \geqslant t_{st}$ для $k = n - 1$ и принятого уровня значимости (α).

Пример 7. Изучали влияние возраста производителей на пол потомства у крупного рогатого скота. Для спаривания с разновозрастными быками подбирали коров одинакового возраста. Результаты испытаний приведены в табл. 39.

Таблица 39

Возраст быков, лет (от—до)	Родилось телят		Доля телок p	$\bar{P} - p$	Ошибка разности s_{d_p}	Критерий $t_ф$
	всего	из них телок				
2—3	141	77	0,55	0,05	0,043	1,16
4—5	89	43	0,48	0,02	0,053	0,38
6—7	88	41	0,46	0,04	0,053	0,75
>7	118	49	0,42	0,08	0,046	1,74

Доля телок в генеральной совокупности принята равной $\bar{P} = 0{,}50$. Ошибку разности между генеральной и выборочной долями определяли по формуле (88): $s_{d_p} = \sqrt{\frac{0{,}5 \cdot 0{,}5}{141}} = \sqrt{0{,}0018} = 0{,}043$ и т. д. В табл. 39 приведены значения $t_ф$-критерия Стьюдента для каждой группы. Поскольку все они не превосходят критическую точку $t_{st} = 2{,}0$ для 5%-ного уровня значимости (см. табл. V Приложений), нулевая гипотеза остается в силе. Вопрос о влиянии возраста производителей на пол потомства в данном исследовании остался открытым.

***F*-критерий Фишера (*F*-распределение).** Для проверки H_0-гипотезы о равенстве генеральных дисперсий ($\sigma^2_1 = \sigma^2_2$) нормально распределяющихся генеральных совокупностей t-критерий оказывается недостаточно точным, особенно при оценке разности дисперсий малочисленных выборок. В поисках лучшего критерия Р. Фишер нашел, что вместо выборочной разности $s_1 - s_2$ удобнее использовать разность между натуральными логарифмами этих величин, т. е. $\ln s_1 - \ln s_2$, где $s_1 \geqslant s_2$. Эта разность, обозначенная Фишером буквой z, распределяется нормально при наличии как больших, так и средних по объему статистических совокупностей.

При определении величины z можно вместо натуральных использовать десятичные логарифмы, так как $z = 2{,}3026 (\lg s_1 - \lg s_2)$ или $z = 2{,}3026 \lg (s_1/s_1)$, а также $z = 1{,}1513 \lg (s^2_1/s^2_2)$, где $s^2_1 \geqslant s^2_2$. Д. Снедекор предложил вместо логарифма отноше-

ний использовать отношения выборочных дисперсий, обозначив этот показатель в честь Фишера буквой F, т. е.

$$F = s_1^2/s_2^2 \quad \text{при} \quad s_1^2 \geqslant s_2^2. \tag{89}$$

Так как принято брать отношение большей дисперсии к меньшей, то критерий $F \geqslant 1$. Если $s^2{}_1 = s^2{}_2$, то $F = 1$. Чем значительнее неравенство между выборочными дисперсиями, тем больше будет и величина F, и, наоборот, чем меньше окажется разница между дисперсиями, тем меньше будет величина F.

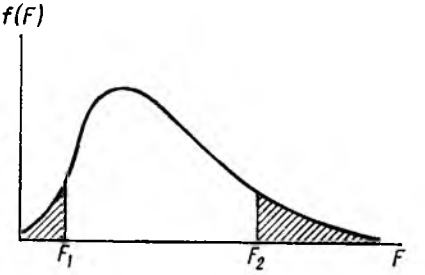


Рис. 21. График плотности вероятности F-распределения для типичных значений степени свободы k_1 и k_2 и критические границы F_1 и F_2 (по Н. В. Смирнову и Дунину-Барковскому, 1965)

Величина F имеет непрерывную функцию распределения и зависит только от чисел степеней свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$. F полностью определяется выборочными дисперсиями и не зависит от генеральных параметров, так как предполагают, что сравниваемые выборки, характеризуемые дисперсиями $s_1{}^2$ и $s_2{}^2$, взяты из генеральных совокупностей с $\sigma_1{}^2 = \sigma_2{}^2$ или из одной и той же генеральной совокупности. Функция распределения возможных значений величины F при небольшом n имеет форму асимметричной кривой, которая по мере увеличения числа испытаний ($n \to \infty$) приближается к кривой нормального распределения (рис. 21).

Функция F-распределения табулирована для 5%-ного ($P = 0{,}05$) и 1%-ного ($P = 0{,}01$) уровней значимости и чисел степеней свободы k_1 для большей дисперсии и k_2 для меньшей. Критические точки для F-критерия содержатся в табл. VI Приложений. В этой таблице степени свободы для большей дисперсии k_1 расположены в верхней строке (по горизонтали), а степени свободы для меньшей дисперсии k_2 — в первой графе (по вертикали). Если сравниваемые выборки извлечены из одной и той же генеральной совокупности или из разных совокупностей с дисперсиями $\sigma^2{}_1$ и $\sigma^2{}_2$, равными друг другу: $\sigma^2{}_1 = \sigma^2{}_2$, то величина F-критерия не превысит критические точки (F_{st}), указанные в табл. VI Приложений для k_1 и k_2, и принятого уровня значимости α. Если же выборки взяты из разных совокупностей с их параметрами $\sigma^2{}_1$ и $\sigma^2{}_2$, не равными друг другу, то $F_\Phi \geqslant F_{st}$ и нулевая гипотеза должна быть отвергнута.

Пример 8. В табл. 35 содержатся данные о влиянии кобальта на массу тела кроликов. Рассчитанные для этих данных дисперсии таковы: в опытной группе $s^2{}_1 = 2596{,}3$, в контроле

$s^2_2=3579{,}0$. Дисперсионное отношение $F=3579{,}0/2596{,}3=1{,}3$. В табл. VI Приложений для 5%-ного уровня значимости ($P=0{,}05$) и чисел степеней свободы $k_1=9-1=8$ (см. верхнюю строку таблицы) и $k_2=8-1=7$ (см. первую графу той же таблицы) находим $F_{st}=3{,}5$. Так как $F_\Phi<F_{st}$, нулевая гипотеза остается в силе ($P>0{,}05$). Это означает, что генеральные параметры сравниваемых групп $\sigma^2_1=\sigma^2_2$ и что применение t-критерия для проверки H_0-гипотезы в отношении оценки разности между выборочными средними $\bar{x}_1$ и $\bar{x}_2$ имеет достаточные основания.

F-критерий можно применить и для оценки разности между долями из неравновеликих выборок. Нулевая гипотеза отвергается при условии, что

$$F=(\varphi_1-\varphi_2)^2\frac{n_1n_2}{n_1+n_2}\geqslant F_{st} \quad \text{для} \quad \begin{cases} k_1=1; \\ k_2=n_1+n_2-2. \end{cases}$$

При этом k_1 находят по горизонтали, а k_2 — в первом столбце табл. VI Приложений. Так, обращаясь к примеру 6 о влиянии эндотоксина на выживаемость подопытных животных (см. табл. 38), имеем $\varphi_1=1{,}047$ и $\varphi_2=1{,}824$. Значения φ_1 и φ_2 найдены по табл. VIII Приложений. Учитывая, что $n_1=14$ и $n_2=36$, находим $F=(1{,}824-1{,}047)^2 14\cdot 36/(14+36)=0{,}604\,(504/50)=6{,}09$. Для $k_1=1$, $k_2=48$ и 5%-ного уровня значимости критическая точка $F_{st}=4{,}0$ (см. табл. VI Приложений). H_0-гипотеза опровергается на 5%-ном уровне значимости ($P<0{,}05$). Таким образом подтверждается ранее сделанный вывод о положительном влиянии эндотоксина на выживаемость подопытных животных.

Если оценивают разность между средними $\bar{x}_1$ и $\bar{x}_2$ выборок, извлеченных из совокупностей, которые распределяются по закону Пуассона, F-критерий строят в виде отношения

$$F_\Phi=\frac{\bar{x}_1}{\bar{x}_2+1}\geqslant F_{st} \quad \text{для} \quad \begin{cases} k_1=2(\bar{x}_2+1); \\ k_2=2\bar{x}_1 \end{cases}$$

и принятого уровня значимости (α).

Оценка разности между коэффициентами вариации. Разность между коэффициентами вариации сравниваемых групп, извлеченных из нормально распределяющихся совокупностей, можно оценить с помощью t-критерия Стьюдента. Приближенной оценкой разности $Cv_1-Cv_2=d_{Cv}$ служит ее отношение к своей ошибке, которая равна корню квадратному из суммы ошибок коэффициентов вариации сравниваемых групп, т. е.

$$s_{d_{Cv}}=\sqrt{s^2_{Cv_1}+s^2_{Cv_2}}. \qquad (90)$$

Нулевую гипотезу отвергают, если $t_\Phi>t_{st}$ для принятого уровня значимости и числа степеней свободы $k=n_1+n_2-2$.

Пример 9. Выше было найдено (см. табл. 35), что опытная ($n_1=8$) и контрольная ($n_2=9$) группы кроликов характеризуются средними $\bar{x}_1=638$ и $\bar{x}_2=526$ г. Соответственно $s_1=\sqrt{2596,3}=50,95$ и $s_2=\sqrt{3597,0}=59,97$. Отсюда $Cv_1=100(50,95/638)=8,0\%$ и $Cv_2=100(59,97/526)=11,4\%$. Разница $d_{Cv}=11,4-8,0=3,4\%$. Определяем ошибки этих показателей по формуле (65):

$$s_{Cv_1}=\sqrt{\frac{(8,0)^2}{8-1}\left[0,5+\left(\frac{8,0}{100}\right)^2\right]}=\sqrt{9,143\cdot 0,506}=\sqrt{4,63}=2,15;$$

$$s_{Cv_2}=\sqrt{\frac{(11,4)^2}{9-1}\left[0,5+\left(\frac{11,4}{100}\right)^2\right]}=\sqrt{16,245\cdot 0,513}=5,04=2,24.$$

Ошибка разности $s_{dCv}=\sqrt{2,15+2,24}=\sqrt{4,39}\approx 2,10$. Критерий $t_ф=3,40/2,10=1,62$. Эта величина не превосходит критическую точку $t_{st}=2,13$ для $k=8+9-2=15$ и $\alpha=5\%$, что не дает основания для отвергания нулевой гипотезы.

Разность между коэффициентами варнации можно оценить путем сопоставления доверительных интервалов, построенных для генеральных параметров сравниваемых групп. При этом границы доверительных интервалов определяют по формулам

$$P_н=\frac{Cv}{1+K\sqrt{1+2Cv^2}}; \qquad (91)$$

$$P_в=\frac{Cv}{1-K\sqrt{1+2Cv^2}}, \qquad (92)$$

где $P_н$ — нижняя, а $P_в$ — верхняя границы доверительного интервала; $K=\frac{t}{2(n-1)}$; t — нормированное отклонение (для $\alpha_1=5\%$ $t=1,96$).

Определим границы доверительного интервала для опытной группы, выразив коэффициент вариации в долях единицы, т. е. $Cv=0,08$:

$$K=\frac{1,96}{\sqrt{2\cdot 7}}=0,52; \quad P_и=\frac{0,08}{1+0,52\sqrt{1+2\cdot 0,08^2}}=\frac{0,08}{1,52}=0,053,$$

или 5,3%; $P_в=\frac{0,08}{1-0,52\sqrt{1+2\cdot 0,08^2}}=\frac{0,08}{0,48}=0,167$, или 16,7 %.

Аналогично определяем границы доверительного интервала для контрольной группы:

$$K=\frac{1,96}{\sqrt{2\cdot 8}}=0,49; \quad P_и=\frac{0,114}{1+0,49\sqrt{1+2\cdot 0,114^2}}=\frac{0,114}{1,496}=$$

$=0{,}076$, или 7,6 %; $P_в = \frac{0{,}114}{1-0{,}49\sqrt{1+2\cdot 0{,}114^2}} = \frac{0{,}114}{0{,}504} = 0{,}226$, или 22,6%.

Итак, в первом случае границы доверительного интервала оказались от 5,3 до 16,7%, во втором — от 7,6 до 22,6%. Таким образом, границы доверительного интервала, построенного для контрольной группы, близки к границам интервала опытной группы кроликов, что указывает на отсутствие существенных различий между коэффициентами вариации этих групп.

V.3. НЕПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ

Правильное применение параметрических критериев для проверки статистических гипотез основано на предположении о нормальном распределении совокупностей, из которых взяты сравниваемые выборки. Однако это не всегда имеет место, так как не все биологические признаки распределяются нормально. Немаловажным является и то обстоятельство, что исследователю приходится иметь дело не только с количественными, но и с качественными признаками, многие из которых выражаются порядковыми номерами, индексами и другими условными знаками. В таких случаях необходимо использовать *непараметрические критерии.*

Известен целый ряд непараметрических критериев, среди которых видное место занимают так называемые *ранговые критерии,* применение которых основано на ранжировании членов сравниваемых групп. При этом сравниваются не сами по себе члены ранжированных рядов, а их порядковые номера, или ранги. Ниже рассмотрены некоторые непараметрические критерии, применяемые для проверки нулевой гипотезы при сравнении как независимых, так и зависимых выборочных групп.

***X*-критерий Ван-дер-Вардена.** Этот критерий относится к группе ранговых критериев, его применяют для проверки нулевой гипотезы при сравнении друг с другом независимых выборок. Техника расчетов *X*-критерия сводится к следующему. Сравниваемые выборки ранжируют в один общий ряд по возрастающим значениям признака. Затем каждому члену ряда присваивают порядковый номер, отмечающий его место в общем ранжированном строю. Далее по порядковым номерам одной из выборок, обычно меньшей по объему, находят отношение $R/(N+1)$, где $N+1=n_1+n_2+1$, т. е. сумма всех членов сравниваемых групп, увеличенная на единицу, а R — порядковый номер членов ряда, их «ранг».

С помощью специальной таблицы (см. табл. IX Приложений) находят значения функции $\psi[R/(N+1)]$ для каждого значения $R/(N+1)$. Суммируя результаты (обязательно с учетом знаков!), получают величину $X_ф=\Sigma\psi[R/(N+1)]$, которую срав-

ивают с критической точкой этого критерия X_{st} для принято-о уровня значимости α и общего числа членов сравниваемых ыборок, т. е. $N=n_1+n_2$. Критические точки X-критерия для 5%-ного и 1%-ного уровней значимости и общего числа членов вух выборок $N=n_1+n_2$ (с учетом разности n_1-n_2) содержат-я в табл. X Приложений.

Нулевая гипотеза сводится к предположению, что сравни-аемые выборки извлечены из генеральных совокупностей с динаковыми функциями распределения. Если окажется, что $X_ф \geqslant X_{st}$, нулевая гипотеза должна быть отвергнута на принятом ровне значимости.

Пример 10. Вернемся к результатам опыта по проверке дей-твия нового препарата на массу тела лабораторных мышей см. пример 2), где сравнивали две группы — опытную ($n_1=9$) контрольную ($n_2=11$) — с попарно не связанными числовыми начениями признака. Применим X-критерий Ван-дер-Вардена оценке результатов этого эксперимента. Расчет X-критерия по ранжированным значениям признака сравниваемых групп риведен в табл. 40.

Таблица 40

Масса тела мышей, г		Порядковый номер R	$\frac{R}{N+1}$	$\psi\left(\frac{R}{N+1}\right)$
опыт	контроль			
	60	1		
	60	2		
	62	3		
64		4	4/21=0,190	−0,88
	66	5		
68		6	6/21=0,286	−0,57
	68	7		
	69	8		
70		9	9/21=0,429	−0,18
	70	10		
	71	11		
72		12	12/21=0,571	+0,18
	73	13		
75		14	14/21=0,667	+0,43
76		15	15/21=0,714	+0,57
	78	16		
79		17	17/21=0,810	+0,88
	80	18		
80		19	19/21=0,905	+1,31
83		20	20/21=0,952	+1,66
$n_1=9$	$n_2=11$	—	—	+3,40

В данном случае $N=n_1+n_2=9+11=20$. Для этого числа $N=20$) и 5%-ного уровня значимости с учетом разности $n_1-n_2=11-9=2$ в табл. X Приложений находим $X_{st}=3{,}84$. Так

как $X_\phi = 3{,}40 < X_{st} = 3{,}84$, нулевую гипотезу не учитывать нельзя; разница между контролем и опытом оказывается статистически недостоверной.

Пример 11. Проанализируем с помощью X-критерия Ван-дер-Вардена результаты опыта о влиянии кобальта на величину массы тела кроликов (см. табл. 35). Расчет величины $X = \Sigma\psi[R/(N+1)]$ приведен в табл. 41.

Таблица 4

Масса тела кроликов, г		Порядковый номер R	$\frac{R}{N+1}$	$\psi\left(\frac{R}{N+1}\right)$
контроль	опыт			
420		1		
470		2		
490		3		
504		4		
530		5		
560		6		
	561	7	7/18=0,389	—0,28
580		8		
	580	9	9/18=0,500	0,00
580		10		
600		11		
	621	12	12/18=0,667	+0,43
	630	13	13/18=0,722	+0,59
	640	14	14/18=0,778	+0,77
	680	15	15/18=0,833	+0,97
	692	16	16/18=0,889	+1,22
	700	17	17/18=0,944	+1,59
$n_1=9$	$n_2=8$	—	—	+5,29

Найденная величина критерия $X_\phi = \Sigma\psi[R/(N+1)] = 5{,}29$ превосходит критическую точку $X_{st} = 4{,}44$ для 1%-ного уровня значимости и $N = 9+8 = 17$ с учетом разности $n_1 - n_2 = 1$ (см. табл. X Приложений), что дает основание отвергнуть нулевую гипотезу на высоком уровне значимости ($P < 0{,}01$) и заключить, что влияние кобальта на величину массы тела кроликов достоверно.

***U*-критерий Уилкоксона (Манна—Уитни).** Гипотезу о принадлежности сравниваемых независимых выборок к одной и той же генеральной совокупности или к совокупностям с одинаковыми параметрами, т. е. H_0-гипотезу, можно проверить с помощью *рангового критерия Уилкоксона (Манна—Уитни)*.

Для расчета U-критерия необходимо: 1. Расположить числовые значения сравниваемых выборок в возрастающем порядке в один общий ряд и пронумеровать члены общего ряда от 1 до $N = n_1 + n_2$. (Эти номера и будут «рангами» членов ря-

да.) 2. Отдельно для каждой выборки найти суммы рангов R и определить величины

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \qquad (93)$$

и

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}, \qquad (94)$$

которые отображают связь между суммами рангов первой и второй выборки. 3. В качестве U-критерия использовать меньшую величину U_Φ, которую сравнить с табличным значением U_{st}. Условием для сохранения принятой H_0-гипотезы служит неравенство $U_\Phi > U_{st}$. Критические точки U-критерия U_{st} для n_1, n_2 и принимаемого уровня значимости α содержатся в табл. XI Приложений.

Пример 12. В примере 2 было показано, что разность в массе тела между опытной ($n_1=9$) и контрольной ($n_2=11$) группами лабораторных мышей является статистически недостоверной. Проверим этот вывод с помощью U-критерия. Для этого обратимся к табл. 40, в которой содержатся расположенные в возрастающем порядке числовые значения сравниваемых выборок и их «ранги». Суммируя «ранги» отдельно для каждой группы, находим $R_1 = 4+6+9+...+20=112$; $R_2 = 1+2+3+...+18=94$. Подставляем эти данные в формулы (93) и (94): $U_1 = 112-9\cdot 10/2=67$; $U_2 = 94-11\cdot 12/2=22$. Меньшую величину $U_\Phi=22$ сравниваем с табличным значением U_{st} для $n_1=9$, $n_2=11$ и уровня значимости $\alpha=1\%$, которое равно $U_{st}=19$ (см. табл. XI Приложений). Поскольку $U_\Phi > U_{st}$, отвергнуть проверяемую H_0-гипотезу нельзя. Следовательно, подтверждается ранее сделанный вывод о статистической недостоверности различий, наблюдаемых между этими выборками.

Критерий знаков z. В тех случаях, когда результаты наблюдений выражаются не числами, а знаками плюс (+) и минус (—), различия между попарно связанными членами сравниваемых выборок оценивают с помощью *критерия знаков* z. Конструкция этого критерия базируется на весьма простых соображениях: если попарно сравниваемые значения двух зависимых выборок существенно не отличаются друг от друга, то число плюсовых и минусовых разностей окажется совершенно одинаковым; если же заметно преобладают плюсы или минусы, это будет указывать на положительное или отрицательное действие изучаемого фактора на результативный признак. Большее число однозначных разностей служит в качестве фактически найденной величины z-критерия знаков. При этом нулевые разности, т. е. случаи, не давшие ни положительного, ни отрицательного результата, обозначаемые цифрой 0, в расчет не

принимают и число парных наблюдений соответственно уменьшается.

Как и всякий другой выборочный показатель, z-критерий знаков является величиной случайной; он служит для проверки H_0-гипотезы, т. е. предположения о том, что совокупности или совокупности, из которых взяты сравниваемые выборки, имеют одну и ту же или одинаковые функции распределения. H_0-гипотеза отвергается, если $z_\Phi \geqslant z_{st}$ для принятого уровня значимости α и числа парных наблюдений n, взятых без нулевых разностей. Критические точки z_{st} для двух уровней значимости и числа парных наблюдений содержатся в табл. XII Приложений.

Пример 13. Изучали влияние туберкулина на состав периферической крови низших обезьян. Результаты наблюдений приведены в табл. 42.

Таблица 42

Номера подопытных животных	Эозинофилия в крови		Эффект воздействия
	до введения туберкулина	после введения туберкулина	
1	+ +	+ +	0
2	+ + +	+ +	+
3	+ +	+	+
4	+ +	+	+
5	+ + +	+ +	+
6	+ +	+ + +	—
7	+ +	+	+
8	+	+ +	—
9	+ + +	+ +	+
10	+ +	+	+
11	+ +	+ +	0
12	+ + +	+	+
13	+ +	+	+
14	+ +	+	+

Из табл. 42 видно, что после введения туберкулина количество эозинофилов в периферической крови у большинства обезьян оказалось пониженным. Так, из 14 наблюдений два оказались нулевыми, т. е. $n=14-2=12$. Из этого числа положительных разностей насчитывается 10. Следовательно, $z_\Phi=10$. По табл. XII Приложений для $n=12$ и $\alpha=5\%$ находим $z_{st}=10$. Равенство $z_\Phi=z_{st}$ дает основание отвергнуть H_0-гипотезу на 5%-ном уровне значимости. Следовательно, с вероятностью 95% можно утверждать, что введение туберкулина (реакция Манту) вызывает заметное снижение эозинофилов в периферической крови обезьян.

Пример 14. В табл. 43 приведены данные о годовых удоях коров и их потомства по второму и третьему отелам.

Таблица 43

Удой коров, кг		Разница выражена		Ранги разницы
материнского поколения	потомства	знаками	числами	
3770	2991	—	779	8
3817	4593	+	776	7
2450	3529	+	1076	10
3463	4274	+	811	9
3500	3103	—	397	4
5544	3949	—	1597	12
3112	3491	+	379	3
3150	3559	+	409	5
3118	2916	—	202	1
3018	4580	+	1562	11
4291	4510	+	219	2
3463	4144	+	681	6

Прибавки в удоях потомства обозначены знаком (+), а удои потомства, оказавшиеся ниже удоев матерей, — знаком (—). В той же таблице приведены и числовые показатели разности между удоями коров материнского поколения и их потомства, а также ранги этих различий. Из табл. 43 следует, что из 12 парных наблюдений положительный эффект обнаружился в 8 случаях, т. е. $z_ф = 12 - 4 = 8$. Эта величина оказалась ниже критической точки $z_{st} = 10$ для 5%-ного уровня значимости и $n = 12$ (см. табл. XII Приложений). Следовательно, нулевую гипотезу отвергнуть нельзя.

***T*-критерий Уилкоксона.** Когда члены сравниваемых выборок связаны попарно некоторыми общими условиями (зависимые выборки), различия между ними с достаточной точностью оцениваются с помощью *рангового критерия Уилкоксона T*. Парный *T*-критерий Уилкоксона является более мощным, чем критерий знаков. *T*-критерий рассчитывают следующим образом. 1. Ранжируют попарные разности, как положительные, так и отрицательные, в один общий ряд. При этом нулевые разности в расчет не принимают, а все остальные независимо от знака ранжируют так, чтобы наименьшая абсолютная разность получила первый ранг, причем одинаковым по величине разностям присваивают один и тот же ранг. 2. Находят отдельно суммы положительных и отрицательных разностей. Меньшую из двух сумм разностей, без учета ее знака, используют в качестве фактически установленной величины *T*-критерия. 3. Сравнивают эту величину $T_ф$ с критическим значением T_{st} для при-

нятого уровня значимости α и числа парных наблюдений n, которое берут без нулевых разностей. Нулевую гипотезу отвергают, если $T_\Phi > T_{st}$. Если же $T_\Phi \geqslant T_{st}$, нулевую гипотезу не учитывать нельзя. Критические значения парного критерия Уилкоксона T_{st} содержатся в табл. XIII Приложений.

Пример 15. Применим парный критерий Уилкоксона для проверки H_0-гипотезы относительно данных о годовых удоях коров материнского поколения и их потомства, приведенных в табл. 43. В последней графе этой таблицы помещены ранги абсолютных разностей между парными значениями годового удоя коров. Определяем суммы плюсовых и минусовых разностей: $T_{(+)}=7+10+9+3+5+11+2+6=53$ и $T_{(-)}=8+4+12+1=25$. Меньшая разность дает $T_\Phi=25$. Сравниваем эту величину с критическим значением $T_{st}=15$ для $n=12$ и $\alpha=5\%$ (см. табл. XIII Приложений). Так как $T_\Phi > T_{st}$, то нулевая гипотеза остается в силе.

Пример 16. В табл. 37 приведены результаты семилетних исследований урожайности ячменя и овса в условиях Нечерноземной зоны РСФСР. В той же таблице приведены парные разности между урожаями этих культур в разные годы. Статистическая оценка результатов опыта привела к выводу о недостоверности средней разности между урожайностью ячменя и овса. Проверим этот вывод с помощью непараметрического критерия Уилкоксона. Выпишем в порядке возрастания разности d_i и их ранги R:

d_i	−0,56	+0,97	+1,05	−1,13	+1,78	+1,83	+2,90
R	1	2	3	4	5	6	7

Меньшую сумму рангов определяют по минусовым разностям, т. е. $T_\Phi=1+4=5$. Эта величина превосходит критическую точку $t_{st}=3$ для $n=7$ и 5%-ного уровня значимости (см. табл. XIII Приложений). Таким образом подтверждается ранее сделанный (с помощью t-критерия Стьюдента) вывод о статистической недостоверности разницы между урожайностью ячменя и овса.

V.4. ОЦЕНКА БИОЛОГИЧЕСКИ АКТИВНЫХ ВЕЩЕСТВ

При испытании инсектицидов, лекарственных, радиоактивных и других биологически активных веществ обнаруживается, что особи однородной группы реагируют на одну и ту же дозу по-разному (индивидуальная изменчивость!) и что разные дозы могут вызывать одинаковый эффект у целой группы индивидов. Отсюда следует, что о силе действия на организм биологически активных веществ можно судить лишь по среднему результату.

Дозы сильнодействующих веществ испытывают на однородных группах (мыши, крысы и другие объекты) по 6—10 особей в группе. На каждой группе изучают одну дозу. Обычно применяют 5—9 доз в возрастающем по силе действия порядке. Опыт проводят одновременно (обычно на протяжении одного дня) на всех группах особей. При этом учитывают число особей, у которых обнаружился эффект, и число тех, у которых видимого эффекта от действия доз не обнаружено. О среднем результате судят по обнаружению эффекта действия доз у 50% подопытных индивидов.

Определить дозу, вызвавшую видимый эффект или летальный исход у 50% подопытных индивидов, можно разными способами — графически и аналитически. Установлено, что индивидуальные реакции подопытных животных на воздействие биологически активных веществ распределяются, как правило, нормально. Зависимость между дозой и эффектом действия графически выражается в виде S-образной кривой, или *кумуляты* (см. рис. 3). Кумулята, называемая *кривой эффекта доз,* может быть получена, если по оси абсцисс откладывают дозы вещества, а по оси ординат — эффект воздействия этих доз на подопытных животных. Центральная точка кумуляты совпадает с центром распределения. Опуская из этой точки перпендикуляры на оси координат, можно определить среднюю дозу эффекта. Проще, однако, среднюю дозу эффекта определить аналитическими способами, один из которых приведен ниже.

Способ Спирмена — Кербера. Достоинство этого способа заключается в том, что он позволяет не только рассчитать среднюю дозу эффекта $\bar{m}$, но и построить доверительный интервал для генеральной средией μ. Среднюю дозу эффекта определяют по формуле

$$\bar{m} = m - d\,(P_1 - 0{,}5), \qquad (95)$$

где m — минимальная доза, вызывающая эффект у 100% подопытных индивидов; d — разница между дозами; P_1 — суммарная доля реагирующих на дозы индивидов.

Среднее квадратическое отклонение вычисляют по следующей формуле:

$$s_m = d\sqrt{2P_2 - P_1^2 - P_1 - 1/12}. \qquad (96)$$

Здесь P_2 — сумма ряда накопленных долей реагирующих на дозы индивидов.

Пример 17. На группе, состоящей из десяти лабораторных мышей, испытывали действие ядовитого вещества. Дозы яда рассчитывали в миллиграммах на 1 кг массы тела подопытных животных. Эффект действия яда учитывали по летальным исходам. Результаты опыта приведены в табл. 44.

В данном случае $n=10$, $d=10$, $m=180$ мг/кг, $P_1=4{,}0$ и $P_2=11{,}8$. Подставляем известные величины в формулы (95) и (96): $m=LD^1_{50}=180-10(4{,}0-0{,}5)=180-35=145$ мг/кг;

$s_m=10\sqrt{2\cdot 11{,}8-4{,}0^2-4{,}0-0{,}083}=10\sqrt{3{,}517}=10\cdot 1{,}875=18{,}75$

Найденные величины $\bar{m}=145$ и $s_m=18{,}75$ позволяют построить доверительный интервал для генерального параметра, т. е. истинной средней дозы эффекта: $\bar{m}\pm\Delta_m$, где $\Delta_m=ts_{\bar{m}}$ — величина предельной ошибки средней $\bar{m}$. В данном случае $s_{\bar{m}}=s_m/\sqrt{n}=18{,}75/\sqrt{10}=18{,}75/3{,}16=5{,}93$. Отсюда для 5%-ного уровня значимости и соответственно $t=1{,}96$ нижняя граница доверительного интервала составляет $145-1{,}96\cdot 5{,}93=145-$

Таблица 4

Доза, мг/кг	110	120	130	140	150	160	170	180	Сумма
Число погибших животных	0	1	3	4	6	7	9	10	—
Доля погибших животных	0	0,1	0,3	0,4	0,6	0,7	0,9	1,0	4,0
Накопленная доля погибших животных	0	0,1	0,4	0,8	1,4	2,1	3,0	4,0	11,8

$-11{,}62=133{,}38\sim 133$ мг/кг и верхняя граница $145+11{,}62=156{,}62\sim 157$ мг/кг. Это означает, что с вероятностью $P=0{,}95$ можно утверждать, что генеральная средняя доза эффекта $\overline{LD}_{50}$ находится в пределах от 133 до 157 мг/кг.

ГЛАВА VI

ПРОВЕРКА ГИПОТЕЗ О ЗАКОНАХ РАСПРЕДЕЛЕНИЯ

VI.1. ПРИМЕНЕНИЕ КОЭФФИЦИЕНТОВ АСИММЕТРИИ И ЭКСЦЕССА ДЛЯ ПРОВЕРКИ НОРМАЛЬНОСТИ РАСПРЕДЕЛЕНИЯ

Эмпирический вариационный ряд и его график — вариационная кривая — не позволяют с полной уверенностью судить о законе распределения совокупности, из которой взята выборка. На величине любого варьирующего признака сказывается влияние многочисленных, в том числе и случайных, факторов, ис-

[1] Символ LD_{50} обозначает дозу вещества, вызывающую летальный исход у 50% подопытных особей.

кажающих четкую картину варьирования. Между тем знаиие закона распределения позволяет избежать возможных ошибок в оценке генеральных параметров по выборочным характеристикам.

Гипотезу о законе распределения можно проверить разными способами, в частности с помощью *коэффициентов асимметрии As* и *эксцесса Ex*. При нормальном распределении эти показатели равны нулю. В действительности такое равенство почти не наблюдается. Выборочные показатели *As* и *Ex*, определяемые по формулам (48) и (49), являются случайными величинами, которые сопровождаются ошибками. В качестве критерия нормальности распределения служат t_{As} и t_{Ex}, являющиеся отношениями выборочных коэффициентов *As* и *Ex* к их ошибкам репрезентативности, которые определяют обычно по следующим приближенным формулам:

$$s_{As}=\sqrt{\frac{6}{n+3}}; \tag{97}$$

$$s_{Ex}=\sqrt{\frac{24}{n+5}}=2\sqrt{\frac{6}{n+5}}\,.^{1} \tag{98}$$

В связи с тем что выборочные распределения коэффициентов асимметрии и эксцесса в случае нормальности распределения признака при не слишком больших объемах выборок (особенно это характерно для *Ex*) могут быть довольно далеки от нормального вида, использование квадратических ошибок для *As* и *Ex* при *n*, меньшем нескольких сотен наблюдений, оказывается рискованным. Поэтому более предпочтительным следует считать проверку нормальности распределения по значениям этих коэффициентов с применением таблиц, приведенных в Приложении (см. табл. XIV и XV). В них указаны критические точки для разных уровней значимости α и объемов выборки *n*. Если коэффициенты *As* и *Ex* превосходят критические точки, содержащиеся в этих таблицах, гипотеза о иормальности распределения должна быть отвергнута.

Так, в примере с изучением формы распределения длины хвоинок сосны были получены значения $As=-0{,}556$ и $Ex=0{,}872$. Для $\alpha=1\%$ и $n=200$ в табл. XIV Приложений находим $As_{st}=0{,}403$, а в табл. XV — $Ex_{st}=0{,}832$. Так как эмпирически определенные величины *As* и *Ex* превышают табличные критические значения, можно сделать вывод о наличии у этого распределения значимых асимметрии и эксцесса.

---

[1] Более точно ошибки коэффициеитов *As* и *Ex* определиют по формулам

$$s_{As}=\sqrt{\frac{6(n-1)}{(n+1)(n+3)}} \quad \text{и} \quad s_{Ex}=\sqrt{\frac{24n(n-2)(n-3)(n-5)}{(n-1)^2(n+3)(n+5)}}\,.$$

VI.2. КРИТЕРИЙ ХИ-КВАДРАТ (χ^2-РАСПРЕДЕЛЕНИЕ)

Проверку гипотез о законах распределения также производят с помощью специально выработанных критериев. Один из них, нашедший широкое применение в биометрии,— *критерий согласия,* или *соответствия* χ^2 (предложен в 1900 г. К. Пирсоном). Этот критерий представляет собой сумму квадратов отклонений эмпирических частот f от вычисленных или ожидаемых частот f', отнесенную к теоретическим частотам, т. е.

$$\chi^2=\sum_{i=1}^{k}\frac{(f-f')^2}{f'}=\sum_{i=1}^{k}\left(\frac{d^2}{f'}\right). \qquad (99)$$

Символ χ^2 не является квадратом какого-то числа, а выражает лишь исходную величину, определяемую данной формулой. Буквой d обозначена разность между эмпирическими и вычисленными частотами.

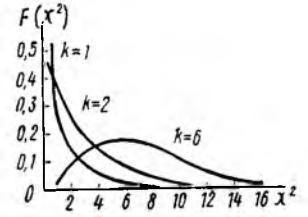


Рис. 22. Функции χ^2-распределении в зависимости от разных чисел степеии свободы k

Величина критерия χ^2 всегда положительна, так как отклонения эмпирических частот от ожидаемых или вычисленных частот возведены в квадрат. Поэтому при определении разности d знаки чисел можно не учитывать, вычитая из больших значений меньшие. При полном совпадении эмпирических частот с вычисленными или ожидаемыми частотами $\Sigma(f_i-f_i')=0$ и $\chi^2=0$.

Распределение вероятных значений случайной величины χ^2 является непрерывным и асимметричным (рис. 22), оно зависит от числа степеней свободы k и приближается к нормальной кривой по мере увеличения числа испытаний n. Поэтому применение критерия χ^2 к оценке дискретных распределений сопряжено с некоторыми погрешностями, которые сказываются на его величине, особенно при малых выборках.

Для того чтобы оценки были более точными, выборка, распределяемая в вариационный ряд, должна содержать не менее 50 вариант. Поэтому часто считают, что применение критерия χ^2 требует, чтобы в крайних классах вариационного ряда содержалось не менее пяти вариант. Если в крайних классах содержится меньше чем пять вариант, то вычисленные и эмпирические частоты объединяются до указанного минимума и соответственно уменьшается число классов вариационного ряда[1].

[1] Существует иная точка зрения на минимальные значения теоретических частот f', которые могут находиться в разных классах вариационного ряда. Согласно ей, при $n>50$ и $k\geqslant 6$ одно из значений f' может быть снижено да-

Число степеней свободы устанавливают по вторичному числу классов с учетом ограничений свободы вариации, которая в разных случаях бывает различной. Так, при оценке эмпирических распределений, следующих нормальному закону, число степеней свободы $k=N-3$ (с учетом трех ограничений свободы вариации: n, $\bar{x}$ и s_x). Если же оценке подлежит распределение, следующее закону Пуассона, число степеней свободы уменьшается на единицу, т. е. $k=N-2$ (с учетом двух ограничений свободы вариации n и s_x^2 или $\bar{x}$). В других случаях число степеней свободы устанавливают особо (см. ниже).

Таблица 45

Частоты		$d=f-f'$	d^2	$\frac{d^2}{f'}$	$f(t)$	$q=1-s(t)$	$f'q$	$\frac{d^2}{f'q}$
эмпирические f	вычисленные f'							
1	2	3	4	5	6	7	8	9
3 } 12	1,6 } 11,6	0,4	0,16	0,01	0,0594	0,9406	10,91	0,01
9 }	10,0 }							
31	34,3	3,3	10,89	0,32	0,1736	0,8264	28,35	0,38
71	67,8	3,2	10,24	0,15	0,3429	0,6571	44,55	0,23
82	77,6	4,4	19,36	0,25	0,3918	0,6082	47,20	0,41
46	51,2	5,2	27,04	0,53	0,2589	0,7411	37,94	0,71
19	19,5	0,5	0,25	0,01	0,0989	0,9011	17,57	0,01
5 } 6	4,4 } 5,0	1,0	1,00	0,20	0,0247	0,9753	4,88	0,20
1 }	0,6 }							
$\Sigma=267$	$\Sigma=267$	—	—	1,47	—	—	—	$\Sigma=1,95$

Примечание. Последние четыре графы понадобятся в дальнейшем (см. пример 7 разд. VI.3).

На величине критерия χ^2 сказывается степень точности, с какой определены теоретически вычисленные или ожидаемые частоты. Поэтому при сопоставлении эмпирических частот с вычисленными частотами последние не следует округлять до целых чисел [1].

Нулевая гипотеза сводится к предположению, что различия, наблюдаемые между эмпирическими и вычисленными или ожидаемыми частотами, носят исключительно случайный характер. Для проверки нулевой гипотезы нужно фактически полученную величину $\chi_ф^2$ сравнить с ее критическим значением χ_{st}^2. Если

...е до 0,5. При $k=2$ минимальное значение f' составляет 2. И только при $k=1$ минимальное значение f' должно быть не менее 4. (*Прим. ред.*)

[1] Технику расчета теоретических частот вариационного ряда см. в табл. 30.

$\chi_\Phi^2 \geqslant \chi^2_{st}$, то нулевая гипотеза должна быть отвергнута на принятом уровне значимости с числом степеней свободы k. Критические точки χ^2_{st} приведены в табл. VII Приложений.

Пример 1. В табл. 28 приведены эмпирические и вычисленные по нормальному закону частоты распределения длины тела у 267 мужчин. Из приведенных данных видно, что между эмпирическими и вычисленными частотами нет полного совпадения. Нужно установить, случайны или закономерны эти различия, т. е. выяснить, следует ли это распределение нормальному закону. Расчет χ^2-критерия, который оказался равным 1,47, приведен в табл. 45.

В данном случае число вторичных классов $N=7$. Число степеней свободы $k=7-3=4$. Исходя из 5%-ного уровня значимости в табл. VII Приложений находим $\chi^2_{st}=9{,}49$. Эта величина значительно превышает $\chi_\Phi^2=1{,}47$, что не позволяет отвергнуть H_0-гипотезу. Следовательно, существуют достаточные основания для утверждения, что данное распределение следует нормальному закону.

Критерий χ^2 применяют и для оценки сходства между вариационными рядами, частоты которых распределяются в границах одних и тех же классов. В таких случаях критерий χ^2 определяют по формулам:

$$\text{при } n_1=n_2 \quad \chi^2=4\left(\sum_{l=1}^{k}\frac{f_1^2}{f_1+f_2}\right)-N; \tag{100}$$

$$\text{при } n_1\neq n_2 \quad \chi^2=\frac{N^2}{n_1n_2}\left(\sum_{l=1}^{k}\frac{f_1^2}{f_1+f_2}-\frac{n^2}{N}\right)^1. \tag{101}$$

В этих формулах f_1 и f_2 — частоты сравниваемых распределений; $n_1=\Sigma f_1$ — объем одного (любого), а $n_2=\Sigma f_2$ — объем другого ряда распределения; $N=n_1+n_2$. Число степеней свободы k определяют по числу классов N без единицы, т. е. $k=N-1$. При этом частоты, меньшие 5, не объединяют, как это принято в отношении теоретически вычисленных частот.

Пример 2. Урожай фасоли, полученный на делянках от посева крупных f_1 и мелких f_2 семян, распределился следующим образом (табл. 46).

С помощью формулы (100) находим $\chi^2=4\cdot 104{,}78-200+200=419{,}12-400=19{,}12$. Эта величина не превышает критическую точку $\chi^2_{st}=20{,}09$ для $k=9-1=8$ и 1%-ного уровня значимости (см. табл. VII Приложений), что не дает оснований для непринятия нулевой гипотезы. Следовательно, наблюдаемые между частотами этих рядов различия носят не систематический, а случайный характер.

[1] См.: *Плохинский Н. А.* Алгоритмы биометрии. М., 1980. С. 101.

Таблица 46

Масса семян, мг	Частоты		f_1^2	f_1+f_2	$\frac{f_1^2}{f_1+f_2}$
	f_1	f_2			
125	1	1	1	2	0,50
175	5	3	25	8	3,12
225	17	7	289	24	12,04
275	45	22	2025	67	30,22
325	70	88	4900	158	31,01
375	51	69	2601	120	21,68
425	10	7	100	17	5,88
475	1	2	1	3	0,33
525	0	1	0	1	0,00
Сумма	200	200	—	—	104,78

Пример 3. Изучали основной вид очанки (*Euphrasia pralcurta* Chitr.) и одну из ее разновидностей (*E. curta* Fr.) — растения, произрастающего в центральных областях европейской части РСФСР. Анализ вида производили по числу растений, зацветающих из того или иного (по счету) узла. Нужно было выяснить, в какой мере основной вид очанки отличается от ее разновидности, что имеет прямое отношение к проблеме видообразования, совершающегося в природе. Результаты наблюдений и их обработка приведены в табл. 47.

Используя формулу (101), находим $\chi_\Phi^2=[177^2/(124\times 53)](91{,}71-124^2/177)=4{,}767(91{,}71-86{,}87)=23{,}07$. Эта величина превосходит критическую точку $\chi^2_{st}=21{,}67$ для $k=10-1=9$ и 1%-ного уровня значимости (см. табл. VII Приложений). Следовательно, на высоком уровне значимости ($P<0{,}01$) можно заключить, что основной вид очанки и ее разновидность достоверно различаются по числу растений, зацветающих из того или иного (по счету) узла. Этот вывод, вероятно, позволяет рассматривать разновидность очанки (*E. curta* Fr.) как зарождающийся вид.

При группировке выборки в четырехпольную или многопольную таблицы критерий χ^2 определяют по следующей формуле:

$$\chi^2=\sum\frac{d_1^2}{f_2'}+\sum\frac{d_2^2}{f_2'}+\ldots+\sum\frac{d_n^2}{f_n'}, \qquad (102)$$

где d — разница между эмпирическими и теоретически вычисленными или ожидаемыми частотами; f' — ожидаемые частоты.

Нулевую гипотезу, или предположение от отсутствии до-

стоверных различий между эмпирическими и ожидаемыми частотами, проверяют с помощью табл. VII Приложеиий, в которой приведеиы критические точки χ^2_{st} для разных уровней значимости α и чисел степеней свободы k. H_0-гипотеза отвергается, если $\chi_\phi^2 \geqslant \chi^2_{st}$ для принятого уровня значимости и числа степеней свободы, определяемого по числу строк и столбцов (без учета итогов таблицы) по следующей формуле:

$$k=(c-1)(r-1),$$

где с — число строк, г — число граф или столбцов таблицы.

Таблица 47

Количество узлов на стеблях растений	Частоты		f_1+f_2	f_2^2	$\frac{f_1^2}{f_1+f_2}$
	f_1	f_2			
6	1	0	1	1	1,00
7	8	0	8	64	8,00
8	23	1	24	529	22,04
9	30	11	41	900	21,95
10	38	18	56	1444	257,8
11	12	14	26	144	5,54
12	7	3	10	49	4,90
13	4	4	8	16	2,00
14	1	1	2	1	0,50
15	0	1	1	0	0,00
Сумма	124	53	177	—	91,71

При этом, как и при распределении выборки в вариационный ряд, правильное применение критерия χ^2 основано на требовании, чтобы в клетках таблицы было не меньше пяти ожидаемых или теоретически вычисленных вариант[1]. Кроме того, критерий χ^2 не следует применять к выборкам, значения которых выражены в процентах или другими относительными числами.

Пример 4. В классических опытах Г. Менделя по многогибридным скрещиваниям разных сортов гороха, отличающихся друг от друга по форме и окраске семян, было получено в первом опыте 330 круглых и 101 угловатых, а во втором — 355 желтых и 123 зеленых семени. Соответствуют ли эти данные ожидаемому по схеме Менделя расщеплению признаков в гибридном потомстве в отношении 3:1? Чтобы ответить на этот вопрос, нужно рассчитать ожидаемые частоты с доминант-

[1] См. примечание редактора к с. 138.

ыми и рецессивными признаками, а затем сравнить их с по-
уученными в опыте. Так, из общего числа 330+101=431 соб-
анных в первом случае семян ожидали (3/4)431=324,25 круг-
ых и (1/4)431=107,75 угловатых. Во втором случае из 355+
+123=478 семян ожидали (3/4)478=358,50 желтых и
1/4)478=119,50 зеленых семян. Сопоставляем эти величины с
оответствующими данными опыта (табл. 48).

Таблица 48

Показатели	Форма семян		Всего семян	Окраска семян		Всего семян
	круглая	угловатая		желтая	зеленая	
f	330	101	431	355	123	478
f'	323,25	107,75	431	358,5	119,5	478
d	6,75	6,75	—	3,5	3,5	—
d^2	45,56	45,56	—	12,25	12,25	—
d^2/f'	0,14	0,42	0,56	0,034	0,103	0,137

Критерий $\chi_1{}^2=0{,}56$ и $\chi_2{}^2=0{,}14$. Эти величины не превыша-
от критическую точку $\chi^2_{st}=3{,}84$ для $k=(2-1)(2-1)=1$ и
5%-ного уровня значимости, что не дает оснований для отвер-
ания нулевой гипотезы. Следовательно, данные опыта не про-
иворечат ожидаемому по схеме Менделя соотношению доми-
антных признаков и рецессивных 3:1.

Таблица 49

Показатели	Желтые гладкие	Желтые морщинистые	Зеленые гладкие	Зеленые морщинистые	Всего семян
f	315	108	101	32	556
f'	313	104	104	35	556
d	2	4	3	3	—
d^2	4	16	9	9	—
d^2/f'	0,01	0,15	0,09	0,26	0,51

Пример 5. В другом опыте Менделя в потомстве F_2 от посе-
а гибридных семян произошло расщепление на 315 желтых
ладких, 108 желтых морщинистых, 101 зеленых гладких и 32
еленых морщинистых семени. Необходимо выяснить, соответ-
твуют ли эти данные ожидаемому по схеме Менделя расщеп-
ению признаков в соотношении 9:3:3:1.

Как и в предыдущем случае, находим ожидаемые частоты: (9/16)556=313; (3/16)556=104 и (1/16)556=35. Сравниваем полученные в опыте и рассчитанные по схеме Менделя частоты (табл. 49). В данном случае $\chi_ф^2=0{,}51$. Для $k=(4-1)(2-1)=3$ и 5%-ного уровня значимости в табл. VII Приложения находим $\chi^2_{st}=7{,}82$. Так как $\chi_ф^2<\chi^2_{st}$, нет оснований возражать против нулевой гипотезы. Это означает, что данные опыта достоверно согласуются с ожидаемым соотношением 9:3:3:1.

В тех случаях, когда результаты наблюдений группируются в четырехпольную таблицу по схеме опыт — контроль, критерий χ^2 определяют по следующей формуле:

$$\chi^2=\frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}, \qquad (103$$

где a, b, c и d — численности групп, помещенные в клетках четырехпольной таблицы, а $n=a+b+c+d$ — общее число наблюдений.

Таблица 5(

Опыление	Число семян в корзинке		Всего семян
	заполненных	пустых	
Естественное (контроль)	$a=113$	$b=42$	$a+b=155$
Добавочное (опыт)	$c=131$	$d=11$	$c+d=142$
Всего	$a+c=244$	$b+d=53$	$n=297$

Пример 6. Испытывали влияние добавочного опыления на урожай подсолнечника. Полученные результаты приведены в табл. 50.

Из табл. 50 видно, что добавочное опыление дает прибавку урожая этой культуры. Применим критерий χ^2 к оценке полученных результатов:

$$\chi^2_ф=\frac{297(|113\cdot11-42\cdot131|-297/2)^2}{244\cdot53\cdot155\cdot142}=17{,}63.$$

Эта величина превосходит критическую точку $\chi^2_{st}=10{,}83$ для $k=1$ и 0,1%-ного уровня значимости, что позволяет отвергнуть нулевую гипотезу на высоком уровне значимости ($P<0{,}001$). Разницу между контролем и опытом следует признать статистически достоверной.

Из приведенного примера видно, что применение формулы (103) осложняет вычислительную работу, особенно при наличии многозначных чисел. Это неудобство можно обойти используя следующую формулу:

$$\chi^2=\sum_{b=1}^{k}\frac{(d-0{,}5)^2}{f'}, \qquad (104)$$

где $d=(f_i-f_i')$ — разность между наблюдаемыми f и ожидаемыми или вычисленными f' численностями групп, которые рассчитывают по формуле

$$f'=\frac{n_c n_r}{N}. \qquad (105)$$

Здесь n_c — итоги частот по строке, а n_r — итоги частот по графам или столбцам четырехпольной или многопольной таблицы; $N=n_c+n_r$ — общее число наблюдений.

Применим формулы (104) и (105) к только что рассмотренному примеру. Расчет величины критерия χ^2 показан в табл. 51.

Таблица 51

Частоты		$f-f'=d$	$d-0{,}5$	$(d-0{,}5)^2$	$\frac{(d-0{,}5)^2}{f'}$
наблюдаемые f	вычисленные f'				
113	127,3	14,3	13,8	190,44	1,5
131	116,7	14,3	13,8	190,44	1,6
42	27,7	14,3	13,8	190,44	6,9
11	25,3	14,3	13,8	190,44	7,5
$\Sigma=297$	$\Sigma=297$	—	—	—	$\Sigma=17{,}5$

Приведенные во второй графе вычисленные частоты (f') получены следующим образом: $f_1'=(244\cdot155)/297=127{,}3$ (см. табл. 51); $f_2'=(244\cdot142)/297=116{,}7$; $f_3'=(53\cdot155)/297=27{,}7$ и $f_4'=(53\cdot142)/297=25{,}3$. Остальные действия понятны из табл. 51. Суммируя последний столбец этой таблицы, находим $\chi_\Phi^2=17{,}5$.

VI.3. КРИТЕРИЙ ЯСТРЕМСКОГО *J*

Задавшись целью исследовать критерий согласия χ^2 Пирсона на его практическую годность, проф. Б. С. Ястремский нашел, что закон распределения χ^2 не дает базы для суждения о степени близости между теоретически вычисленными и эмпи-

рически наблюденными частотами. Критерий χ^2 указывает не на степень сходства между эмпирическими и вычисленными частотами, а лишь на вероятностную оценку расхождения между ними. Имея в виду эту особенность χ^2-критерия, Ястремский (1948) построил другой критерий согласия, который в общем виде записывается так:

$$J=\frac{|C-N|}{\sqrt{2N+4\theta}}, \qquad (106)$$

где $C=\sum\frac{(f-f')^2}{f'q}$; N — число групп или классов вариационного ряда (разность $|C-N|$ берут без учета знака); θ — величина, зависящая от числа групп N; при $N\leqslant 20\,\theta$ не превосходит 0,6. Так как число классов или групп обычно не превышает 20, то величину 4θ можно считать равной 2,4; $q=1-p$, где $p=f(t)$, т. е. функция нормированного отклонения, а f и f' — соответственно эмпирические и вычисленные частоты ряда.

Величина J имеет непрерывную функцию распределеиия и подчинена нормальному закону. Следовательно, с вероятностью $P=99{,}5\%$ можно утверждать, что различия, наблюдаемые между эмпирическими f и вычисленными f' частотами носят случайный характер, если $J\leqslant 3{,}0$.

Пример 7. Применим критерий согласия Ястремского к оценке ряда распределения длины тела у 267 мужчин. Необходимые данные содержатся в табл. 45. В последних четырех столбцах этой таблицы показан расчет величины C, входящей в состав формулы Ястремского, которая оказалась равной 1,95. Эта величина рассчитана следующим образом. В табл. 45 приведены значения функции $f(t)$, соответствующие нормированным отклонениям членов ряда $t=(x_i-\bar{x})/s_x$ (см. табл. 28). Так как частоты классов объединены, то и значения $f(t)$ тоже объединяются. Так, первая варианта x_1 отклоняется от средней на $t=-2{,}77$. Этой величине отвечает $f(t)=0{,}0086$ (см. табл. II Приложений). Соседняя варианта x_2 отклоняется от средней на $t=-2{,}03$. Этой величине соответствует $f(t)=0{,}0508$, поэтому $f(t)=0{,}0086+0{,}0508=0{,}0594$. Эта величина и записана в 6-м столбце табл. 45. Остальные действия понятны из той же таблицы.

Число вторичных классов равно 7. Подставляя известные величины в формулу (106), находим $J=(1{,}95-7)/\sqrt{2\cdot 7+2{,}4}=5{,}05/4{,}05=1{,}25$. Эта величина не превышает даже 5%-ного уровня значимости, которому отвечает $t=1{,}96$, что не дает оснований для отвергания H_0-гипотезы. Следовательно, можно утверждать, что вычисленные по нормальному закону частоты хорошо согласуются с частотами данного эмпирического ряда.

При использовании J-критерия частоты классов ряда, мень-

шие пяти, можно не объединять. Принято также вычислять -критерий упрощенным и приближенным способом, придавая величине C, входящей в состав формулы (106), следующее значение: $C=\sum\frac{(f-d')^2}{f'}$. Такого рода преобразования сокращают затраты времени и объем вычислительной работы при определении величины J-критерия Ястремского.

Пример 8. В табл. 27 содержатся эмпирические и вычисленные частоты распределения 517 случаев поражения клеток альфа-частицами. Проверим, отвечает ли это распределение закону Пуассона. Расчет вспомогательных величин приведен в табл. 52 (вычисление частот см. в табл. 26).

Таблица 52

Количество поражений клеток α-частицами m	Частоты		Разница $d=f-f'$	d^2	$\frac{d^2}{f'}$
	наблюдаемые f	вычисленные f'			
0	112	115,34	3,34	11,1556	0,10
1	168	173,04	5,04	25,4016	0,15
2	130	129,77	0,23	0,0529	0,00
3	68	64,88	3,12	9,7344	0,15
4	32	24,35	7,65	58,5225	2,40
5	5	7,29	2,29	5,2441	0,72
6	1	1,81	0,81	0,6561	0,36
7	1	0,41	0,59	0,3481	0,85
Сумма	517	516,90	—	—	4,73

В данном случае число классов равно 8, а $C=4{,}73$. Подставляя эти величины в формулу (106), находим

$$J=\frac{|4{,}73-8|}{\sqrt{2\cdot 8\cdot 2{,}1}}=\frac{3{,}27}{\sqrt{18{,}4}}=\frac{3{,}27}{4{,}29}=0{,}76.$$

Так как $J<3$, расхождения между эмпирическими и вычисленными по формуле Пуассона частотами данного ряда можно считать случайными.

Пример 9. Применим критерий согласия Ястремского к оценке расхождения между эмпирическими и вычисленными по формуле Шарлье (50) частотами распределения 200 хвоинок сосны обыкновенной. Необходимые данные содержатся в табл. 33,34. Расчет величины C приведен в табл. 53.

В данном случае $C=15{,}04$, число классов равно 13. Отсюда

$$J=\frac{|15{,}04-13|}{\sqrt{18+2{,}4}}=\frac{2{,}04}{\sqrt{20{,}4}}=\frac{2{,}04}{4{,}32}=0{,}47.$$

Таблица 53

Класс x, мм	Частоты		$d=f-f'$	d^2	$\frac{d^2}{f'}$
	f	f'			
125	2	0,6	1,4	1,96	3,27
175	2	2,4	0,4	0,16	0,07
225	4	3,9	0,1	0,01	0,00
275	5	7,6	2,6	6,76	0,89
325	7	13,3	6,3	39,69	2,98
375	25	22,1	2,9	8,41	0,38
425	39	32,8	6,2	38,44	1,17
475	46	39,6	6,4	40,96	1,03
525	31	37,4	6,4	40,96	1,10
575	23	25,4	2,4	6,76	0,23
625	13	12,0	1,0	1,00	0,08
675	2	3,5	1,5	2,25	0,64
725	1	0,2	0,8	0,64	3,20
Сумма	200	200,8	—	—	15,04

Так как $J<3$, можно считать, что формула (50) выбрана правильно и что между эмпирическими и вычисленными по этой формуле частотами существует полное согласие.

VI.4. ПРИЧИНЫ АСИММЕТРИИ ЭМПИРИЧЕСКИХ РАСПРЕДЕЛЕНИЙ

А. Кетле и Ф. Гальтои считали, что биологические признаки распределяются нормально. Вскоре, однако, К. Пирсон показал, что существует не один, а несколько типов распределений. Возникла необходимость выяснить причины различных отклонений от нормальной кривой. Решение этого вопроса имеет свою историю, останавливаться на которой не входит в задачу данного пособия.

В настоящее время можно указать иа следующие причины возникиовения асимметричных распределеиий. Одна из них — чисто *механическая,* связаиная с «неправильной» группировкой выборочиых данных в вариационный ряд. Впервые на это указал В. Иогансен. Отобрав 1522 фасолины и измерив каждую в сантиграммах (сг), он распределил их в вариационный ряд:

x_q	8,75	9,75	10,75	11,75	12,75	13,75	14,75	15,75	16,75
f	2	43	314	809	316	30	6	2	

Характеристики этого симметричного ряда следующие: $\bar{x}=12{,}25$ сг; $s_x=0{,}82$; $As=0{,}17\pm0{,}063$.

Затем границы классов были изменены путем округления дробей до целых чисел и частоты вариант распределились по классам следующим образом:

Классы x_i	9	10	11	12	13	14	15	16	17
Частоты f	7	67	466	761	201	15	4	1	

Характеристики этого ряда оказались следующие: $\bar{x}=12{,}25$ сг; $s_x=0{,}82$; $As=0{,}75\pm0{,}063$. Из приведенных данных видно, что в результате незначительных изменений границ классов средняя арифметическая и среднее квадратическое отклонение остались без изменения, тогда как коэффициент асимметрии As увеличился более чем в четыре раза.

Таблица 54

Количество растений на 1 м²	Измерено колосьев n	Средняя длина колосьев $\bar{x}$, мм	Среднее квадратическое отклонение s_x	Коэффициент	
				вариации Cv	асимметрии As
197	132	64,0	19,0	27,7	+0,004
222	115	63,3	18,5	29,2	+0,410
364	182	49,3	18,2	36,9	+0,620

Такого рода асимметрию Иогансен назвал *кажущейся* или *ложной,* возникшей исключительно вследствие технических причин. Отсюда следует, насколько важно соблюдение выработанных правил распределения результатов наблюдений в интервальной вариационный ряд.

Другая причина возникновения асимметрии — это *модифицирующие условия внешней среды,* в которой происходит развитие организма и формируются количественные признаки. Известно, например, что засуха, невыравненность почвенного питания растений и т. п. факторы среды вызывают асимметрию в распределении хозяйственно полезных признаков. Замечено также, что прн прочих равных условиях характеристики ряда распределения размеров колосьев у озимой ржи находятся в зависимости от густоты стояния растений в посеве этой культуры (табл. 54).

Из табл. 54 видно, что с увеличением численности растений на единице площади средняя длина колосьев уменьшается и, как следствие, происходит увеличение коэффициента вариации при более или менее стабильной величине среднего квадратического отклонения. Особенно сильно увеличивается коэффициент асимметрии.

Существует и еще одна причина возникновения асимметрии — *генетическая,* обусловленная взаимодействием аллельных и неаллельных активных генов. Известно, что количественные признаки наследуются полигенно. При отсутствии доминирования, эпистаза и других способов взаимодействия неаллельных генов действие аддитивных, т. е. однозначных или сходных по силе действия на признак, генов обусловливает промежуточный тип наследования, их нормальное распределение. Если же в процессе формирования признака имеет место взаимодействие генов, при котором одни активные гены подавляют или ограничивают активность других, то промежуточного типа наследования признака наблюдаться не будет и кривая распределения такого признака окажется асимметричной.

Каждая из названных причин действует на признак не изолированно, а, как правило, суммарно, поэтому без строгого статистического и генетического анализа нельзя выявить конкретную причину, вызывающую отклонение эмпирического распределения от нормальной кривой. Во всяком случае отклонения вариационных кривых от нормального закона, если они не случайны, могут указывать на постоянно действующую причину или причины, вызывающие асимметрию изучаемого признака, выяснение которых является больше задачей биологии, чем биометрии.

VI.5. ОЦЕНКА ТРАНСГРЕССИИ РЯДОВ

При распределении независимых выборок, взятых из разных генеральных совокупностей, нередко случается, что некоторая часть членов этих выборок оказывается в одних и тех же классах вариационного ряда. Такие ряды, у которых часть классов оказывается общей (несмотря на то что между средними арифметическими этих рядов может быть статистически достоверная разница), называют *трансгрессирующими,* а сам факт неполного разобщения вариационных рядов и их графиков — *трансгрессией.*

Вариационные кривые трансгрессирующих рядов выглядят так, что правая сторона одной кривой и левая сторона другой взаимно проникают друг в друга, так что под ними образуется часть общей площади в системе прямоугольных координат, показывающая величину трансгрессии.

В табл. 55 приведены ряды распределения кальция в сыворотке крови обезьян, страдавших припадками тетании (больных) и клинически здоровых. Из той же таблицы видно, что часть членов каждой из выборок распределилась по одним и тем же классам. Между тем разница $\bar{x}_2 - \bar{x}_1 = 11{,}90 - 8{,}92 = (2{,}98 \pm 0{,}21)$ мг% оказывается статистически достоверной.

Это пример трансгрессирующих рядов выборок из разных генеральных совокупностей, каждая из которых характеризуется своими параметрами. Наличие гипокальцемии в сыворотке крови, взятой у больных обезьян,— характерный признак их патологического состояния, связанного с гипофункцией паращитовидных желез. Однако, учитывая, что ряды распределения трансгрессируют, нельзя утверждать, что гипокальцемия — единственная причина возникновения припадков тетании у обезьян.

Таблица 55

Классы по уровню кальция в сыворотке крови, мг %	Срединные значения класса x_i	Частоты вариационных рядов	
		больных f_1	здоровых f_2
6,45—7,14	6,8	2	
7,15—7,84	7,5	6	
7,85—8,54	8,2	7	
8,55—9,24	8,9	12	2
9,25—9,94	9,6	9	3
9,95—10,64	10,3	2	9
10,65—11,34	11,0	3	17
11,35—12,04	11,7	1	25
12,05—12,74	12,4		23
12,75—13,44	13,1		10
13,45—14,14	13,8		7
14,15—14,84	14,5		4
Сумма	—	42	100
Средняя арифметическая $\bar{x}$		8,92	11,90
Среднее квадратическое отклонение s_x		1,13	1,20

Величину трансгрессии можно измерить, выразив сумму трансгрессирующих вариант в процентах от общей численности обеих выборок. Так, из табл. 55 следует, что из 142 вариант трансгрессируют в первом ряду $1+3+2+9+12=27$, а во втором — $2+3+9+17+25=56$ вариант, всего трансгрессируют $27+56=83$ варианты, что составляет 58,5% от общего числа наблюдений. Это довольно значительная трансгрессия: больше половины членов двух выборок распределяется по одним и тем же классам вариационного ряда. Понятно, что этот способ измерения величины трансгрессии не точен и дает лишь приближенное представление о ее размере. Более точно рассчитать величину трансгрессии Tr для нормально распределяющихся совокупностей можно с помощью следующей формулы:

$$Tr = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}, \qquad (107)$$

где n_1 и n_2 — объемы сопоставляемых распределений; $P_1 = 0{,}5 + 0{,}5\Phi(t_1)$ и $P_2 = 0{,}5 + 0{,}5\Phi(t_2)$. Здесь $t_1 = \frac{\min_2 - \bar{x}_1}{s_1}$ и $t_2 = \frac{\max_1 - \bar{x}_2}{s_2}$, а $\min_2 = \bar{x}_2 - 3s_2$ и $\max_1 = \bar{x}_1 + 3s_1$. Значения $\min_2$, $\max_1$, $\Phi(t_1)$ и $\Phi(t_2)$ легче уяснить из рис. 23, на котором изображены выровненные кривые распределения кальция в сыворотке крови больных и здоровых обезьян. Остальные символы объяснены выше.

Показатель трансгрессии выражают в долях единицы или процентах. Значения функции $\Phi(t)$ приведены в табл. I Приложений; эти значения нужно брать с отрицательным знаком, если $\min_2 > \bar{x}_1$ и $\max_1 < \bar{x}_2$, т. е. исходить из $P = 0{,}5 - 0{,}5\Phi(t)$.

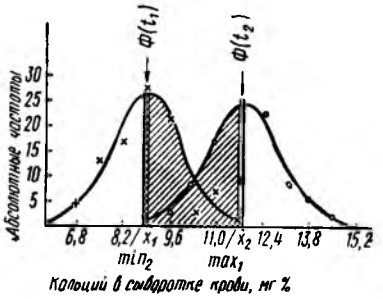


Рис. 23. Трансгрессия рядов распределения кальция в сыворотке крови больных ($\bar{x}_1$) и клинически здоровых ($\bar{x}_2$) обезьян

Пример 10. Рассчитать величину трансгрессии между рядами распределения кальциия (мг%) в сыворотке крови больных и здоровых обезьян. В данном случае $\bar{x}_1 = 8{,}92$ мг%; $\bar{x}_2 = 11{,}90$ мг%; $s_1 = 1{,}13$; $s_2 = 1{,}20$; $n_1 = 42$; $n_2 = 100$. Определяем $\min_2 = 11{,}90 - 3 \cdot 1{,}20 = 8{,}30$; $\max_1 = 8{,}92 + 3 \cdot 1{,}13 = 12{,}31$. Находим $t_1 = (8{,}30 - 8{,}92)/1{,}13 = -0{,}55$; $t_2 = (12{,}31 - 11{,}90)/1{,}20 = 0{,}34$. Вычисляем значения $P_1 = 0{,}5 + 0{,}5\Phi(-0{,}55) = 0{,}5 + 0{,}20885 = 0{,}70885$; $P_2 = 0{,}5 + 0{,}5\Phi(0{,}34) = 0{,}5 + 0{,}13305 = 0{,}63305$. Подставляем известные величины в формулу:

$$Tr = \frac{42 \cdot 0{,}70885 + 100 \cdot 0{,}63305}{42 + 100} = \frac{93{,}0767}{142} = 0{,}655, \text{ или } 65{,}5\,\%.$$

Пример 11. В табл. 56 приведены ряды распределения массы фасолин, которые были отобраны из урожая f_1 и из посевного материала f_2. Из данных таблицы видно, что эти ряды трансгрессируют, хотя и незначительно: из общего объема двух выборок ($n = 300$) трансгрессирует лишь 21 варианта, что составляет 7% от их общего числа. Вычислим показатель трансгрессии для этих распределений.

В данном случае $\bar{x}_1 = 319{,}0$ мг; $\bar{x}_2 = 573{,}5$ мг; $s_1 = 58{,}3$; $s_2 = 59{,}3$. Определяем $\min_2 = 573{,}5 - 3 \cdot 59{,}3 = 395{,}6$; $\max_1 = 319{,}0 + 3 \cdot 58{,}3 = 493{,}9$; $t_1 = (395{,}6 - 319{,}0)/58{,}3 = 1{,}31$; $t_2 = (493{,}9 - 573{,}5)/59{,}3 = -1{,}34$. Отсюда $P_1 = 0{,}5 - 0{,}5\Phi(1{,}31) = 0{,}5 - 0{,}5 \cdot 0{,}8098 = 0{,}095$; $P_2 = 0{,}5 - 0{,}5\Phi(-1{,}34) = 0{,}5 - 0{,}5 \cdot 0{,}8198 =$

$=0{,}090$[1]. Подставляем найденные величины в формулу:

$$Tr = \frac{200 \cdot 0{,}095 + 100 \cdot 0{,}090}{200 + 100} = 0{,}093, \text{ или } 9{,}3\,\%.$$

Как и следовало ожидать, трансгрессия этих рядов оказалась незначительной.

VI.6. ПРОВЕРКА СОМНИТЕЛЬНЫХ ВАРИАНТ

Отдельные варианты, попавшие в состав выборки, иногда сильно отличаются от остальных ее членов, так что возникает сомнение в их принадлежности к генеральной совокупности, из которой взята эта выборка. Сомнительная варианта может по-

Таблица 56

Классы для массы фасолин, мг	Срединные значения классов x_i	Частоты	
		урожая f_1	семян f_2
100—149,9	125	1	
150—199,9	175	5	
200—249,9	225	17	
250—299,9	275	45	
300—249,9	325	70	
350—399,9	375	51	
400—449,9	425	10	1
450—499,9	475	1	9
500—549,9	525		29
550—599,9	575		26
600—649,9	625		25
650—699,9	675		8
700—749,9	725		2
Сумма	—	200	100
Средняя арифметическая $\bar{x}$		319,0	573,5
Среднее квадратическое отклонение s_x		58,3	59,3

пасть в выборку ошибочно, что отразится на выводах, которые делают на основании выборочных данных. Такая варианта должна быть исключена при вычислении обобщающих характеристик статистической совокупности. Однако произвольно отбрасывать сомнительные варианты нельзя, так как они могут принадлежать к той же генеральной совокупности, из которой извлечены и другие члены выборки. Возникает задача статис-

[1] Величины $0{,}5\Phi(t_1)$ и $0{,}5\Phi(t_2)$ взяты с отрицательным знаком на том основании, что в данном примере $\min_2 > \bar{x}_1$ и $\max_1 < \bar{x}_2$.

тической проверки сомнительных вариант. При этом исходят из предположения, что сомнительные варианты принадлежат к одной и той же нормально распределяющейся генеральной совокупности.

Для проверки этой (нулевой) гипотезы применяют особые критерии. В качестве одного из них служит *нормированное отклонение t*, которое уже упоминалось ранее [см. формулу (20)]. Нулевую гипотезу отвергают, если $t_\Phi \geqslant t_{st}$ (критические значения t_{st} для 5%-ного и 1%-ного уровней значимости с учетом объема выборки n приведены в табл. XVI Приложений).

Пример 12. Собранный с шести опытных делянок урожай зерна озимой ржи варьировал следующим образом:

Номера делянок	1	2	3	4	5	6
Урожай, кг	21,9	24,6	20,8	25,1	30,8	23,2

Из приведенных данных выделяется варианта $x_5=30{,}8$, сильно отличающаяся от остальных членов выборки. Нужно проверить гипотезу H_0 о принадлежности этой варианты к данной генеральной совокупности. Находим характеристики выборки: $\bar{x}=\frac{21{,}9+24{,}6+\ldots+123{,}3}{6}=\frac{146{,}4}{6}=24{,}40$ кг; $s_x=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n=1}}=\sqrt{\frac{62{,}14}{5}}=\sqrt{12{,}428}=3{,}53$. Нормируем варианту 30,8: $t=(30{,}8-24{,}4)/3{,}53=1{,}81$. В табл. XVI для $\alpha=5\%$ и $n=6$ находим $t_{st}=2{,}07$. Так как $t_\Phi=1{,}81<2{,}07$, то нулевую гипотезу отвергнуть нельзя.

Другие критерии для проверки нулевой гипотезы основаны на использовании разностей между сомнительными и соседними членами ранжированного ряда. Для этого служат формулы

$$t_1=\frac{x_2-x_1}{x_{n-1}-x_1};\quad t_2=\frac{x_n-x_{n-1}}{x_n-x_2}. \qquad (108)$$

Вычисление t_1 применяют для проверки наименьших x_1, а t_2 — для проверки наибольших x_n членов ранжированного ряда. Нулевую гипотезу отвергают, если $t_\Phi \geqslant t_{st}$ для принятого уровня значимости α и объема выборки n. Критические точки для t_1 приведены в табл. XVII, а для t_2 — в табл. XVIII Приложений.

Пример 13. Проверить с помощью критерия t_2 вывод, который был описан при оценке варианты $x_5=30{,}8$. Для этого ранжируем выборку в порядке возрастания числовых значений признака:

Номера по порядку (ранги)	1	2	3	4	5	6
Варианты x_i	20,8	21,9	23,2	24,6	25,1	30,8

Подставляем нужные величины в формулу (108): $t_2=(30{,}8-25{,}1)/(30{,}8-21{,}9)=5{,}7/8{,}9=0{,}64$. В табл. XVIII При-

ложений для $\alpha=5\%$ и $n=6$ находим $t_{st}=0{,}69$. Так как $t_\Phi=0{,}64<0{,}69$, это не дает оснований для непринятия нулевой гипотезы. Следовательно, все члены данной выборки принадлежат к одной и той же генеральной совокупности и при расчете среднего урожая нельзя отбрасывать варианту 30,8.

Пример 14. Рассмотрим выборку следующего состава: 37, 40, 38, 28, 9, 32, 25, 26, 23 — всего 9 вариант. В данном случае вызывает сомнение варианта 9. Проверим ее принадлежность к данной генеральной совокупности. Ранжируем выборку следующим образом:

Номера по порядку . . .	1	2	3	4	5	6	7	8	9
Варианты x_i	9	23	25	26	28	32	37	38	40

Подставляя нужные величины в формулу, находим $t_1=(23-9)/(38-9)=14/29=0{,}48$.

Эта величина несколько превышает критическую точку $t_{st}=0{,}44$ для $\alpha=5\%$ и $n=9$, что опровергает нулевую гипотезу на 5%-ном уровне значимости (см. табл. XVII Приложений). Однако на уровне $\alpha=1\%$ нулевая гипотеза сохраняется. Как же поступить в таком случае с вариантой 9? Прежде чем принять окончательное решение по этому вопросу, следует проверить H_0-гипотезу с помощью более мощного параметрического критерия — t-нормированного отклонения. Характеристики рассматриваемой выборки: $\bar{x}=28{,}67$ и $s_x=9{,}59$. Отсюда нормированное отклонение $t_1=(9-28{,}67)/9{,}59=2{,}07$. По табл. XVI Приложений для $\alpha=5\%$ и $n=9$ находим $t_{st}=2{,}35$. Так как $t_\Phi=2{,}07<2{,}35$, нулевую гипотезу отбросить нельзя. Следовательно, при расчете обобщающих характеристик выборки оцениваемую варианту 9 исключать не следует.

ГЛАВА VII

ДИСПЕРСИОННЫЙ АНАЛИЗ

Сущность метода. Наряду с относительно простыми способами сравнения одной выборки с другой в исследовательской работе встречаются и более сложные задачи, когда приходится сравнивать одновременно несколько выборок, объединяемых в единый статистический комплекс. В таких случаях метод попарных сравнений выборочных характеристик оказывается обременительным, требующим большой вычислительной работы. Учитывая это обстоятельство, Р. Фишер (1925) предложил метод комплексной оценки сравниваемых средних, получивший название *дисперсионного анализа*. Этот метод основан на разложении общей дисперсии статистического комплекса на составляющие ее компоненты (отсюда и название метода), сравнивая

которые друг с другом посредством F-критерия можно определить, какую долю общей вариации учитываемого (результативного) признака обусловливает действие на него как регулируемых, так и не регулируемых в опыте факторов.

Так, если регулируемый фактор (например, доза удобрений) оказывает существенное влияние на результативный признак (урожай культуры), оно непременно скажется на величине групповых средних, которые будут заметно отличаться друг от друга. Таким образом, здесь происходит варьирование групповых средних, причиной которого является влияние регулируемого фактора.

Внутри каждой группы, входящей в статистический (дисперсионный) комплекс, тоже обнаружится варьирование, вызванное влиянием на признак не регулируемых в опыте факторов. Зависимость между этими источниками варьирования выразится равенством $D_y = D_x + D_e$, где D_x — межгрупповая девиата, представляющая собой сумму квадратов отклонений групповых средних $\bar{x}_i$ (их общее число — a) от общей средней $\bar{x}$ комплекса, взвешенную на численность вариант в группах n, т. е. $D_x = \sum_{i=1}^{a} \frac{n(\bar{x}_i - \bar{x})^2}{N}$ при $N = \Sigma n$; D_e (от англ. eггог — ошибка) — внутригрупповая девиата, представляющая сумму из сумм квадратов отклонений отдельных вариант x_j от их групповых средних $\bar{x}_i$, т. е. $D_e = \sum_{i}^{a} \left[\sum_{j=1}^{n} (x_j - \bar{x}_i)^2 \right]$; D_y — общая девиата или сумма квадратов отклонений вариант x_i (дат, по терминологии Фишера) от общей средней $\bar{x}$ комплекса, т. е. $D_y = \sum_{j=1} (x_j - \bar{x})^2$.

Деление сумм квадратов отклонений (девиат) на числа степеней свободы k дает *выборочные дисперсии* $s_y^2 = D_y/k_y$; $s_x^2 = D_x/k_x$; $s_e^2 = D_e/k_e$, которые служат оценками соответствующих генеральных параметров: s_y^2 является оценкой общей дисперсии всего комплекса σ_y^2; s_x^2 — оценкой межгрупповой дисперсии σ_x^2; s_e^2 — оценкой внутригрупповой, или остаточной, дисперсии σ_e^2.

Отношение межгрупповой дисперсии (называемой также факториальной дисперсией, так как она зависит от действия регулируемых факторов) к внутригрупповой, или остаточной, дисперсии служит критерием оценки влияния регулируемых в опыте факторов на результативный призиак, т. е. $F = s_x^2/s_e^2$ (при $s_x^2 \geqslant s_e^2$).

Нулевая гипотеза сводится к предположению, что генеральные межгрупповые средние и дисперсии равны между собой и различия, наблюдаемые между выборочными показателями,

вызваны случайными причинами, а не влиянием на признак регулируемых факторов. Нулевую гипотезу отвергают, если $F_ф \geqslant F_{st}$ для принятого уровня значимости α и чисел степеней свободы k_x и k_e, и принимают, если $F_ф < F_{st}$; при этом различия, наблюдаемые между групповыми средними комплекса, признают статистически недостоверными.

После того как действие регулируемого фактора, нескольких факторов или их совместного действия на признак будет доказано, т. е. окажется статистически достоверным, переходят, когда это необходимо, к сравнительной оценке групповых средних. Заключительным этапом дисперсионного анализа является оценка силы влияния отдельных факторов или их совместного действия на признак.

Будучи методом одновременных сравнений выборочных средних, дисперсионный анализ предъявляет определенные требования к группировке выборочных данных и к планированию наблюдений. Результаты наблюдений, подлежащие дисперсионному анализу, группируют с учетом градаций каждого регулируемого фактора, воздействующего на признак, например по дозам удобрений, по срокам или способам внесения их в почву, по породной принадлежности экспериментальных животных, их возрастному составу и т. д.

Правильное применение дисперсионного анализа предполагает также нормальное или близкое к нормальному распределение совокупности, из которой взяты выборки, объединяемые в дисперсионный комплекс. При этом важно, чтобы дисперсии выборочных групп были одинаковыми или не очень сильно отличались друг от друга. Не менее важным является и то, чтобы при планировании наблюдений и особенно при обработке их результатов в группах дисперсионного комплекса содержались одинаковые или пропорциональные числа вариант (дат), что значительно облегчает дисперсионный анализ.

Дисперсиоиный анализ возник в процессе усовершенствования методики сельскохозяйственного опытного дела. Однако вскоре его стали применять не только в биологии и смежных с ней областях знания, но и в технике, а затем в психологии и педагогике.

Дисперсионный анализ характеризуется строгой логичностью и последовательностью вычислительных операций. Ценность этого метода заключается в том, что он позволяет выявить и суммарное действие факторов, и действие каждого регулируемого в опыте фактора в отдельности, и действие различных сочетаний факторов друг с другом на результативный признак.

Основные понятия и символы. Признаки, изменяющиеся под воздействием тех или иных причин, называют *результативными*. Причины, вызвавшие изменение величины результативного

признака или признаков, принято называть *факторами*. Например, масса тела, его линейные размеры, физическое развитие организма; урожайность той или иной культуры; успеваемость учащихся и т. д.— все это признаки, на которые оказывают влияние самые различные факторы: элементы или режим питания, физические или умственные упражнения, дозы лекарственных или токсических веществ, удобрений и т. д. Факторы обозначают прописными начальными буквами латинского алфавита (*A, B, C,...*), учитываемые признаки — конечными (*X, Y, Z*).

Существует много факторов, воздействующих на один и тот же признак. В опыте регулируются лишь некоторые из них; их называют *регулируемыми* или *организованными факторами* в отличие от факторов, которые не подвергаются регулированию, хотя и оказывают воздействие на величину результативного признака. Обычно каждый регулируемый фактор испытывают серийно, т. е. в виде нескольких независимых друг от друга групп или градаций.

Градации принято обозначать теми же буквами, что и факторы. Например, градации фактора A обозначают через A_1, A_2, A_3 и т. д., а градации фактора B — соответственно через B_1, B_2, B_3 и т. д. Число градаций того или иного фактора определяется условиями опыта, например испытываемыми дозами удобрений, количеством сортов, подвергаемых испытанию на урожайность, и т. д. Результативные признаки также могут иметь свои градации, на которых испытывают действие регулируемых факторов.

Выше было указано, что дисперсионный анализ позволяет учитывать различные сочетания действия регулируемых факторов на результативный признак. Эти свойства не распространяются на не регулируемые в опыте факторы, действие которых на признак учитывают не дифференцированно, а суммарно. При этом дисперсионный анализ позволяет выражать учитываемые признаки не только абсолютными единицами измерения и счета, но и в баллах, индексах и других относительных и условных единицах.

Условия образования и виды дисперсионных комплексов. Статистические, или дисперсионные, комплексы могут формироваться как в планах намечаемых исследований, так и на основании уже собранных данных, подвергаемых дисперсионному анализу. При образовании дисперсионных комплексов необходимо соблюдать по крайней мере два важных *условия, гарантирующих правильное применение дисперсионного анализа.* 1. Действующие на признак регулируемые факторы должны быть независимы друг от друга. 2. Выборки, группируемые в статистический комплекс, должны производиться по принципу

рандомизации, т. е. способом случайного отбора из нормально распределяющейся совокупности.

Структуру дисперсионного комплекса определяет число градаций регулируемого фактора или факторов, а также число подразделений или групп, образуемых по результативному признаку. Форму дисперсионного комплекса задают таблицей, в которой число строк соответствует числу подразделений результативного признака, а число столбцов равно числу градаций регулируемого фактора или нескольких факторов с их градациями.

Если испытывают действие на признак одного регулируемого фактора, дисперсионный комплекс будет *однофакторным,* если одновременно исследуют действие на признак двух, трех или большего числа регулируемых факторов, комплекс называют *двух-, трех-* и *многофакторным.* Числовые значения результативного признака, т. е. варианты или даты, могут распределяться по градациям комплекса равномерно, пропорционально и неравномерно, поэтому дисперсионные комплексы называют *равномерными, пропорциональными* и *неравномерными.* Равномерные и пропорциоиальные комплексы носят общее название *ортогональные,* а неравномерные комплексы называют *неортогональными.*

В ортогональных комплексах осуществляется равенство $D_y = D_x + D_e$; в двухфакторных — $D_y = D_A + D_B + D_{AB} + D_e$; в неортогональных комплексах это равенство нарушается. Эту особенность следует учитывать при планировании опытов, а при проведении дисперсиоиного анализа — стремиться к тому, чтобы в градациях многофакторного комплекса были одинаковые или пропорциоиальные числа вариант, что значительно облегчает и упрощает вычислительную работу.

VII.1. АНАЛИЗ ОДНОФАКТОРНЫХ КОМПЛЕКСОВ

Равночисленные комплексы. Однофакторные дисперсионные комплексы могут быть равномерными и неравномерными. Независимо от этого техника дисперсионного анализа однофакторных комплексов сводится главным образом к расчету показателей варьирования, которыми в области дисперсионного анализа служат средние квадраты отклонений, или дисперсии, а также к расчету групповых средних $\bar{x}_i$ и общей средней арифметической для всего комплекса $\bar{x}$. Обычно дисперсионный анализ проводят по определенной схеме. Дисперсионный анализ однофакторных равномерных комплексов удобно проводить по следующей схеме.

1. Первичные данные, подлежащие дисперсионному анализу, группируют в виде комбинационной таблицы, в которой градации организованного (регулируемого) фактора A обычно

располагают по горизонтали в верхней части таблицы, а числовые значения признака X, т. е. варианты x, или даты, размещают соответственно по градациям фактора A (см. для примера табл. 59). Можно избрать и другую форму группировки (см. ниже), но предлагаемая здесь форма очень удобна при вычислении вспомогательных величин, необходимых для расчета девиат.

2. Сгруппировав исходные данные, как указано в п. 1, приступают к расчету вспомогательных величин Σx_i, $\Sigma(\Sigma x_i)^2$ и Σx_i^2.

3. Затем переходят к расчету девиат:

$$D_y = \sum_{l=1}^{N} x_l^2 - H; \qquad (109)$$

$D_A = \sum^{a} \frac{(\Sigma x_l)^2}{n} - H$ или в случае равномерного комплекса

$$D_A = \frac{\sum^{a}(\sum x_l)^2}{n} - H \text{ или } D_A = n\sum^{a}\bar{x}_l^2 - H; \qquad (110)$$

$$D_e = \sum_{l=1}^{N} x_l^2 - \frac{\Sigma(\Sigma x_l)^2}{n} \text{ или } D_e = D_y - D_A. \qquad (111)$$

Повторяемая в этих формулах величина $H = (\Sigma x_i)^2/N$, где x_i — варианты, или даты, входящие в состав комплекса; $N = \Sigma n$ — общее число наблюдений, или объем комплекса; n — численность вариант x_i в каждой из градаций дисперсионного комплекса. D_A — факториальная девиата, характеризующая межгрупповое варьирование не вообще (как девиата D_x), а применительно к конкретному фактору, который здесь обозначен буквой A. Между D_x и D_A существует принципиальная разница, хотя в однофакторных комплексах она неощутима.

4. Закончив расчет девиат, переходят к определению чисел степеней свободы k, которые равны:

$k_y = N - 1$ для общего варьирования;

$k_A = a - 1$ для факториального варьирования;

$k_e = (N-1) - (a-1) = N - a$ для остаточной вариации.

Через a обозначено число градаций фактора A.

Как и равенство $D_y = D_x + D_e$, числа степеней свободы находятся между собой в определенных количественных соотношениях: $k_y = k_x + k_e$. Эти равенства позволяют контролировать правильность расчета как девиат, так и чисел степеней свободы.

5. Делением девиат на соответствующие числа степеней свободы получают выборочные дисперсии:

$s_y^2 = \frac{D_A}{N-1}$ общая для всего комплекса;

$s_A^2 = \frac{D_A}{a-1}$ межгрупповая, или факториальная;

$s_e^2 = \frac{D_e}{N-a}$ внутригрупповая, или остаточная.

6. Наконец определяют дисперсионное отношение $F = s_A^2 / s_e^2$ (при $s_A^2 \geqslant s_e^2$), по которому судят о действии фактора A на результативный признак. Так как фактически полученное дисперсионное отношение ($F_\Phi = s_A^2 / s_e^2$) является величиной случайной, его необходимо сравнить с табличным (стандартным) значением критерия Фишера F_{st} для принятого уровня значимости α и чисел степеней свободы k_A и k_e. При этом число степеней свободы для большей дисперсии находят в верхней строке, а для меньшей — в первом столбце таблицы Фишера (см. табл. VI Приложений).

Таблица 57

Вариация	Числа степеней свободы k	Суммы квадратов отклонений, или девиаты D	Средние квадраты отклонений, или дисперсии s^2	Дисперсионное отношение F_Φ
Факториальная	$k_A = a-1$	D_A	$s_A^2 = D_A / k_A$	$F_\Phi = s_A^2 / s_e^2$
Остаточная	$k_e = N-a$	D_e	$s_e^2 = D_e / k_e$	—
Общая	$k_y = N-1$	D_y	$s_y^2 = D_y / k_y$	—

Нулевую гипотезу отвергают и эффективность действия фактора A на результативный признак X признают статистически достоверной, если $F_\Phi \geqslant F_{st}$; в противном случае отвергать нулевую гипотезу нельзя.

Обычно результаты дисперсионного анализа сводят в таблицу, общий вид которой представлен в табл. 57.

Пример 1. На учебно-опытном участке агростанции изучали влияние различных способов внесения в почву органических удобрений на урожай зеленой массы кукурузы. Опыт проводили на десятиметровых делянках в трех вариантах, не считая контроля. Каждый вариант опыта имел трехкратную повторность. Результаты опыта приведены в табл. 58.

Из данных табл. 58 видно, что полученные результаты варьируют как по вариантам, так и по повторностям. Чтобы установить, случайны или достоверны различия между средними

Таблица 5

Варианты опыта	Урожай по повторностям, кг			Средний урожай $\bar{x}_i$
	1	2	3	
Контроль	21,2	28,0	31,2	26,8
Удобрения помещали:				
ниже семян на 4 см	23,6	22,6	28,0	24,7
в стороне от семян на 4 см	24,0	30,0	29,2	27,7
выше заделки семян на 4 см	29,2	28,0	27,0	28,1

арифметическими групп, подвергнем эти данные дисперсионному анализу. Обозначим фактор, регулируемый в опыте, через A, а его градации (варианты опыта) — соответственно через A_1, A_2, A_3 и A_4. Для упрощения расчетов вспомогательных ве-

Таблица 5

Урожай по повторностям X	Градации фактора A (варианты опыта)				Суммы
	A_1	A_2	A_3	A_4	
x_i—20	1,2 8,0 11,2	3,6 2,6 8,0	4,0 10,0 9,2	9,2 8,0 7,0	$a=4$
n	3	3	3	3	$\Sigma n=N=12$
Σx_i	20,4	14,2	23,2	24,2	$\Sigma x_i=82$
$(\Sigma x_i)^2$	416,16	201,64	538,24	585,64	$\Sigma(\Sigma x_i)^2=1741,68$
Σx_i^2	190,88	83,72	200,64	197,64	$\Sigma x_i^2=672,88$

личин уменьшим каждую варианту комплекса на 20, т. е. вместо x_i будем оперировать значениями $x_i—20=x_i^*$, что не повлияет на конечные результаты дисперсионного анализа. Для удобства расчетов вспомогательных величин сгруппируем преобразованные данные так, чтобы градации фактора A помещались в верхней части комбинационной таблицы (табл. 59).

Переходим к определению девиат. Предварительно найдем величину $H=82^2/12=560{,}33$. Затем определяем общую девиату: $D_y=\Sigma x_i^2-H=672{,}88-560{,}33=112{,}55$; факториальную девиату:

$$D_A=\frac{\Sigma(\Sigma x_l)^2}{n}-H=\frac{1741{,}68}{3}-560{,}33=580{,}56-560{,}33=20{,}23;$$

наконец, остаточную девиату: $D_e=D_y-D_A=112{,}55-20{,}23=92{,}32$.

Определяем числа степеней свободы. Так как комплекс содержит 12 вариант, число степеней свободы для общей дисперсии $k_y=N-1=12-1=11$. Фактор A содержит четыре градации (три варианта опыта и контроль); следовательно, число степеней свободы для факториальной дисперсии $k_A=a-1=4-1=3$. Для внутригрупповой, или остаточной, дисперсии число степеней свободы $k_e=k_y-k_A=11-3=8$ (или $k_e=N-a=12-4=8$). Проверим правильность расчета: $k_A+k_e=k_y=3+8=11$. Расчет произведен правильно.

Переходим к определению дисперсий: факториальной $s_A^2=D_A/k_A=20{,}23/3=6{,}74$ и остаточной $s_e^2=D_e/k_e=92{,}32/8=11{,}54$. Общую дисперсию вычислять нет необходимости, поскольку при выяснении влияния фактора A на результативный признак X используется отношение факториальной дисперсии к остаточной дисперсии; общая дисперсия в таком случае применения не находит.

В данном случае оказалось $s_A^2<s_e^2$. Это означает, что межгрупповая вариация не превышает внутригруппового случайного уровня и, следовательно, считать достоверным влияние фактора на исследуемый признак нет оснований. С другой стороны, обнаруженное соотношение двух дисперсий может вызывать недоумение, так как по теории должно быть $s_A^2\geqslant s_e^2$. Однако в связи с влиянием случайностей выборок при справедливости нулевой гипотезы может наблюдаться не только небольшое (незначимое) превышение уровня s_A^2 над s_e^2, но и обратное соотношение, как это обнаружено в рассматриваемом примере. С другой стороны, значительное уменьшение s_A^2 по сравнению с s_e^2 может свидетельствовать о серьезных нарушениях требований случайностей образования выборок или о других нарушениях условий корректности методики получения экспериментальных данных. Какова ситуация в данном случае? Если уменьшение s_A^2 по сравнению с s_e^2 имеет случайный в рамках справедливости нулевой гипотезы характер, то применение F-критерия ($F=s_e^2/s_A^2$) должно дать незначимые результаты. Действительно, $F_\Phi=11{,}54/6{,}74=1{,}71$, что значительно меньше $F_{st}=8{,}85$ для $k_e=8$ и $k_A=3$. Поэтому можно считать, что различия в двух величинах дисперсий отсутствуют и проверяемая гипотеза сохраняется.

Пример 2. На одной из опытных станций испытывали урожайность шести местных сортов пшеницы. Опыт проводили в четырехкратной повторности по каждому сорту. Полученные результаты приведены в табл. 60.

Таблица 60

Номера сортов	Урожай по повторностям, ц/га				Средний урожай $\bar{x}_i$
	1	2	3	4	
1	26,1	29,2	30,0	27,3	28,2
2	25,0	24,3	28,5	29,0	26,7
3	27,2	26,4	31,0	26,4	27,8
4	23,6	27,2	25,2	24,8	25,2
5	30,0	33,0	36,0	29,8	32,2
6	23,0	26,0	26,0	24,8	25,0

Из табл. 60 видно, что на одни и те же условия выращивания различные сорта реагируют по-разному. Подвергнем эти данные дисперсному анализу. Как и в предыдущем примере, чтобы облегчить вычислительную работу, преобразуем дро-

Таблица 6

Урожай по повторностям	Сорта пшеницы (градация фактора A)						Суммы
	1	2	3	4	5	6	
x_i*	41 72 80 53	30 23 65 70	52 44 90 44	16 52 32 28	80 110 140 78	10 40 40 28	$a=6$
n	4	4	4	4	4	4	$N=24$
Σx_i	246	188	230	128	408	118	1318
$(\Sigma x_i)^2$	60516	35344	52900	16384	166464	13924	345532
Σx_i^2	16074	10554	14676	4768	44184	4084	94340

ные числа следующим образом: каждую варианту комплекса уменьшим на одно и то же число $A=22$, близкое к $x_{min}=23,0$. Затем полученные разности умножим на $K=10$, что позволит избавиться от дробей. В результате получим преобразованные

числовые значения результативного признака (26,1—22)10=
=41, (25,0—22)10=30, (27,2—22)10=52 и т. д. Переходим к
расчету вспомогательных величин (табл. 61).

Рассчитываем суммы квадратов отклонений (девиаты), приводя их делением на $K^2=10^2=100$ к исходным величинам:
$D_y=\sum x_i^2-H=\frac{1}{100}\left(94\,340-\frac{1318^2}{24}\right)=\frac{1}{100}(94\,340-72380{,}2)=$
$=21\,959{,}8/100=219{,}598$; $D_A=\frac{\Sigma(\Sigma x_l)^2}{n}-H=\frac{1}{100}\left(\frac{345\,532}{4}-\right.$
$\left.-72380{,}2\right)=\frac{1}{100}(86\,383-72380{,}2)=14002{,}8/100=140{,}028$;
$D_e=D_y-D_A=219{,}598-140{,}028=79{,}570$. Переходим к определению чисел степеней свободы: $k_y=24-1=23$; $k_A=6-1=5$ и $k_e=24-6=18$. Наконец находим величины факториальной и остаточной дисперсий и сводим результаты дисперсионного анализа в заключительную таблицу (табл. 62).

Таблица 62

Варьирование	Степени свободы	Девиаты D	Дисперсии s^2	F_Φ	F_{st}	
					5%	1%
По фактору A	5	140,0	$\frac{140}{5}=28{,}0$	6,4	2,8	4,2
Остаточное	18	79,6	4,4	—	—	—
Общее	23	219,6	—	—	—	—

Последние графы этой таблицы содержат критические (процентные) точки F_{st}, которые содержатся в таблице Фишера (см. табл. VI Приложений) для двух уровней значимости и чисел степеней свободы $k_1=k_A=5$ (находят по горизонтали табл. VI Приложений) и $k_2=k_e=18$ (находят в первой графе той же таблицы). Поскольку $F_\Phi>F_{st}$, нулевую гипотезу отвергают на 1%-ном уровне значимости ($P<0{,}01$). Следовательно, с вероятностью более 99% можно заключить, что разница в урожайности между сортами не случайна.

Неравночисленные комплексы. Дисперсионный анализ однофакторных неравномерных комплексов, т. е. комплексов, в градациях которых содержатся разные числа вариант x_i, принципиально не отличается от анализа равномерных комплексов. Однако в связи с тем, что групповые средние неравномерных

комплексов имеют разный статистический вес n_j, факториальную девиату следует вычислять по формуле

$$D_A = \sum_j^a \frac{(\sum x_l)^2}{n_j} - H \text{ или} \qquad (112)$$

$$D_A = \sum_j^a (n_j \bar{x}_j^2) - H; \quad D_A = \sum_j^a [n_j (\bar{x}_j - \bar{x})^2]. \qquad (113)$$

Пример 3. Испытывали влияние различных доз минеральных удобрений на урожайность озимой ржи. Результаты испытаний приведены в табл. 63.

Таблица 63

Дозы удобрений, кг/га	Урожай по повторностям, ц/га						n_l	Средний урожай $\bar{x}_l$
	1	2	3	4	5	6		
15	8,0	8,4	9,0	8,6			4	8,5
20	8,2	9,0	10,0	10,0	9,2	10,0	6	9,4
25	11,0	13,0		12,0			3	12,0
30	7,5	8,5					2	8,0

Здесь результативным признаком X является урожайность ржи, а регулируемым фактором A — дозы удобрений. Фактор A имеет четыре градации, т. е. $a=4$. Подвергнем эти данные дисперсионному анализу. Предварительно рассчитаем вспомогательные величины, построив таблицу таким образом, чтобы градации фактора A располагались по вершинам столбцов, а значения результативного признака X распределялись по градациям фактора A (табл. 64).

Рассчитав вспомогательные величины, переходим к определению девиат и чисел степеней свободы: $D_y = \Sigma x_i^2 - H = 1384{,}90 - (142{,}4)^2/15 = 1384{,}90 - 1351{,}85 = 33{,}05$; $D_A = \sum_j^a \frac{(\sum x_l)^2}{n} - H = 1379{,}16 - 1351{,}85 = 27{,}31$; $D_e = D_y - D_A = 33{,}05 - 27{,}31 = 5{,}74$. $k_y = 15 - 1 = 14$; $k_A = 4 - 1 = 3$; $k_e = 15 - 4 = 11$.

Находим значения дисперсий: $s_A^2 = D_A/k_A = 27{,}31/3 = 9{,}1$; $s_e^2 = D_e/k_e = 5{,}74/11 = 0{,}52$. Дисперсионное отношение $F_\Phi = s_A^2/s_e^2 = 9{,}1/0{,}52 = 17{,}5$. Эта величина значительно превышает критическую точку $F_{st} = 6{,}2$ для $k_A = 3$ (находим по горизонтали таблицы Фишера), $k_e = 11$ (находим в первом столбце той же таблицы) и 1%-ного уровня значимости, что дает основание для отвергания нулевой гипотезы. Следовательно, с вероятностью, большей 99%, можно утверждать, что различия между

групповыми средними комплекса не являются случайными, они вызваны действием испытываемых доз удобрений на урожай озимой ржи.

Применение корреляционных таблиц. Довольно часто, особенно в выборках большого объема, отдельные варианты неоднократно повторяются, что позволяет распределять такие выборки в вариационный ряд или в ряд ранжированных значений признака. В подобных случаях удобной формой группировки

Таблица 64

Урожай по повторностям	Дозы удобрений (градации фактора A)				Суммы
	1 (15)	2 (20)	3 (25)	4 (30)	
x_i	8,0 8,4 9,0 8,6	8,2 9,0 10,0 10,0 9,2 10,0	11,0 13,0 12,0	7,5 8,5	$a=4$
n_j	4	6	3	2	$N=15$
Σx_i	34,0	56,4	36,0	16,0	142,4
$(\Sigma x_i)^2$	1156,00	3180,96	1296,00	256,00	—
$(\Sigma x_i)^2/n_j$	289,00	530,16	432,00	128,00	1379,16
Σx_i^2	289,52	532,88	434,00	128,50	1384,90

исходных данных, подвергаемых дисперсионному анализу, будет *корреляционная решетка,* образуемая сочетанием строк и столбцов, число которых равно числу групп или классов сопряженных рядов. Классы располагаются в верхней строке и в первом (слева) столбце корреляционной таблицы; общие частоты, обозначаемые символом f_{xy}, распределяются по ячейкам решетки.

Классы, или значения признаков, помещаемые в верхней строке таблицы, располагаются обычно слева направо в возрастающем порядке, а в первом столбце таблицы — в убывающем порядке, т. е. сверху вниз. При этом промежутки между классами могут быть равио- и неравиовеликими. При наличии нерав-

комплексов имеют разный статистический вес n_j, факториальную девиату следует вычислять по формуле

$$D_A = \sum_j^a \frac{(\sum x_l)^2}{n_j} - H \text{ или} \qquad (112)$$

$$D_A = \sum_j^a (n_j \bar{x}_j^2) - H; \quad D_A = \sum_j^a [n_j (\bar{x}_j - \bar{x})^2]. \qquad (113)$$

Пример 3. Испытывали влияние различных доз минеральных удобрений на урожайность озимой ржи. Результаты испытаний приведены в табл. 63.

Таблица 63

Дозы удобрений, кг/га	Урожай по повторностям, ц/га						n_l	Средний урожай $\bar{x}_l$
	1	2	3	4	5	6		
15	8,0	8,4	9,0	8,6			4	8,5
20	8,2	9,0	10,0	10,0	9,2	10,0	6	9,4
25	11,0	13,0		12,0			3	12,0
30	7,5	8,5					2	8,0

Здесь результативным признаком X является урожайность ржи, а регулируемым фактором A — дозы удобрений. Фактор A имеет четыре градации, т. е. $a=4$. Подвергнем эти данные дисперсионному анализу. Предварительно рассчитаем вспомогательные величины, построив таблицу таким образом, чтобы градации фактора A располагались по вершинам столбцов, а значения результативного признака X распределялись по градациям фактора A (табл. 64).

Рассчитав вспомогательные величины, переходим к определению девиат и чисел степеней свободы: $D_y = \Sigma x_i^2 - H = 1384{,}90 - (142{,}4)^2/15 = 1384{,}90 - 1351{,}85 = 33{,}05$; $D_A = \sum_j^a \frac{(\sum x_l)^2}{n} - H = 1379{,}16 - 1351{,}85 = 27{,}31$; $D_e = D_y - D_A = 33{,}05 - 27{,}31 = 5{,}74$. $k_y = 15 - 1 = 14$; $k_A = 4 - 1 = 3$; $k_e = 15 - 4 = 11$.

Находим значения дисперсий: $s_A^2 = D_A/k_A = 27{,}31/3 = 9{,}1$; $s_e^2 = D_e/k_e = 5{,}74/11 = 0{,}52$. Дисперсионное отношение $F_\Phi = s_A^2/s_e^2 = 9{,}1/0{,}52 = 17{,}5$. Эта величина значительно превышает критическую точку $F_{st} = 6{,}2$ для $k_A = 3$ (находим по горизонтали таблицы Фишера), $k_e = 11$ (находим в первом столбце той же таблицы) и 1%-ного уровня значимости, что дает основание для отвергания нулевой гипотезы. Следовательно, с вероятностью, большей 99%, можно утверждать, что различия между

Подставляя найденные величины в формулы (114) и (115), определяем девиаты: $D_y = 1384{,}90 - (142{,}4)^2/15 = 1384{,}90 - 1351{,}85 = 33{,}05$; $D_A = 1379{,}16 - 1351{,}85 = 27{,}31$; $D_e = 33{,}05 - 27{,}31 = 5{,}74$. Получился тот же результат, что и выше. Дальнейший ход анализа понятен из предыдущего примера.

Если межклассовые промежутки рядов X и Y одинаковы, девиаты проще определять по следующим формулам:

$$D_y = \sum (f_x a_x^2) - H; \tag{116}$$

$$D_A = \sum \frac{(f_{xy} a_x)^2}{f_y} - H, \tag{117}$$

где $H = \frac{(\sum f_x a_x)^2}{N}$; a_x — отклонения классов от условного нуля; остальные символы те же, что и в формулах (114) и (115).

Таблица 66

Урожай X, ц/га	Удобрения Y, т/га — фактор A					f_x	a_x	$f_x a_x$	$f_x a^2_x$
	1	2	3	4	5				
20			1		5	6	4	24	96
19		1	3	5	4	13	3	39	117
18	2	2	8	4		16	2	32	64
17	1	3	5		1	10	1	10	10
10	2		3			5	0	0	0
f_y	5	6	20	9	10	50	—	105	287
$f_{xy} a_x$	5	10	34	23	33	$\Sigma (f_{xy} a_x) = 105$			
$\frac{(f_{xy}\, a_x)^2}{f_y}$	5,0	16,67	57,80	58,78	108,90	$\Sigma \frac{(f_{xy}\, a_x)^2}{f_y} = 247{,}15$			

Пример 5. Испытывали влияние различных доз вносимых в почву органических удобрений на урожай пшеницы. Полученные результаты и их обработка приведены в табл. 66.

Как и в предыдущем примере, крылья этой таблицы служат для вычисления вспомогательных величин. Подставляя эти величины в формулы (116) и (117), находим: $D_y = 287 - 105^2/50 = 287 - 220{,}50 = 66{,}50$; $D_A = 247{,}15 - 220{,}50 = 26{,}55$; $D_e = 66{,}50 - 26{,}65 = 39{,}85$. Определяем числа степеней свободы: $k_y = 50 - 1 = 49$; $k_A = 5 - 1 = 4$; $k_e = 50 - 5 = 45$. Находим значения дис-

персий: $S_A^2=26{,}65/4=6{,}66$; $s_e^2=39{,}85/45=0{,}89$. Отсюда $F_\Phi=6{,}66/0{,}89=7{,}5>F_{st}=3{,}76$ для $k_A=4$, $k_e=45$ и $\alpha=0{,}01$ (см. табл. VI Приложений). Нулевая гипотеза отвергается на высоком уровне значимости ($P<0{,}01$).

Ранговый анализ. Равночисленные (по объему) комплексы. Правильное применение дисперсионного анализа основано на предположении о нормальном распределении совокупностей, из которых извлечены выборки, входящие в дисперсионный комплекс. Если это условие не выполняется или о характере распределения нет сведений, применяют *непараметрические методы анализа.* Этот метод не требует, чтобы исходные данные были представлены абсолютными величинами; здесь допустимо использование относительных величин.

Метод основан на *сравнении сумм рангов в градациях дисперсионного комплекса.* Исходные данные ранжируют, т. е. располагают (в пределах градаций) в ряд по возрастающим значениям признака. Затем каждому значению признака присваивают порядковый номер, его ранг. (Понятие ранжирования уже было описано в гл. V, например применение ранговых критериев, в частности U-критерия Уилкоксона.)

При наличии равномерных дисперсионных комплексов ранговый анализ проводят с помощью критерия Фридмана по формуле

$$\chi^2_R=\frac{12\sum(\sum R_i)^2}{na(n+1)}-3n(a+1), \qquad (118)$$

где ΣR_i — сумма рангов в каждой градации; n — численность вариант в каждой градации; a — число градаций. Полученное значение χ^2_R сравнивают с критическим значением этого критерия по табл. XIX Приложений в случаях, когда $a=3$ и $2\leqslant n\leqslant 9$ или $a=4$ и $2\leqslant n\leqslant 4$. При больших значениях a и n $\chi_R^2\to\chi^2$ и можно сравнивать $\chi^2_{R\Phi}$ с критическими значениями по табл. VII Приложений для принятого уровня значимости α и числа степеней свободы $k=a-1$. H_0 — гипотеза, или предположение о том, что суммы рангов в градациях равны, а их различия случайны, отвергается, если $\chi^2_R\geqslant\chi^2_{st}$.

Пример 6. В трех разновозрастных группах детей со здоровыми зубами определяли гигиенический индекс (ГИ), выражаемый в условных единицах. Полученные результаты и их обработка приведены в табл. 67.

Как видно из данных табл. 67, значения ГИ ранжируются по пробам (по строкам), а ранги суммируются по градациям (по столбцам). Например, ранжируем величину ГИ в первой строке: наименьшей величине ГИ=1 (3 года) соответствует ранг, равный единице, а одинаковым значениям ГИ в 4 и 5 лет соответствуют и одинаковые ранги $(2+3)/2=2{,}5$. Сумма рангов комплекса должна быть равна $na(a+1)/2$, т. е. в данном прн-

мере $6 \cdot 3 \cdot 4/2 = 36$. Сумма рангов комплекса $\Sigma(\Sigma R_l) = 8{,}5 + 12 + 15{,}5 = 36$.

Следовательно, расчеты произведены правильно. Теперь по формуле (118) вычислим $\chi^2 = \frac{12 \cdot 8{,}5^2 + 12^2 + 15{,}5^2}{6 \cdot 3 \cdot 4} - 3 \cdot 6 \cdot 4 =$ $= 76{,}08 - 72 = 4{,}08$.

По табл. XIX Приложений для $\alpha = 0{,}05$, $n = 6$ и $a = 3$ находим $\chi^2_{st} = 7{,}00$. Так как эта величина превосходит фактически полученное значение критерия Фридмана (4,08), то H_0-гипотеза сохраняется. Таким образом, разница в гигиеническом индексе у детей разного возраста оказалась в приведенном примере недостоверной.

Таблица 67

Пробы	Возраст детей, лет					
	3		4		5	
	A_1	R_1	A_2	R_2	A_3	R_3
1	1,0	1	3,0	2,5	3,0	2,5
2	1,0	1	1,2	2	1,3	3
3	1,0	1	2,0	2	2,2	3
4	1,0	1,5	1,0	1,5	1,2	3
5	2,6	2	2,7	3	1,3	1
6	2,7	2	1,3	1	3,0	3
Суммы	$\Sigma R_1 = 8{,}5$		$\Sigma R_2 = 12$		$\Sigma R_3 = 15{,}5$	

Неравночисленные комплексы. Если в градациях дисперсионного комплекса число вариант не одинаково (неортогональный комплекс), то при обработке собранных данных следует использовать непараметрический критерий Краскелла — Уоллиса

$$H = \frac{12}{N(N+1)} \sum \frac{(\Sigma R_l)^2}{n_l} - 3(N+1), \qquad (119)$$

где N — общее число наблюдений, объем комплекса; n_l — численность вариант в отдельных градациях комплекса; R_l — ранги вариант, ранжированных в *общем* порядке, т. е. не по отдельным градациям, как в предыдущем примере, а путем расположения всех вариант комплекса в один общий ряд (см. ниже).

Условием для отвергания нулевой гипотезы на принятом уровне значимости α будет $H_\Phi \geqslant H_{st}$. Последняя величина находится в табл. XX Приложений. При $a > 3$ или $n \geqslant 5$ $H \to \chi^2$, поэтому величину H_Φ можно сравнивать с табличным значением χ^2_{st} (см. табл. VII Приложений). H_0-гипотезу отвергают, если

$H_\phi \geqslant \chi^2_{st}$ для принятого уровня значимости (α) и числа степеней свободы $(k)=a-1$.

Пример 7. Изучали глазо-сердечный рефлекс (рефлекс Данини — Ашнера) у детей дошкольного и школьного возраста. Нужно было выяснить, существуют ли различия между этими группами детей по длительности латентного периода исследуемого рефлекса. Изучение проводили методом прикосновения (пальцем) к глазному яблоку с последующей регистрацией времени наступления сердечного рефлекса. Результаты исследования приведены в табл. 68.

Таблица 68

Возраст, лет	Длительность латентного периода, с							
	<0,5	R_1	0,5	R_2	2	R_3	4	R_4
1—2	12	16	7	10	2	3,5	1	1,5
3—4	9	13	5	8	4	6,5	1	1,5
5—6	11	15	8	11,5	3	5	2	3,5
7—8			8	11,5	6	9	4	6,5
9—10			10	14				
Сумма	3	44	5	55	4	24	4	13

В этой таблице 16 вариант, которые варьируют в пределах от 1 до 12 с. Расположим все варианты по возрастающим значениям в один общий ряд и определим их ранги. Если бы отдельные варианты не повторялись, их рангами были бы соответствующие порядковые числа от 1 до 12. При наличии в выборке повторяющихся вариант им присваивают одинаковые ранги, при этом ряд последовательных значений признака удлиняется. В данном примере это выглядит следующим образом:

Номера вариант	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$x_i=$	1	1	2	2	3	4	4	5	6	7	8	8	9	10	11	12
$R_i=$	1,5	1,5	3,5	3,5	5	6,5	6,5	8	9	10	11,5	11,5	13	14	15	16

Проделав эту операцию, помещаем ранги на присущие им места в градациях комплекса. Так, в первом столбце табл. 68 три варианты: 12, 9 и 11. Им соответствуют ранги 16, 13, 15, сумма этих рангов равна 44. Определив таким образом суммы рангов для каждой градации комплекса в отдельности, подставляем известные величины в формулу (119):

$$H=\frac{12}{16\cdot 17}\left(\frac{44^2}{3}+\frac{55^2}{5}+\frac{24^2}{4}+\frac{14^2}{4}\right)-3\cdot 17=$$

$$=\frac{12}{272}(645{,}33+605{,}00+144{,}00+49{,}00)-51=\frac{3}{68}\,1443{,}33-51=$$

$$=12{,}68.$$

Известно, что при $a>3$ или $n\geqslant 5$ $H\rightarrow\chi^2$ [1]. Поэтому величину H_Φ можно сравнивать с табличным значением χ^2_{st} (см. табл. VII Приложений). H_0-гипотезу отвергают, если $H_\Phi\geqslant\chi^2_{st}$. В данном примере $a>3$ и $n=5$. В табл. VII Приложений для $k=a-1=4-1=3$ и 1%-ного уровня значимости находим $\chi^2_{st}=11{,}34$. Эта величина меньше фактически полученной $H_\Phi=12{,}68$. Следовательно, с вероятностью $P>99\%$ можно утверждать, что между детьми дошкольного и школьного возраста существуют различия в скорости протекания глазо-сердечного рефлекса.

Таблица 69

Пробы	Возраст, лет					
	3		4		5	
	A_1	R_1	A_2	R_2	A_3	R_3
1	2	3	3	9,5	2	3
2	3	9,5	3	9,5	2	3
3	3	9,5	3	9,5	2	3
4	3	9,5	3	9,5	2	3
5	3	9,5	4	14	—	—
Суммы	$\Sigma R_1=41$		$\Sigma R_2=52$		$\Sigma R_3=12$	

Когда дисперсионные комплексы имеют не больше трех градаций, т. е. при $a=3$ с числом вариант в градациях $n<5$, оценку достоверности фактически полученной величины H_Φ производят по специальным таблицам критических значений критерия I (см. табл. XX Приложений).

Пример 8. В трех разновозрастных группах детей со здоровыми зубами был проведен тест резистентности эмали (ТЭР-тест). Величины ТЭР-индекса выражаются в баллах и могут колебаться от 1 до 4. Результаты исследования сведены в табл. 69. В этой таблице содержится 14 вариант, которые варьируют в пределах от 2 до 4.

Как и в предыдущем примере, определяем ранги вариант следующим образом:

x_j	2	2	2	2	2	3	3	3	3	3	3	3	3	4
R_j	3	3	3	3	3	9,5	9,5	9,5	9,5	9,5	9,5	9,5	9,5	14

Затем находим суммы рангов членов ряда для каждой группы в отдельности. Подставляем известные величины в форму-

[1] См.: *Терентьев П. В., Ростова Н. С.* Практикум по биометрии. Л., 1977. С. 116—117, 150.

лу (119):

$$H=\frac{12}{14\cdot 15}\left(\frac{41^2}{5}+\frac{52^2}{5}+\frac{12^2}{4}\right)-3\cdot 15=\frac{12\cdot 913}{14\cdot 15}-45=52{,}17-45=7{,}17.$$

Так как в данном примере $a=3$ и $n\leqslant 5$, то необходимо воспользоваться табл. XX Приложений. Находим для $n_1=5$, $n_2=5$, $n_3=4$ критические значения: при $\alpha=0{,}05$ $H_{st}=5{,}66$ и при $\alpha=0{,}01$ $H_{st}=7{,}79$. Сравнивая эти величины с $H_\Phi=7{,}17$, заключаем, что H_0-гипотеза отвергается на 5%-ном уровне значимости ($0{,}01<P<0{,}05$). Таким образом, установлено, что кислотная резистентность здоровых зубов в обследованных группах детей с возрастом достоверно изменяется.

Оценка силы влияния факторов. Метод Плохинского. После того как достоверно установлено действие регулируемого фактора, можно измерить *силу его влияния на результативный признак*. Последнюю определяют как долю межгрупповой вариации в общем варьировании результативного признака. Для измерения силы влияния предложено несколько методов. Наибольшее признание получили методы Плохинского (1966, 1970) и Снедекора (1961). Метод Плохинского базируется на равенстве девиат $D_y=D_x+D_e$, которое осуществляется в любом дисперсионном комплексе. *Показатель силы влияния,* обозначаемый символом h_x^2, строят следующим образом: все члены указанного равенства делят на D_y, что дает

$$\frac{D_x}{D_y}+\frac{D_e}{D_y}=1.$$

Отсюда вычисляют показатель силы влияния:

$$h_x^2=\frac{D_x}{D_y} \text{ или } h_x^2=1-\frac{D_e}{D_y}. \qquad (120)$$

Критерием достоверности этого показателя служит его отношение к своей ошибке, которую определяют по следующей приближенной формуле:

$$s_{h_x^2}=(1-h_x^2)\frac{a-1}{N-a}, \qquad (121)$$

где a — число градаций регулируемого фактора; N — объем дисперсионного комплекса. Нулевую гипотезу отвергают, если $F=h_x^2/s_{h_x^2}\geqslant F_{st}$ для принятого уровня значимости α и чисел степеней свободы $k_1=a-1$ (находится в верхней строке табл. V Приложений) и $k_2=N-a$ (находится в первом столбце той же таблицы).

Пример 9. Определить силу влияния сорта на урожайность пшеницы. Исходные данные: $D_y=219{,}60$; $D_A=140{,}03$; $s_A^2=28{,}0$; $s_e^2=4{,}4$; $n=4$ (см. табл. 61 и 62). Отсюда сила влияния сорта

(фактор А) на урожайность пшеницы (результативный признак X) выражается следующей величиной:

$$h_A^2 = \frac{D_A}{D_y} = \frac{140{,}03}{219{,}50} = 0{,}638, \text{ или } 63{,}9\%.$$

Это означает, что примерно около 64% от общего варьирования признака X обусловлено природой сорта, его генетическим разнообразием и около 36% приходится на долю воздействующих на признак других (модифицирующих) факторов.

Метод Снедекора. В отличие от метода Плохинского этот метод основан на применении *не девиат, а дисперсий,* причем показатель силы влияния строят с учетом действия на признак не только регулируемых, но и нерегулируемых в опыте факторов, оценкой которых служит внутригрупповая дисперсия, т. е.

$$h_x^2 = \frac{\hat{s}_x^2}{\hat{s}_x^2 + s_e^2}, \qquad (122)$$

где $\hat{s}_x{}^2 = (s_x{}^2 - s_e{}^2)/n$ — «исправленная» межгрупповая дисперсия, равная разности между дисперсиями межгрупповой (неисправленной) и внутригрупповой, или остаточной, отнесенной к числу членов n в градациях комплекса. Если комплекс неравномерный, величину n определяют по формуле

$$\bar{n} = \frac{1}{a-1}\left(N - \frac{\Sigma (n_l)^2}{N}\right), \qquad (123)$$

где a — число градаций регулируемого фактора; n — численность вариант x_l в отдельных градациях фактора; $N = \Sigma n$ — общее число вариант, или объем комплекса.

Формулу (122) можно преобразовать таким образом, чтобы при вычислении показателя силы влияния $h_x{}^2$ не прибегать к «исправлению» межгрупповой дисперсии:

$$h_x^2 = \frac{(s_x^2 - s_e^2)/n}{(s_x^2 - s_e^2)/n + s_e^2} = \frac{s_x^2 - s_e^2}{s_x^2 - s_e^2 + ns_e^2}. \qquad (124)$$

Очень удобной при определении величины $h_x{}^2$ является формула

$$h_x^2 = \frac{s_x^2 - s_e^2}{s_x^2 + (n-1)\, s_e^2}. \qquad (125)$$

Достоверность оценок силы влияния, определяемых по методу Снедекора, устанавливают обычным в дисперсионном анализе способом — посредством F-критерия Фишера ($F = s_x{}^2/s_e{}^2$). Величину F-критерия сравнивают с критическим значением этого показателя для принятого уровня значимости α и чисел сте-

пеней свободы k_x и k_e, причем последние определяют так же, как и при оценке критерия Плохинского.

Пример 9а. Определить силу влияния сорта на урожайность пшеницы по методу Снедекора. Необходимые исходные данные приведены в примере 9:

$$h_A^2 = \frac{28{,}0 - 4{,}4}{28{,}0 + (4-1)\,4{,}4} = \frac{23{,}6}{41{,}2} = 0{,}573, \text{ или } 57{,}32\%.$$

Нетрудно заметить, что показатель силы влияния, определенный по методу Плохинского ($h_A^2 = 0{,}638$), оказался больше, чем тот же показатель, вычисленный по методу Снедекора ($h_A^2 = 0{,}573$). Это не случайное явление: как правило, определение показателя силы влияния по Плохинскому и по Снедекору не приводит к идентичному результату. Поэтому при определении силы влияния следует указывать, каким методом вычислен этот показатель [1].

Достоверность оценок силы влияния $h_A^2 = 0{,}638$ (по Плохинскому) и $h_A^2 = 0{,}573$ (по Снедекору) устанавливают следующим образом: по Плохинскому, $h_A^2/s_{h^2_A} = 0{,}638/0{,}101 = 6{,}32$, где $s_{h^2_A}$ — ошибка показателя силы влияния, определяемая по формуле (121). По Снедекору, $F_\Phi = s_A^2/s_e^2 = 28{,}0/4{,}4 = 6{,}36$. В обоих случаях $F_{st} = 4{,}25$ для $k_1 = a - 1 = 6 - 1 = 5$; $k_2 = N - a = 24 - 6 = 18$ и $\alpha = 0{,}1\%$. Нулевую гипотезу отвергают на высоком уровне значимости ($P < 0{,}001$). Это означает, что в данном случае оценки хорошо репрезентируют генеральные параметры и вполне достоверны.

Пример 10. В табл. 63 приведены данные о влиянии различных доз минеральных удобрений на урожай озимой ржи. В данном случае представлен неравномерный дисперсионный комплекс с его характеристиками: $D_y = 33{,}05$; $D_A = 27{,}31$; $s_A^2 = 9{,}10$; $s_e^2 = 0{,}52$; $n = 4$; $N = 15$. Из-за рассматриваемой неравномерности комплекса необходимо рассчитать усредненное значение $\bar{n}$ по формуле (123):

$$\bar{n} = \frac{1}{4-1}\left(15 - \left(\frac{4^2 + 6^2 + 3^2 + 2^2}{15}\right)\right) = \frac{1}{3}\left(15 - \frac{65}{15}\right) = \frac{10{,}67}{3} = 3{,}56.$$

Отсюда, по Снедекору, $h_A^2 = \frac{9{,}10 - 0{,}52}{9{,}10 + 2{,}56^{0{,}52}} = \frac{8{,}58}{10{,}43} = 0{,}82$;

по Плохинскому, $h_A^2 = \frac{h_A^2}{s_{h^2_A}} = \frac{27{,}31}{33{,}05} = 0{,}83$.

Переходим к определению критерия достоверности оценок (по Фишеру):

[1] Известно, что показатель силы влияния фактора, найденный по методу Плохинского, оказывается весьма смещенной оценкой. Поэтому лучше пользоваться аналогичным показателем Снедекора, (*Прим. ред.*)

по Снедекору, $F_\Phi = s_h^2/s_e^2 = 9{,}10/0{,}52 = 17{,}5$;

по Плохинскому, $F_\Phi = h_A^2/s_{h_A^2}^2 = 0{,}83/0{,}05 = 16{,}6$.

Оба показателя значительно превосходят критическую точку $F_{st} = 6{,}22$ для $k_1 = 4-1 = 3$ (находим по горизонтали табл. VI Приложений), $k_2 = 15-4 = 11$ (находим в первом столбце упомянутой таблицы) и уровня значимости $\alpha = 0{,}1\%$. Так как $F_\Phi > F_{st}$, нулевая гипотеза опровергается на высоком уровне значимости ($P < 0{,}001$). Таким образом, доказано, что различные дозы минеральных удобрений по-разному влияют на урожай озимой ржи.

Сравнение групповых средних дисперсионного комплекса. После того как достоверно установлено влияние регулируемого фактора или факторов на результативный признак, при необходимости прибегают к сравнению групповых средних друг с другом или с какой-либо другой величиной, например с контролем, стандартом, установленной нормой и т. п.

Разность между средними величинами, как описано выше, оценивают по t-критерию Стьюдента, т. е. по отношению указанной разности к ее ошибке. Этот способ, однако, неприменим к сравнительной оценке средних в дисперсионном комплексе, так как наряду с межгрупповой дисперсией на величине ошибки разности s_d между групповыми средними комплекса сказывается и влияние внутригрупповой дисперсии s_e^2, величина которой зависит и от численности вариант x_i в группах, и от количества групп a, входящих в данный комплекс. Эти обстоятельства ограничивают применение критериев Стьюдента и Фишера. Поэтому в качестве ошибки разности между групповыми средними дисперсионного комплекса принят корень квадратный из отношения внутригрупповой, или остаточной, дисперсии к числу вариант, входящих в состав градаций фактора A, т. е.

$$s_d = \sqrt{\frac{s_e^2}{n}}. \qquad (126)$$

Для оценки разности между групповыми средними дисперсионного комплекса применяют специальные методы, созданные на базе критериев Стьюдента и Фишера. Из них наиболее подходящими считают методы множественных сравнений, разработанные Дж. Тьюки (1949) и Г. Шеффе (1953).

Метод Тьюки. Этот метод применяют для проверки нулевой гипотезы при сравнении групповых средних $\bar{x}_1$ и $\bar{x}_2$ равновеликих групп, т. е. при $n_1 = n_2 = n$. Критерием оценки служит отношение разности сравниваемых средних к своей ошибке:

$$t_Q = \frac{|\bar{x}_1 - \bar{x}|}{\sqrt{s_e^2/n}}.$$

Величину t_Q сравнивают с критической точкой Q_{st} для k_e и 5%-ного уровня значимости с учетом числа групп или градаций a регулируемого фактора A. Критические значения Q_{st} содержатся в табл. XXIV Приложений. Нулевую гипотезу отвергают, если $t_Q \geqslant Q_{st}$ или $|\bar{x}_1 - \bar{x}_2| \geqslant s_{\bar{x}} Q_{st}$.

Пример 11. В табл. 60 приведены групповые средние, характеризующие урожайность шести сортов пшеницы в местных условиях возделывания этой культуры. Из этих данных видно, что наиболее урожайным оказался пятый сорт ($\bar{x}_5 = 32{,}2$ ц/га), а наименее урожайным — шестой сорт ($\bar{x}_6 = 25{,}0$ ц/га). Если в качестве стандарта условно принять урожайность шестого сорта, то разница по урожайности между пятым и шестым сортами пшеницы составит $32{,}2 - 25{,}0 = 7{,}2$ ц/га.

Проверим достоверность этой разности. Здесь средние вычислены на равных по объему группах вариант ($n = 4$); число испытываемых сортов равно шести, т. е. $a = 6$; объем комплекса $N = 24$; внутригрупповая дисперсия $s_e^2 = 4{,}4$. Отсюда

$$t_Q = \frac{32{,}2 - 25{,}0}{\sqrt{\frac{4{,}4}{4}}} = \frac{7{,}2}{1{,}05} = 6{,}86.$$

В табл. XXIV Приложений для $k_e = N - a = 24 - 6 = 18$ и $a = 6$ находим $Q_{st} = 4{,}5$. Так как $t_Q > Q_{st}$, нулевую гипотезу отвергают на 5%-ном уровне значимости. Разницу между сравниваемыми средними дисперсионного комплекса следует признать статистически достоверной.

Метод Шеффе. В отличие от метода Тьюки этот метод множественных сравнений одинаково применим и к равно-, и к неравновеликим по составу группам. Критерием достоверности различий, наблюдаемых между групповыми средними дисперсионного комплекса, служат следующие отношения:

$$F = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_e^2}} \sqrt{\frac{n}{2}} \quad (\text{при } n_1 = n_2);$$

$$F = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_e^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad (\text{при } n_1 \neq n_2).$$

Нулевую гипотезу отвергают, если $F \geqslant \sqrt{(a-1)F_{st}}$, где a — число градаций фактора A; F_{st} определяют с помощью таблицы Фишера (табл. VI Приложений) для принятого уровня значимости α и чисел степеней свободы $k_1 = a - 1$ и $k_2 = N - a$, где N — объем дисперсионного комплекса.

Пример 12. В табл. 63 приведены данные о влиянии различных доз минеральных удобрений на урожай озимой ржи. В той же таблице содержатся групповые средние, характеризующие

зависимость урожая ржи от внесения в почву различных доз удобрений. Из результатов видно, что более высокий урожай ржи (12,0 ц/га) получен при внесении удобрений в количестве 25 кг/га. Сравним эту величину с урожаем данной культуры (9,4 ц/га), полученным при внесении в почву 20 кг/га удобрений.

В данном случае здесь представлен неравномериый комплекс: средняя $\bar{x}_2=9{,}4$ рассчитана по шести, а средняя $\bar{x}_3=12{,}0$ — по трем измерениям. Число групп, входящих в комплекс, равно четырем, т. е. $a=4$; объем комплекса $N=15$; внутригрупповая дисперсия оказалась равной $s_e^2=0{,}52$. Отсюда

$$F_\Phi=\frac{d}{\sqrt{s_e^2}}\sqrt{\frac{n_1n_2}{n_1+n_2}}=\frac{12{,}0-9{,}4}{0{,}52}\sqrt{\frac{3\cdot6}{3+6}}=$$

$$=\frac{2{,}60\cdot1{,}414}{0{,}72}=\frac{3{,}68}{0{,}72}=5{,}11.$$

Для $k_1=a-1=4-1=3$; $k_2=N-a=15-4=11$ и 1%-ного уровня значимости в табл. VI Приложений находим $F_{st}=6{,}2$. Отсюда $F_{st}=\sqrt{(4-1)\,6{,}2}=\sqrt{18{,}6}=4{,}31$. Таким образом, $F_\Phi>F_{st}$, что позволяет отвергиуть нулевую гипотезу на высоком уровне значимости ($P<0{,}01$).

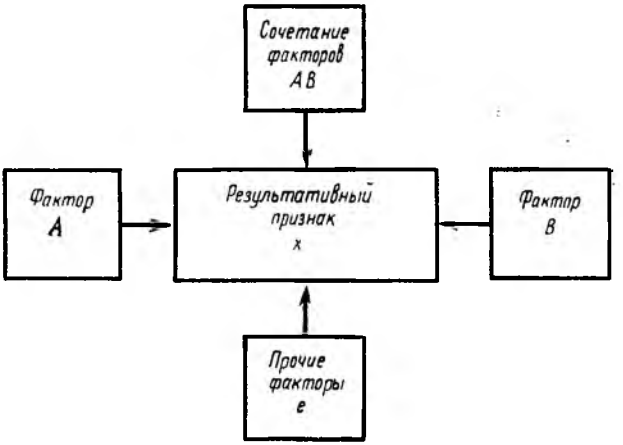
Рассмотренный метод Шеффе выгодно отличается от метода Тьюки, поскольку последний неприменим к оценке групповых средних дисперсионного комплекса, вычисленных на разных объемах групп. Однако при сравнении средних $\bar{x}_1$ и $\bar{x}_2$ равновеликих групп предпочтение следует отдавать методу Тьюки.

VII.2. АНАЛИЗ ДВУХФАКТОРНЫХ КОМПЛЕКСОВ

Ортогональные комплексы. Приступая к рассмотрению двухфакторных равномерных и пропорциональных комплексов, следует заметить, что при их образоваиии, как и вообще при образовании многофакторных комплексов, иеобходимо, чтобы регулируемые факторы были независимы друг от друга. Выполнение этого требования — *независимости факторов* — одно из важнейших условий правильного применения дисперсионного анализа. Нельзя подвергать дисперсионному аналнзу корреляционно связанные признаки, такие, например, как масса тела и его линейные размеры и т. п.

Общие схемы дисперсионного анализа двухфакторных ортогональных комплексов в принципе не отличаются от описаниых выше схем однофакторного дисперсионного анализа. Анализ двухфакторных комплексов не меняет, а лишь несколько усложняет общие схемы, поскольку наряду с действием каждого фактора в отдельности приходится учитывать и их совместное

действие на результативный признак. Так, изучая два регулируемых фактора A и B, можно влияние на результативный признак из прочих факторов изобразить в виде следующей схемы:



Из этой схемы следует, что общая сумма квадратов отклонений D_y содержит четыре компонента варьирования: $D_y=D_A+D_B+D_{AB}+D_e$, а общая факториальная сумма квадратов отклонений D_x состоит из трех компонентов: $D_x=D_A+D_B+D_{AB}$.

Если учитывать не два, а три регулируемых фактора A, B и C, то наряду с их индивидуальным действием возможно влияние на признак трех попарных сочетаний (AB, AC, BC), их совместное действие (ABC), а также влияние неорганизованных (случайных) факторов. Таким образом, общий компонент варьирования будет содержать восемь элементов:

$$D_y=D_A+D_B+D_C+D_{AB}+D_{AC}+D_{BC}+D_{ABC}+D_e.$$

При большем числе учитываемых факторов число их возможных сочетаний будет еще больше. В изучении влияния на результативный признак всех учитываемых факторов и их возможных комбинаций и заключается *основная задача дисперсионного анализа.* При этом не всегда необходимо учитывать все возможные сочетания организованных факторов. Этот вопрос исследователь решает в зависимости от цели исследования и принятой полноты дисперсионного анализа.

Дисперсионный анализ двухфакторных ортогональных комплексов проводят по следующей примерной схеме.

1. Рассчитывают девиаты (как и при обработке однофакторных комплексов): общую для всего комплекса D_y, межгрупповую D_x и остаточную D_e. Для этого служат формулы (109), (110) и (111), причем $D_x=\frac{\Sigma(\Sigma x_i)^2}{n}-H$.

2. Затем определяют факториальные девиаты:

$$D_A = \frac{\sum^{a} (\Sigma x_A)^2}{n_A} - H; \tag{127}$$

$$D_B = \frac{\sum^{b} (\Sigma x_B)^2}{n_B} - H. \tag{128}$$

3. Анализ двухфакторных пропорциональных комплексов тоже начинается с определения девиат (общей, межгрупповой и остаточной) по указанным выше формулам (109), (110) и (111), причем при определении D_x формула (110) выглядит так:

$$D_x = \sum \frac{(\Sigma x_i)^2}{n} - H.$$

Факториальные девиаты определяют по следующим формулам:

$$D_A = \sum_{i=1}^{a} \frac{(\Sigma x_A)^2}{n_A} - H; \tag{129}$$

$$D_B = \sum_{i=1}^{b} \frac{(\Sigma x_B)^2}{n_B} - H. \tag{130}$$

Девиату совместного действия факторов в обоих случаях определяют по формуле

$$D_{AB} = D_x - (D_A + D_B). \tag{131}$$

Как и в предыдущих случаях, в этих формулах повторяется величина $H = (\Sigma x_i)^2/N$, где x_i — варианты, входящие в состав дисперсионного комплекса; $N = \Sigma n = nab$ — общая численность вариант, или объем комплекса; a — число градаций фактора A; b — число градаций фактора B; n — количество вариант в отдельных градациях комплекса; $n_A = nb$ — общая численность вариант в каждой градации фактора A; $n_B = na$ — общая численность вариант в каждой градации фактора B; Σx_A — сумма вариант в градациях фактора A; Σx_B — сумма вариант в градациях фактора B.

4. Определив значения девиат, переходят к установлению чисел степеней свободы, которые равны: $k_y = N - 1$ для общей дисперсии; $k_x = ab - 1$ для межгрупповой дисперсии, характеризующей влияние обоих факторов A и B на результативный признак X; $k_e = N - ab$ для внутригрупповой, или остаточной, дисперсии; $k_A = a - 1$ для факториальной дисперсии A; $k_B = b - 1$ для факториальной дисперсии B; $k_{AB} = (a-1)(b-1) = k_A k_B$ для дисперсии совместного действия факторов A и B.

При этом, как и в предыдущих случаях, числа степеней свободы должны находиться в таких же количественных соотношениях, как и соответствующие девиаты, т. е. $D_y = D_A + D_B + D_{AB} + D_e$ и соответственно $k_y = k_A + k_B + k_{AB} + k_e$. Равенству $D_x = D_A + D_B + D_{AB}$ должно соответствовать равенство $k_x = k_A + k_B + k_{AB}$, а равенству $D_y = D_x + D_e$ — равенство $k_y = k_x + k_e$. Эти равенства могут служить для проверки правильности расчета девиат и чисел степеней свободы.

Таблица 7(

Вариация	Степени свободы	Девиаты	Дисперсии	Дисперсионные отношения
По фактору A	$a-1$	D_A	$s_A^2 = \frac{D_A}{a-1}$	s_A^2/s_e^2
По фактору B	$b-1$	D_B	$s_B^2 = \frac{D_B}{b-1}$	s_B^2/s_e^2
Совместно AB	$(a-1)(b-1)$	D_{AB}	$s_{AB}^2 = \frac{D_{AB}}{(a-1)(b-1)}$	s_{AB}/s_e^2
Остаточная	$N-ab$	D_e	$s_e^2 \frac{D}{N-ab}$	—
Общая	$N-1$	D_y	—	—

5. Отнесением девиат к соответствующим числам степеней свободы определяют значения дисперсий, а по их отношениям к величине остаточной дисперсии устанавливают дисперсионные отношения F, которые сравнивают с критическими точками F_{st} для соответствующих чисел степеней свободы и принимаемого уровня значимости α. H_0-гипотезу отвергают, если $F_\Phi \geqslant F_{st}$.

Заключительным этапом дисперсионного анализа является сведение результатов в таблицу, которая содержит следующие показатели (табл. 70).

Эта таблица одновременно служит и *схемой двухфакторного дисперсионного анализа.* Обычно к ней добавляют еще два столбца, в которых приводят критические точки дисперсионного отношения F_{st} для 5%-ного и 1%-ного уровней значимости F

ответствующих чисел степеней свободы, что облегчает выво-ы относительно проверки нулевой гипотезы.

Пример 13. Испытывали влияние трех видов микроэлементов на жирномолочность коров. Эксперимент проводили на четырех

Таблица 71

Группы коров B	Процент жира в молоке при трехкратных пробах (градации фактора A)									Групповые средние $\bar{x}_B$
	A_1			A_2			A_3			
B_1	2,1	2,0	3,4	2,8	2,6	3,0	2,4	2,1	2,8	2,6
B_2	4,0	3,2	4,1	3,9	4,1	4,5	3,0	3,9	4,2	3,9
B_3	3,0	2,8	2,7	3,5	4,0	2,8	4,8	3,1	2,9	3,3
B_4	3,4	3,0	2,9	3,0	2,9	3,0	3,3	2,8	3,0	3,0
$\bar{x}_A$	3,0			3,3			3,2			—

Таблица 72

Градации		Варианты x_i			Σx_i	$(\Sigma x_i)^2$	Σx_i^2
A	B	1	2	3			
A_1	B_1	21	20	34	75	5 625	1 997
	B_2	40	32	41	113	12 769	4 305
	B_3	30	28	27	85	7 225	2 413
	B_4	34	30	29	93	8 649	2 897
A_2	B_1	28	26	30	84	7 056	2 360
	B_2	39	41	45	125	15 625	5 227
	B_3	35	40	28	103	10 609	3 609
	B_4	30	29	30	89	7 921	2 641
A_3	B_1	24	21	28	73	5 329	1 801
	B_2	30	39	42	111	12 321	4 185
	B_3	48	31	29	108	11 664	4 106
	B_4	33	28	30	91	8 281	2 773
Сумма		—	—	—	1150	113 074	38 314

группах одновозрастных животных различных пород. Каждое испытание имело трехкратную повторность. Полученные результаты приведены в табл. 71.

Здесь через A обозначен фактор воздействия, т. е. микроэлементы, а через B — группы коров различных пород. Число гра-

даций фактора A равно трем, т. е. $a=3$, а число групп фактора B равно четырем, т. е. $b=4$. Видно, что групповые средние $\bar{x}_A$ варьируют по своей величине как для группы коров, так и по градациям фактора A. Необходимо выяснить, случайны или достоверны различия, наблюдаемые между групповыми средними.

Чтобы облегчить обработку этих данных, прежде всего избавимся от дробей, увеличив каждую варианту комплекса в $K=10$ раз. Затем сгруппируем выборку так, чтобы градации фактора A и подразделения или группы фактора B располагались по строкам комбинационной таблицы (можно распределять их и по столбцам таблицы), что облегчит расчет вспомогательных величин, необходимых для определения девиат D_y, D_x и D_e, а затем и расчет других (факториальных) девиат. Такая группировка данных и расчет Σx_i, $(\Sigma x_i)^2$ и Σx_i^2 приведены в табл. 72.

Для определения общих девиат D_y, D_x и D_e в двухфакторном равномерном комплексе применим формулы (109), (110) и (111), уменьшая результаты расчетов в $K^2=100$ раз (так как каждая варианта комплекса была увеличена в $K=10$ раз):

$$D_y=\sum x_i^2-H=\frac{1}{100}\left(38\cdot 314-\frac{1150^2}{36}\right)=\frac{1}{100}(38\cdot 314-36\cdot 736{,}1)=$$

$$=\frac{1577{,}9}{100}=15{,}78;\quad D_x=\frac{\sum(\sum x_i)^2}{100}-H=\frac{1}{100}\left(\frac{113\cdot 0{,}74}{3}-H\right)=$$

$$=\frac{1}{100}(37\cdot 691{,}3-36\cdot 736{,}1)=\frac{955{,}2}{100}=9{,}55;\quad D_e=$$

$$=15{,}78-9{,}55=6{,}23.$$

Определяем числа степеней свободы: $k_y=36-1=35$; $k_x=ab-1=3\cdot 4-1=11$; $k_e=N-ab=36-12=24$. Находим дисперсии: $s_x^2=9{,}55/11=0{,}87$ и $s_e^2=6{,}23/24=0{,}26$. Отсюда $F_\Phi=0{,}87/0{,}26=3{,}35$. В табл. VI Приложений для 1%-ного уровня значимости и чисел степеней свободы $k_x=11$ и $k_e=24$ находим $F_{st}=3{,}1$. Так как $F_\Phi>F_{st}$, нулевая гипотеза отвергается на высоком уровне значимости ($P<0{,}01$).

Переходим к расчету факториальных девиат. Предварительно вычислим $\Sigma(\Sigma x_A)^2$ и $\Sigma(\Sigma x_B)$ (табл. 73). Напомним, что $n=3$, $a=3$ и $b=4$. Отсюда $n_A=nb=3\cdot 4=12$ и $n_B=na=3\cdot 3=9$. Тогда

$$D_A=\frac{1}{100}\left(\frac{441\,446}{12}-H\right)=\frac{1}{100}(36787{,}3-36736{,}1)=\frac{51{,}1}{100}=0{,}51;$$

$$D_B=\frac{1}{100}\left(\frac{337\,770}{9}-H\right)=\frac{1}{100}(37530{,}0-36736{,}1)=\frac{739{,}9}{100}=7{,}94;$$

$D_{AB}=D_x(D_A+D_B)=9{,}55-0{,}51+7{,}94=1{,}10$. Проверим правильность расчета девиат: $D_y=D_A+D_B+D_{AB}+D_e=0{,}51+7{,}94+1{,}10+6{,}23=15{,}78$. Расчет произведен правильно.

Определяем числа степеней свободы: $k_y=N-1=36-1=35$; $k_A=a-1=3-1=2$; $k_B=b-1=4-1=3$; $k_{AB}=k_Ak_B=2\cdot 3=6$ и

$k_e = N - ab = 36 - 12 = 24$. Проверяем правильность расчета чисел степеней свободы: $k_y = k_A + k_B + k_{AB} + k_e = 2 + 3 + 6 + 24 = 35$. Расчет произведен правильно.

Таблица 73

Расчет Σx_A	$(\Sigma x_A)^2$	Расчет Σx_B	$(\Sigma x_B)^2$
75+113+85+93=366	133956	75+84+73=232	53824
84+125+103+89=401	160801	113+125+111=349	121801
73+111+108+91=383	146689	85+103+108=296	87616
		93+89+91=273	74529
Сумма —	441446	Сумма —	337770

Относим девиаты к соответствующим числам степеней свободы и находим значения дисперсий. Затем определяем дисперсионные отношения факториальных дисперсий к дисперсии остаточной F_Φ, которые сравниваем с критическими точками F_{st}. Результаты дисперсионного анализа сводим в заключительную таблицу (табл. 74).

Таблица 74

Вариация	Степени свободы	Девиаты	Дисперсии	F_Φ	F_{st}	
					5%	1%
По фактору A	2	0,51	0,26	1,0	3,4	5,6
По фактору B	3	7,94	2,65	10,2	3,0	4,7
Совместно AB	6	1,10	0,18	1,4	3,8	7,3
Остаточная	24	6,23	0,26	—	—	—
Общая	35	15,78	—	—	—	—

Из табл. 74 видно, что нулевая гипотеза опровергается только в отношении фактора B, действие которого на признак оказалось в высшей степени достоверным ($P<0{,}01$). Это означает, что жирномолочность коров связана со свойствами их породы, т. е. контролируется наследственностью и не зависит от влияния на этот признак испытываемых препаратов микроэлементов. Видимо, поэтому и взаимодействие факторов AB существенно не сказалось на величине результативного признака.

Пример 14. В одном из опытных хозяйств испытывали урожайность разных сортов крыжовника и их устойчивость против вредоносного действия крыжовникового пилильщика. Полученные результаты приведены в табл. 75. Из этой таблицы видно,

что испытываемые сорта крыжовника обладают разной устойчивостью к поражаемости растений пилильщиком и что независимо от сортовой принадлежности растений этот вредитель снижает урожай плодов этой культуры. Чтобы установить, достоверны или случайны различия, наблюдаемые между групповыми средними, подвергнем эти данные дисперсионному анализу. Как и в предыдущем примере, сгруппируем выборку в

Таблица 75

Собрано плодов с отдельных кустов крыжовника, кг	Сорта крыжовника A				Σx_B
	Английский желтый A_1		Малахитовый A_2		
	наблюдения	средний урожай x_1	наблюдения	средний урожай x_2	
Не поврежденных пилильщиком B_1	6,3 6,1 5,3 4,9 5,2 4,2 3,8	5,1	4,4 4,3 3,6 4,3 3,5 3,9 4,0	4,0	14
Поврежденных пилильщиком B_2	5,2 4,3 3,5 4,5 5,3 4,0	4,7	3,9 3,2 4,3 3,0 3,8 2,7	3,5	12
Σx_A	13		13		26

таблицу, удобную для расчета вспомогательных величин, предварительно умножив каждую варианту на $K=10$, что избавит нас от дробных чисел и облегчит вычислительную работу. Преобразованные таким образом варианты и расчет вспомогательных величин приведены в табл. 76.

Так как анализируется двухфакторный пропорциональный комплекс, рассчитываем общие девиаты: $D_y=\Sigma x_i^2-H=$ $=49767-\frac{1115^2}{26}=49767-47816{,}34=1950{,}66$. Уменьшаем эту величину в $K^2=100$ раз: $D_y=\frac{1950{,}66}{100}=19{,}51$; $D_x=\Sigma\frac{(\Sigma x_l)^2}{n}-$ $-H=\frac{1}{100}(48759{,}98-47816{,}34)=\frac{943{,}64}{100}=9{,}44$; $D_e=19{,}51-9{,}44=$ $=10{,}07$. Затем находим факториальные девиаты: $D_A=$ $=\sum\frac{(\Sigma x_A)^2}{n_A}-H=\frac{1}{100}(48538{,}28-47816{,}34)=\frac{721{,}89}{100}=7{,}22$; $D_B=\sum\frac{(\Sigma x_B)^2}{n_B}-H=\frac{1}{100}(48035{,}32-47816{,}34)=\frac{218{,}98}{100}=2{,}19$ $D_{AB}=D_x-(D_A+D_B)=9{,}44-(7{,}22+2{,}19)=0{,}03$.

Таблица 76

A, B / x	A_1		A_2		Сумма
	B_1	B_2	B_1	B_2	
x_i	63 61 53 49 52 42 38	52 43 45 35 53 40	44 43 36 43 35 39 40	39 32 43 30 38 27	$a=2$ $b=2$
n	7	6	7	6	$N=26$
Σx_i	358	268	280	209	1115
$(\Sigma x_i)^2$	128164	71824	78400	43681	—
$\frac{(\Sigma x_i)^2}{n}$	18309,14	11970,67	11200,00	7280,17	48759,98
Σx_i^2	18812	12212	11276	7467	49767
n_A	7+6=13		7+6=13		26
Σx_A	358+268=626		280+209=489		—
$\frac{(\Sigma x_A)^2}{n_A}$	30144,31		18393,92		48538,23
n_B	7+7=14		6+6=12		26
Σx_B	358+280=638		268+209=477		—
$\frac{(\Sigma x_B)^2}{n_B}$	29074,57		18960,75		48035,32

Определяем числа степеней свободы: $k_y=N-1=26-1=25$; $k_A=a-1=2-1=1$; $k_B=b-1=2-1=1$; $k_{AB}=k_Ak_B=1$; $k_e=N-ab=26-4=22$. Относим девиаты к числам степеней свободы, что дает значение дисперсий, и сводим результаты анализа в заключительную таблицу (табл. 77).

Из данных табл. 77 видно, что нулевая гипотеза опровергается как в отношении фактора A ($P<0{,}01$), так и фактора B ($P<0{,}05$), хотя и на разных уровнях значимости. Следовательно, с определенной уверенностью можно считать статистически доказанным, что на урожай крыжовника оказывают влияние и сорт (A) и вредоносное действие крыжовникового пилильщика (B).

Неортогональные комплексы. Для двухфакторных ортогональных комплексов характерно равенство $D_x=D_A+D_B+D_{AB}$. В неортогональных комплексах, т. е. таких, в градациях которых содержатся неодинаковые и непропорциональные числа вариант, это равенство нарушается, т. е. $D_x\neq D_A+D_B+D_{AB}$, а следовательно, и $D_{AB}\neq D_x-(D_A+D_B)$. Сохраняется лишь равенство $D_y=D_x+D_e$. Поэтому общие девиаты рассчитывают по тем же

формулам, которые используют при анализе равномерных и пропорциональных комплексов.

Факториальные девиаты (D_A, D_B и D_{AB}) рассчитывают в два этапа. Сначала находят значения некорректированных девиат обозначаемых здесь символами D'_A, D'_B и D'_{AB}. Сумма этих девиат равна некорректированной общей девиате D'_x. Корректируя неисправленные девиаты, получают девиаты исправленные т. е. не смещенные относительно равенства $D_x = D_A + D_B + D_{AB}$. Коррекцию девиат производят умножением их на поправочный коэффициент $K = D_x/D'_x$. Дальнейший ход анализа проводят по обычной для двухфакторных комплексов схеме, описанной выше.

Таблица 7

Варьирование	Степени свободы k	Девиаты D	Дисперсии s^2	Дисперсионное отношение		
				F_Φ	F_{st}	
					5%	1%
По фактору A	1	7,22	7,22	$\frac{7,22}{0,46} = 15,7$	4,3	7,9
По фактору B	1	2,19	2,19	$\frac{2,19}{0,46} = 4,8$	4,3	7,9
Совместно AB	1	0,03	0,03	$\frac{0,03}{0,46} = 0,06$	4,3	7,9
Остаточное	22	10,07	0,46	—	—	—
Общее	25	19,51	—	—	—	—

Неисправленные девиаты определяют по следующим рабочим формулам:

$$D'_x = N\left(\frac{\Sigma \bar{x}_l^2}{ab} - H\right); \qquad (132$$

$$D'_A = N\left(\frac{h_A}{a} - H\right); \qquad (133$$

$$D'_B = N\left(\frac{h_B}{b} - H\right); \qquad (134$$

$$D'_{AB} = D'_x - (D'_A + D'_B). \qquad (135$$

В этих формулах

$$H = \left(\frac{\Sigma \bar{x}_l}{ab}\right)^2, \quad h_A = \sum_j^a \left(\frac{\Sigma \bar{x}_A}{b}\right) \text{ и } \quad h_B = \left(\frac{\Sigma \bar{x}_B}{a}\right)^2,$$

де $\bar{x}_i = \Sigma x_i/n$ — групповые средние; $\Sigma\bar{x}_A$ — сумма групповых средних для каждой из градаций фактора A; $\Sigma\bar{x}_B$ — сумма групповых средних для каждой из градаций фактора B; a — число градаций фактора A; b — число градаций фактора B (в группах A); n — численность вариант в отдельных градациях комплекса; $N = \Sigma n$ — объем комплекса.

Пример 15. Изучали действие сока и паров чеснока, лука и перца на заживление гноящихся ран. Исследование проводили на одновозрастной группе подопытных животных. Эффект оценивали в условных единицах (в баллах). Результаты опыта приведены в табл. 78.

Таблица 78

Способ воздействия	Чеснок	Лук	Перец	Среднее
Сок	7 8 6 7	7 6 5 2	3 2 6 5	5,33
Пары	5 4 6	5 3 6 4	2 4 3 3	4,09
Среднее	6,14	4,75	3,50	—

Таблица 79

A, B / X	A_1		A_2		A_3		Сумма
	B_1	B_2	B_1	B_2	B_1	B_2	
x_i	7 8 6 7	5 4 6	7 6 5	2 5 3 6 4	2 3 6 5	2 4 3 3	$a=3$ $b=2$
n	4	3	3	5	4	4	$N=23$
Σx_i	28	15	18	20	16	12	109
$\Sigma x_i/n = \bar{x}_i$	7,0	5,0	6,0	4,0	4,0	3,0	29
$\bar{x}_i^2$	49	25	36	16	16	9	151
$(\Sigma x_i)^2/n$	196	75	108	80	64	36	559
Σx_i^2	198	77	110	90	74	38	587
$\Sigma\bar{x}_A$	7+5=12		6+4=10		4+3=7		—
$(\Sigma\bar{x}_A/b)^2$	$(12/2)^2=$ $=36{,}00$		$(10/2)^2=$ $=25{,}00$		$(7/2)^2=$ $=12{,}25$		$h_A=73{,}25$
$\Sigma\bar{x}_B$	7+6+4=17			5+4+3=12			—
$(\Sigma\bar{x}_B/a)^2$	32,14			16,00			$h_B=48{,}14$

Обозначим изучаемые факторы через A, а способы их воздействия — через B и подвергнем эти данные дисперсионному анализу. Расчет вспомогательных величин приведен в табл. 79.

Для определения общих девиат сначала находим величину $H=\frac{(\sum x_i)^2}{N}=\frac{109^2}{23}=516{,}56$. Рассчитываем общие девиаты:

$$D_y=\Sigma x_i^2-H=587-516{,}56=70{,}44;\quad D_x=\sum_{j}^{ab}\frac{(\sum x_l)^2}{n}-H=559-$$

$-516{,}56=42{,}44$ и $D_e=70{,}44-42{,}44=28{,}00$.

Числа степеней свободы: $k_y=N-1=23-1=22$; $k_x=ab-1=3\cdot2-1=5$; $k_e=N-ab=23-6=17$. Дисперсии: $s_x^2=42{,}44/5=8{,}49$; $s_e^2=28{,}00/17=1{,}65$. Отсюда $F_\Phi=8{,}49/1{,}65=5{,}1$. Эта величина превосходит критическую точку $F_{st}=4{,}34$ для $\alpha=1\%$. Следовательно, нулевую гипотезу отвергают на высоком уровне значимости ($P<0{,}01$).

Переходим к расчету некорректированных девиат. Предварительно определяем величину $H=\left(\frac{\sum \bar{x}_l}{ab}\right)^2=\left(\frac{29}{3\cdot2}\right)^2=23{,}3$. Подставляем известные величины в формулы (132), (133), (134) и (135):

$$D'_x=N\left(\frac{\sum \bar{x}_l^2}{ab}-H\right)=23\left(\frac{151}{6}-23{,}36\right)=23\cdot1{,}807=41{,}56;$$

$$D'_A=N\left(\frac{h_A}{a}-H\right)=23\left(\frac{73{,}25}{3}-23{,}36\right)=23\cdot1{,}507=24{,}31;$$

$$D'_B=N\left(\frac{h_B}{b}-H\right)=23\left(\frac{48{,}14}{2}-23{,}36\right)=23\cdot0{,}71=16{,}33;$$

$$D'_{AB}=D'_x-(D'_A+D'_B)=41{,}56-24{,}31+16{,}33=0{,}92.$$

Проверяем правильность расчетов: $D'_x=D'_A+D'_B+D'_{AB}=24{,}31+16{,}33+0{,}92=41{,}56$. Расчет девиат произведен правильно. Находим поправочный коэффициент: $K=D_x/D'_x=42{,}44/41{,}56=1{,}0212$. Исправляем факториальные девиаты: $D_A=D'_AK=24{,}31\cdot1{,}0212=24{,}82$; $D_B=D'_BK=16{,}33\cdot1{,}0212=16{,}68$; $D_{AB}=D'_{AB}K=0{,}92\cdot1{,}0212=0{,}94$. Проверяем правильность расчетов: $24{,}82+16{,}68+0{,}94=42{,}44=D_x$. Расчет девиат произведен правильно.

Определяем числа степеней свободы для факториальных дисперсий: $k_A=a-1=3-1=2$; $k_B=b-1=2-1=1$; $k_{AB}=k_Ak_B=2\cdot1=2$; $k_e=17$ (см. выше). Относим девиаты к числам степеней свободы и сводим результаты анализа в заключительную таблицу (табл. 80).

Из данных табл. 80 ясно, что нулевая гипотеза отвергается на высоком уровне значимости ($P<0{,}01$) как в отношении фак-

ора A, так и в отношении фактора B; недоказанным остается совместное влияние этих факторов на результативный признак. Следовательно, достоверно установлено, что лечебный эффект от использования препаратов лука, чеснока и перца при лечении гнойных ран различен. Доказано, что эффективность лечения этими препаратами зависит и от способов их применения.

Таблица 80

Варьирование	Степени свободы k	Девиаты D	Дисперсии s^2	F_Φ	F_{st}	
					5%	1%
По фактору A	2	24,82	12,41	7,5	3,59	6,11
По фактору B	1	16,68	16,69	10,1	4,45	8,40
Совместно AB	2	0,94	0,47	0,28	3,59	6,11
Остаточное	17	28,00	1,65	—	—	—
Общее	22	70,44	—	—	—	—

В тех случаях, когда выборка распределяется в вариационный ряд удобной формой группировки исходных данных, подлежащих дисперсионному анализу, служит решетчатая (корреляционная) таблица.

Пример 16. Изучали продуктивность пчелиных маток трех групп различных пород (фактор A) в зависимости от условий их расплода (фактор B). Продуктивность маток (результативный признак X) оценивали по числу отложенных яиц (в сотнях штук) в среднем за два года. Полученные результаты и их обработка приведены в табл. 81.

Здесь через f обозначены частоты классов (групп), распределенные по ячейкам таблицы; a_x — отклонения равноотстоящих групп от условного нуля. Остальные символы объяснены выше.

Как и в предыдущем примере, предварительно находим величину $H=\frac{(\Sigma fa_x)^2}{N}=\frac{116^2}{50}=269{,}12$ и определяем общие девиаты: $D_y=\Sigma(\Sigma fa_x^2)-H=334-269{,}12=64{,}88$; $D_x=\sum\frac{(\Sigma fa_x)^2}{n}-H=293{,}3-269{,}12=24{,}18$; $D_e=64{,}88-24{,}18=40{,}70$. Числа степеней свободы: $k_y=50-1=49$; $k_x=3\cdot2-1=5$; $k_e=50-6=44$. Дисперсии: $s_x^2=24{,}18/5=4{,}836$; $s_e^2=40{,}70/44=0{,}925$. Отсюда $F_\Phi=4{,}836/0{,}925=5{,}23$. В табл. VI Приложений для $\alpha=1\%$; $k_x=5$ и $k_e=44$ находим $f_{st}=3{,}5$. Нулевую гипотезу отвергают на высоком уровне значимости ($P<0{,}01$).

Находим $H=\left(\frac{\Sigma\bar{x}_l}{ab}\right)^2=\frac{(13{,}94)^2}{(3\cdot2)}=5{,}40$. Рассчитываем факториальные (некорректированные) девиаты по формулам (132)—(135).

Таблица 8

A, B / X	a_x	A_1		A_2		A_3		Сумма
		B_1	B_2	B_1	B_2	B_1	B_2	
20—21,9	5							
18—19,9	4	1	1	2	1		1	$a=3$
16—17,9	3	4	2	1	2	2	3	$b=2$
14—15,9	2	3	4	3	4	3	3	
12—13,9	1	1	1	1	3	2	2	
10—11,9	0							
n		9	8	7	10	7	9	$N=50$
$\Sigma f a_x$		32	19	18	21	14	12	116
$\Sigma f a_x^2$		120	51	54	53	32	24	334
$\frac{(\Sigma f a_x)^2}{n}$		113,8	45,1	46,3	44,1	28,0	16,0	293,3
$\frac{\Sigma f a_x}{n}=\bar{x}_i$		3,56	2,38	2,57	2,10	2,00	1,33	13,94
$\bar{x}_i^2$		12,67	5,66	6,60	4,41	4,00	1,77	35,11
$\Sigma \bar{x}_A$		$3,56+2,38=5,94$		$2,57+2,1=4,67$		$2,0+1,33=3,33$		—
$\left(\frac{\Sigma \bar{x}_A}{b}\right)^2$		$\left(\frac{5,94}{2}\right)^2=8,82$		$\left(\frac{4,67}{2}\right)^2=5,45$		$\left(\frac{3,33}{2}\right)^2=2,77$		$h_A=17,04$
$\Sigma \bar{x}_B$		$3,56+2,57+2,00=8,13$				$2,38+2,10+1,33=5,81$		—
$\left(\frac{\Sigma \bar{x}_B}{a}\right)^2$		$\left(\frac{8,13}{3}\right)^2=7,34$				$\left(\frac{5,81}{3}\right)^2=3,75$		$h_B=11,09$

Таблица 8.

Варьирование	Степени свободы k	Девиаты D	Дисперсии s^2	F_Φ	F_{st} 5%	F_{st} 1%
По фактору A	2	15,05	7,52	8,1	3,21	5,12
По фактору B	1	8,06	8,06	8,7	4,06	7,25
Совместно AB	2	1,07	0,54	0,58	3,21	5,12
Остаточное	44	40,70	0,925	—	—	—
Общее	49	64,88	—	—	—	—

$$D'_x = 50\left(\frac{35,11}{6} - H\right) = 50\,(5,85 - 5,40) = 50 \cdot 0,45 = 22,5;$$

$$D'_A = 50\left(\frac{17,04}{3} - H\right) = 50\,(5,68 - 5,40) = 50 \cdot 0,28 = 14,0;$$

$$D'_B = 50\left(\frac{11,09}{2} - H\right) = 50\,(5,55 - 5,40) = 50 \cdot 0,15 = 7,5;$$

$$D'_{AB} = 22,5 - (14,0 + 7,5) = 1,0.$$

Поправочный коэффициент $K = D_x/D'_x = 24,18/22,50 = 1,0747$. Корректируем неисправленные девиаты: $D_A = 14,0 \cdot 1,0747 = 15,05$; $D_B = 7,5 \cdot 1,0747 = 8,06$; $D_{AB} = 1,0 \cdot 1,0747 = 1,07$. Определив числа степеней свободы, сводим результаты анализа в таблицу (табл. 82).

Нулевую гипотезу отвергают на высоком уровне значимости в отношении как фактора A, так и фактора B ($P < 0,01$). Совместное влияние факторов AB не установлено.

Для облегчения вычислительной работы желательно, где это возможно, переводить неортогональные комплексы в ортогональные путем исключения «лишних» наблюдений из соответствующих градаций. При этом исключение должно быть случайным, нетенденциозным.

Оценка силы влияния факторов. Силу влияния того или иного фактора или их совместного действия на результативный признак определяют с помощью следующих показателей:

$$h^2_A = \hat{s}^2_A / s^2_y \; ; \tag{136}$$

$$h^2_B = \hat{s}^2_B / s^2_y ; \tag{137}$$

$$h^2_{AB} = \hat{s}^2_{AB} / s^2_y , \tag{138}$$

где $\hat{s}^2_A = (s^2_A - s^2_e)/bn$, $\hat{s}^2_B = (s^2_B - s^2_e)/an$ и $\hat{s}^2_{AB} = (s^2_{AB} - s^2_e)/n$ — факториальные дисперсии, определяемые по значениям межгрупповых («неисправленных») и остаточной дисперсий с учетом числа групп a в градациях фактора A и числа групп b в градациях фактора B, а также численности вариант в группах n. Если комплекс неравномерный или пропорциональный, величину n определяют по формуле

$$\bar{n} = \frac{1}{ab - 1}\left\{N - \frac{\Sigma (n_l)^2}{N}\right\}. \tag{139}$$

Знаменателем в формулах (136)—(138) служит величина $s^2_y = \hat{s}^2_A + \hat{s}^2_B + \hat{s}^2_{AB} + s^2_e$. Причем, если влияние одного из регулируемых факторов или их совместное действие на результативный признак не установлено, т. е. статистически недостоверно, этот компонент из знаменателя исключают.

Пример 17. При выяснении влияния микроэлементов (фак-

тор A) и породных свойств (фактор B) на жирномолочность коров (признак X) достоверным оказалось лишь влияние фактора B. Определить силу влияния этого фактора на признак. Выше было найдено: $s^2_B=2{,}65$; $s^2_e=0{,}26$; $n=3$ и $a=3$. Определяем $\hat{s}^2_B=(2{,}65-0{,}26)/(3\cdot 3)=2{,}39/9=0{,}266$. Подставляем нужные данные в формулу (137): $h^2_b=0{,}266/(0{,}266+0{,}26)=0{,}266/0{,}526=0{,}50$. Тот же показатель, определяемый по способу Плохинского, оказывается несколько выше: $h^2_B=D_B/D_y=7{,}94/15{,}78=0{,}53$. Проверим достоверность этого показателя.

1. По Снедекору, $h^2_B=0{,}50$. Исходные данные: $s^2_B=2{,}65$; $s^2_e=0{,}26$; $N=36$; $b=4$ (см. табл. 71). Отсюда $F_\Phi=2{,}65/0{,}26=10{,}2$. По табл. VI Приложений для $k_B=b-1=4-1=3$; $k_e=N-b=36-4=32$ и $\alpha=1\%$ находим $F_{st}=4{,}46$. Так как $F_\Phi>Fst$, нулевую гипотезу отвергают на высоком уровне значимостн ($P<0{,}01$).

2. По Плохинскому, $h^2_B=0{,}53$. Ошибка показателя силы влияния $sh^2_B=(1-0{,}53)(4-1)/(36-4)=1{,}41/32=0{,}044$. Отсюда $E_\Phi=0{,}53/0{,}044=12{,}05$. Эта величина превышает $Fst=4{,}46$, что позволяет отвергнуть нулевую гипотезу на высоком уровне значимости ($P<0{,}01$).

Пример 18. Определить силу влияния фитонцидов лука, перца и чеснока (фактор A), а также способов их воздействия (фактор B) на заживление гнойных ран (признак X). В данном случае комплекс неравномерный. Здесь $N=23$; $a=3$; $b=2$ (см. табл. 79); $s^2_A=12{,}41$; $s^2_B=16{,}68$; $s^2_e=1{,}65$ (см. табл. 80).

Находим усредненную величину n (по Снедекору):

$$n=\frac{1}{3\cdot 2-1}-\left(23-\frac{4^2+3^2+3^2+5^2+4^2+4^2}{23}\right)=3{,}81.$$

Рассчитываем факториальные дисперсии: $\hat{s}^2_A=\frac{s^2_A-s^2_e}{nb}=\frac{12{,}41-1{,}65}{3{,}81\cdot 2}=\frac{10{,}76}{7{,}62}=1{,}412$; $\hat{s}^2_B=\frac{s^2_a-s^2_e}{nQ}=\frac{16{,}68-1{,}65}{3{,}81\cdot 3}=\frac{15{,}03}{11{,}43}=1{,}315$.

Знаменатель $s^2_y=1{,}412+1{,}315+1{,}650=4{,}377$, откуда $h^2_A=\frac{1{,}412}{4{,}377}=0{,}323$; $h^2_B=\frac{1{,}315}{4{,}377}=0{,}300$; $h^2_e=\frac{1{,}650}{4{,}377}=0{,}377$.

Отсюда следует, что доля общей вариации признака, определяемая влиянием фактора A, равна 32,3%, тогда как доля общей вариации, связанная с влиянием на признак фактора B, равна 30,0%. Остальные 37,7% общей вариации признака вызваны влиянием неорганизованных (случайных) факторов.

Те же показатели, вычисленные по методу Плохинского, оказались равными $h^2_A=D_A/D_y=24{,}82/70{,}44=0{,}352$; $h^2_B=D_B/D_y=16{,}68/70{,}44=0{,}237$; $h^2_e=D_e/D_y=28{,}00/70{,}44=0{,}398$. Сумма

$h^2_A+h^2_B+h^2_e=0{,}352+0{,}237+0{,}398=0{,}987$, т. е. не $=1$. Это и понятно, если учесть, что здесь взяты не исправленные величины, а просто девиаты и величина h^2_{AB} не вычислялась; при таких обстоятельствах сумма всегда будет меньше единицы.

Проверим достоверность этих показателей.

1. По Снедекору. Исходные данные: $N=23$; $a=3$; $b=2$; $s^2_A=12{,}41$; $s^2_B=16{,}68$; $s^2_e=1{,}65$. Отсюда для $h^2_A=0{,}323$ критерий $F_A=12{,}41/1{,}65=7{,}52$. Эта величина превосходит критическую точку $F_{st}=5{,}85$ для $k=a-1=3-1=2$; $k_e=N-a==23-3=20$ и $\alpha=1\%$; $F_B=16{,}68/1{,}65=10{,}11$. Эта величина превосходит критическую точку $F_{st}=8{,}02$ для $k_B=a-1=2-1=1$; $k_e=23-2=21$ и $\alpha=1\%$. H_0-гипотезу отвергают на 1%-ном уровне значимости.

2. По Плохинскому. Ошибка для $h^2_A=0{,}352$ составит $s_{h^2_A}=(1-0{,}352)(3-1)/(23-3)=1{,}296/20=0{,}065$. Отсюда $F_A==0{,}352/0{,}065=5{,}42$. Эта величина превосходит $F_{st}=3{,}49$ для $k_1=2$; $k_2=20$ и $\alpha=5\%$. Ошибка для $h^2_B=0{,}237$ составит $s_{h^2_B}==(1-0{,}237)(2-1)/(23-2)=0{,}763/21=0{,}036$. Отсюда $F_B==0{,}237/0{,}036=6{,}58$. Эта величина превосходит $F_{st}=4{,}32$ для $k_1=1$ и $k_2=21$, а также $\alpha=5\%$. Таким образом, H_0-гипотеза отвергается по Снедекору на 1%-ном, а по Плохинскому — на 5%-ном уровне значимости ($0{,}01<P<0{,}05$).

VII.3. АНАЛИЗ ТРЕХФАКТОРНЫХ КОМПЛЕКСОВ

Выше уже было показано, что с увеличением числа организованных факторов, воздействующих на результативный признак, увеличивается и число их возможных сочетаний, усложняется символика, особенно при определении девиат. В остальном организация и анализ многофакторных комплексов принципиально не отличаются от простых комплексов. Схему анализа трехфакторного равномерного комплекса можно представить в виде следующей заключительной таблицы (табл. 83).

Здесь A, B, C — организованные факторы, воздействующие на результативный признак X; $n_A=nac$ — количество вариант в отдельных градациях фактора A; $n_B=nbc$ — количество вариант в каждой градации фактора B; $n_C=nab$ — количество вариант в каждой градации фактора C; a, b, c — число градаций или групп факторов A, B, C; n — численность вариант в отдельных градациях комплекса; $\Sigma n=nabc=N$ — общее число вариант, входящих в дисперсионный комплекс, его объем; $h_{AB}=\dfrac{\Sigma(\Sigma x_{AB})^2}{nc}-H$; $h_{AC}=\dfrac{\Sigma(\Sigma x_{AC})^2}{nb}-H$; $h_{BC}=\dfrac{\Sigma(\Sigma x_{BC})^2}{na}-H$ — вспомогательные величины; $H=\dfrac{(\Sigma x_i)^2}{N}$ — общая для всех девиат величина.

Анализ трехфакторных комплексов начинают, как обычно с определения Σx_i, $\Sigma(\Sigma x_i)^2$ и $\Sigma x^2{}_i$. Затем рассчитывают девиаты общую D_y, факториальную, или межгрупповую, D_x и остаточную D_e; устанавливают числа степеней свободы k_y, k_x и k_e

Таблица 8

Вариация	Степени свободы k	Девиаты D
Общая	$k_y = N-1$	$D_y = \Sigma(\Sigma x_i^2) - H$
Межгрупповая	$k_x = abc-1$	$D_x = \dfrac{\Sigma(\Sigma x_i)^2}{n} - H$
Внутригрупповая, или остаточная	$k_e = k_y - k_x$	$D_e = D_y - D_x$
По фактору A	$k_A = a-1$	$D_A = \dfrac{\Sigma(\Sigma x_A)^2}{n_A} - H$
» B	$k_B = b-1$	$D_B = \dfrac{\Sigma(\Sigma x_B)^2}{n_B} - H$
» C	$k_C = c-1$	$D_C = \dfrac{\Sigma(\Sigma x_C)^2}{n_C} - H$
Совместного действия AB	$k_{AB} = k_A k_B$	$D_{AB} = h_{AB} - (D_A + D_B)$
» AC	$k_{AC} = k_A k_C$	$D_{AC} = h_{AC} - (D_A + D_C)$
» BC	$k_{BC} = k_B k_C$	$D_{BC} = h_{BC} - (D_B + D_C)$
» ABC	$k_{ABC} = k_A k_B k_C$	$D_{ABC} = D_x - (D_A + D_B + D_C + D_{AB} + D_{AC} + D_{BC})$

Делением девиат на числа степеней свободы определяют дисперсии $s^2{}_x$ и $s^2{}_e$. Дисперсионное отношение находят по факториальной дисперсии, отнесенной к остаточой дисперсии, т. е. $F_\Phi = s^2{}_x/s^2{}_e$ (при $s^2{}_x \geqslant s^2{}_e$). Если $F_\Phi \geqslant F_{st}$ для k_x, k_e и α, то H_0-гипотезу отвергают, что дает основание для перехода к расчету факториальных девиат D_A, D_B, D_C и девиат совместного действия D_{AB}, D_{AC}, D_{BC}, D_{ABC}. Результаты вычислений сводят в заключительную таблицу с последующими выводами. Таким образом, как и в рассмотренных выше случаях, наибольших усилий и внимания требуют расчеты девиат. Другие действия сравнительно просты и не требуют дополнительных разъяснений.

Пример 19. Описанный ход анализа легче усвоить из соответствующего примера. В табл. 84 сгруппированы и частично обработаны данные о влиянии трех независимых факторов A, B и C на результативный признак X. Здесь $n=4$, $a=2$, $b=2$, $c=4$ Σx_i — сумма вариант в отдельных группах или градациях комплекса, а $\bar{x}_i$ — групповые средние арифметические. Остальные символы понятны из табл. 84. Объем комплекса $N = 4 \cdot 2 \cdot 2 \cdot 4 = 64$. Величина $H = 1728^2/64 = 46656$. Определяем девиаты: $D_y = \Sigma x^2 -$

$-H=48\,698-46\,656=2042$; $D_x=\frac{\Sigma(\Sigma x_i)^2}{n}-H=\frac{194\,092}{4}-$

$-H=1867$; $D_e=D_y-D_x=175$. Числа степеней свободы: $k_y=$ $=N-1=63$; $k_x=(abc)-1=16-1=15$; $k_e=k_y-k_x=$ $=63-15=48$. Дисперсия: $s^2{}_x=1867/15=124{,}47$ и $s^2{}_e=175/48=$ $=3{,}65$. Критерий $F_\Phi=124{,}47/3{,}65=34{,}10>F_{st}=2{,}48$ для $k_x=15$, $k_e=48$ и $\alpha=0{,}01$. H_0-гипотезу отвергают на высоком уровне значимости ($P<0{,}01$).

Таблица 84

Градации факторов			Состав градаций (x_i)				Σx_i	$(\Sigma x_i)^2$	Σx_i^2	$\overline{x_i}$
A	B	C	1	2	3	4				
A_1	B_1	C_1	17	18	21	20	76	5 776	1454	19,0
		C_2	19	21	22	22	84	7 056	1770	21,0
		C_3	20	22	23	23	88	7 744	1942	22,0
		C_4	20	23	24	25	92	8 464	2130	23,0
	B_2	C_1	19	22	24	23	88	7 744	1950	22,0
		C_2	21	22	25	24	92	8 464	2126	23,0
		C_3	21	24	26	25	96	9 216	2318	24,0
		C_4	23	23	25	25	96	9 216	2308	24,0
A_2	B_1	C_1	24	26	29	28	107	11 449	2877	26,8
		C_2	27	28	32	30	117	13 689	3437	29,3
		C_3	26	30	31	32	119	14 161	3561	29,8
		C_4	30	29	33	32	124	15 376	3854	31,0
	B_2	C_1	29	30	32	33	124	15 376	3854	31,0
		C_2	31	32	35	34	132	17 424	4366	33,0
		C_3	34	35	38	37	144	20 736	5194	36,0
		C_4	36	38	39	36	149	22 201	5557	37,3
Сумма			—	—	—	—	1728	194 092	48698	—

Переходим к расчету факториальных девиат D_A, D_B, D_C. Сначала определяем суммы вариант Σx_i для каждой градации комплекса (см. табл. 84). Затем, суммируя числовые значения всех возможных сочетаний независимых факторов A, B и C, находим Σx_A, Σx_B и Σx_C. Начнем с определения Σx_A и Σx_B (табл. 85).

Отсюда $D_A=\frac{\Sigma(\Sigma x_A)^2}{nac}-H=\frac{7122+10162}{32}-H=48\,100-$

$-46\,656=1444$;

Таблица 8

A \ B	B_1	B_2	ΣX_A
A_1	76+84+88+92=340	88+92+96+96=372	712
A_2	107+117+119+124=467	124+132+144+149=549	1016
Σx_B	807	921	

Таблица 8

A \ C	C_1	C_2	C_3	C_4
A_1	76+88=164	84+92=176	88+96=184	92+96=188
A_2	107+124=231	117+32=249	119+144=263	124+149=273
Σx_C	395	425	447	461

$$D_B = \frac{\Sigma(\Sigma x_B)^2}{nbc} - H = \frac{807^2 + 921^2}{32} - H = 46859{,}06 - H = 203{,}06.$$

Сочетая значения факторов A_1 и A_2 по C_{1-4}, находим Σx_c (табл. 86).

Определяем $D_C = \frac{\Sigma(\Sigma x_C)^2}{nab} - H = \frac{395^2 + 425^2 + 447^2 + 461^2}{4 \cdot 2 \cdot 2} -$

$$- H = \frac{748\,980}{16} - H = 46811{,}25 - 46\,656 = 155{,}25.$$

Расчет девиат совместного действия D_{AB}, D_{AC}, D_{BC} и D_{ABC} совершается последовательно в два приема: сначала находят величины h_{AB}, h_{AC} и h_{BC}, затем определяют девиаты совместного действия. Исходные данные для определения h_{AB} и h_{AC} приведены в табл. 85 и 86, а для определения h_{BC} — в табл. 87, содержащей сочетание числовых значений факторов B_1 и B_2 по C_{1-4}.

Отсюда $h_{AB} = \frac{\Sigma(\Sigma x_{AB})^2}{nc} - H = \frac{340^2 + 372^2 + 467^2 + 549^2}{4 \cdot 4} - H =$

$$= \frac{773\,474}{16} - H = 48342{,}13 - 46\,656 = 1686{,}13; \quad h_{AC} = \frac{\Sigma(\Sigma x_{AC})^2}{nb} -$$

$$- H = \frac{164^2 + 176^2 + 184^2 + 188^2 + 231^2 + 249^2 + 263^2 + 273^2}{4 \cdot 2} - H =$$

$$=\frac{386\,132}{8}-H=48266{,}50-H=1610{,}50;\quad h_{BC}=\frac{\Sigma(\Sigma x_{BC})^2}{na}-H=$$

$$=\frac{183^2+201^2+207^2+216^2+212^2+224^2+240^2+245^2}{4\cdot 2}-H=$$

$$=\frac{376\,140}{8}-H=47017{,}50-46656{,}00=361{,}50.$$

Таблица 87

B \ C	C_1	C_2	C_3	C_4
B_1	76 + 107 = 183	84 + 117 = 201	88 + 119 = 207	92 + 124 = 216
B_2	88 + 124 = 212	92 + 132 = 224	96 + 144 = 240	96 + 149 = 245
Σx_C	395	425	447	461

Переходим к определению девиат совместного действия: $D_{AB}=h_{AB}-(D_A+D_B)=1686{,}13-1444+203{,}06=39{,}07$; $D_{AC}=h_{AC}-(D_A+D_C)=1610{,}50-1444+155{,}25=11{,}25$; $D_{BC}=$

Таблица 88

Вариация	Степени свободы k	Девиаты D	Дисперсии s^2_x	Дисперсионные отношения F_Φ	F_{st} 0,05	F_{st} 0,01
По фактору A	1	1444,00	1444,00	$F_A=395{,}6$	4,04	7,20
» B	1	203,06	203,06	$F_B=55{,}6$	4,04	7,20
» C	3	155,25	51,75	$F_C=14{,}2$	2,80	4,22
Совместная AB	1	39,07	39,07	$F_{AB}=10{,}7$	4,04	7,20
» AC	3	11,25	3,75	$F_{AC}=1{,}0$	2,80	4,22
» BC	3	3,19	1,06	$F_{BC}=0{,}29$	2,80	4,22
» ABC	3	11,18	3,73	$F_{ABC}=1{,}0$	2,80	4,22
Остаточная	48	175,00	3,65	—	—	—
Общая	63	2042,00	—	—	—	—

$=h_{BC}-(D_B+D_C)=361{,}50-203{,}06+155{,}25=3{,}19$; $D_{ABC}=D_x-(D_A+D_B+D_C+D_{AB}+D_{AC}+D_{BC})=1867-(1444+203{,}06+155{,}25+39{,}07+3{,}19+11{,}25)=1867-1855{,}82=11{,}18$.

Устанавливаем числа степеней свободы. В данном комплексе $a=2$, $b=2$, $c=4$. Отсюда $k_A=a-1=2-1=1$; $k_B=b-1=$

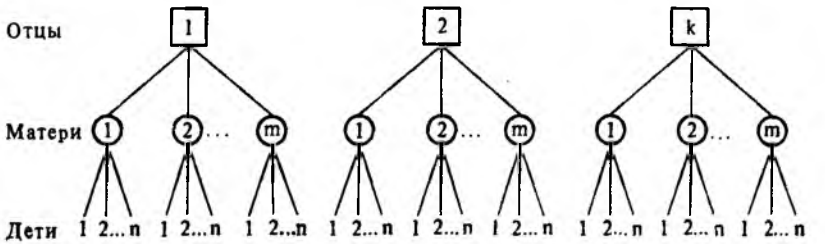
$=2-1=1$; $k_C=c-1=4-1=3$; $k_{AB}=k_Ak_B=1$; $k_{AC}=k_Ak_C=$ $=1\cdot3=3$; $k_{BC}=k_Bk_C=1\cdot3=3$; $k_{ABC}=k_Ak_Bk_C=1\cdot1\cdot3=3$.

Делим девиаты на числа степеней свободы и определяем дисперсии; сводим результаты анализа в заключительную таблицу (табл. 88). Из табл. 88 следует, что нулевая гипотеза отвергается на 1%-ном уровне значимости лишь в отношении действия на признак факторов A, B, C и совместного действия факторов AB. Влияние на результативный признак остальных сочетаний факторов остается статистически недоказанным.

VII.4. АНАЛИЗ ИЕРАРХИЧЕСКИХ КОМПЛЕКСОВ

Наряду с рассмотренными выше схемами, где возможны любые комбинации факторов, воздействующих на признак, в практике встречаются и такие дисперсионные комплексы, в которых свободное комбинирование факторов друг с другом исключено. Такие комплексы называют *иерархическими*. Они организуются, например, при изучении наследственного влияния родительских поколений на продуктивность или поведение потомства, при выяснении взаимоотношений между родственными в систематическом отношении группами живых существ и в других подобных случаях.

Характерной особенностью таких комплексов является определенная *иерархическая соподчиненность* их структурных компонентов, когда группы относительно низкого ранга находятся в строгой зависимости от связанных с ними групп более высокого положения. Наглядно эту зависимость можно выразить в виде следующей примерной схемы:



Анализ иерархических комплексов имеет свои особенности, обусловленные невозможностью свободного комбинирования различных групп по фактору B из разных градаций фактора A, занимающего более высокое положение в общей схеме иерархического комплекса. При обработке таких дисперсионных комплексов не вычисляют дисперсию s^2_{AB} совместного действия факторов AB, несколько по-другому выглядят дисперсионные отношения F_Φ, иначе по сравнению с обычными многофакторными комплексами определяют факториальные дисперсии.

Как и прочие, иерархические комплексы могут быть равномерными, пропорциональными и неравномерными. Структура иерархического комплекса зависит от количества учитываемых факторов и их градаций. Простейшей иерархической схемой является схема двухфакторного дисперсионного анализа. Ее можно представить в виде следующей таблицы (табл. 89).

Таблица 89

Варьирование	Число степеней свободы k	Девиаты D	Средние квадраты или дисперсии s^2_x	F_Φ	Факторальные дисперсии	Сила влияния факторов
По фактору A	$k_A=a-1$	D_A	$s_1^2=\frac{D_A}{k_A}$	$F_A=\frac{s_1^2}{s_2^2}$	$s_A^2=\frac{s_1^2-s_2^2}{bn}$	$h_A^2=\frac{s_A^2}{s_y^2}$
По фактору B	$k_B=a(b-1)$	D_B	$s_2^2=\frac{D_B}{x_B}$	$F_B=\frac{s_2^2}{s_e^2}$	$s_B^2=\frac{s_2^2-s_e^2}{n}$	$h_B^2=\frac{s_B^2}{s_y^2}$
Остаточное	$k_e=ab(n-1)$	D_e	$s_e^2=\frac{D_e}{k_e}$	—	$s_e^2=\frac{D_e}{k_e}$	$h_e^2=\frac{s_e^2}{s_y^2}$
Общее	$k_y=N-1$	D_y	—	—	$s_y^2=s_A^2+s_B^2+s_e^2$	$h_A^2+h_B^2+h_e^2=1$

Общие суммы квадратов отклонений, или девиаты D, определяют по общим для всех комплексов формулам (109), (110) и (111). Факториальные девиаты определяют следующим образом: D_A — по формуле (127) для равномерных комплексов или по формуле (129) для неравномерных и пропорциональных комплексов, а девиату D_B — по формуле

$$D_B=\frac{\Sigma(\Sigma x_i)^2}{n}-\frac{\Sigma(\Sigma x_A)^2}{n_A}, \qquad (140)$$

а для неравномерных и пропорциональных комплексов — по иной формуле:

$$D_B=\sum_{j}^{B}\frac{(\Sigma x_i)^2}{n}-\sum\frac{(\Sigma x_A)^2}{n_A}. \qquad (141)$$

При этом $D_y=D_A+D_B+D_e$, равно как и $k_y=k_A+k_B+k_e$. Здесь x_i — варианты, находящиеся в градациях комплекса AB; x_A — варианты, находящиеся в градациях фактора A (занимающего самое высокое положение в иерархической схеме); n_i — численность вариант в отдельных градациях комплекса; n_A — количество вариант в каждой из градаций фактора A; $\Sigma n_i=\Sigma n_A=N$ — общее число вариант, входящих в состав данного комплекса, его объем.

При неодинаковой численности вариант в градациях комплекса в качестве знаменателя в формулах для определения факториальных дисперсий $s^2{}_A=(s^2{}_1-s^2{}_2)/bn$ и $s^2{}_B=(s^2{}_2-s^2{}_e)/n$ берут усредненные величины $\overline{bn}$ и $\bar{n}$, вычисляемые по следующим формулам:

$$\overline{bn}=\frac{1}{a-1}\left(N-\frac{\Sigma(n_A^2)}{N}\right); \qquad (142)$$

$$n=\frac{n_A+n_B}{2}, \qquad (143)$$

где $n_A=\left(\frac{1}{a-1}\sum\frac{\Sigma n_A^2}{n_A}\right)-\frac{\Sigma n_i^2}{N}$ и $n_B=\frac{1}{b-1}\left(N-\sum\frac{\Sigma n_A^2}{n_A}\right)$.

В этих формулах a — число градаций фактора A; b — число градаций фактора B; n — численность вариант в отдельных градациях комплекса; n_A — численность вариант в каждой из градаций фактора A и $N=\Sigma n=\Sigma n_A$ — объем всего дисперсионного комплекса.

Формулы для определения чисел степеней свободы k_B и k_e, приведенные в табл. 89, применяют к комплексам с равночисленными группами фактора B, находящимися в градациях фактора A, т. е. здесь b обозначает численность групп фактора B в отдельных градациях фактора A.

Можно, однако, определять число степеней свободы k_B и k_e исходя из учета общего числа групп фактора B, входящих в дисперсионный комплекс b', по формулам $k_B=b'-a$; $k_e=N-b'$. Эти формулы универсальны, пригодны для определения k_B и k_e при наличии равночисленных и неравночисленных групп фактора B, находящихся в градациях фактора A дисперсионного комплекса.

Рассмотрим иерархическую схему двухфакторного равномерного и неравномерного комплекса на конкретных примерах.

Пример 20. Изучали влияние породных свойств хряков Барона A_1 и Сокола A_2 на плодовитость их дочернего потомства X,

юлученных от трех свиноматок B. Плодовитость дочерних осо-ей учитывали по числу живых поросят в их пометах. Результа-ы испытаний приведены в табл. 90.

Таблица 90

Отцовское поколение A	Барон			Сокол		
Материнское поколение B	B_1	B_2	B_3	B_1	B_2	B_3
Число поросят в пометах дочерних особей x_i	7 6 8 7	8 9 8 10	9 7 9 9	10 8 11 10	9 8 7 8	12 10 9 10
$\bar{x}_B$	7,00	8,75	8,50	9,75	8,00	10,25
$\bar{x}_A$	8,08			9,33		

Из табл. 90 видно, что групповые средние $\bar{x}_B$, характеризующие плодовитость дочерних особей каждой свиноматки в отдельности, а также средние показатели плодовитости дочерних осо-ей по отцам $\bar{x}_A$ варьируют как внутри групп A_1 и A_2, так и между группами. Нужно выяснить, существенны ли расхождения между средними показателями и какова сила влияния органи-ованных и неорганизованных факторов на величину варьирова-ия результативного признака, т. е. на плодовитость дочерних особей.

Как обычно, начинаем с расчета вспомогательных величин, необходимых для определения сумм квадратов отклонений табл. 91). Подставляя известные величины в формулы, нахо-им: $D_y = \Sigma x^2_i - H = 1867 - 209^2/24 = 1867 - 1820{,}04 = 46{,}96$;

$D_x = \frac{\Sigma(\Sigma x_l)^2}{n} - H = \frac{7391}{4} - H = 1847{,}75 - 1820{,}04 = 27{,}71$; $D_e =$

$= D_y - D_x = 46{,}96 - 27{,}71 = 19{,}25$; $D_A = \frac{\Sigma(\Sigma x_A)^2}{n_A} - H = \frac{21\,953}{12} -$

$- H = 1829{,}42 - 1820{,}04 = 9{,}38$; $D_B = \frac{\Sigma(\Sigma x_l)^2}{n} - \frac{\Sigma(\Sigma x_A)^2}{n_A} =$

$= 1847{,}75 - 1829{,}42 = 18{,}33$.

Определяем числа степеней свободы: $k_y = N - 1 = 24 - 1 = 23$; $k_A = a - 1 = 2 - 1 = 1$; $k_B = a(b - 1) = 2(3 - 1) = 4$ или $k_B = b' - a = 6 - 2 = 4$; $k_e = ab(n - 1) = 2 \cdot 3(4 - 1) = 18$ или $k_e = N - b' = 24 - 6 = 18$.

Таблица 9

Хряки	A_1			A_2			Сумма
Свиноматки	B_1	B_2	B_3	B_1	B_2	B_3	
Дочерние особи x_i	7 6 8 7	8 9 8 10	9 7 9 9	10 8 10 11	9 8 7 8	12 10 10 9	$a=2$ $b=3$
n	4	4	4	4	4	4	$N=24$
Σx_i	28	35	34	39	32	41	209
$(\Sigma x_i)^2$	784	1225	1156	1521	1024	1681	7391
Σx_i^2	198	309	292	385	258	425	1867
n_A	12			12			24
Σx_A	97			112			—
$(\Sigma x_A)^2$	9409			12544			21953

Делим суммы квадратов отклонений (девиаты) на числа степеней свободы и сводим результаты анализа в таблицу (табл. 92).

Нулевая гипотеза отвергается на 5%-ном уровне значимости только в отношении фактора B (влияние различий материнских особей). Факториальная дисперсия $s^2_B=(s^2_2-s^2_e)/n=(4{,}58-1{,}07)/4=3{,}51/4=0{,}88$. Общая дисперсия $s^2_y=s^2_B+s^2_e=0{,}88+1{,}07=1{,}95$. Отсюда показатели силы влияния факторов: $h^2_B=s^2_B/s^2_y=0{,}88/1{,}95=0{,}451$; $h^2_e=s^2_e/s^2_y=1{,}07/1{,}95=0{,}549$. Это означает, что 54,9% общего варьирования результативного признака (плодовитость дочерних особей) обусловлены влиянием неорганизованных (случайных) факторов и 45,1% общей вариации признака определяется компонентом материнских особей [1].

[1] Разумеется, эта доля общей вариации результативного признака определяется не только генотипическим разнообразием матерей, но и совместным влиянием родителей (хряки×свиноматки) на плодовитость дочерних особей.

Рассчитанные таким образом показатели силы влияния факторов есть не что иное, как *коэффициенты внутриклассовой корреляции* r_w; в селекционно-генетических исследованиях их используют в качестве показателей наследуемости h^2 в широком смысле [1].

Таблица 92

Варьирование	Степени свободы k	Девиаты D	Дисперсии s^2	Дисперсионное отношение F_Φ	F_{st} 5%	F_{st} 1%
По фактору A (между хряками)	1	9,38	9,38	$F_A = 9{,}38/4{,}58 = 2{,}0$	7,71	21,20
По фактору B (внутри групп хряков)	4	18,33	4,58	$F_B = 4{,}58/1{,}07 = 4{,}3$	2,93	4,58
Остаточное	18	19,25	1,07	—	—	—
Общее	23	46,96	—	—	—	—

Пример 21. На основе данных родословных записей была составлена выборка по такому признаку: процент жира в молоке коров дочернего поколения по второму и третьему отелам (табл. 93).

Подвергнем двухфакторный неравномерный иерархический комплекс дисперсионному анализу. Предварительно уменьшим каждую варианту x_i на три единицы, что облегчит расчет вспомогательных величин, нужных для определения девиат (табл. 94).

Определяем общие девиаты комплекса: $D_y = \Sigma x^2_i - H = 34{,}50 - 36{,}6^2/40 = 34{,}50 - 33{,}49 = 1{,}01$; $D_x = \sum \frac{(\Sigma x_l)^2}{n} - H = 33{,}89 - 33{,}49 = 0{,}40$; $D_e = D_y - D_x = 1{,}01 - 0{,}40 = 0{,}61$. Переходим к определению факториальных девиат: $D_A = \sum \frac{(\Sigma x_A)^2}{n_A} - H = 33{,}79 - 33{,}49 = 0{,}30$; $D_B = \sum \frac{(\Sigma x_l)^2}{n} - \frac{(\Sigma x_A)^2}{n_A} = 33{,}89 - 33{,}79 = 0{,}10$.

[1] Если наследуемость в широком смысле определяют как отношение дисперсии, характеризующей генетическое разнообразие s_g^2, к сумме генетической и паратипической, или средовой s_e^2, дисперсий, т. е. $h^2 = \frac{s_g^2}{s_g^2 + s_e^2}$, то в узком смысле наследуемость h^2 определяют отношением аддитивной дисперсии s_a^2 к сумме трех компонентов изменчивости: генетической дисперсии, средовой и аддитивной, т. е. $h^2 = \frac{s_a^2}{s_g^2 + s_a^2 + s_e^2}$.

Таблица 9:

Отцовское поколение A	Материнское поколение B	Дочерние поколения							Групповые средние	
		1	2	3	4	5	6	7	$\bar{x}_B$	$\bar{x}_A$
I	1	4,0	3,8	3,6	3,8				3,80	3,76
	2	3,9	3,7	3,8	3,7	3,5			3,72	
II	3	4,0	4,1	3,9	4,0				4,00	
	4	4,2	4,0	4,0	3,9	4,0	4,1	3,8	4,00	3,97
	5	3,9	3,9	4,0	3,8				3,90	
II	6	4,1	4,2	4,0	3,9	4,0	3,8		4,00	
	7	4,0	4,1	4,1	3,8	3,9			3,98	3,95
	8	3,9	3,9	3,8	4,1	3,6			3,86	

Устанавливаем числа степеней свободы: $k_y = N-1 = 40-1 = 39$; $k_A = a-1 = 3-1 = 2$; $k_B = b-a = 8-3 = 5$; $k_e = N-b = 40-8 = 32$.

Делим суммы квадратов отклонений (девиаты) на числа степеней свободы и сводим результаты дисперсионного анализа в таблицу (табл. 95). Статистически достоверным оказалось влияние фактора A ($P<0{,}05$).

Переходим к расчету факториальных девиат. Предварительно находим усредненные значения $\bar{bn}$ и $\bar{n}$: $\bar{bn} = \frac{1}{3-1}\left(40 - \frac{9^2+15^2+16^2}{40}\right) = \frac{1}{2}\cdot 40 - \frac{562}{40} = 12{,}98 \approx 13$; $\bar{n}_A = \frac{1}{3-1}\left(\frac{4^2+5^2}{9} + \frac{4^2+7^2+4^2}{15} + \frac{6^2+5^2+5^2}{16} - \frac{4^2+5^2+4^2+7^2+4^2+6^2+5^2+5^2}{40}\right) = \frac{15{,}33}{2} - \frac{208}{40} = 7{,}66 - 5{,}20 = 2{,}46$; $\bar{n}_B = \frac{1}{8-1}(40 - \mathbf{15{,}33}) = \frac{24{,}67}{7} = 3{,}52$.

Отсюда $\bar{n} = \frac{2{,}46+3{,}52}{2} = 2{,}99 \approx 3$.

Определяем факториальные дисперсии:

$$s_A^2 = \frac{s_1^2 - s_2^2}{bn} = \frac{0{,}150 - 0{,}020}{\mathbf{13}} = 0{,}01;$$

$$s_B^2 = \frac{s_2^2 - s_e^2}{n} = \frac{0{,}020 - 0{,}019}{3} = 0{,}00033.$$

Общая дисперсия $s^2_y = 0{,}010 + 0{,}00033 + 0{,}019 = 0{,}0293$. Отсюда сила влияния факторов: $h^2_A = 0{,}010/0{,}0293 = 0{,}341$; $h^2_B = 0{,}00033/0{,}0293 = 0{,}011$; $h^2_e = 0{,}0190/0{,}0293 = 0{,}648$.

Таблица 94

Отцовское поколение	A_1		A_2			A_3			Сумма
Материнское поколение	B_1	B_2	B_1	B_2	B_3	B_1	B_2	B_3	
Процент жира в молоке дочерей x_i	1,0 0,8 0,6 0,8	0,9 0,7 0,8 0,7 0,5	1,0 1,1 0,9 1,0	1,2 1,0 1,0 0,9 1,0 1,1 0,8	0,9 0,9 1,0 0,8	1,1 1,2 1,0 0,9 1,0 0,8	1,0 1,1 1,1 0,8 0,9	0,9 0,9 0,8 1,1 0,6	$a=3$ $b=8$
n	4	5	4	7	4	6	5	5	$N=40$
Σx_i	3,2	3,6	4,0	7,0	3,6	6,0	4,9	4,3	36,6
$(\Sigma x_i)^2$	10,24	12,96	16,0	49,0	12,96	36,0	24,01	18,49	—
$\frac{(\Sigma x_i)^2}{n}$	2,56	2,59	4,0	7,0	3,24	6,0	4,80	3,70	33,89
Σx_i^2	2,64	2,68	4,02	7,10	3,26	6,10	4,87	3,83	34,50
n_A	9		15			16			40
Σx_A	6,8		14,6			15,2			36,6
$(\Sigma x_A)^2$	46,24		213,16			231,04			—
$\frac{(\Sigma x_A)^2}{n_A}$	5,14		14,21			14,44			33,79

Таблица 95

Варьирование	Степени свободы k	Девиаты D	Дисперсии s^2	Дисперсионные отношения F_Φ	F_{st} 5%	F_{st} 1%
По фактору A	2	0,30	0,150	$F_A=0{,}150/0{,}020=7{,}5$	5,79	13,27
По фактору B	5	0,10	0,020	$F_B=0{,}020/0{,}019=1{,}1$	2,51	3,65
Остаточное	32	0,61	0,019	—	—	—
Общее	39	1,01	—	—	—	—

Из приведенных расчетов факториальных дисперсий и показателей силы влияния факторов A и B (хотя действие B на признак и не было доказано) становится ясным, каким образом можно разложить общую дисперсию комплекса на составляющие ее компоненты, выявить силу влияния каждого компонента на общее варьирование результативного признака.

ГЛАВА VIII

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Функциональная зависимость и корреляция. Еще Гиппократ в VI в. до н. э. обратил внимание на наличие связи между телосложением и темпераментом людей, между строением тела и предрасположенностью к тем или иным заболеваниям. Определенные виды подобной связи выявлены также в животном и растительном мире. Так, существует зависимость между телосложением и продуктивностью у сельскохозяйственных животных; известна связь между качеством семян и урожайностью культурных растений и т. д. Наличие связей между варьирующими признаками обнаруживается на всех уровнях организации живого. Поэтому естественно стремление использовать эту закономерность в интересах человека, придать ей более или менее точное количественное выражение.

Для описания связей между переменными величинами применяют математическое понятие функции f, которая ставит в соответствие каждому определенному значению независимой переменной X, называемой аргументом, определенное значение зависимой переменной Y: $y=f(x)$. Здесь x — аргумент, а y — соответствующее ему значение функции $f(x)$. Такого рода однозначные зависимости между переменными величинами Y и X называют *функциональными*. Примеров функциональной зависимости между переменными величинами много. Известно, что повышение температуры на 10 °С ускоряет химическую реакцию в два раза, объем куба однозначно определяется по длине одного из его ребер и т. д.

Однако такого рода однозначные, или функциональные, связи между переменными величинами встречаются далеко не всегда. Известно, например, что между ростом и массой тела у человека существует положительная связь: более высокие индивиды имеют обычно и бо́льшую массу тела, чем индивиды низкого роста. То же наблюдается и в отношении качественных признаков: блондины, как правило, имеют голубые глаза, а брюнеты — карие. Однако из этого правила существуют исключения, когда сравнительно низкорослые индивиды оказываются тяжелее высокорослых, и среди населения, хотя и не часто, встречаются кареглазые блондины и голубоглазые брюнеты.

Причиной таких «исключений» является тот факт, что каждый биологический признак представляет собой функцию многих переменных: на него влияют и генетические, и средовые факторы, что и обусловливает варьирование признаков. Поэтому зависимость между биологическими признаками имеет не функциональный, а статистический характер, когда в массе однородных индивидов определенному значению одного признака, рассматриваемого в качестве аргумента, соответствует не одно и то же числовое значение, а целая гамма распределяющихся в вариационный ряд числовых значений другого признака, рассматриваемого в качестве зависимой переменной, или функции. Такого рода зависимость между переменными величинами называется корреляционной или *корреляцией* [1].

Функциональные связи легко обнаружить и измерить на единичных и групповых объектах, однако этого нельзя проделать с корреляционными связями, которые можно изучать только на групповых объектах методами математической статистики. Корреляционная связь между признаками бывает линейной и нелинейной, положительной и отрицательной. Задача корреляционного анализа сводится к установлению направления и формы связи между варьирующими признаками, измерению ее тесноты и, наконец, к проверке достоверности выборочных показателей корреляции.

Зависимость между переменными Y и X можно выразить аналитически (с помощью формул и уравнений) и графически (как геометрическое место точек в системе прямоугольных координат). График корреляционной зависимости строят по уравнению функции $\bar{y}_x = f(x)$ или $\bar{x}_y = f(y)$, которая со времен Гальтона получила название *регрессии*. Здесь $\bar{y}_x$ и $\bar{x}_y$ — средние арифметические, найденные при условии, что X или Y примут некоторые значения x или y. Эти средние называются *условными*. Регрессионному анализу посвящена следующая глава. Здесь же будут рассмотрены параметрические и непараметрические способы анализа линейных и нелинейных статистических связей.

VIII.1. ПАРАМЕТРИЧЕСКИЕ ПОКАЗАТЕЛИ СВЯЗИ

Коэффициент корреляции. Сопряженность между переменными величинами Y и X можно установить, сопоставляя числовые

[1] Этот термин (от лат. correlatio — соотношение, связь) впервые применил Ж. Кювье в труде «Лекции по сравнительной анатомии» (1806). Математическое обоснование метода измерения корреляции было дано в 1846 г. другим французским ученым О. Браве. Обосновывая метод, Браве имел в виду «теорию ошибок в плоскости», перенося закон ошибок Гаусса на случаи двух переменных Y и X в область кристаллографии, которой он занимался. Разработка и применение метода корреляции к измерению зависимости между биологическими признаками были сделаны Ф. Гальтоном и К. Пирсоном. Гальтону принадлежит и введение термина «корреляция» в биометрию (1886).

значения одной из них с соответствующими значениями другой. Если при увеличении одной переменной увеличивается другая, это указывает на *положительную связь* между этими величинами, и, наоборот, когда увеличение одной переменной сопровождается уменьшением значений другой, это указывает на *отрицательную связь*. Подобную взаимосвязь устанавливают при наличии однозначных отношений между переменными Y и X, когда речь идет о приращении или уменьшении функции по заданным значениям аргумента. Иная ситуация наблюдается в случае варьирующих признаков. Здесь приходится исследовать собственно не приращение или уменьшение функции, а сопряженную вариацию (ковариацию), выражая ее в виде взаимно связанных отклонений вариант от их средних $\bar{y}$ и $\bar{x}$.

Ковариация (cov) есть усредненная величина произведений $(x_i - \bar{x})(y_i - \bar{y})$ отклонений каждой пары наблюдений от их средних, т. е. $cov = (1/n)[\Sigma (x_i - \bar{x})(y_i - \bar{y})]$. Очевидно, что величина этого показателя будет в значительной мере зависеть от того, насколько часто в общем ряду произведение $(x_i - \bar{x})(y_i - \bar{y})$ будет иметь один знак — плюс или минус. В первом случае пары вариант должны отклоняться от своих средних в одном направлении (т. е. $x_i > \bar{x}$ и $y_i > \bar{y}$ или $x_i < \bar{x}$ и $y_i < \bar{y}$). В другом случае, если $x_i > \bar{x}$, то $y_i < \bar{y}$ или наоборот. При этом преобладание величин одного знака в принципе способствует большему абсолютному значению коэффициента ковариации, так как величины с разными знаками в сумме дают меньшую абсолютную величину. Среднее значение всех произведений указывает, в какой мере бо́льшим (или меньшим) значениям x_i соответствуют бо́льшие (или меньшие) значения y_i.

Недостаток коэффициента ковариации заключается в том, что этот коэффициент не учитывает случаи, когда коррелируемые признаки выражаются разными единицами измерения. Например, масса тела может коррелировать с его линейными размерами, длина колосьев — с массой содержащихся в них зерен и т. д. Недостаток, присущий ковариации, устраняется, если вместо отклонений $(x_i - \bar{x})(y_i - \bar{y})$ использовать их отношения к средним квадратическим отклонениям s_x и s_y. В результате получается показатель, который называют *эмпирическим коэффициентом корреляции r:*

$$r_{xy} = \frac{\frac{1}{n}\left[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right]}{s_x s_y}. \qquad (144)$$

Коэффициент корреляции можно вычислить, не прибегая к расчету средних квадратических отклонений, что упрощает вычислительную работу, по следующей аналогичной формуле:

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(\bar{y}_i-y)^2}}. \qquad (145)$$

Коэффициент корреляции — отвлеченное число, лежащее в пределах от —1 до +1. При независимом варьировании признаков, когда связь между ними полностью отсутствует, $r=0$. Чем сильнее сопряженность между признаками, тем выше значение коэффициента корреляции. Следовательно, при $|r|>0$ этот показатель характеризует не только наличие, но и степень сопряженности между признаками. При положительной или прямой связи, когда бóльшим значениям одного признака соответствуют бóльшие же значения другого, коэффициент корреляции имеет положительный знак и находится в пределах от 0 до +1, при отрицательной или обратной связи, когда бóльшим значениям одного признака соответствуют меньшие значения другого, коэффициент корреляции сопровождается отрицательным знаком и находится в пределах от 0 до —1.

Коэффициент корреляции нашел широкое применение в практике, но он не является универсальным показателем корреляционных связей, так как способен характеризовать только линейные связи, т. е. выражаемые уравнением линейной регрессии (см. гл. IX). При наличии нелинейной зависимости между варьирующими признаками применяют другие показатели связи, о которых речь пойдет ниже.

Вычисление коэффициента корреляции. Это вычисление производят разными способами и по-разному в зависимости от числа наблюдений (объема выборки). Рассмотрим отдельно специфику вычисления коэффициента корреляции при наличии малочисленных выборок и выборок большого объема.

Малые выборки. При наличии малочисленных выборок коэффициент корреляции вычисляют непосредственно по значениям сопряженных признаков, без предварительной группировки выборочных данных в вариационные ряды. Для этого служат приведенные выше формулы (144) и (145). Более удобными, особенно при наличии многозначных и дробных чисел, которыми выражаются отклонения вариант x_i и y_i от средних $\bar{x}$ и $\bar{y}$, служат следующие рабочие формулы:

$$r_{xy}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}}; \qquad (146)$$

$$r_{xy}=\frac{\sum_{i=1}^{n}x_iy_i-\frac{\sum x_i\sum y_i}{n}}{\sqrt{D_xD_y}}; \qquad (147)$$

$$r_{xy}=\frac{D_x+D_y-D_d}{2\sqrt{D_xD_y}}, \qquad (148$$

где $D_x=\Sigma(x_i-\bar{x})^2=\Sigma x^2{}_i-(\Sigma x_i)^2/n$; $D_y=\Sigma(y_i-\bar{y})^2=\Sigma y^2{}_i-(\Sigma y_i)^2/n$ и $D_d=\Sigma d^2{}_i-(\Sigma d_i)^2/n$. Здесь x_i и y_i — парные варианты сопряженных признаков X и Y; $\bar{x}$ и $\bar{y}$ — средние арифметические; $d=(x_i-y_i)$ — разность между парными вариантами сопряженных признаков X и Y; n — общее число парных наблюдений, или объем димерной выборочной совокупности.

Таблица 9

Масса тела гамадрилов-матерей x_i, кг	Масса тела детенышей y_i, кг	x_iy_i	$x^2{}_i$	$y^2{}_i$
10,0	0,70	7,000	100,00	0,4900
10,8	0,73	7,884	116,64	0,5329
11,3	0,75	8,475	127,69	0,5625
10,0	0,70	7,000	100,00	0,4900
10,1	0,65	6,565	102,01	0,4225
11,1	0,65	7,215	123,21	0,4225
11,3	0,70	7,910	127,69	0,4900
10,2	0,61	6,222	104,04	0,3721
13,5	0,70	9,450	182,25	0,4900
12,3	0,63	7,749	151,29	0,3969
14,5	0,70	10,150	210,25	0,4900
11,0	0,65	7,150	121,00	0,4225
12,0	0,72	8,640	144,00	0,5184
11,8	0,69	8,142	139,24	0,4761
13,4	0,78	10,452	179,56	0,6084
11,4	0,70	7,980	129,96	0,4900
12,0	0,60	7,200	144,00	0,3600
15,6	0,85	13,260	243,36	0,7225
13,0	0,80	10,400	169,00	0,6400
12,1	0,75	9,075	146,41	0,5625
$\Sigma=237,4$	14,06	167,919	2861,60	9,9598

Из этих формул следует, что для вычисления коэффициента корреляции необходимо предварительно рассчитать величины Σx_i, Σy_i, Σx_iy_i, $\Sigma x_i{}^2$ и $\Sigma y_i{}^2$, а также (при использовании формулы (146) еще и Σd_i (обязательно с учетом знаков!) и $\Sigma d_i{}^2$.

Пример 1. Изучали зависимость между массой тела гамадрилов-матерей и их новорожденных детенышей. Под наблюдением находилось 20 обезьян. Результаты наблюдений и их обработка приведены в табл. 96.

Определяем суммы квадратов отклонений (девиаты): $D_x=2861,60-\frac{237,4^2}{20}=43,662$; $D_y=9,9598-\frac{14,06^2}{20}=0,076$.

Подставляем эти величины в формулу (147): $r_{xy}=\frac{167{,}919-(1/20)(237{,}4\cdot 14{,}06)}{\sqrt{43{,}662\cdot 0{,}076}}=\frac{167{,}919-166{,}892}{\sqrt{3{,}318}}=\frac{1{,}027}{1{,}822}=+0{,}564.$

Полученная величина ($r_{xy}=0{,}564$) указывает на наличие положительной средней силы корреляционной связи между массой тела гамадрилов-матерей и массой тела их детенышей.

Пример 2. На основании накопленных в хозяйстве данных о жирномолочности коров и их 12 одновозрастных дочерних особей была составлена следующая выборка (табл. 97).

Таблица 97

Процент жира в молоке					
коров x_i	дочерних особей y_i	x_i^2	y_i^2	$x_i-y_i=d$	d_i^2
3,10	3,65	9,6100	13,3225	+0,55	0,3025
3,17	3,11	10,0489	9,6721	—0,06	0,0036
3,76	3,57	14,1376	12,7449	—0,19	0,0361
3,61	3,61	13,0321	13,0321	±0,00	0,0000
3,27	3,44	10,6929	11,8336	+0,17	0,0289
3,61	3,71	13,0321	13,7641	+0,10	0,0100
3,80	3,61	14,4400	13,0321	—0,19	0,0361
3,65	3,98	13,3225	15,8404	+0,33	0,1089
3,34	3,36	11,1556	11,2896	+0,02	0,0004
3,65	3,89	13,3225	15,1321	+0,24	0,0576
3,45	3,45	11,9025	11,9025	0,00	0,0000
4,05	3,79	16,4025	14,3641	—0,26	0,0676
$\Sigma=42{,}46$	43,17	151,0992	155,9301	0,71	0,6517

Вычислим коэффициент корреляции между жирномолочностью коров и их дочерних особей. Расчет вспомогательных величин, нужных для определения девиат, приведен в табл. 97. Находим значения девиат: $D_x=151{,}0992-(42{,}46)^2/12=151{,}0992-150{,}2376=0{,}8616$; $D_y=155{,}9301-(43{,}17)^2/12=155{,}9301-155{,}3041=0{,}6260$; $D_d=0{,}6517-(0{,}71)^2/12=0{,}6517-0{,}0420=0{,}6097$. Подставляем известные величины в формулу (148):

$$r_{xy}=\frac{0{,}8616+0{,}6260-0{,}6097}{2\sqrt{0{,}8616\cdot 0{,}6260}}=\frac{0{,}8779}{2\sqrt{0{,}53936}}=\frac{0{,}8779}{1{,}4688}=0{,}598.$$

Корреляция между жирномолочностью родительских особей и их потомства у крупного рогатого скота оказалась положительной и довольно высокой.

Эмпирический коэффициент корреляции, как и любой другой выборочный показатель, служит оценкой своего *генерального*

параметра ρ и как величина случайная сопровождается ошибкой:

$$s_r = \sqrt{\frac{1-r^2}{n-2}}. \qquad (149)$$

Отношение выборочного коэффициента корреляции к своей ошибке служит критерием для проверки нулевой гипотезы — предположения о том, что в генеральной совокупности этот показатель равен нулю, т. е. $\rho = 0$. Нулевую гипотезу отвергают на принятом уровне значимости α, если

$$t_\Phi = r\sqrt{\frac{n-2}{1-r^2}} \geqslant t_{st}.$$

Значения критических точек t_{st} для разных уровней значимости и чисел степеней свободы $k = n-2$ приведены в табл. V Приложений.

Так, значимость коэффициента корреляции между массой тела гамадрилов-матерей и их новорожденных детенышей оценивается следующим образом $t_\Phi = 0{,}564\sqrt{\frac{20-2}{1-(0{,}564)^2}} = 0{,}564 \times$ $\times \sqrt{26{,}397} = 0{,}564 \cdot 5{,}138 = 2{,}90$.

В табл. V Приложений для $k = 20-2 = 18$ и $\alpha = 1\%$ находим $t_{st} = 2{,}88$. Нулевую гипотезу отвергают на высоком уровне значимости ($P < 0{,}01$).

Достоверность выборочного коэффициента корреляции можно проверить по специальной таблице, в которой содержатся значения критических точек r_{st} для уровней значимости $\alpha = 5\%$ и $\alpha = 1\%$ с учетом числа степеней свободы $k = n-2$. Так, для $k = 18$ и $\alpha = 1\%$ в табл. XVI Приложений находим $r_{st} = 0{,}56$.

Установлено, что при обработке малочисленных выборок (особенно когда $n < 30$) расчет коэффициента корреляции по приведенным выше формулам дает несколько заниженные оценки генерального параметра ρ. В таких случаях лучшую оценку ρ получают при использовании поправки $\left[1 + \frac{1-r^2}{2(n-3)}\right]$, на которую умножают эмпирический коэффициент корреляции, т. е.

$$r^*_{xy} = r_{xy}\left[1 + \frac{1-r^2}{2(n-3)}\right]. \qquad (150)$$

Применим эту формулу к расчету коэффициента корреляции между жирномолочностью коров и их дочернего потомства (табл. 97). В данном случае объем выборки $n = 12$ и $r_{xy} = 0{,}598$. Отсюда $r^*_{xy} = 0{,}598\left[1 + \frac{1-(0{,}598)^2}{2(12-3)}\right] = 0{,}598 \cdot 1{,}036 = 0{,}620$. Опре-

деляем критерий достоверности: $t_ф = 0{,}620\sqrt{\frac{12-2}{1-(0{,}620)^2}} = 0{,}620\sqrt{16{,}2} = 0{,}620 \cdot 4{,}0125 = 2{,}56$.

В табл. V Приложений для $k = 12-2 = 10$ и $\alpha = 5\%$ находим $t_{st} = 2{,}23$. Так как $t_ф = 2{,}56 > t_{st} = 2{,}23$, нулевую гипотезу отвергают на 5%-ном уровне значимости ($0{,}01 < P < 0{,}05$). Этот вывод подтверждается и при оценке величины $r_{xy}{}^* = 0{,}620$ с помощью табл. XXI Приложений, в которой для $k = 10$ и $\alpha = 5\%$ находим $r_{st} = 0{,}58$.

z-преобразование Фишера. Правильное применение коэффициента корреляции предполагает нормальное распределение двумерной совокупности сопряженных значений случайных величин Y и X. Из математической статистики известно, что при наличии значительной корреляции между переменными величинами X и Y ($r_{xy} > 0{,}5$) выборочное распределение коэффициента корреляции для большого числа малых выборок, взятых из нормально распределяющейся генеральной совокупности, значительно отклоняется от нормальной кривой. Об этом наглядно свидетельствует рис. 24, на котором изображены кривые распределения эмпирического коэффициента корреляции при $n = 12$ для значений генерального параметра $\rho = 0$; 0,4 и 0,8. Видно, что при значениях ρ, приближающихся к единице, кривая распределения эмпирического коэффициента корреляции становится все более асимметричной. Следовательно, выборочный коэффициент корреляции не будет точной оценкой генерального параметра, если он вычислен на малочисленной выборке и его абсолютное значение превышает 0,5.

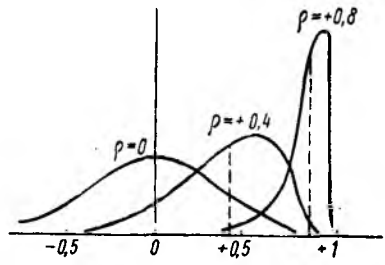


Рис. 24. Кривые распределения эмпирического коэффициента корреляции при $n = 12$ для разных значений генерального параметра (ρ) (по А. К. Митропольскому, 1971)

Учитывая это обстоятельство, Р. Фишер нашел более точный способ оценки генерального параметра по значению выборочного коэффициента корреляции. Этот способ сводится к замене r_{xy} преобразованной величиной z, которая связана с эмпирическим коэффициентом корреляции следующим образом:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} \quad \text{или} \quad z = 1{,}15129\lg\frac{1+r}{1-r}.$$

Распределение величины z является почти неизменным по форме, так как мало зависит от объема выборки и от значения коэффициента корреляции в генеральной совокупности. Если

эмпирический коэффициент корреляции меняет свое значение от -1 до $+1$, то величина z меняет свое значение от $-\infty$ до $+\infty$, а ее распределение быстро приближается к нормальному (рис. 25) со средним значением $\bar{z}=\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}+\frac{\rho}{2(n-1)}\right)$ и дисперсией $\sigma_z^2=\frac{1}{n-3}$.

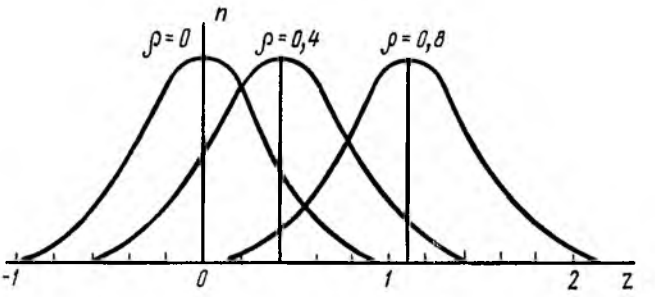


Рис. 25. Распределение величины z при $n=12$ (по А. К. Митропольскому, 1971)

Преобразование коэффициента корреляции в величину z производят по специальной таблице (см. табл. XXII Приложений). В этой таблице содержатся величины z, соответствующие значениям эмпирического коэффициента корреляции r. Критерием достоверности показателя z является следующее отношение:

$$t_z=\frac{z}{s_z}=z\sqrt{n-3}.$$

Этот критерий используют в тех случаях, когда вместо коэффициента корреляции берут соответствующее ему число z. Нулевая гипотеза отвергается на принятом уровне значимости α и числе степеней свободы $k=n-2$. Значения критических точек t_{st} приведены в табл. V Приложений.

Применение z-преобразования позволяет с большей уверенностью оценивать статистическую значимость выборочного коэффициента корреляции, а также и разность между эмпирическими коэффициентами корреляции $r_1-r_2=d_r$, когда в этом возникает необходимость.

Проверим нулевую гипотезу в отношении z-преобразованного коэффициента корреляции между жирномолочностью коров и их дочернего потомства, который с поправкой на малый объем выборки оказался равным $r=0{,}620$. В табл. XXII Приложений для этой величины находится значение $z=0{,}725$, откуда критерий $t_z=0{,}725\sqrt{12-3}=0{,}725\cdot 3{,}0=2{,}18$. Эта величина оказывается

ниже критической точки $t_{st}=2{,}23$ для $k=12-2=10$ и $\alpha=5\%$ (см. табл. V Приложений). Следовательно, отвергнуть нулевую гипотезу на этом уровне значимости нельзя. Этот вывод не согласуется с ранее сделанным заключением о том, что $r=0{,}620$ можно считать величиной достоверной на 5%-ном уровне значимости.

Минимальный объем выборки для точной оценки коэффициента корреляции. Только что рассмотренный пример показывает, с какой осторожностью следует делать заключение о статистической значимости выборочного коэффициента корреляции, вычисленного на малообъемных выборках. Очевидно, предпочтение в таких случаях следует отдавать оценке r по преобразованной величине z.

Можно рассчитать объем выборки для заданного значения коэффициента корреляции, который был бы достаточен для опровержения нулевой гипотезы (если корреляция между признаками Y и X действительно существует). Для этого служит следующая формула:

$$n=\frac{t^2}{z^2}+3, \tag{151}$$

где n — искомый объем выборки; t — величина, заданная по принятому уровню значимости (лучше для $\alpha=1\%$); z — преобразованный эмпирический коэффициент корреляции.

Так, в отношении только что рассмотренного примера для $\alpha=1\%$, которому соответствуют $t=2{,}58$; $r=0{,}598$ и $z=0{,}693$, находим

$$n=\frac{(2{,}58)^2}{(0{,}693)^2}+3=13{,}9+3=16{,}9=17.$$

Это означает, что для окончательного решения вопроса о статистической значимости эмпирического коэффициента корреляции между жирномолочностью коров и их дочернего потомства необходимо увеличить число наблюдений по меньшей мере до $n=17$.

Большие выборки. При наличии многочисленных исходных данных их приходится группировать в вариационные ряды и, построив корреляционную решетку, разносить по ее клеткам (ячейкам) общие частоты сопряженных рядов. Корреляционная решетка, как уже было показано в гл. VII, образуется пересечением строк и столбцов, число которых равно числу групп или классов коррелируемых рядов. Классы располагаются в верхней строке и в первом (слева) столбце корреляционной таблицы, а общие частоты, обозначаемые символом f_{xy}, — в клетках корреляционной решетки, составляющей основную часть корреляционной таблицы.

Классы, помещаемые в верхней строке таблицы, обычно располагаются слева направо в возрастающем порядке, а в первом столбце таблицы — сверху вниз в убывающем порядке. При таком расположении классов вариационных рядов их общие частоты (при наличии положительной связи между признаками X и Y) будут распределяться по клеткам решетки в виде эллипса по диагонали от нижнего левого угла к верхнему правому углу решетки или (при наличии отрицательной связи между признаками) в направлении от верхнего левого угла к нижнему правому углу решетки. Если же частоты f_{xy} распределяются по клеткам корреляционной решетки более или менее равномерно, не образуя фигуры эллипса, это будет указывать на отсутствие корреляции между признаками.

Таблица 98

Объем выборки	Значение K
$50 \geqslant n > 30$	$K = 1 + 3{,}32 \lg n$
$100 \geqslant n > 50$	$K = 5 \lg n$
$200 \geqslant n \geqslant 100$	$K = 7 \lg n$
$300 \geqslant n \geqslant 200$	$K = 8 \lg n$

Распределение частот f_{xy} по клеткам корреляционной решетки дает лишь общее представление о наличии или отсутствии связи между признаками. Судить о тесноте или силе связи, ее направлеиии можно более или менее точно лишь по значению и знаку *коэффициента корреляции*. При вычислении коэффициента корреляции с предварительной группировкой выборочных данных в интервальные вариационные ряды не следует брать слишком широкие классовые интервалы. Грубая группировка гораздо сильнее сказывается на значении коэффициента корреляции, чем это имеет место при вычислении средних величин и показателей вариации.

Учитывая это обстоятельство, авторы известного руководства «Теория статистики» Дж. Юл и М. Кендэл (1960) рекомендуют избирать величину классового интервала не крупнее $^1/_{20}$ вариационного размаха коррелируемых признаков, группировать димерную совокупность наблюдеиий не менее чем в 15—25 классов. Эта ценная рекомендация имеет, однако, один существенный недостаток: она не согласуется с объемом выборки, не учитывает то, что между числом классов и величиной классового интервала λ существует определенное соотношение, в общем виде оно выражается формулой (1), в которой знаменатель K находится в зависимости от объема выборки n. Опыт показал, что в области корреляционного анализа величину K можно поставить в зависимость от объема выборки примерно следующим образом (табл. 98).

Эта таблица позволяет подойти к определению величины классового интервала дифференцированно в зависимости от

объема выборки, что важно в практической работе исследователя, так как довольно часто приходится группировать сравнительно небольшие выборки ($n \leqslant 50$), что не учитывает рекомендация Юла — Кендэла [1].

Как и другие статистические характеристики, вычисляемые с предварительной группировкой исходных данных в вариационные ряды, коэффициент корреляции определяют разными способами, дающими совершенно идентичные результаты.

Способ произведений. Коэффициент корреляции можно вычислить используя основные формулы (144) или (145), внеся в них поправку на повторяемость вариант в димерной совокупности. При этом, упрощая символику, отклонения вариант от их средних обозначим через a, т. е. $a_x = (x_i - \bar{x})$ и $a_y = (y_i - \bar{y})$. Тогда формула (145) с учетом повторяемости отклонений примет следующее выражение:

$$r_{xy} = \frac{\sum_{i=1}^{K} f_{xy} a_x a_y}{\sqrt{\sum f_x a_x^2 \sum f_y a_y^2}}. \qquad (152)$$

Очевидно, что для определения коэффициента корреляции этим способом необходимо предварительно рассчитать средние арифметические $\bar{x}$ и $\bar{y}$, а также величины $\Sigma f_x a^2_x$, $\Sigma f_y a^2_y$ и $\Sigma f_{xy} a_x a_y$.

Пример 3. При проведении осеннего кросса на 10-километровую дистанцию и лыжных гонок спортсменов на дистанцию 5 км были получены результаты, которые сведены в табл. 99. Эти данные, выраженные в минутах, указывают на положительную зависимость между переменными Y и X, где через Y обозначены результаты лыжных гонок, а через X — результаты 10-километрового кросса.

Вычислим коэффициент корреляции для этих данных. Предварительно рассчитаем вспомогательные величины (см. «крылья» таблицы). Находим средние арифметические рядов X и Y: $\bar{x} = \Sigma x f_x / n = 2391/60 = 39{,}85$ мин и $\bar{y} = \Sigma y f_y / n = 1088/60 = 18{,}13$ мин. Умножая квадраты отклонений вариант от их средних на соответствующие частоты, получаем величины $f_x a^2_x$ и $f_y a^2_y$. Например, величина $f_x a^2_x = 26{,}52$ получена следующим об-

[1] По данному вопросу существует и другое мнение. Согласно ему, единственным соображением, которое следует учитывать при построении корреляционной таблицы для вычислительных целей, является возможность выигрыша в объеме вычислений при сохранении их достаточной точности. Этому требованию удовлетворяет условие $15 \leqslant K \leqslant 25$, которое близко к рекомендации Д. Юла и М. Кендэла. При $K > 25$ объем вычислений по корреляционной таблице необоснованно возрастает, не приводя к существенному повышению их точности, при $K < 15$ может значительно уменьшиться точность вычислений. (*Прим. ред.*)

разом: в первой верхней строке корреляционной решетки находятся $f_x=1$ и $a_x=5{,}15$, откуда $f_x a^2_x=1\cdot 5{,}15^2=26{,}52$ и т. д.

Наибольшего внимания и усилий требует расчет значений $f_{xy}a_x a_y$. Эти значения получаются в результате перемножения частот f_{xy} на соответствующие отклонения по рядам Y и X (обязательно с учетом знаков отклонений!). Например, величина $\Sigma f_{xy}a_x a_y=40{,}68$ получена следующим образом:

$$\begin{array}{l} 2\cdot(-1{,}85)\cdot(-2{,}13)=\ 7{,}881 \\ 2\cdot(-2{,}85)\cdot(-2{,}13)=12{,}141 \\ 2\cdot(-4{,}85)\cdot(-2{,}13)=20{,}662 \\ \hline f_{xy}a_x a_y=40{,}683\approx 40{,}68 \end{array}$$

Таблица 9

Y \ X	16	17	18	19	20	21	f_x	xf_x	a_x	$f_x a^2_x$
45						1	1	45	+5,15	26,52
44				1	1		2	88	+4,15	34,45
43			1	2	1	2	6	258	+3,15	59,54
42			1	1	3	2	7	294	+2,15	32,36
41		12	1	3	1		7	287	+1,15	9,28
40		3	4	1			8	320	+0,15	0,18
39		3	8	1			12	468	−0,85	8,67
38	2	4	1	1			8	304	−1,85	27,38
37	2	3					5	185	−2,85	40,61
36		2					2	72	−3,85	29,65
35	2						2	70	−4,85	47,04
f_y	6	17	16	10	6	5	60	2391	—	315,66
yf_y	96	289	288	190	120	105	1088			
a_y	−2,13	−1,13	−0,13	+0,87	+1,87	+2,87	—			
$f_y a^2_y$	27,22	21,71	0,27	7,57	20,98	41,18	118,93			
$f_{xy}a_x a_y$	40,68	26,50	0,21	11,75	27,86	45,20	152,20			

Аналогичным способом рассчитана следующая величина $f_{xy}a_x a_y$:

$$\begin{array}{l} 2\cdot(+1{,}15)\cdot(-1{,}13)=-2{,}5990 \\ 3\cdot(+0{,}15)\cdot(-1{,}13)=0{,}5085 \end{array}\Big\}\ -3{,}1075$$

$$\begin{array}{l} 3\cdot(-0{,}85)\cdot(-1{,}13)=+2{,}8815 \\ 4\cdot(-1{,}85)\cdot(-1{,}13)=+8{,}3620 \\ 3\cdot(2.85)\cdot(-1{,}13)=+9{,}6615 \\ 2\cdot(-3{,}85)\cdot(-1{,}13)=+8{,}7010 \end{array}\Big\}\ +29{,}6060$$

$$f_{xy}a_x a_y=+26{,}4985\approx 26{,}50$$

и так далее, пока не будут определены все значения $f_{xy}a_x a_y$ корреляционной таблицы по столбцам или по ее строкам. Суммируя

значения $f_{xy}a_xa_y$, получаем величину $\Sigma f_{xy}a_xa_y=152{,}20$. Подставляя известные величины в формулу, находим

$$r_{xy}=\frac{152{,}20}{\sqrt{315{,}66\cdot 118{,}93}}=\frac{152{,}20}{193{,}70}=+0{,}786.$$

Это довольно высокий показатель связи между переменными Y и X. Достоверность этого показателя оценивается с помощью критерия Стьюдента, который представляет отношение выборочного коэффициента корреляции к своей ошибке, определяемой по формуле

$$s_r=\frac{\sqrt{1-r^2}}{\sqrt{n}}. \qquad (153)$$

Так, в данном случае $r_{xy}=0{,}786$ и $n=60$. Ошибка $s_r=\frac{\sqrt{1-(0{,}786)^2}}{\sqrt{60}}=0{,}080$. Отсюда $t_ф=0{,}786/0{,}080=9{,}83$. Эта величина значительно превышает $t_{st}=3{,}46$ для $k=58$ и $\alpha=0{,}1\%$ (см. табл. V Приложений), что опровергает H_0-гипотезу на высоком уровне значимости ($P<0{,}001$).

Способ условных средних. При вычислении коэффициента корреляции отклонения вариант («классов») можно находить не только от средних арифметических $\bar{x}$ и $\bar{y}$, но и от условных средних A_x и A_y. При этом способе в числитель формулы (145) вносят поправку и формула приобретает следующий вид:

$$r_{xy}=\frac{\sum_{i=1}^{K} f_{xy}a_xa_y-nb_xb_y}{ns_xs_y}, \qquad (154)$$

где f_{xy} — частоты классов одного и другого рядов распределения; $a_x=(x_i-A_x)/\lambda_x$ и $a_y=(y_i-A_y)/\lambda_y$, т. е. отклонения классов от условных средних, отнесенные к величине классовых интервалов λ; n — общее число парных наблюдений, или объем выборки; $b_x=\Sigma f_xa_x/n$ и $b_y=\Sigma f_ya_y/n$ — условные моменты первого порядка, где f_x — частоты ряда X, а f_y — частоты ряда Y; s_x и s_y — средние квадратические отклонения рядов X и Y; они вычисляются по способу условных средних, но без умножения на величину классового интервала λ и без внесения поправки $n/(n-1)$.

Способ условных средних имеет преимущество перед способом произведений, так как позволяет избегать операции с дробными числами и придавать один и тот же (положительный) знак отклонениям a_x и a_y, что упрощает технику вычислительной работы, особенно при наличии многозначных чисел.

Пример 4. Воспользуемся только что рассмотренными данными из примера 3 и вычислим коэффициент корреляции между результатами осеннего 10-километрового кросса и 5-километро-

вого пробега спортсменов. Как и в предыдущем случае, предварительно рассчитываем вспомогательные величины (табл. 100.

Подсчитав частоты f_x и f_y по каждому ряду, намечаем условные средние A_x и A_y. В качестве этих величин могут быть взяты любые классовые варианты. Удобнее брать в качестве условных средних A_x и A_y значения первых (начальных) классов, тогда все отклонения a_x и a_y получают положительный знак.

Таблица 10

X \ Y	16	17	18	19	20	21	f_x	a_x	$f_x a_x$	$f_x a^2_x$
45						1	1	10	10	100
44				1	1		2	9	18	162
43			1	2	1	2	6	8	48	384
42			1	1	3	2	7	7	49	343
41		2	1	3	1		7	6	42	252
40		3	4	1			8	5	40	200
39		3	8	1			12	4	48	192
38	2	4	1	1			8	3	24	72
37	2	3					5	2	10	20
38		2					2	1	2	2
35	2						2	0	0	0
f_y	6	17	16	10	6	5	60	—	−291	1727
a_y	0	1	2	3	4	5	—			
$f_y a_y$	0	17	32	30	24	25	128			
$f_y a_y^2$	0	17	64	90	96	125	392			
$f_{xy} a_x a_y$	0	59	152	186	176	200	773			

В данном случае намечаем: $A_x = 35$ и $A_y = 16$. От этих условных средних, где $a_x = 0$ и $a_y = 0$, откладываем отклонения классов, как показано в табл. 100. Перемножая частоты рядов на отклонения классов, находим $f_x a_x$ и $f_y a_y$, а также $f_x a_x^2$ и $f_y a_y^2$. Значения $f_{xy} a_x a_y$ рассчитываем так же, как и в предыдущем примере. Так, величина $f_{xy} a_x a_y = 59$ получена следующим образом:

$f_{xy} a_x a_y$:
$2 \cdot 6 \cdot 1 = 12$
$3 \cdot 5 \cdot 1 = 15$
$3 \cdot 4 \cdot 1 = 12$
$4 \cdot 3 \cdot 1 = 12$
$3 \cdot 2 \cdot 1 = 6$
$2 \cdot 1 \cdot 1 = 2$

Сумма = 59 и т. д.

Определяем $b_x = 291/60 = 4{,}850$ и $b_y = 128/60 = 2{,}133$, а также средние квадратические отклонения: $s_x = \sqrt{1727/60 - (4{,}850)^2} = \sqrt{5{,}26} = 2{,}294$ и $s_y = \sqrt{392/60 - (2{,}133)^2} = \sqrt{1{,}984} = 1{,}408$. Подставляем известные величины в формулу (154) и определяем коэффициент корреляции:

$$r_{xy} = \frac{773 - 60 \cdot 4{,}850 \cdot 2{,}133}{60 \cdot 2{,}294 \cdot 1{,}408} = \frac{152{,}30}{193\,680} = +0{,}786.$$

Пример 5. Из Госплемкниги крупного рогатого скота горбатовской породы была извлечена выборка, включающая 100 попарно связанных значений двух признаков — годового удоя первотелок и коров по второму и третьему отелам и их масса тела (табл. 101).

Таблица 101

Масса тела, кг	Удой, кг	Масса тела, кг	Удой, кг	Масса тела, кг	Удой, кг	Масса тела, кг	Удой, кг
327	2325	440	3219	287	1396	360	2696
302	1761	405	1806	337	1819	324	2510
327	2310	323	2803	295	2523	245	2615
294	2035	411	2385	339	2133	345	2715
410	2172	434	2826	400	1918	368	2103
342	2277	352	1832	306	1302	397	2023
409	2784	295	2413	335	2372	405	2162
311	1523	369	2625	341	2688	368	2403
297	1838	444	2614	343	2131	418	2483
364	1984	319	2297	411	2901	371	2016
377	1775	303	1946	316	2151	382	2715
358	2700	352	2278	314	1734	410	2878
284	2241	344	2111	396	2537	443	2431
314	1954	361	3082	339	1979	385	3048
352	2046	303	2478	332	2142	285	1791
387	2323	328	2801	328	1917	321	2554
375	1710	344	2248	314	1873	351	2281
311	1868	284	1085	409	2630	331	2292
332	2166	295	2293	367	2100	355	2340
262	1384	360	2282	396	2493	396	2609
333	2288	244	1736	384	2632	390	2499
381	2249	279	1446	356	2043	426	3013
320	1520	303	2376	446	2358	430	2933
295	2389	329	1937	338	2309	386	2682
345	2012	323	1999	300	1442	331	1689

Обозначим удой через X, а массу тела коров — через Y и вычислим коэффициент корреляции между этими признаками. В совокупности значений признаков находим минимальные и максимальные варианты: $x_{min} = 1085$ и $x_{max} = 3219$; $y_{min} = 244$ и

$y_{max}=446$. В этих границах намечаем классовые интервалы, принимая $K=7$:

$$\lambda_x=\frac{3219-1085}{7\lg 100}=\frac{2134}{14}=152{,}4\approx 152 \quad \text{и} \quad \lambda_y=\frac{446-244}{14}\approx 14.$$

Затем определяем нижние границы первых классов: $x_н=1085-152/2=1009$ и $y_н=244-14/2=237$. Отсюда получаются следующие классовые интервалы — по удою (Y): 1009 — 1161 — 1313 — 1465 — 1617 — 1769 — 1921 — 2073 — 2225—2377—2529 — 2681 — 2833 — 2985 — 3137 — 3289; по массе коров (X): 237 — 251 — 265 — 279 — 293 — 307 — 321 — 335 — 349 — 363 — 377 — 391 — 405 — 419 — 433 — 447. В каждом ряду образуется по 15 классов, что отвечает рекомендации Юла — Кендэла и гарантирует доста-

Масса коров X / Удой Y	Центральные значения классовых интервалов	237—250	251—264	265—278	279—292	293—306	307—320	321—334	335—348
		244	258	272	286	300	314	328	342
3137—3288	3213								
2985—3136	3061								
2833—2984	2909								
2681—2832	2757							2	2
2529—2680	2605	1						1	
2377—2528	2453					4		1	
2225—2376	2301				1	2	1	4	4
2073—2224	2149						1	2	3
1921—2072	1997					2	1	2	2
1769—1920	1845				1	1	2	1	1
1617—1768	1693	1				1	1	1	
1465—1616	1541						2		
1313—1464	1389		1		2	1			
1161—1312	1237					1			
1009—1160	1085				1				
f_x	2	1	0	5	12	8	14	12	10
a_x	0	1	2	3	4	5	6	7	8
$f_x a_x$	0	1	0	15	48	40	84	84	80
$f_x a_x^2$	0	1	0	45	192	200	50	588	640
$f_{yx} a_y a_x$	0	2	0	51	304	205	648	644	672
$\bar{y}_x$	2149,0	1389,0	—	1601,8	2047,7	1864,0	2257,6	2250,3	2361,8

точную точность коэффициента корреляции, вычисляемого для этой выборки.

Разграничиваем классы, уменьшая их верхние границы на единицу, строим (по числу классов) корреляционную таблицу и по ее клеткам разносим все 100 вариант данной выборочной совокупности. В результате получаем распределение общих частот (f_{xy}) двух сопряженных рядов распределения X и Y (табл. 102).

«Крылья» табл. 102 содержат расчет вспомогательных величин, нужных для вычисления коэффициента корреляции способом условных средних. В качестве последних приняты наименьшие классовые варианты, т. е. срединные значения классовых интервалов (A_x=1085 и A_y=244), для того чтобы все отклоне-

Таблица 102

349— —362	363— —376	377— —390	391— —404	405— —418	419— —432	433— —446	f_y	a_y	$f_y a_y$	$f_y a_y^2$	$\overline{x}_y$
356	370	384	398	412	426	440					
						1	1	14	14	196	440,0
1		1			1		3	13	39	507	388,7
				2	1		3	12	36	432	416,7
2		2		1		1	10	11	110	1210	367,2
	1	1	2	1		1	8	10	80	800	341,3
	1	1	1	2		1	11	9	99	891	358,5
4		2				1	19	8	152	1216	342,7
	2			2			10	7	70	490	357,0
2	2		1				12	6	72	432	342,0
1		1	1	1			10	5	50	250	343,4
	1						5	4	20	80	311,2
							2	3	6	18	314,0
							4	2	8	16	282,5
							1	1	1	1	300,0
							1	0	0	0	286,0
7	8	5	9	2	5	100	—		757		6539
9	10	11	12	13	14	—	$\overline{y} = A + \lambda \frac{f_x a_x}{n} = 1085 +$				
63	80	55	108	26	70	754					
567	800	605	1296	338	980	6756	$+ 152 \frac{757}{100} = 2235{,}64$ кг;				
441	750	440	984	325	728	6194					
2149,0	2510,0	2301,0	2469,9	2985,0	2665,8		$\overline{x} = 244 + 14 \frac{754}{100} = 349{,}56$ кг				

ния классов от условиых средних (где эти отклонеиия равны иулю) имели положительиый знак.

Значеиия $f_{xy}a_xa_y$, как и в предыдущих случаях, рассчитаиы перемиожеиием отклоиеиий классовых вариант от условиых средиих иа соответствующие частоты f_{xy}. Например, $f_{xy}a_xa_y=51$, что иаходится внизу табл. 102, получена так:

$f_{xy}a_xa_y$:
$1\cdot8\cdot3=24$
$1\cdot5\cdot3=15$
$2\cdot2\cdot3=12$
$1\cdot0\cdot3=0$

Сумма=51 и т. д.

Определяем условиые момеиты первого порядка: $b_x=757/100=7{,}57$ и $b_y=754/100=7{,}54$, а также средиие квадратические отклоиеиия: $s_x=\sqrt{6539/100-(7{,}57)^2}=2{,}843$ и $s_y=\sqrt{6756/100-(7{,}54)^2}=3{,}272$. Как и в предыдущих примерах, s_x и s_y вычисляют без виесения поправочного коэффициеита $n/(n-1)$ и ие умиожают на λ_x и λ_y, так как эти величины входят и в числитель и в зиамеиатель формулы, взаимно сокращаясь.

Определив величииы $\Sigma f_{xy}a_xa_y$, b_x и b_y, а также s_x и s_y с учетом объема выборки $n=100$, иаходим значеиие коэффициента корреляции между этими призиаками:

$$r_{xy}=\frac{6194-100\cdot7{,}57\cdot7{,}54}{100\cdot2{,}843\cdot3{,}272}=\frac{486{,}22}{930{,}23}=0{,}523.$$

Полученная величина r_{xy} указывает на иаличие положительной средней силы связи между массой тела данной группы коров и их годовым удоем.

Применим t-критерий для проверки H_0-гипотезы в отношении величиины $r_{xy}=0{,}523$. Ошибка этого показателя $s_r=\frac{\sqrt{1-(0{,}523)^2}}{\sqrt{100}}=\frac{\sqrt{0{,}7265}}{10}=0{,}085$. Отсюда $t_\Phi=\frac{0{,}523}{0{,}085}=6{,}14$.

В табл. V Приложеиий для $k=n-2=100-2=98$ и уровня значимости $\alpha=0{,}1\%$ иаходим $t_{st}=3{,}37$. Поскольку $t_\Phi=6{,}14>3{,}37$, имеются достаточиые осиоваиия для неприиятия нулевой гипотезы на высоком уровне значимости ($P<0{,}001$).

Оценка разности между коэффициентами корреляции. При сравнении коэффициентов корреляции двух независимых выборок нулевая гипотеза сводится к предположению о том, что в генеральной совокупиости разиица между этими показателями равна нулю. Иными словами, следует исходить из предположеиия, что разиица, наблюдаемая между сравниваемыми эмпирическими коэффициентами корреляции, возиикла случайио.

Для проверкн этой (нулевой) гипотезы служит t-критерий Стьюдента, т. е. отношеиие разности между эмпирическими ко-

эффициентами корреляции r_1 и r_2 к своей статистической ошибке, определяемой по формуле

$$s_d = \sqrt{s_{r_1}^2 + s_{r_2}^2}, \qquad (155)$$

где s_{r_1} и s_{r_2} — ошибки сравниваемых коэффициентов корреляции.

Нулевая гипотеза опровергается при условии, что $t_\Phi = \frac{r_1 - r_2}{s_d} \geqslant t_{st}$ для принятого уровня значимости α и числа степеней свободы $k = (n_1 - 2)(n_2 - 2) = n_1 + n_2 - 4$.

Пример 6. Сравнить коэффициенты $r_1 = 0{,}762$ и $r_2 = 0{,}603$, вычисленные на независимых выборках $n_1 = 80$ и $n_2 = 86$. Разность между этими показателями $d_{(r_1 - r_2)} = 0{,}762 - 0{,}603 = 0{,}159$. Чтобы оценить достоверность этой разности, нужно рассчитать ее статистическую ошибку. Предварительно находим квадраты ошибок для каждого коэффициента корреляции:

$$s_{r_1}^2 = \frac{1 - (0{,}76)^2}{80 - 2} = \frac{0{,}4194}{28} = 0{,}0054;$$

$$s_{r_2}^2 = \frac{1 - (0{,}603)^2}{86 - 2} = \frac{0{,}6364}{84} = 0{,}0076.$$

Переходим к определению ошибки разности ($d_r = r_1 - r_2$). $s_d = \sqrt{0{,}0054 + 0{,}0076} = 0{,}114$. Критерий достоверности $t_\Phi = 0{,}159/0{,}114 = 1{,}39$. Эта величина не превосходит критической точки $t_{st} = 1{,}96$ для $k = 80 + 86 - 4 = 162$ и $\alpha = 5\%$ (см. табл. V Приложений), что не позволяет отвергнуть нулевую гипотезу.

Известно, что более точную оценку достоверности коэффициента корреляции получают при переводе r_{xy} в число z. Это связано с особенностями распределения r_{xy}, которое не всегда следует нормальному закону (см. рис. 25). Не является исключением и оценка разности между выборочными коэффициентами корреляции r_1 и r_2, особенно в тех случаях, когда последние вычислены на выборках сравнительно небольшого объема ($n < 100$) и по своему абсолютному значению значительно превышают 0,50.

Разность $z_1 - z_2 = d$ оценивают с помощью t-критерия Стьюдента, который строят по отношению этой разности к своей ошибке, вычисляемой по формуле

$$s_d = \sqrt{\frac{1}{n - 3} + \frac{1}{n_2 - 3}}. \qquad (156)$$

Нулевую гипотезу отвергают, если $t_z = \frac{d_z}{s_d} \geqslant t_{st}$ для $k = n_1 + n_2 - 4$ и принятого уровня значимости α.

Так, оценку разности между $r_1 = 0{,}762$ и $r_2 = 0{,}603$ с переводом этих значений в числа z_1 и z_2 производят следующим образом.

В табл. XXII Приложений для $r_1=0{,}762$ и $r_2=0{,}603$ находим числа $z_1=0{,}996$ и $z_2=0{,}693$, откуда $t_z=\frac{0{,}996-0{,}693}{\sqrt{1/80+1/86}}=$ $=\frac{0{,}303}{\sqrt{0{,}0241}}=\frac{0{,}303}{0{,}155}=3{,}95$. Эта величина не превосходит критическую точку $t_{st}=1{,}96$ для $\alpha=5\%$ и $k=162$. Следовательно, нулевая гипотеза остается в силе; разница между сравниваемыми коэффициентами корреляции оказывается статистически недостоверной.

Корреляционное отношение. Как было показано, коэффициент корреляции служит для измерения только линейной связи. Для измерения нелинейной зависимости между переменными X и Y используют предложенный К. Пирсоном показатель, который называют *корреляционным отношением.* Если коэффициент корреляции характеризует связь между признаками с точки зрения прямой пропорциональности, то корреляционное отношение, обозначаемое греческой буквой η (эта), описывает ее двусторонне. Поясним это на следующем примере. Возьмем ряд сопряженных (парных) значений двух переменных величин X и Y:

Значения X	2	4	6	8	4	4	2	6
Значения Y	4	8	8	7	6	10	6	12

Ранжируем эти значения по X:

X	2	2	4	4	6	6	6	8
Y	4	6	4	8	10	8	12	7

Видно, что некоторые значения X повторяются, что позволяет распределить эту выборку следующим образом:

X	2	4	6	8
$\bar{y}_x$	5	6	10	7

Здесь $\bar{y}_x$ — частные или групповые средние из соответствующих значений переменной Y. Например, значению $x_i=2$ соответствует $\bar{y}_x=(4+6):2=5$; значению $x_i=6$ соответствует $\bar{y}_x=(10+8+12):3=10$ и т. д.

Если же данную совокупность ранжировать по Y, получается следующий результат:

Y	4	4	6	7	8	8	10	12
X	2	4	2	8	6	4	6	6

Этот ряд состоит ие из четырех, как в первом случае, а из шести групп, представленных значениями переменной $Y=4$; 6; 7; 8; 10; 12, которым соответствуют следующие частные средние:

Y	4	6	7	8	10	12
$\bar{x}_y$	3	2	8	5	6	6

Конструкция корреляционного отношения предполагает сопоставление двух видов вариации: изменчивости отдельных наблю-

дений по отношению к частным средним и вариации самих частных средних по сравнению с общей средней величиной. Чем меньшую часть составит первый компонент по отношению ко второму, тем теснота связи окажется большей. В пределе, когда никакой вариации отдельных значений признака возле частных средних не будет наблюдаться, теснота связи окажется предельно большой. Аналогичным образом, при отсутствии изменчивости частных средних теснота связи окажется минимальной. Так как это соотношение вариации может быть рассмотрено для каждого из двух признаков, получается два показателя тесноты связи.

Таким образом, связь между переменными случайными величинами X и Y выражается по-разному в зависимости от того, по значениям какой величины ранжируется совокупность. Этот пример объясняет, почему корреляционное отношение характеризует связь между признаками X и Y двусторонне, т. е. Y по X и X по Y; отсюда два коэффициента этого показателя: h_{yx} и h_{xy}. Коэффициент корреляции, как и корреляционное отношение,— величина относительная. Но в отличие от коэффициента корреляции корреляционное отношение всегда является величиной положительной, способной принимать значения от 0 до 1. Коэффициент корреляции — равнозначная мера для обоих корреляционно связанных признаков X и Y, тогда как коэффициенты корреляционного отношения обычно не равны друг другу, т. е. $h_{xy} \neq h_{yx}$. Равенство между этими показателями осуществимо только при строго линейной зависимости между признаками. Корреляционное отношение является универсальным показателем: оно позволяет характеризовать любую форму корреляционной связи — и линейную, и нелинейную.

Коэффициенты корреляционного отношения h_{xy} и h_{yx} определяют рассмотренными выше способами, т. е. способом произведений и способом условных средних.

Способ произведений. Коэффициенты корреляционного отношения Y по X и X по Y определяют по следующим формулам:

$$h_{yx}=\frac{s^2_{yx}}{s^2_y} \quad \text{и} \quad h_{xy}=\sqrt{\frac{s^2_{xy}}{s^2_x}}, \tag{157}$$

где $s^2_{yx}=\frac{1}{n}\sum_{l=1}^{K} f_x(\bar{y}_x-\bar{y})^2$ и $s^2_{xy}=\frac{1}{n}\sum_{l=1}^{K} f_y(\bar{x}_y-\bar{x})^2$ — групповые, а $s^2_y=\frac{1}{n}\sum_{l=1}^{K} f_y(y_l-\bar{y})^2$ и $s^2_x=\frac{1}{n}\sum_{l=1}^{K} f_x(x_l-\bar{x})^2$ — общие дисперсии. Нетрудно заметить, что величина n входит и в числитель, и в знаменатель. А поскольку она взаимно сокращается, коэффициенты корреляционного отношения можно представить в

виде корня из отношений групповых девиат к общим девиатам, т. е.

$$h_{yx}=\sqrt{\frac{\sum f_x(\bar{y}_x-\bar{y})^2}{\sum f_y(y_l-\bar{y})^2}} \quad \text{и} \quad h_{xy}=\sqrt{\frac{\sum f_y(\bar{x}_y-\bar{x})^2}{\sum f_x(x_l-\bar{x})^2}}. \qquad (158)$$

Здесь $\bar{y}$ и $\bar{x}$ — общие, а $\bar{y}_x$ и $\bar{x}_y$ — групповые средние арифметические; f_y — частоты ряда Y, а f_x — частоты ряда X.

Следовательно, чтобы вычислить корреляционное отношение Y по X или X по Y описываемым способом, необходимо: 1) сгруппировать первичные данные в форме корреляционной таблицы; 2) определить общие ($\bar{y}$ и $\bar{x}$) и групповые ($\bar{y}_x$ и $\bar{x}_y$) средние арифметические; 3) возвести в квадрат отклонения групповых средних от общей средней данного ряда распределения, умножить на соответствующие частоты f_i и результаты сложить; 4) умножить суммы квадратов отклонений классовых вариант от их средних на частоты этих отклонений и результаты сложить; 5) подставить полученные данные в формулу (158) и рассчитать корреляционное отношение Y по X или X по Y, а по необходимости и оба эти показателя.

Пример 7. Определить корреляционное отношение годового удоя Y коров горбатовской породы по массе их тела X. Сначала находим среднюю арифметическую годового удоя коров $\bar{y}$, а затем и групповые средние $\bar{y}_x$. Необходимые данные помещены в табл. 103: $\bar{y}=A+\lambda_x\frac{\sum f_x a_x}{n}=1085+153\frac{757}{100}=2235{,}64$ кг.

Групповые средние $\bar{y}_x$ представляют суммы произведений груп-

Y \ X	244	258	272	286	300	314	328	342	356
3213									
3061									1
2909									
2757							2	2	2
2605	1						1		
2453					4		1		
3201				1	2	1	4	4	4
2149						1	2	3	
1997					2	1	2	2	2
1845				1	1	2	1	1	1
1693	1				1	1	1		
1541						2			
1389		1		2	1				
1237					1				
1085				1					
f_x	2	1	0	5	12	8	14	12	10
$\bar{y}_x$	2149,00	1389,00	0	1601,80	2047,67	1864,00	2257,57	50,33	2361,80
$\bar{y}_x-\bar{y}$	86,64	846,64	0	633,84	187,97	371,64	21,93	14,69	126,16
$f_x(\bar{y}_x-\bar{y})^2$	15013,0	716799,3	0	2008765,7	423992,7	1104930,3	6732,9	2589,6	159163,5

повых частот f_{xy} ряда X на соответствующие срединные значения классовых интервалов, отнесенные к сумме частот данного интервала по другому ряду f_y. Например, частная средняя $\bar{y}_x = 2149{,}00$, что находится в первом столбце табл. 103, получена умножением частот f_{xy}, расположенных в этом столбце корреляционной решетки, на соответствующие срединные значения классовых интервалов (ряд X) с последующим делением суммы на $f_x = 2$, т. е. $\bar{y}_x = (1 \cdot 1693 + 1 \cdot 2605)/2 = 2149{,}00$ и т. д.

Определив групповые средние, находим разности между ними и общей средней $\bar{y}$ данного ряда, а также разности между отдельными классовыми вариантами y_i и общей средней $\bar{y}$, которые возводим в квадрат и умножаем на соответствующие частоты рядов распределения. Полученные результаты суммируем. Описанные операции помещены в табл. 103.

Подставляя найденные величины $\Sigma f_x(\bar{y}_x - \bar{y})^2$ и $\Sigma f_y(y_i - \bar{y})^2$ в формулу (158), определяем корреляционное отношение годового удоя коров Y по массе их тела X:

$$h_{yx} = \sqrt{\frac{7656205{,}8}{18679828{,}0}} = \sqrt{0{,}410} = 0{,}640.$$

Таким же образом рассчитываем и корреляционное отношение X по Y, т. е. массы тела коров по их годовому удою: $h_{xy} = 0{,}581$ (читателю предлагается рассчитать эту величину).

Способ условных средних. Определяя коэффициенты корреляционного отношения по формулам (157), отклонения классовых вариант x_i и y_i можно брать не только от средних

Таблица 103

370	384	398	412	426	440	f_y	$y_i - \bar{y}$	$f_y(y_i - \bar{y})^2$
					1	1	977	954 529
	1			1		3	825	2 041 875
			2	1		3	673	1 358 787
	2		1		1	10	521	2 714 410
1	1	2	1		1	8	369	1 089 288
1	1	1	2		1	11	217	517 979
	2				1	19	65	80 275
2						10	87	75 690
2			2			12	239	685 452
		1				10	391	1 528 810
1	1	1	1			5	543	1 474 245
						2	695	966 050
						4	847	2 869 636
						1	999	998 001
						1	1151	1 324 801
7	8	5	9	2	5	100	2	18 679 828
2149,00	2510,00	2301,00	2469,89	2985,00	2665,80	—		
86,64	274,36	65,36	234,25	749,36	430,16	—		
52545,4	602187,3	21359,6	493857,6	1123080,8	925188,1	7656205,8		

Таблица 104

Y \ X	244	258	272	286	300	314	328	342	356	370	384	398	412	426	440	f_y	a_y	$f_y a_y$	$f_y a_y^2$	$\frac{f_{xy}}{a_x}$	$f_{xy} a_x^2 / f_y$
3213															1		+7	+7	49	+7	49,00
3061									1		1		2	1		3	+6	+18	108	+10	33,33
2909											2		1	1		3	+5	+15	75	+16	85,33
2757							2	2	2		1	2	1			10	+4	+40	160	+18	36,13
2605	1				4		1				1	1	2		1	8	+3	+24	72	+17	36,13
2453				1	2	1	1				2				1	11	+2	+22	44	+13	15,36
2301						1	4	4	4						1	19	+1	+19	19	+1	0,05
2149					2	1	2	3		2			2		1	10	0	0	0	+10	10,00
1997				1	1	2	2	2	2	2		1				12	−1	−12	12	0	0,00
1845					1	1	1	1	1		1	1	1			10	−2	−20	40	+1	0,10
1693	1					2	1			1						5	−3	−15	45	−11	24,20
1541				2	1											2	−4	−8	32	−4	8,00
1389		1			1											4	−5	−20	100	−17	72,25
1237																1	−6	−6	36	−3	9,00
1085				1												1	−7	−7	49	−4	16,00
f_x	2	1	0	5	12	8	14	12	10	7	8	5	9	2	5	100	—	+57	841	—	391,15
a_x	−7	−6	−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5	+6	+7	—					
$f_x a_x$	−14	−6	0	−20	−36	−16	−14	0	+10	+14	+24	+20	+45	+12	+35	+54					
$f_x a_x^2$	98	36	0	80	108	32	14	0	10	28	72	80	225	72	245	1100					
$f_{xy} a_y$	0	−5	0	−18	−8	−15	+10	+8	+14	0	+19	+5	+19	+11	+17	—					
$(f_{xy} \times \times a_y)^2 / f_x$	0	25,0	0	64,8	5,3	28,1	7,2	5,3	19,6	0	45,1	5,0	40,1	60,5	57,8	363,8					

арифметических x и y, но и от условных средних $\bar{x}$ и $\bar{y}$. В таких случаях групповые и общие девиаты рассчитывают по формулам

$$D_{yx}=\sum\frac{(f_{xy}a_y)^2}{f_x}-H_y \quad \text{и} \quad D_{xy}=\sum\frac{(f_{xy}a_x)^2}{f_y}-H_x, \quad \text{а также,}$$

$$D_y=\sum f_y a_y^2-H_y \text{ и } D_x=\sum f_x a_x^2-H_x, \text{ где } H_y=\frac{(\sum f_y a_y)^2}{n} \text{ и}$$

$$H_x=\frac{(\sum f_x a_x)^2}{n}.$$

В развернутом виде формулы (157) выглядят следующим образом:

$$h_{yx}=\sqrt{\left[\sum\frac{(f_{xy}a_x)^2}{f_x}-\frac{(\sum f_y a_y)^2}{n}\right]\Big/ D_y};$$

$$h_{xy}=\sqrt{\left[\sum\frac{(f_{xy}a_x)^2}{f_y}-\frac{(\sum f_x a_x)^2}{n}\right]\Big/ D_x}. \qquad (159)$$

В этих формулах $a_y=(y_i-A_y)/\lambda_y$ и $a_x=(x_i-A_x)/\lambda_x$ — отклонения классов от условных средних, сокращенные на величину классовых интервалов; значения a_y и a_x выражаются числами натурального ряда: 0, 1, 2, 3, 4, ... Остальные символы объяснены выше.

Пример 8. Рассчитать способом условных средних корреляционное отношение между удоем Y и массой тела X и между массой тела и годовым удоем коров горбатовской породы. Расчет вспомогательных величин, необходимых для вычисления коэффициентов корреляционного отношения h_{xy} и h_{yx}, приведен в табл. 104.

Значения $f_{xy}a_y$ и $f_{xy}a_x$ рассчитывают так же, как и при вычислении коэффициента корреляции. Например, величина $f_{xy}a_x=-18$ (табл. 104) получена перемножением частот f_{xy} на соответствующие отклонения a_x ряда X, а именно:

$1\cdot(+1)=+1$
$1\cdot(-2)=-2$
$2\cdot(-5)=-10$
$1\cdot(-7)=-7$

Сумма $=-18$ и т. д.

Остальные действия понятны из табл. 104.

Закончив расчет вспомогательных величин, переходим к определению общих и групповых девиат: $D_y=1100-54^2/100=1100-29{,}16=1070{,}846$; $D_x=841-54^2/100=841-32{,}49=808{,}51$; $D_{yx}=391{,}15-29{,}16=361{,}99$; $D_{xy}=363{,}80-32{,}49=331{,}31$. Подставляем эти величины в формулы:

$$h_{yx}=\sqrt{\frac{361{,}99}{1070{,}84}}=\sqrt{0{,}338}=0{,}581; \qquad h_{xy}=\sqrt{\frac{331{,}31}{808{,}51}}=$$

$=\sqrt{0{,}410}=0{,}640$. Получился такой же результат, что и выше.

Сравнивая способ произведений со способом условных средних, нельзя не заметить преимущество первого способа, особенно в тех случаях, когда приходится иметь дело с многозначными числами. Как и другие выборочные показатели, корреляционное отношение является оценкой своего генерального параметра и, как величина случайная, сопровождается ошибкой, определяемой по формуле

$$s_h = \frac{\sqrt{1-h^2}}{\sqrt{n-2}}. \qquad (160)$$

Достоверность оценки корреляционного отношения можно проверить по t-критерию Стьюдента или F-критерию Фишера. H_0-гипотеза исходит из предположения, что генеральный параметр равен нулю. Так, в приведенном выше примере $t_\Phi = \frac{0{,}581\sqrt{100-2}}{\sqrt{1-(0{,}581)^2}} = \frac{5{,}752}{\sqrt{0{,}662}} = 7{,}07 > 3{,}37$ для $k = n-2 = 98$ и $\alpha = 0{,}1\%$ (см. табл. V Приложений), что позволяет отвергнуть H_0-гипотезу на 0,1%-ном уровне значимости ($P < 0{,}001$).

К такому же выводу приводит и проверка H_0-гипотезы по F-критерию Фишера: $F_\Phi = \frac{h^2}{1-h^2}\frac{N-a}{a-2}$, где N — объем выборки, a — число классов вариационного ряда. Именно: $F_\Phi = \frac{(0{,}581)^2}{1-(0{,}581)^2}\frac{100-15}{15-2} = \frac{28{,}73}{8{,}61} = 3{,}34 > 2{,}42$ $k_1 = a-2 = 15-2 = 13$ (см. табл. VI Приложений по горизонтали); $k_2 = N-a = 100-15 = 85$ (см. в той же таблице первый столбец по вертикали). Нулевую гипотезу отвергают на 0,1%-ном уровне значимости ($P < 0{,}001$). Следовательно, можно считать доказанным, что между годовым удоем Y и массой тела X коров горбатовской породы существует положительная связь.

Коэффициенты детерминации. Для истолкования значений, принимаемых показателями тесноты корреляционной связи, используют так называемые *коэффициенты детерминации*, которые показывают, какая доля вариации одного признака зависит от варьирования другого признака. При наличии линейной связи коэффициентом детерминации служит квадрат коэффициента корреляции r^2_{xy}, а при нелинейной зависимости между признаками Y и X — квадрат корреляционного отношения h^2_{yx}. Так, коэффициент детерминации между массой тела коров X и их годовым удоем Y составляет $r^2_{xy} = (0{,}523)^2 = 0{,}274$, или 27,4%. Это означает, что лишь 27,4% вариации признака X определяется варьированием признака Y.

Корреляционное отношение является универсальным показателем корреляционных связей, поэтому в качестве коэффициента детерминации обычно применяют квадрат корреляционного отно-

шения. Именно корреляционное отношение между массой тела коров X и их годовым удоем Y составляет $h_{yx}=0{,}581$ (см. выше). Отсюда $h^2_{yx}=(0{,}581)^2=0{,}338$, или 33,8%.

Коэффициенты детерминации дают основание построить следующую примерную шкалу, позволяющую судить о тесноте связи между признаками: при $r=0{,}5\div0{,}6$ связь считается средней; $r<0{,}5$ указывает на слабую связь и лишь при $r\geqslant0{,}7$ можно судить о сильной связи, когда около 50% вариации признака Y зависит от вариации признака X. Разумеется, шкала эта весьма условна, но она необходима при сравнительной оценке показателей корреляционных связей.

Можно показать, что коэффициенты детерминации имеют прямое отношение к показателям силы влияния факторов на результативный признак. Это особенно хорошо видно на примере вычисления коэффициента детерминации $h^2_{yx}=(0{,}581)^2=0{,}338$, или 33,8% и показателя силы влияния, определяемого по методу Плохинского. Так, если массу тела коров X рассматривать как фактор, воздействующий на их годовой удой Y, то сила влияния этого фактора на результативный признак X определяется следующим образом: используя данные табл. 104, находим $H=54^2/100=29{,}16$. Девиаты: $D_y=\sum f_y a_y^2-H=1100-29{,}16=1070{,}84$; $D_x=\dfrac{(\sum f_{x_i}a_y)^2}{f_x}-H=391{,}15-29{,}16=361{,}99$; $D_e=D_y-D_x=708{,}85$. Отсюда показатель силы влияния (по Плохинскому) $h^2_{yx}=\dfrac{D_x}{D_y}=\dfrac{391{,}15}{1070{,}84}=0{,}338$, или 33,8%, т. е. оба показателя равны друг другу.

Тот же показатель, определяемый по методу Снедекора (читателю предлагается вычислить его), оказывается равным 0,247, или 24,7%. Эта величина оказалась близкой к квадрату коэффициента корреляции ($r^2_{yx}=0{,}274$). Как и следовало ожидать, показатель силы влияния, вычисленный по методу Плохинского, оказался выше, чем тот же показатель, вычисленный по методу Снедекора. Преимущество метода Плохинского заключается в его универсальности и в простоте вычисления по сравнению с показателем силы влияния, определяемым по методу Снедекора [1].

Оценка формы связи. При строго линейной зависимости между переменными величинами Y и X осуществляется равенство $h_{yx}=h_{xy}$. В таких случаях коэффициенты корреляционного отношения совпадают со значением коэффициента корреляции. Совпадут при этом по своему значению и коэффициенты детерминации, т. е. $h^2_{yx}=r^2_{xy}$. Следовательно, по разности между этими

[1] Следует, впрочем, помнить о том, что показатель силы влияния фактора конструкции Плохинского отличается смещениостью своих оценок по отношению к генеральному параметру. (*Прим. ред.*)

величинами можно судить о форме корреляционной зависимости между переменными Y и X:

$$\gamma = h^2 - r^2. \qquad (161)$$

Очевидно, при линейной связи между переменными Y и X показатель γ будет равен нулю; если же связь между переменными Y и X нелинейна, то $\gamma > 0$.

Показатель γ является *оценкой генерального параметра* и, как величина случайная, нуждается в проверке достоверности. При этом исходят из предположения (H_0) о том, что связь между величинами Y и X линейна. Проверить эту гипотезу позволяет F-критерий Фишера:

$$F = \frac{\gamma}{1 - h^2} \frac{N - a}{a - 2},$$

где a — численность групп, или классов вариационного ряда; N — объем выборки. Нулевую гипотезу отвергают, если $F_\Phi > F_{st}$ для $k_1 = a - 2$ (находят по горизонтали табл. VI Приложений), $k_2 = N - a$ (находят в первом столбце той же таблицы) и принятого уровня значимости α.

Применим F-критерий для проверки гипотезы о линейной зависимости массы тела X от годового удоя Y коров горбатовской породы. Исходные данные: $N = 100$; $h_{xy} = 0{,}581$; $h^2_{yx} = 0{,}338$, число классов вариационного ряда $n = 15$. Отсюда

$$F_\Phi = \frac{0{,}338 - 0{,}274}{1 - 0{,}338} \frac{100 - 15}{15 - 2} = \frac{0{,}064 \cdot 85}{0{,}662 \cdot 13} = \frac{5{,}44}{8{,}606} = 0{,}632.$$

Следует иметь в виду, что F-критерий не универсален и не во всех случаях пригоден для получения вполне надежной информации о форме связи между коррелируемыми признаками. Поэтому наряду с F-критерием Фишера при проверке гипотезы о форме связи между переменными величинами применяют довольно простой и строгий *критерий Блекмана:* $B = N(h^2 - r^2) \gtrless 11{,}37$.

При наличии линейной связи этот показатель не превышает 11,37. Если же связь между признаками нелинейна, то $N(h^2 - r^2) > 11{,}37$. Так, применительно к рассмотренному выше примеру о связи между массой тела X и годовым удоем Y коров горбатовской породы имеем $h_{xy} = 0{,}581$ и $h^2_{xy} = 0{,}338$; $r_{xy} = 0{,}523$; $r^2_{xy} = 0{,}274$, а также $N = 100$ и $a = 15$ (см. табл. 104). Отсюда $B_{xy} = 100\,(0{,}338 - 0{,}274) = 100 \cdot 0{,}064 = 6{,}40 < 11{,}37$, что подтверждает гипотезу о линейной зависимости между этими признаками.

Другой результат получается при проверке гипотезы о форме связи между годовым удоем Y и массой тела X коров той же породы. Так, применение F-критерия Фишера приводит к следующему результату:

$$F_{\text{ф}} = \frac{(0{,}640)^2 - 0{,}274}{1 - (0{,}640)^2} \frac{85}{13} = \frac{0{,}410 - 0{,}274}{1 - 0{,}410} \frac{85}{13} =$$

$$= \frac{11{,}526}{7{,}675} = 1{,}51 < F_{st} = 1{,}82$$

для $k_1 = 13$, $k_2 = 85$ и $\alpha = 5\%$. Это означает, что гипотезу о линейной связи между этими признаками не учитывать нельзя.

Однако применение критерия Блекмана приводит к иному результату: $B_{yx} = 100(0{,}410 - 0{,}274) = 100 \cdot 0{,}136 = 13{,}60 > 11{,}37$, что не подтверждает вывод о линейности связи Y и X. Итак, два способа проверки — два противоположных результата. Какой же из них ближе к истине? Ответить на этот вопрос позволит более точный метод дисперсионного анализа (см. гл. IX).

VIII.2. НЕПАРАМЕТРИЧЕСКИЕ ПОКАЗАТЕЛИ СВЯЗИ

Коэффициент корреляции Фехнера. Наряду с параметрическими показателями для измерения корреляционной зависимости между признаками применяют и *непараметрические показатели.* Одним из них является *коэффициент корреляции,* предложенный Г. Фехнером (1897):

$$r_{\text{ф}} = \frac{\text{С} - \text{Н}}{\text{С} + \text{Н}}. \qquad (162)$$

Этот показатель основан на учете знаков отклонений вариант от их средних арифметических. Здесь С — число совпадений одинаковых, как положительных, так и отрицательных, знаков разностей $(x_i - \bar{x})$ и $(y_i - \bar{y})$, а Н — число несовпадающих знаков.

Как и пирсоновский коэффициент корреляции r_{yx}, основанный на учете не знаков отклонений, а их абсолютных значений, коэффициент корреляции Фехнера может принимать значения от —1 до +1. При положительной корреляции он имеет положительный, а при отрицательной — отрицательный знак.

Применяя коэффициент корреляции Фехнера, следует иметь в виду, что по сравнению с параметрическими показателями непараметрические являются лишь приближенными оценками связи. Поэтому вычисление последних можно ограничивать сотыми долями единицы. Также нужно иметь в виду то, что закон распределения коэффициента корреляции Фехнера неизвестен, поэтому вопрос об оценке достоверности $r_{\text{ф}}$ остается открытым.

Пример 9. В табл. 97 содержатся данные о содержании жира в молоке коров и их дочернего потомства. Воспользуемся этими данными и вычислим (по Фехнеру) коэффициент корреляции между жирномолочностью сравниваемых животных. Сначала находим средние арифметические жирномолочности коров $\bar{x} = 42{,}46/12 = 3{,}54$ и их дочерних особей $\bar{y} = 43{,}17/12 = 3{,}60$ (табл. 105).

Затем подсчитываем число совпадающих и несовпадающих знаков, которыми отмечены разности между значениями вариант и их средними арифметическими. Число совпадающих, как положительных, так и отрицательных, знаков оказывается равным С=10, а число несовпадающих знаков Н=2. Отсюда коэффициент корреляции $r_\Phi = (10-2)/(10+2) = 0{,}667 \approx 0{,}68$. Этот показатель оказался несколько выше, чем пирсоновский коэффициент корреляции ($r_{xy}=0{,}620$).

Таблица 10

Номера наблюдений	Процент жира в молоке		Отклонения от средних	
	коров материнского поколения x	коров дочериего поколения y	$(x_i-\bar{x})$	$(y_i-\bar{y})$
1	3,10	3,65	−	+
2	3,17	3,11	−	−
3	3,76	3,57	+	−
4	3,61	3,61	+	+
5	3,27	3,44	−	−
6	3,61	3,71	+	+
7	3,80	3,61	+	+
8	3,65	3,98	+	+
9	3,34	3,36	−	−
10	3,65	3,89	+	+
11	3,45	3,45	−	−
12	4,05	3,79	+	+
Сумма	42,46	43,17		

Пример 10. Вычислим коэффициент корреляции Фехнера между годовыми удоями тех же коров материнского поколения и их одновозрастного потомства. Необходимые данные и их обработка приведены в табл. 106.

Средние арифметические: $\bar{x}=42696/12=3558{,}00$; $\bar{y}=45639/12=3903{,}25$. В данном случае число совпадающих знаков С=8, а число несовпадающих знаков Н=4. Отсюда $r_\Phi = (8-4)/(8+4) = 0{,}33$. Следовательно, можно утверждать, что между годовым удоем коров материнского поколения и их одновозрастного потомства существует более слабая связь, чем в отношении жирномолочности между теми же группами коров.

Коэффициент корреляции рангов. Из непараметрических показателей связи наиболее широкое применение нашел *коэффициент корреляции* рангов, предложенный К. Спирменом (1904):

$$r_S = 1 - \frac{6\sum d^2}{n(n^2-1)}, \qquad (163$$

где $d = R_x - R_y$ — разность между рангами сопряженных значений признаков X и Y; n — число парных членов ряда, или объем выборки [1].

В основу конструкции этого показателя положены весьма простые соображения. Ранжируя попарно связанные значения признаков, можно видеть, как они распределяются относительно друг друга. Если возрастающим значениям одного признака X соответствуют возрастающие значения другого Y, то между ними существует положительная связь. Если же при возрастании значений одного признака значения другого последовательно уменьшаются, это указывает на наличие отрицательной связи между ними. При отсутствии корреляции ранжированным значениям одного признака будут соответствовать самые различные значения другого.

Таблица 106

Годовые удои коров, кг		$(x_i - \bar{x})$	$(y_i - \bar{y})$
материнского поколения X	дочернего поколения Y		
3 770	2 991	+	−
3 817	4 593	+	+
2 450	3 529	−	−
3 463	4 274	−	+
3 500	3 103	−	−
5 544	3 949	+	+
2 112	3 491	−	−
3 150	3 559	−	−
3 118	2 916	−	−
3 018	4 580	−	+
4 291	4 510	+	+
3 463	4 144	−	+
$\Sigma = 42\,696$	45 639	−	−

Обозначив ранжированные значения признаков порядковыми числами 1, 2, 3, 4, ..., нетрудно определить ранги этих значений и по их разности судить о степени зависимости одного признака от изменений другого. Очевидно, при полной связи ранги коррелируемых признаков совпадут и разность между ними будет равна нулю. В таких случаях коэффициент корреляции рангов окажется равным единице. Если же признаки варьируют неза-

[1] Эквивалентная формула коэффициента корреляции рангов:

$$r_s = \frac{3}{n-1}\left[\frac{4\Sigma R_x R_y}{n(n+1)} - (n+1)\right].$$

висимо друг от друга, то величина $\frac{6\sum d^2}{n(n^2-1)}=1$, и коэффициент корреляции рангов будет равен нулю. Таким образом, как и пирсоновский коэффициент корреляции и коэффициент корреляции Фехнера, коэффициент корреляции рангов выражается в долях единицы и может принимать значения от -1 до $+1$, т. е. сопровождается положительным или отрицательным знаком.

Как и другие выборочные показатели, эмпирический коэффициент корреляции рангов служит оценкой генерального параметра ρ_S и, как величина случайная, меняет свои значения при повторных выборках вариант из одной и той же генеральной совокупности. Значимость этого показателя, имеющего распределение со средней $\rho_S=0$ и дисперсией $\sigma^2_{\rho_S}=1/(n-1)$, оценивают путем сравнения выборочного коэффициента r_S с критической точкой r_{st}, которую можно определить по формуле

$$r_{st}=\frac{t}{\sqrt{n-1}}\left(1-\frac{m}{n-1}\right),$$

где n — объем выборки; t и m — величины, связанные с уровнем значимости α следующим образом: для $\alpha=5\%$ $t=1{,}96$ и $m=0{,}16$; для $\alpha=1\%$ $t=2{,}58$ и $m=0{,}69$. Нулевую гипотезу отвергают, если эмпирически найденная величина r_S превзойдет или окажется равной критическому значению r_{st} для принятого уровня значимости α и объема выборки n. Чтобы каждый раз не рассчитывать критические точки r_{st}, составлена специальная таблица, которая приводится в Приложениях (см. табл. XXIII).

Приведенный способ оценки значимости выборочного r_S не единственный. При $n\geqslant 10$ значимость эмпирического коэффициента корреляции рангов можно оценить с помощью t-критерия Стьюдента, т. е. по отношению этого показателя к своей статистической ошибке $t_\Phi=|r_s|\sqrt{\frac{n-2}{1-r_S^2}}\geqslant t_{st}$ для $n-2$ и принятого уровня значимости (α). Рассмотрим применение r_S на конкретных примерах.

Пример 11. Изучали зависимость между массой живого тела и содержанием гемоглобина (по Сали) в крови павианов-гамадрилов. Результаты наблюдений и их обработка приведены в табл. 107.

Если бы отдельные члены ряда не повторялись, их рангами были бы порядковые числа. Но так как некоторые варианты повторяются, их рангами будут средние арифметические из соответствующих чисел натурального ряда. У одинаковых членов ряда должны быть и одинаковые ранги. Так, в ряду X вариаиты 18 и 19 повторяются дважды и их ранги равны полусуммам соответствующих порядковых чисел: $(2+3)/2=2{,}5$ и $(4+5)/2=4{,}5$

и т. д. В последнем столбце табл. 107 показан расчет рангов для членов ряда Y.

Если ранги рассчитаны правильно, их суммы должны быть одинаковыми, т. е. $R_x = R_y$ и $\Sigma d = 0$. Если же $\Sigma d \neq 0$, следует искать ошибку в присвоении рангов или в их разностях. Поэтому, прежде чем рассчитывать Σd^2, следует проверить Σd, которая должна быть равна нулю. Так, в данном примере $R_x = 55$ и $R_y = 55$, а $\Sigma d = +9 - 9 = 0$, что указывает на отсутствие ошибки в расчете рангов. Подставляя $\Sigma d^2 = 54{,}00$ в формулу (163), находим

$$r_S = 1 - \frac{6 \cdot 54{,}00}{10(10^2 - 1)} = 1 - \frac{324}{990} = 1 - 0{,}33 = 0{,}67.$$

Таблица 107

Номера исследований	Масса X, кг	Y	Ранги рядов		$R_x - R_y = d$	d^2	Расчет рангов Y	
			R_x	R_y			Y	R_y
1	17	70	1	1	0,0	0,00	70	1
2	18	74	2,5	3	—0,5	0,25	72	2
3	18	78	2,5	7	—4,5	20,25	74	3
4	19	72	4,5	2	+2,5	6,25	76	4
5	19	77	4,5	5,5	—1,0	1,00	77	5,5
6	20	76	6	4	+2,0	4,00	77	5,5
7	21	88	7	10	—3,0	9,00	78	7
8	22	80	8	8	0,0	0,00	80	8
9	23	77	9	5,5	+3,5	12,25	86	9
10	25	86	10	9	+1,0	1,00	88	10
Сумма		—	55	55	0,0	54,00	—	—

Полученная величина (0,67) превосходит критическую точку (0,64) для $n = 10$ и 5%-ного уровня значимости (см. табл. XXIII Приложений), что позволяет отвергнуть нулевую гипотезу ($0{,}01 < P < 0{,}05$). К такому же выводу приводит и оценка значимости $r_S = 0{,}67$ по величине t-критерия Стьюдента:

$$t_\Phi = \frac{0{,}67\sqrt{10-2}}{1-(0{,}67)^2} = \frac{0{,}67 \cdot 2{,}83}{0{,}551} = \frac{1{,}896}{0{,}742} = 2{,}57.$$

В табл. V Приложений $k = n - 2 = 8$ и $\alpha = 5\%$ находим $t_{st} = 2{,}23$. Так как $t_\Phi > t_{st}$, нулевую гипотезу отвергают на 5%-ном уровне значимости. Следовательно, с вероятностью $P > 0{,}95$ можно утверждать, что между массой тела и количеством гемоглобина в крови у павианов-гамадрилов существует положительная корреляционная связь.

Рассчитывая коэффициент корреляции рангов, следует иметь в виду, что на его значении сказывается наличие групп с одинаковыми рангами, и тем сильнее, чем больше таких групп среди сопряженных значений признаков X и Y. Чтобы получить более- или менее точную оценку генерального параметра ρ_S, нужно при наличии указанных групп вносить поправку в формулу (163). Эту поправку, обозначаемую буквой T, прибавляют к числителю формулы, т. е.

$$r_S^* = 1 - \frac{6\sum d^2 + T}{n(n^2-1)}, \qquad (163а)$$

где $T = V_x + V_y$, а V_x — поправка для одного признака (ряд X) V_y — для другого (ряд Y). Для определения V_x и V_y составлена специальная таблица, в которой l обозначает число групп с одинаковыми рангами, а t — число рангов в этих группах (табл. 108)

Таблица 108

t \ l	1	2	3	4	5	6	7
2	0,5	1,0	1,5	2,0	2,5	3,0	3,5
3	2,0	4,0	6,0	8,0	10,0	12,0	14,0
4	5,0	10,0	15,0	20,0	25,0	30,0	35,0
5	10,0	20,0	30,0	40,0	50,0	60,0	70,0

Так, в отношении только что рассмотренного примера (табл. 107) в ряду X — две группы с одинаковыми рангами, т. е. $l=2$ в каждой группе — по два ранга, т. е. $t=2$. В табл. 108 для этого ряда находим $V_x=1{,}0$. В ряду Y — одна группа с одинаковыми рангами, т. е. $l=1$; в ней два ранга, т. е. $t=2$. В табл. 108 для ряда Y находим $V_y=0{,}5$. Всего $T=1{,}0+0{,}5=1{,}5$. Эту поправку вносим в формулу (163) и определяем коэффициент корреляции рангов:

$$r_S^* = 1 - \frac{6\cdot 54 + 1{,}5}{10(10^2-1)} = 1 - \frac{325{,}5}{990} = 1 - 0{,}329 = 0{,}671 \approx 0{,}67.$$

В данном случае число групп с одинаковыми вариантами невелико, поэтому поправка практически не сказалась на величине r_S.

Пример 12. Воспользуемся данными табл. 105 и вычислим коэффициент корреляции рангов между жирномолочностью коров материнской линии и их дочернего потомства. Предварительно освободимся от дробей, уменьшив каждую варианту на три единицы и умножив сотые доли на 100. Тогда вместо 3,10 получим 10, вместо 3,65—65 и т. д. Такое преобразование чисел никак не скажется на конечном результате, а вычисление Σd^2 значительно упростится (табл. 109).

Как и в предыдущем примере, здесь поправка $T=1{,}5$, откуда

$$r_S^* = 1 - \frac{6 \cdot 129 + 1{,}5}{12\,(12^2-1)} = 1 - \frac{775{,}5}{1716} = 1 - 0{,}452 = 0{,}548 \approx 0{,}55.$$

Эта величина (0,55) для $n=12$ и $\alpha=5\%$ не превосходит критическую точку (0,58) (см. табл. XXI Приложений). Такой же результат дает оценка $r_S=0{,}55$ по величине t-критерия Стьюдента:

$$t_\Phi = 0{,}55 \sqrt{\frac{12-2}{1-(0{,}55)^2}} = 0{,}55\sqrt{14{,}3} = 0{,}55 \cdot 3{,}76 = 2{,}08.$$

Эта величина не превосходит критическую точку $t_{st}=2{,}23$ для $k=12-2=10$ и $\alpha=5\%$ (см. табл. V Приложений). Обе оценки не дают основания для отвергания нулевой гипотезы. Полученный результат не согласуется с оценкой пирсоновского коэффициента корреляции ($r_{xy}=0{,}598$), который оказался статистически значимым на 5%-ном уровне ($0{,}01<P<0{,}05$).

Таблица 109

Номера исследований	Жирномолочность коров		Ранги рядов		d	d^2
	материнской линии X	дочерних особей Y	R_x	R_y		
1	10	65	1	8	−7	49,00
2	17	11	2	1	+1	1,00
3	27	44	3	3	0	0,00
4	34	36	4	2	+2	4,00
5	45	45	5	4	+1	1,00
6	61	71	6,5	9	−2,5	6,25
7	61	61	6,5	6,5	0	0,00
8	65	98	8,5	12	−3,5	12,25
9	65	89	8,5	11	−2,5	6,25
10	76	57	10	5	+5	25,00
11	80	61	11	6,5	+4,5	20,25
12	105	79	12	10	+2	4,00
Сумма	—	—	78	78	0	129,0

Какому показателю следует отдать предпочтение? Ответ на этот вопрос не может быть однозначным. Дело в том, что параметрический пирсоновский коэффициент корреляции достаточно точно характеризует линейную связь, когда коррелируемые признаки X и Y имеют нормальное или лог-нормальное распределение, т. е. такое, при котором не сама случайная величина, а логарифмы ее значений распределяются нормально. Примеры та-

кого рода приведены в гл. IX. Коэффициент корреляции рангов характеризует корреляционную связь независимо от закона распределения. И все же, если коррелируемые признаки распределяются нормально, предпочтение следует отдавать пирсоновскому коэффициенту корреляции, как более мощному показателю связи между переменными Y и X по сравнению с коэффициентом Спирмена. В тех случаях, когда коррелируемые признаки не распределяются нормально, следует исследовать непараметрические показатели связи.

Коэффициент ранговой корреляции Спирмена и другие непараметрические показатели независимы от закона распределения, и в этом их большая ценность. Они позволяют измерять тесноту сопряженности между такими признаками, которые не поддаются непосредственному измерению, но могут быть выражены баллами или другими условными единицами, позволяющими ранжировать выборку. Ценность коэффициента корреляции рангов заключается также в том, что он позволяет быстро оценивать взаимосвязь между признаками независимо от закона распределения.

Коэффициент ассоциации. Тесноту связи между качественными признаками Y и X, группируемыми в четырехпольную корреляционную таблицу, измеряют с помощью *коэффициента ассоциации,* или *тетрахорического показателя связи,* предложенного К. Пирсоном в 1901 г. В простейшем виде формула, по которой рассчитывают этот показатель, обозначаемый символом r_A, выглядит следующим образом:

$$r_A = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \qquad (164)$$

Здесь a, b, c и d — численности коррелируемых групп (вариант), распределяемых по клеткам четырехпольной таблицы.

Коэффициент ассоциации, как и другие подобные показатели, имеет прямое отношение к пирсоновскому критерию χ^2, на котором он основан; в данном случае $r_A = \sqrt{\chi^2/n}$. Коэффициент ассоциации, как и пирсоновский коэффициент корреляции, изменяется от -1 до $+1$. Значимость выборочного коэффициента ассоциации оценивают по величине критерия Пирсона χ^2. Нулевая гипотеза сводится к предположению, что в генеральной совокупности этот показатель ρ_A равен нулю. H_0-гипотезу отвергают, если $\chi^2 = nr^2_A \geqslant \chi^2_{st}$ для принятого уровня значимости (α) и числа степеней свободы $k = (2-1)(2-1) = 1$.

Значимость r_A можно проверить и с помощью t-критерия Стьюдента. Нулевую гипотезу отвергают, если

$$t_\Phi = \frac{r_A\sqrt{n-2}}{\sqrt{1-r_A^2}} \geqslant t_{st}$$

ля принятого уровня значимостн (α) и числа степеней свободы $=n-2$.

Пример 13. От скрещивания самцов плодовой мушки дрозофилы, имеющих окраску тела и зачаточные крылья (рецессивные признаки), с нормальными самками того же вида, гетерозиготными по генам этих признаков, в потомстве оказались мухи:

Серые с нормальнымн крыльямн	75
Серые с зачаточными крыльями	16
Черные с нормальными крыльямн	14
Черные с зачаточными крыльями	68

Выяснить, имеется ли связь между окраской тела и развитием крыльев у дрозофилы. Группируем эти данные и подсчитываем численность мух по столбцам и строкам четырехпольной таблицы (табл. 110).

Таблица 110

Окраска тела X	Крылья Y		Сумма
	нормальные	зачаточные	
Серая	$a=75$	$b=16$	$a+b=91$
Черная	$c=14$	$d=68$	$c+d=82$
Сумма	$a+c=89$	$b+d=84$	$n=173$

Подставляя известные значения в формулу (164), находим

$$r_A=\frac{75\cdot 68-14\cdot 16}{\sqrt{89\cdot 84\cdot 91\cdot 82}}=\frac{4876}{\sqrt{55\,785\,912}}=\frac{4876}{7469}=0{,}653.$$

Ранее было показано, что распределение вероятных значений критерия χ^2 является непрерывным (см. рис. 22). Качественные же признаки дискретны, их числовые значения не распределяются непрерывно. Учитывая эту особенность, в формулу (164) принято вносить поправку Йейтса на непрерывность вариации, равную половине объема выборки. Эту поправку вычитают из разности $(ad—bc)$, и формула (164) принимает следующий вид:

$$r_A=\frac{(\,|\,ad-bc\,|\,)-0{,}5n}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \qquad (165)$$

Применим эту формулу к только что рассмотренному примеру:

$$r_A=\frac{(75\cdot 68-14\cdot 16)-0{,}5\cdot 173}{\sqrt{89\cdot 94\cdot 91\cdot 82}}=\frac{47\,895}{7469{,}0}=0{,}641.$$

Полученная величина указывает на наличие тесной связи между окраской тела и развитием крыльев у дрозофилы. Значимость этого показателя ($\chi^2_ф=173\ (0{,}641)^2=71{,}08$) значительно превышает критический уровень $\chi^2_{st}=10{,}83$ для $\alpha=0{,}1\%$ и $k=1$ (см. табл. VII Приложений). К такому же заключению приводит и оценка достоверности (значимости) коэффициента $r_A=0{,}641$ по величине t-критерия Стьюдента:

$$t_ф=0{,}641\sqrt{\frac{173-2}{1-(0{,}641)^2}}=0{,}641\sqrt{\frac{171}{0{,}589}}=$$

$$=0{,}641\sqrt{290{,}3}=0{,}641\cdot 17{,}0838=10{,}92.$$

В табл. V Приложений для $k=173-2=171$ и $\alpha=0{,}1\%$ находим $t_{st}=3{,}37$. Так как $t_ф>t_{st}$, нулевая гипотеза опровергается на высоком уровне значимости ($P<0{,}001$). Следовательно, с вероятностью $P\approx 99\%$ можно считать доказанным наличие тесной связи между окраской тела и развитием крыльев у дрозофилы.

Коэффициент ассоциации Юла. Этот непараметрический показатель связи между качественными признаками, группируемыми в четырехпольную таблицу, определяют по формуле

$$r_Q=\frac{ad-bc}{ad+bc}. \qquad (166)$$

Как величина случайная, коэффициент ассоциации Юла сопровождается статистической ошибкой

$$s_{r_Q}=\frac{1-r_Q^2}{2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}. \qquad (167)$$

В формулах (166) и (167) символы a, b, c и d имеют то же значение, что и в формуле (164). Достоверность этого выборочного показателя проверяют по величине t-критерия Стьюдента. Вычислим r_Q для данных из примера 13. Необходимые данные содержатся в табл. 110:

$$r_Q=\frac{75\cdot 68-14\cdot 16}{75\cdot 68+14\cdot 16}=\frac{5100-224}{5100+224}=\frac{4876}{5324}=0{,}916.$$

Полученная величина ($r_Q=0{,}916$), как и можно было ожидать, значительно выше той, которая была найдена ранее [см. формулу (164)], что связано с конструкцией коэффициента ассоциации Юла. Находим ошибку этого показателя:

$$s_{r_Q}=\frac{1-(0{,}916)^2}{2}\sqrt{\frac{1}{75}+\frac{1}{68}+\frac{1}{14}+\frac{1}{16}}=$$

$$=\frac{0{,}161}{2}\sqrt{0{,}1599}=0{,}0805\cdot 0{,}40=0{,}032.$$

Отсюда $t_ф = 0{,}916/0{,}032 = 28{,}6$. Для $k = 173 - 2 = 171$ и $\alpha = 0{,}1\%$ критическая точка $t_{st} = 3{,}37$ (см. табл. V Приложений). Нулевая гипотеза отвергается на 0,1%-ном уровне значимости ($P < 0{,}001$), что подтверждает сделанный выше вывод о наличии связи между окраской тела и развитием крыльев у дрозофилы.

Коэффициент взаимной сопряженности. Для определения степени сопряженности между качественными признаками с числами вариант, большими двух, служит *коэффициент взаимной сопряженности* или *полихорический показатель связи,* предложенный К. Пирсоном:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{\varphi^2}{\varphi^2 + 1}}, \qquad (168)$$

где $\varphi^2 = \left(\sum_{l=1}^{n} \frac{f_{xy}^2}{\Sigma f_x \Sigma f_y} \right) - 1$ — величина, в которой f_{xy} обозначает частоты в клетках многопольной корреляционной таблицы, а Σf_x и Σf_y — суммы частот по строкам и столбцам той же таблицы; $N = \Sigma f_x + \Sigma f_y$ — общая сумма частот, или объем выборки.

Пирсоновский коэффициент взаимной сопряженности (C) имеет один существенный недостаток: его значение значительно зависит от количества вариант коррелируемых качественных признаков.

Учитывая этот недостаток, А. А. Чупров внес поправки в формулу (168), которые приняли следующие выражения:

$$K^2 = \frac{\chi^2}{N\sqrt{(n_x - 1)(n_y - 1)}} = \frac{\varphi^2}{\sqrt{(n_x - 1)(n_y - 1)}}. \qquad (169)$$

Здесь K — коэффициент взаимной сопряженности Чупрова; n_x и n_y — численность групп по строкам и столбцам многопольной таблицы; N — объем выборки. Остальные символы объяснены выше. Нулевую гипотезу отвергают, если $\chi^2_ф = N\varphi^2 \geqslant \chi^2_{st}$ для принятого уровня значимости и числа степеней свободы

Пример 14. Изучали зависимость между цветом волос и цветом глаз у человека. Результаты наблюдений сведены в табл. 111.

Определим коэффициент взаимной сопряженности между этими признаками, предварительно рассчитав величину φ^2:

$$\varphi^2 = \frac{170^2}{255 \cdot 308} + \frac{80^2}{255 \cdot 572} + \frac{5^2}{255 \cdot 20} + \frac{70^2}{230 \cdot 308} + \frac{152^2}{230 \cdot 572} +$$

$$+ \frac{8^2}{230 \cdot 20} + \frac{68^2}{415 \cdot 308} + \frac{340^2}{415 \cdot 308} + \frac{7^2}{415 \cdot 20} - 1 = 1{,}205 - 1 = 0{,}205.$$

Подставляем известные значения в формулу (169):

$$K = \sqrt{\frac{0{,}205}{\sqrt{(3-1)(3-1)}}} = \sqrt{\frac{0{,}205}{4}} = 0{,}226.$$

Найденная величина $K=0{,}226$ указывает на наличие слабой связи между цветом глаз и цветом волос у человека. Критерий $\chi^2_{\Phi}=N\varphi^2=900\cdot 0{,}205=184{,}5>\chi^2_{st}=18{,}47$ для $\alpha=0{,}1\%$ и $k=(3-1)(3-1)=4$. Так как $\chi^2_{\Phi}>\chi^2_{st}$, нулевая гипотеза отвергается на весьма высоком уровне значимости ($P<0{,}001$).

Необходимо помнить, что правильное применение критерия χ^2 основано на требовании, чтобы в клетках корреляционной таблицы содержалось не менее пяти вариант и чтобы общее число наблюдений не было меньше 50. Несоблюдение этих требова-

Таблица 11

Цвет глаз	Цвет волос			Всего
	блондины	шатены	рыжие	
Голубые	170	80	5	255
Серые	70	152	8	230
Карие	68	340	7	415
Всего	308	572	20	900

ний не гарантирует получение достаточно точных оценок генерального параметра ρ_{χ}^2, а следовательно, и правильных выводов, которые делают на основании выборочных показателей.

Коэффициент корреляции знаков. Иногда коррелируемые признаки выражают не числами, а знаками: наличие признака — знаком плюс, отсутствие — знаком минус. Такие случаи встречаются, например, в психоло-педагогических исследованиях, когда выясняют зависимость между поведенческими признаками. Для измерения корреляции между такими признаками предложена формула

$$R_{xy}=\frac{P(XY)-P(X)P(Y)}{\sqrt{P(X)P(Y)(1-P(X)(1-P(Y))}}, \qquad (170$$

где $P(XY)$ — число совпадений положительных знаков в общей серии испытаний, отнесенное к их числу n, т. е. $P(XY)=\Sigma(\overset{+}{+})/n$; $P(X)$ и $P(Y)$ — частости положительных знаков для каждого признака отдельно, т. е. $P(X)=\Sigma_X(+)/n$ и $P(Y)=\Sigma_Y(+)/n$.

Коэффициент корреляции знаков принимает значения от нуля до единицы. Чем сильнее связь между признаками, тем этот показатель ближе к единице, и, наоборот, чем слабее зависимость одного признака от другого, тем меньше будет и коэффициент корреляции знаков.

Пример 15. Изучали зависимость между увлеченностью знаниями X и склонностью учащихся к математике Y. Под наблюде-

ием находилось десять мальчиков ($n=10$). Результаты наблюдений приведены в табл. 112.

В этой таблице наличие признака обозначено положительным, а отсутствие — отрицательным знаком. Общее число совпадений положительных знаков $P(XY)=4/10=0{,}4$; частости для каждого признака в отдельности $P(X)=6/10=0{,}6$ и $P(Y)=5/10=0{,}5$; разности: $1-0{,}6=0{,}4$ и $1-0{,}5=0{,}5$. Подставляем известные значения в формулу (170):

$$R_{XY}=\frac{0{,}4-0{,}6\cdot 0{,}5}{\sqrt{0{,}6\cdot 0{,}5\cdot 0{,}4\cdot 0{,}5}}=\frac{0{,}10}{\sqrt{0{,}06}}=\frac{0{,}100}{0{,}245}=0{,}408\approx 0{,}41.$$

Таблица 112

Признаки	Номера исследуемых										Число +
	1	2	3	4	5	6	7	8	9	10	
X	−	+	+	−	−	+	−	+	+	+	6
Y	−	+	−	−	−	+	+	+	+	−	5

Пример 16. Выясняли зависимость между упрямством детей X и строгостью требований родителей Y. Под наблюдением находилось 15 учащихся и их родителей из разных семей. Результаты наблюдений приведены в табл. 113.

Таблица 113

Признаки	Номера испытуемых															Число +
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
X	+	−	−	+	+	−	−	−	+	−	+	+	+	−	−	7
Y	+	−	−	−	+	−	−	−	+	−	+	−	+	+	−	6

В данном случае $P(XY)=5/15=0{,}33$; $P(X)=7/15=0{,}47$ и $P(Y)=6/15=0{,}40$. Отсюда

$$R_{XY}=\frac{0{,}33-0{,}47\cdot 0{,}40}{\sqrt{0{,}47\cdot 0{,}40\cdot 0{,}53\cdot 0{,}60}}=\frac{0{,}142}{\sqrt{0{,}06}}=\frac{0{,}142}{0{,}245}=0{,}580.$$

Если сравнивать первый результат со вторым, то можно сказать, что в первом случае сопряженность между признаками несколько слабее, чем во втором.

Бисериальный коэффициент корреляции. Для измерения тесноты связи между качественными признаками с двумя вариан-

тами и количественными признаками используют *бисериальный коэффициент корреляции*

$$r_{bs}=\frac{\bar{x}_1-\bar{x}_2}{s_x}\sqrt{\frac{n_1n_2}{N(N-1)}}, \qquad (171$$

где $\bar{x}_1$ и $\bar{x}_2$ — средние арифметические из отдельных значений альтернативных групп с их объемами n_1 и n_2; $N=(n_1+n_2)$ — общее число наблюдений, или объем выборки; s_x — среднеквадратическое отклонение для всей выборки.

Бисериальный коэффициент корреляции изменяется от — до +1; при $\bar{x}_1=\bar{x}_2$ он равен нулю. Знак для этого показателя не имеет, однако, смыслового значения.

Таблица 11

Показания эстезиометра X, мм	Пол подростков		f_i	f_ix_i	$f_ix_i^2$
	мужской	женский			
1,5	1	5	6	9,0	13,50
1,6	—	2	2	3,2	5,12
1,7	—	2	2	3,4	5,78
1,8	1	1	2	3,6	6,48
1,9	1	1	2	3,8	7,22
2,0	2	2	4	8,0	16,00
2,1	—	1	1	2,1	4,41
2,2	1	—	1	2,2	4,84
2,3	2	—	2	4,6	10,58
2,4	3	—	3	7,2	17,28
2,5	2	1	3	7,5	18,75
Сумма	$n_1=13$	$n_2=15$	$N=28$	54,6	109,96

Значимость выборочного r_{bs} оценивают по величине t-критерия Стьюдента с числом степеней свободы $k=N-2$ и принимаемым уровнем значимости.

Пример 17. Изучали зависимость между полом подростков 16—17-летнего возраста и их тактильной чувствительностью. Единицей признака служило расстояние между ножками эстезиометра, при котором ощущение двух прикосновений к концу среднего пальца левой руки воспринималось как одно прикосновение, т. е. сливалось. Результаты опыта и расчет вспомогательных величин приведены в табл. 114.

Вычислим коэффициент корреляции между этими признаками; начнем с определения средних арифметических для женской (ж) и мужской (м) групп: $\bar{x}_ж=(1/15)(5\cdot1,5+2\cdot1,6+2\times1,7+1\cdot1,8+1\cdot1,9+2\cdot2,0+1\cdot2,1+1\cdot2,5)=26,4/15=1,76$ мм $\bar{x}_м=(1/13)(1,5+1,8+1,9+2\cdot2,0+2,2+2\cdot2,3+3\cdot2,4+2\cdot2,5)=$

$=28{,}2/13=2{,}17$ мм. Затем определяем среднее квадратическое отклонение: $s_x^2=\frac{1}{N-1}\left[\sum f_i x_i^2-\frac{(\sum f_i x_i)^2}{N}\right]=\frac{1}{27}\left[109{,}96-\right.$ $\left.-\frac{(54{,}6)^2}{28}\right]=\frac{3{,}49}{27}=0{,}1292$ и $s_x=0{,}36$.

Подставляем найденные величины в формулу (171):

$$r_{bs}=\frac{2{,}17-1{,}76}{0{,}36}\sqrt{\frac{13\cdot 15}{28\cdot 27}}=\frac{0{,}41\cdot 0{,}508}{0{,}36}=\frac{0{,}208}{0{,}36}=0{,}578.$$

Критерий достоверности $t=\frac{0{,}578\sqrt{(28-2)}}{\sqrt{1-(0{,}578)^2}}=\frac{2{,}947}{0{,}816}=3{,}61$. Эта величина превосходит критический уровень $t_{st}=1\%$ для $\alpha=0{,}01$ и $k=28-2=26$. Нулевая гипотеза отвергается на 1%-ном уровне значимости ($0{,}001<P<0{,}01$). Можно считать установленным, что между полом подростков и тактильной чувствительностью конца среднего пальца левой руки существует определенная связь: девушки оказываются более чувствительными к прикосновению эстезиометра, чем юноши того же возраста.

VIII.3. МНОЖЕСТВЕННАЯ И ЧАСТНАЯ КОРРЕЛЯЦИЯ

Множественная корреляция. Наряду с анализом двумерных совокупностей в биологии широкое применение находит *статистический анализ многомерных корреляционных связей*. Простейшим случаем множественной корреляции является зависимость между тремя признаками: X, Y и Z. Тесноту связи одного из них (X) с двумя другими признаками (Y и Z) измеряют с помощью *коэффициента множественной корреляции*:

$$r_{x(yz)}=\sqrt{\frac{r_{xy}^2+r_{xz}^2-2r_{xy}r_{xz}r_{yz}}{1-r_{yz}^2}}, \qquad (172)$$

где r_{xy}, r_{xz} и r_{yz} — коэффициенты линейной корреляции между парами признаков X и Y, X и Z, Y и Z.

Коэффициент множественной корреляции принимает значения от нуля до единицы ($0\leqslant r\leqslant 1$). Значимость этого совокупного показателя корреляции оценивают по величине t-критерия Стьюдента с числом степеней свободы $k=n-3$ и принятым уровнем значимости.

Пример 18. Из снопа озимой ржи случайным способом было отобрано 10 колосьев. Затем измерили длину каждого колоса X, подсчитали число колосков Y и количество зерен Z в каждом колосе. Собранные данные и их первичная обработка приведены в табл. 115.

Чтобы определить коэффициент множественной корреляции между этими признаками, необходимо сначала рассчитать пар-

ные коэффициенты корреляции. Используя итоги табл. 115, находим суммы квадратов отклонений вариант от их средних арифметических, т. е. девианты:

$$\sum (x_i - x)^2 = \sum x_i^2 - (\sum x_i)^2/n = 34\,469 - 575^2/10 =$$

$$= 34\,469 - 33062{,}5 = 1406{,}5;$$

$$\sum (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n = 2891 - 165^2/10 = 2891 - 2722{,}5 = 168{,}5$$

$$\sum (z_i - \bar{z})^2 = \sum z_i^2 - (\sum z_i)^2/n = 9456 - 294^2/10 = 9456 - 8643{,}6 =$$

$$= 812{,}4.$$

Таблица 117

X	Y	Z	X^2	Y^2	Z^2	XY	YZ	XZ
70	18	36	4900	324	1296	1260	648	2520
60	17	29	3600	289	841	1020	493	1740
70	22	40	4900	484	1600	1540	880	2800
46	10	12	2116	100	144	460	120	552
58	16	31	3364	256	961	928	496	1798
69	18	32	4761	324	1024	1242	576	2208
32	9	13	1024	81	169	288	117	416
62	18	35	3844	324	1225	1116	630	2170
46	15	30	2116	225	900	690	450	1380
62	22	36	3844	484	1296	1364	792	2232
575	165	294	34469	2891	9456	9908	5202	17816

Отсюда $s_x = \sqrt{1406{,}5/10} = 11{,}86$; $s_y = \sqrt{168{,}5/10} = 4{,}10$; $s_z = \sqrt{812{,}4/10}$. Затем рассчитываем величины сопряженной вариации:

$$\sum (y_i - \bar{y})(x_i - \bar{x}) = \sum yx - \sum y \sum x/n =$$

$$= 9908 - 575 \cdot 165/10 = 420{,}5;$$

$$\sum (y_i - \bar{y})(z_i - \bar{z}) = \sum yz - \sum y \sum z/n =$$

$$= 5202 - 165 \cdot 294/10 = 351{,}0;$$

$$\sum (x_i - \bar{x})(z_i - \bar{z}) = \sum xz - \sum x \sum z/n =$$

$$= 17\,816 - 575 \cdot 294/10 = 911{,}0.$$

Наконец, определяем парные коэффициенты корреляции:

$$r_{xy}=\frac{\sum(y_l-\bar{y})(\bar{x}_l-x)}{ns_xs_y}=\frac{420{,}5}{10\cdot11{,}86\cdot4{,}10}=\frac{420{,}5}{486{,}3}=0{,}865;$$

$$r_{yz}=\frac{\sum(y_l-\bar{y})(z_l-\bar{z})}{ns_ys_z}=\frac{351{,}0}{10\cdot4{,}10\cdot9{,}01}=\frac{351{,}0}{364{,}4}=0{,}950;$$

$$r_{xz}=\frac{\sum(x_l-\bar{x})(z_l-\bar{z})}{ns_xs_z}=\frac{911{,}0}{10\cdot11{,}86\cdot9{,}01}=\frac{911{,}0}{1068{,}6}=0{,}853.$$

Подставляем известные величины в формулу (172):

$$r_{x(yz)}=\sqrt{\frac{0{,}865^2+0{,}853^2-2\cdot0{,}865\cdot0{,}853\cdot0{,}950}{1-(0{,}950)^2}}=$$

$$=-\sqrt{\frac{0{,}0739}{0{,}0975}}=\sqrt{0{,}758}=0{,}871.$$

Критерий достоверности $t_\Phi=\frac{0{,}871\sqrt{10-3}}{\sqrt{1-(0{,}871)^2}}=\frac{2\ 304}{\sqrt{0{,}241}}=\frac{2{,}304}{0{,}491}=$ $=4{,}69$; $t_{st}=3{,}50$ для $k=10-3=7$ и $\alpha=1\%$ (см. табл. V Приложений). Нулевая гипотеза отвергается на 1%-ном уровне значимости ($0{,}001<P<0{,}01$).

Частная корреляция. Если известна связь между признаками X, Y и Z, можно определить *частные* или *парциальные коэффициенты корреляции*, показывающие корреляционную зависимость между двумя варьирующими признаками при постоянной величине третьего признака. Для определения частного коэффициента корреляции между признаками X и Y при постоянной величине признака Z применяют формулу

$$r_{xy(z)}=\frac{r_{xy}-r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}. \qquad (173)$$

Заключение знака Z в скобки обозначает, что влияние признака Z на корреляцию между X и Y исключено.

Соответственно формула для определения частного коэффициента корреляции между признаками X и Z при исключении влияния на эту связь признака Y будет выглядеть так:

$$r_{xy(y)}=\frac{r_{xz}-r_{xy}r_{yz}}{\sqrt{(1-r_{xy}^2)(1-r_{yz}^2)}}. \qquad (174)$$

И наконец, частный коэффициент корреляции между признаками Y и Z при постоянной величине признака X определяется по формуле

$$r_{yz(x)}=\frac{r_{yz}-r_{xy}r_{xz}}{\sqrt{(1-r_{xy}^2)(1-r_{xz}^2)}}. \qquad (175)$$

Рассчитаем частные коэффициенты корреляции для выборки из примера 18:

$$r_{xy(z)}=\frac{0{,}865-0{,}853\cdot 0{,}950}{(1-0{,}853^2)(1-0{,}950^2)}=\frac{0{,}055}{\sqrt{0{,}027}}=\frac{0{,}055}{0{,}164}=0{,}335;$$

$$r_{xz(y)}=\frac{0{,}853-0{,}865\cdot 0{,}950}{(1-0{,}865^2)(1-0{,}950^2)}=\frac{0{,}032}{\sqrt{0{,}024}}=\frac{0{,}031}{0{,}155}=0{,}200;$$

$$r_{yz(x)}=\frac{0{,}950-0{,}865\cdot 0{,}853}{\sqrt{(1-0{,}865^2)(1-0{,}853^2)}}=\frac{0{,}212}{\sqrt{0{,}0686}}=\frac{0{,}212}{0{,}262}=0{,}809.$$

Наиболее высоким оказался коэффициент корреляции между числом колосков Y и количеством зерен в колосьях Z при исключенном влиянии на эту связь признака X, т. е. длины колосьев.

Критерий достоверности $t_ф=0{,}809\sqrt{\frac{10-2}{1-0{,}809^2}}=0{,}809\times$ $\times\sqrt{23{,}15}=0{,}809\cdot 4{,}81=3{,}89$. Эта величина превосходит критическую точку $t_{st}=3{,}36$ для $k=10-2=8$ и $\alpha=1\%$ (см. табл. V Приложений). Нулевая гипотеза отвергается на 1%-ном уровне значимости ($P<0{,}01$).

Рассмотренные коэффициенты множественной и частной корреляции применяют лишь для измерения линейных связей. Анализ множественных нелинейных связей описан в специальной литературе.

ГЛАВА IX

РЕГРЕССИОННЫЙ АНАЛИЗ

Понятие регрессии. Зависимость между переменными величинами X и Y может быть описана разными способами. В частности, любую форму связи можно выразить уравнением общего вида $y=f(x)$, где y рассматривают в качестве зависимой переменной, или *функции* от другой — независимой переменной величины x, называемой *аргументом*. Соответствие между аргументом и функцией может быть задано таблицей, формулой, графиком и т. д. Изменение функции в зависимости от изменений одного или нескольких аргументов называется *регрессией*[1].

[1] Термин «регрессия» (от лат. regressio — движение назад) ввел в биологию Ф. Гальтон, изучавший наследование количественных признаков. Он обнаружил, что потомство высокорослых и низкорослых родителей возвращается (регрессирует) на $^1/_3$ в сторону среднего уровня этого признака в данной популяции. С развитием биометрии этот термин утратил свое буквальное значение и стал применяться для обозначения и корреляционной зависимости между переменными величинами Y и X.

Весь арсенал средств, применяемых для описания корреляционных связей, составляет содержание регрессионного анализа.

Как было показано в гл. VIII, отличие статистической связи от функциональной заключается в том, что в последнем случае между аргументом и функцией существует однозначное соответствие, т. е. каждому определенному значению аргумента x соответствует определенное значение функции $y=f(x)$. При статистической связи разным значениям одной переменной соответствуют различные распределения другой переменной, в которых могут быть найдены частные средние $\bar{y}_x$. Поэтому форма статистической связи может быть описана не как зависимость отдельных значений y от величин x, а как зависимость частных средних $\bar{y}_x$ от значений x.

Для выражения регрессии служат корреляционные уравнения, или уравнения регрессии, эмпирические и теоретически вычисленные ряды регрессии, их графики, называемые линиями регрессии, а также коэффициенты линейной и нелинейной регрессии.

Показатели регрессии выражают корреляционную связь двусторонне, учитывая изменение усредненных значений $\bar{y}_x$ признака Y при изменении значений x_i признака X, и, наоборот, показывают изменение средних значений $\bar{x}_y$ признака X по измененным значениям y_i признака Y. Исключение составляют временные ряды, или ряды динамики, показывающие изменение признаков во времени. Регрессия таких рядов является односторонней.

Различных форм и видов корреляционных связей много. Задача сводится к тому, чтобы в каждом конкретном случае выявить форму связи и выразить ее соответствующим корреляционным уравнением, что позволяет предвидеть возможные изменения одного признака Y на основании известных изменений другого X, связанного с первым корреляционно.

IX.1. ЛИНЕЙНАЯ РЕГРЕССИЯ

Уравнение регрессии. Результаты наблюдений, проведенных над тем или иным биологическим объектом по корреляционно связанным признакам Y и X, можно изобразить точками на плоскости, построив систему прямоугольных координат. В результате получается некая диаграмма рассеяния, позволяющая судить о форме и тесноте связи между варьирующими признаками. Довольно часто эта связь выглядит в виде прямой или может быть аппроксимирована прямой линией.

Линейная зависимость между переменными Y и X описывается уравнением общего вида $\bar{y}_x=a+bx_1+cx_2+dx_3+\ldots$, где $a, b, c, d, \ldots$ — параметры уравнения, определяющие соотношения между аргументами $x_1, x_2, x_3, \ldots, x_m$ и функций $\bar{y}_x$. В прак-

тике учитывают не все возможные а лишь некоторые аргументы, в простейшем случае — всего один:

$$\bar{y}_x = a + bx. \qquad (176)$$

В этом уравнении линейной регрессии a — свободный член, а параметр b определяет наклон линии регрессии по отношению к осям прямоугольных координат. В аналитической геометрии этот параметр называют *угловым коэффициентом,* в биометрии — *коэффициентом регрессии.* Наглядное представление об этом параметре и о положении линий регрессии Y по X и X по Y в системе прямоугольных координат дает рис. 26 .

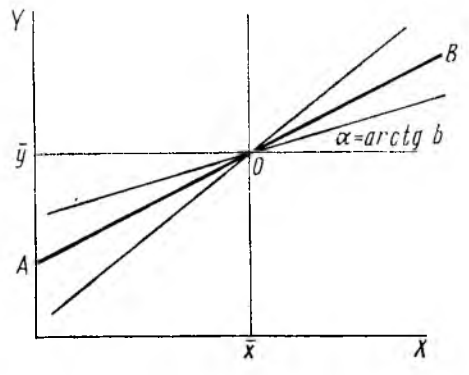


Рис. 26. Линии регрессии Y по X и X по Y в системе прямоугольных координат

Линии регрессии, как показано на рис. 26, пересекаются в точке O $(\bar{x}, \bar{y})$, соответствующей средним арифметическим значениям корреляционно связанных друг с другом признаков Y и X. При построении графиков регрессии по оси абсцисс откладывают значения независимой переменной X, а по оси ординат — значения зависимой переменной, или функции Y. Линия AB, проходящая через точку O $(\bar{x}, \bar{y})$ соответствует полной (функциональной) зависимости между переменными величинами Y и X, когда коэффициент корреляции $r_{xy} = 1$. Чем сильнее связь между Y и X, тем ближе линии регрессии к AB, и, наоборот, чем слабее связь между этими величинами, тем более удаленными оказываются линии регрессии от AB. При отсутствии связи между признаками линии регрессии оказываются под прямым углом (90°) по отношению друг к другу и $r_{xy} = 0$.

Поскольку показатели регрессии выражают корреляционную связь двусторонне, уравнение регрессии (176) следует записывать так:

$$\bar{y}_x = a_{yx} + b_{yx}x \quad \text{и} \quad \bar{x}_y = a_{xy} + b_{xy}y. \qquad (177)$$

По первой формуле определяют усредненные значеиия $\bar{y}_x$ при изменении признака X на единицу меры, по второй — усредненные значения $\bar{y}_x$ при изменении иа единицу меры признака Y.

Коэффициент регрессии. Коэффициент регрессии показывает, насколько в среднем величина одного признака y изменяется при изменении на единицу меры другого, корреляционно

связанного с Y признака X. Этот показатель определяют по формуле

$$b_{yx}=r_{xy}\frac{s_y}{s_x} \text{ или } b_{xy}=r_{xy}\frac{s_x}{s_y}. \quad (178)$$

Здесь значения s домножают на размеры классовых интервалов λ, если их находили по вариационным рядам или корреляционным таблицам.

Пример 1. В гл. VIII было показано, что корреляция между годовым удоем Y и массой тела X коров горбатовской породы характеризуется величиной $r_{xy}=0{,}523$. Установлено также, что между этими признаками имеет место линейная связь. Имея в виду значения средних квадратических отклонений ($s_x=2{,}843$ и $s_y=3{,}272$) и величины классовых интервалов ($\lambda_x=152$ и $\lambda_y=14$), определим коэффициент регрессии годового уровня по массе тела коров:

$$b_{yx}=0{,}523\frac{3{,}272\cdot 14}{2{,}843\cdot 152}=\frac{23{,}958}{432{,}136}=0{,}0554.$$

Аналогичным способом находим коэффициент регрессии массы тела коров по их годовому удою:

$$b_{xy}=0{,}523\frac{2{,}843\cdot 152}{3{,}272\cdot 14}=\frac{226{,}007}{45{,}808}=4{,}934.$$

Увеличение годового удоя коров этой группы на 1 кг связано (при прочих равных условиях) с повышением их живой массы тела в среднем на 0,055 кг, тогда как увеличение массы тела коров на 1 кг в тех же условиях сопряжено с повышением годового удоя в среднем на 4,934 кг. Если же судить о соотношении живой массы тела коров и их годового удоя по средним арифметическим для стада, которые равны по удою $\bar{x}=2235{,}6$ кг, а по массе тела коров $\bar{y}=349{,}6$ кг, то получаются следующие результаты: на 1 кг массы тела коров приходится в среднем $2235{,}6/349{,}6=6{,}395$ кг молока, а прибавка годового удоя на 1 кг связана с увеличением массы тела коров в среднем на $349{,}6/2235{,}6=0{,}156$ кг.

При сравнении этих величин с коэффициентами регрессии видно, что они оказываются более высокими, чем b_{yx} и b_{xy}. Причина заключается в том, что отношения средних $\bar{x}$ и $\bar{y}$ не учитывают корреляцию между признаками, поэтому и не могут служить точными показателями изменчивости одного признака при изменении на единицу меры другого. Этот пример показывает, какое значение имеет коэффициент линейной регрессии в области анализа статистических связей.

Коэффициент регрессии можно вычислить минуя расчет средних квадратических отклонений s_y и s_x по формуле

$$b_{yx}=r_{xy}\sqrt{\frac{\sum(y_l-\bar{y})^2}{\sum(x_l-\bar{x})^2}} \text{ или } b_{xy}=r_{xy}\sqrt{\frac{\sum(x_l-\bar{x})^2}{\sum(y_l-\bar{y})^2}}. \quad (179)$$

Если же коэффициент корреляции неизвестен, коэффициент регрессии определяют следующим образом:

$$b_{yx}=\frac{\Sigma(y_l-\bar{y})(x_l-\bar{x})}{\Sigma(x_l-\bar{x})^2} \text{ или } b_{xy}=\frac{\Sigma(y_l-\bar{y})(x_l-\bar{x})}{\Sigma(y_l-\bar{y})^2}. \quad (180)$$

Связь между коэффициентами регрессии и корреляции. Сравнивая формулы (180) и (144), видим: в нх числителе одна и та же величина $\Sigma(y_i-\bar{y})(x_i-\bar{x})$, что указывает на наличие связи между этими показателями. Эта связь выражается равенством $r^2_{xy}=b_{yx}b_{xy}$, или

$$r_{xy}=\sqrt{b_{yx}b_{xy}}. \quad (181)$$

Коэффициент корреляции равен средней геометрической из коэффициентов b_{yx} и b_{xy}. Формула (181) позволяет, во-первых, по известным значениям коэффициентов регрессии b_{yx} и b_{xy} определять коэффициент корреляции r_{xy}, а во-вторых, проверять правильность расчета этого показателя корреляционной связи r_{xy} между варьирующими признаками X и Y.

Так, используя известные коэффициенты регрессии удоя коров Y по массе их тела X и массы тела коров по их удою ($b_{yx}=0{,}0554$ и $b_{xy}=4{,}934$), определяем коэффициент корреляции между этими признаками: $r_{xy}=\sqrt{4{,}934\cdot 0{,}0554}=\sqrt{0{,}273}=0{,}523$. Полученная величина совпадает с той, которая была вычислена по формуле (152).

Как и коэффициент корреляции, коэффициент регрессии характеризует только линейную связь и сопровождается знаком плюс при положительной и знаком минус при отрицательной связи.

Определение параметров линейной регрессии. Известно, что сумма квадратов отклонений вариант x_l от их средней $\bar{x}$ есть величина наименьшая, т. е. $\Sigma(x_i-\bar{x})^2=\min$ (см. гл. III). Эта теорема составляет основу метода наименьших квадратов (см. ниже). В отношении линейной регрессии [см. формулу (176)] требованию этой теоремы удовлетворяет некоторая система уравнений, называемых *нормальными*:

$$an+b\Sigma x=\Sigma y;$$
$$a\Sigma x+b\Sigma x^2=\Sigma xy.$$

Совместное решение этих уравнений относительно параметров a и b приводит к следующим результатам:

$$D=\begin{vmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{vmatrix}=n\Sigma x^2-(\Sigma x)^2;\quad A=\begin{vmatrix} \Sigma y & \Sigma x \\ \Sigma xy & \Sigma x^2 \end{vmatrix}=\Sigma y\Sigma x^2-\Sigma x\Sigma xy;\quad B=\begin{vmatrix} n & \Sigma y \\ \Sigma x & \Sigma xy \end{vmatrix}=n\Sigma xy-\Sigma y\Sigma x,$$

откуда $a=A/D$ и $b=B/D$.

Учитывая двусторонний характер связи между переменными Y и X, формулу для определения параметра a следует выразить так:

$$a_{yx}=\frac{\Sigma y\Sigma x^2-\Sigma x\Sigma yx}{n\Sigma x^2-(\Sigma x)^2};\quad a_{xy}=\frac{\Sigma x\Sigma y^2-\Sigma y\Sigma xy}{n\Sigma y^2-(\Sigma y)^2}\quad\text{или}\quad(182)$$

$$a_{yx}=\bar{y}-b_{yx}\bar{x}\ \text{и}\ a_{xy}=\bar{x}-b_{xy}\bar{y}^1.\qquad(183)$$

Параметр b, или коэффициент регрессии, определяют по следующим формулам:

$$b_{yx}=\frac{n\Sigma xy-\Sigma x\Sigma y}{n\Sigma x^2-(\Sigma x)^2};\quad b_{xy}=\frac{n\Sigma xy-\Sigma x\Sigma y}{n\Sigma y^2-(\Sigma y)^2};\qquad(184)$$

$$b_{yx}=\frac{\Sigma xy-\Sigma x\Sigma y/n}{\Sigma x^2-(\Sigma x)^2/n};\quad b_{xy}=\frac{\Sigma xy-\Sigma x\Sigma y/n}{\Sigma y^2-(\Sigma y)^2/n};\qquad(185)$$

$$b_{yx}=\frac{\Sigma xy-n\bar{x}\bar{y}}{\Sigma x^2-n\bar{x}^2};\quad b_{xy}=\frac{\Sigma xy-n\bar{x}\bar{y}}{\Sigma y^2-n\bar{y}^2}.\qquad(186)$$

Пример 2. В табл. 96 содержатся данные о корреляционной зависимости между массой тела гамадрилов-матерей X и массой тела их новорожденных детенышей Y. Воспользуемся этими данными и найдем эмпирическое уравнение регрессии Y по X. Здесь аргументом, или независимой переменной, служит масса тела матерей, а зависимой переменной — масса тела новорожденных детенышей. В табл. 96 приведены нужные значения: $n=20$; $\Sigma y=14{,}06$; $\Sigma x=237{,}4$; $\Sigma xy=167{,}919$ и $\Sigma x^2=2861{,}60$. Определим параметры линейной регрессии Y по X:

$$b_{yx}=\frac{20\cdot167{,}919-237{,}4\cdot14{,}06}{20\cdot2861{,}6-(237{,}4)^2}=\frac{20{,}536}{873{,}240}=0{,}0235.$$

$$a_{yx}=\bar{y}-b_{yx}\bar{x}=\frac{\Sigma y}{n}-b_{yx}\frac{\Sigma x}{n}=\frac{14{,}06}{20}-0{,}0235\,\frac{237{,}4}{20}=$$

$$=0{,}703-0{,}0235\cdot11{,}87=0{,}703-0{,}279=0{,}424.$$

Отсюда эмпирическое уравнение регрессии массы тела детенышей $\bar{y}_x$ по значениям массы тела их матерей x_i оказывается следующим: $\bar{y}_x=0{,}42+0{,}024\,x$.

Это означает, во-первых, что с увеличением массы тела матерей-гамадрилов на 1 кг масса тела новорожденных детенышей увеличивается в среднем на 0,024 кг. Во-вторых, подставляя в это уравнение вместо x конкретные значения, т. е. массу тела гамадрилов-матерей, можно определить вероятную (среднюю) массу новорожденных детенышей. Так, если масса тела

[1] Формулы (183) применимы лишь для определения свободного члена a линейной регрессии. При наличии нелинейной регрессии эти формулы применять нельзя.

Рост стоя X, см	Обхват груди Y, см	72,5— —74,4	74,5— —76,4	76,5— 78,5—	78,5— —80,4	80,5— —82,4
	Центральные точки классовых интервалов	73,5	75,5	77,5	79,5	81,5
175,5—177,4	176,5					
173,5—175,4	174,5			1	1	4
171,5—173,4	172,5				1	3
169,5—171,4	170,5			1	2	3
167,5—169,4	168,5			1	3	10
165,5—167,4	166,5			2	2	11
163,5—165,4	164,5		1	2	8	13
161,5—163,4	162,5		2	3	12	9
159,5—161,4	160,5			2	8	18
157,5—159,4	158,5		1	4	7	8
155,5—157,4	156,5	2	1	3	7	7
153,5—155,4	154,5	1	1		8	3
151,5—153,4	152,5		2	1	2	1
149,5—151,4	150,5				1	1
147,5—149,4	148,5			1		
f_y		3	8	21	62	91
Средний рост для обхвата груди $\bar{x}_y$		155,8	158,0	161,4	160,8	163,3

Таблица 116

82,5— —84,4	84,5— —86,4	86,5— —88,4	88,5— —90,4	90,5— —92,4	92,5— —94,4	94,5— —96,4	f_x	Средний обхват груди для роста $\bar{y}_x$
83,5	85,5	87,5	89,5	91,5	93,5	95,5		
	1	1	**3**	1			6	88,8
3	4	2	5	2	1		23	85,8
11	9	13	7	7	1	1	53	86,9
9	13	19	14	4	2	2	69	87,0
9	14	20	10	5	7	1	80	86,7
16	16	27	10	7	2	3	96	86,3
14	14	17	14	10	2		95	85,6
24	21	14	7	3			95	84,2
20	13	11	5	1			78	83,9
14	10	12	2	2			60	83,8
3	8	5					36	82,1
1	1	1		1			17	81,0
4	3						13	81,5
2							4	82,0
	1						2	81,5
130	128	142	77	43	15	7	727	
163,9	164,6	165,8	167,6	167,2	168,6	168,8	—	

самки гамадрила равна 12 кг, то ожидаемая масса тела новорожденного детеныша будет следующей: $\bar{y}_x = 0{,}42 + 0{,}024 \cdot 12 = 0{,}71$ кг. Для самки с массой тела в 14 кг ожидаемая масса тела новорожденного детеныша будет составлять $\bar{y}_x = 0{,}42 + 0{,}024 \cdot 14 = 0{,}76$ кг и т. д.

Построение эмпирических рядов регрессии. При наличии большого числа наблюдений регрессионный анализ начинается с построения эмпирических рядов регрессии. *Эмпирический ряд регрессии* образуется путем вычисления по значениям одного варьирующего признака X средних значений $\bar{y}_x$ другого, связанного корреляционно с X признака Y. Иными словами, построение эмпирических рядов регрессии сводится к нахождению групповых средних $\bar{y}_x$ и $\bar{x}_y$ из соответствующих значений признаков Y и X. Рассмотрим технику построения рядов регрессии на соответствующем примере.

Пример 3. На численно большой группе мужчин ($n = 727$) изучали корреляцию между длиной тела и обхватом груди. Собранные данные (по А. А. Малиновскому, 1948) сгруппированы в виде корреляционной табл. 116. Эллипсоидный характер распределения частот f_{xy} по ячейкам корреляционной решетки указывает на наличие положительной хотя и не очень тесной, связи между этими признаками. В нижней строке этой таблицы помещен ряд регрессии роста мужчин X по обхвату их груди Y, а в правом крайнем столбце той же таблицы содержится эмпирический ряд регрессии обхвата груди Y по росту мужчин. Эти ряды есть не что иное, как групповые средние $\bar{x}_y$ и $\bar{y}_x$, вычисленные для каждого столбца и каждой строки корреляционной таблицы. Так, величина $\bar{y}_x = 81{,}5$, что находится внизу последнего столбца табл. 116, получена следующим образом:

$$\bar{y}_x = \frac{1 \cdot 77{,}5 + 1 \cdot 85{,}5}{2} = 81{,}5.$$

Стоящая над ней величина $\bar{y}_x = 82{,}0$ вычислена аналогичным способом: $\bar{y}_x = \frac{1 \cdot 79{,}5 + 1 \cdot 81{,}5 + 2 \cdot 83{,}5}{4} = \frac{328}{4} = 82{,}0$. Так же рассчитаны и групповые средние роста мужчин по обхвату их груди, помещенные в нижней строке табл. 116. Например, величина $\bar{x}_y = 155{,}8$ (первая в нижней строке таблицы) вычислена так: $\bar{x}_y = \frac{1 \cdot 154{,}3 + 2 \cdot 156{,}5}{3} = 155{,}8$. Следующая величина $\bar{x}_y = 158{,}0$ получена таким же способом: $\bar{x}_y = (1/8) \cdot (2 \cdot 152{,}5 + 1 \cdot 154{,}5 + 1 \cdot 156{,}5 + 1 \cdot 158{,}5 + 2 \cdot 162{,}5 + 1 \cdot 164{,}5) = 158{,}0$ и т. д.

Из табл. 116 видно, что эмпирический ряд регрессии — это двойной ряд чисел, которые можно изобразить точками на плоскости, а затем, соединив эти точки отрезками прямой, получить эмпирическую линию регрессии. Эмпирические ряды регрессии, особенно их графики, называемые *линиями регрессии,*

дают наглядное представление о форме и тесноте корреляционной зависимости между варьирующими признаками. На рис. 27 изображены эмпирические и выровненные по уравнению (177) линии регрессии окружности груди y по росту x и роста по окружности груди мужчин. Видно, что они неплохо согласуются между собой.

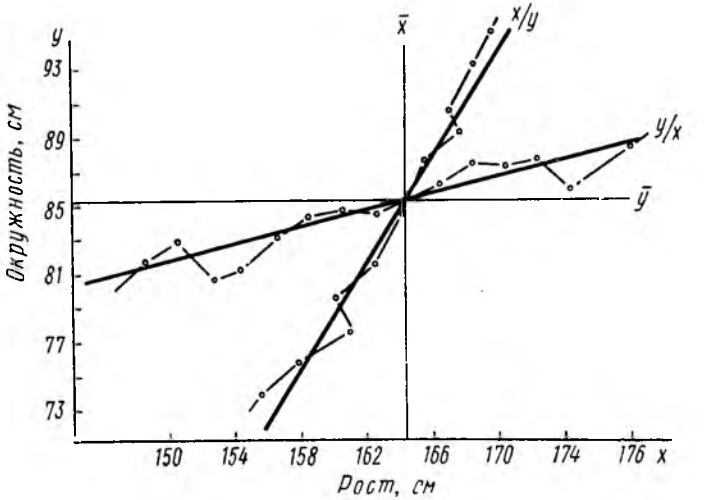


Рис. 27. Эмпирические и вычисленные по способу наименьших квадратов линии регрессии окружности груди мужчин Y по росту X и роста по окружности груди

Пример 4. Выше было найдено (пример 1), что между годовым удоем Y и массой тела X коров горбатовской породы существует положительная связь. Рассчитав усредненные значения $\bar{y}_x$ и $\bar{x}_y$ эмпирических рядов регрессии Y по X и X по Y, можно построить график аналогично тому, как это показано в примере 3. Значения $\bar{y}_x$ и $\bar{x}_y$ содержатся в табл. 102. На основании этих данных построена эмпирическая (ломаная) линия регрессии. Наглядное представление о ней дает рис. 28, на котором наряду с эмпирической изображена и выровненная (плавно идущая) линия регрессии. Последняя рассчитана по уравнению $\bar{y}_x = 4{,}934y + 510{,}98$. Читателю предлагается рассчитать это уравнение, используя предварительно найденные величины: $r_{xy} = 0{,}523$; $s_x = 3{,}27$; $s_y = 2{,}843$; $\lambda_x = 14$; $\lambda_y = 152$.

Выравнивание эмпирических рядов регрессии. Графики эмпирических рядов регрессии оказываются, как правило, не плавно идущими, а ломаными линиями (см. рис. 27 и 28). Это объясняется тем, что наряду с главными причинами, определяющими общую закономерность в изменчивости коррелируемых признаков, на их величине сказывается влияние многочисленных

второстепенных причин, вызывающих случайные колебания узловых точек регрессии. Чтобы выявить основную тенденцию (тренд) сопряженной вариацни коррелируемых признаков, нужно заменить ломаные линии на гладкие, плавно идущие линии регрессии. Процесс замены ломаных линий на плавно идущие называют *выравниванием эмпирических рядов и линий регрессии.*

Графический способ выравнивания. Это наиболее простой способ, не требующий вычислительной работы. Его сущность сводится к следующему. Эмпирический ряд регрессии изображают в виде графика в системе прямоугольных координат. Затем *визуально* намечаются срединные точки регрессии, по которым с помощью линейки или лекала проводят сплошную линию. Недостаток этого способа очевиден: он не исключает влияние индивидуальных свойств исследователя на результаты выравниваиия эмпирических линий регрессии. Поэтому в тех случаях, когда необходима более высокая точность при замене ломаных линий регрессии на плавно идущие, используют другие способы выравнивания эмпирических рядов.

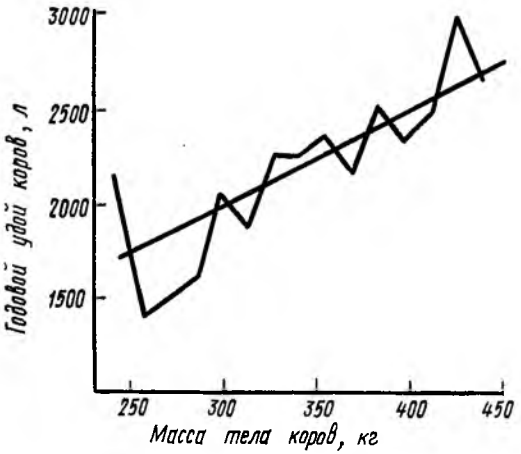


Рис. 28. Эмпирическая и вычислеиная линии регрессии годового удоя по массе тела коров горбатовской породы

Способ скользящей средней. Суть этого способа сводится к последовательиому вычислению средних арифметических из двух или трех соседних членов эмпирического ряда. Этот способ особенно удобен в тех случаях, когда эмпирический ряд представлен большим числом членов, так что потеря двух из них — крайних, что неизбежно при этом способе выравнивания, заметно не отразится на его структуре.

Пример 5. Изучали зависимость между содержанием жира и массой зерен у овса. Результаты приведены ниже:

Классы по содержанию жира в зериах x, % . .	4,5 — 5,0	5,0 — 5,5	5,5 — 6,0	6,0 — 6,5	6,5 — 7,0	7,0 — 7,5	7,5 — 8,0	8,0 — 8,5
Масса зерен, $\bar{y}_x$, мг	45,0	45,8	44,3	41,9	40,1	39,0	37,5	37,5

Чтобы выровнять этот ряд, находим сумму первых двух членов: 45,0+45,8=90,8. Затем определяем сумму следующих двух членов: 45,8+44,3=90,1; 44,3+41,9=86,2 и так до конца ряда. Затем каждую полученную таким образом сумму делим на число слагаемых, в данном случае на два, и находим усредненные значения членов ряда: 45,4 45,0 43,1 41,0 39,6 38,2 37,5. Получился выровненный ряд, более наглядно свидетельствующий о наличии отрицательной корреляции между этими признаками.

Разумеется, в разных случаях способ скользящей средней применяют по-разному, вычисляя средние не из двух или трех, но и большего числа членов ряда.

Метод наименьших квадратов. Этот способ предложен в начале XIX столетия А. М. Лежандром и независимо от него К. Гауссом. Он позволяет наиболее точно выравнивать эмпирические ряды. Этот метод, как было показано выше, основан на предположении, что сумма квадратов отклонений вариант x_i от их средней $\bar{x}$ есть величина минимальная, т. е. $\sum_{i=1}^{n}(x_i-\bar{x})^2=\min$. Отсюда и название метода, который применяют не только в биологии, но и в технике. Метод наименьших квадратов объективен и универсален, его применяют в самых различных случаях при отыскании эмпирических уравнений рядов регрессии и определении их параметров.

Требование метода наименьших квадратов заключается в том, что теоретические точки линии регрессии $\bar{y}_x'$ должны быть получены таким образом, чтобы сумма квадратов отклонений от этих точек для эмпирических наблюдений y_i была минимальной, т. е.

$$Q=\sum(y_i-y_x)^2=\sum(y_i-f(x))^2=Q_{\min}.$$

Вычисляя в соответствии с принципами математического анализа минимум этого выражения и определенным образом преобразуя его, можно получить систему так называемых *нормальных уравнений,* в которых неизвестными величинами оказываются искомые параметры уравнения регрессии, а известные коэффициенты определяются эмпирическими величинами признаков, обычно суммами их значений и их перекрестных произведений. В частности, такая система нормальных уравнений, полученная для прямолинейного уравнения регрессии, приведена выше, в разд. IX.1.

Сущность метода практически уже раскрыта на конкретных примерах, которые рассмотрены выше — при отыскании параметров a и b линейной регрессии. При этом расчет параметров

производили непосредственно по значениям варьирующих признаков Y и X (малые выборки).

Теперь следует выяснить, как применяется этот метод к выборкам, группируемым в вариационные ряды и корреляционные таблицы (большие выборки). Начнем с отыскания эмпирических уравнений регрессии обхвата груди Y по росту X и роста по обхвату груди мужчин. Чтобы решить эту задачу, необходимо предварительно рассчитать средние арифметические $\bar{y}$ и $\bar{x}$, средние квадратические отклонения s_y и s_x и вычислить коэффициент корреляции r_{xy} между этими признаками. Читателю предлагается (по примеру расчета r_{xy} между массой тела и годовым удоем коров горбатовской породы) вычислить эти величины, которые оказались равными $\bar{y}=85{,}17$; $\bar{x}=164{,}62$; $s_y=2{,}02$; $s_x=2{,}73$ и $r_{xy}=0{,}391$.

Переходим к определению параметров регрессии обхвата груди Y по росту X и роста по обхвату груди мужчин. Так как $\lambda_y=\lambda_x=2$, то эти величины можно не учитывать при определении параметров b_{yx} и b_{xy} [см. формулу (178)]:

$$b_{yx}=r_{xy}\frac{s_y}{s_x}=0{,}391\frac{2{,}02}{2{,}73}=0{,}289 \text{ и}$$

$$b_{xy}=r_{xy}\frac{s_x}{s_y}=0{,}391\frac{2{,}73}{2{,}02}=0{,}528;$$

$$a_{yx}=\bar{y}-b_{yx}\bar{x}=85{,}17-0{,}289\cdot 164{,}62=85{,}17-47{,}58=37{,}59$$

и $$a_{xy}=\bar{x}-b_{xy}\bar{y}=164{,}62-0{,}528\cdot 85{,}17=164{,}62-44{,}97=119{,}65.$$

Отсюда эмпирическое уравнение регрессии обхвата груди по росту $\bar{y}_x=0{,}289x+37{,}59$, а эмпирическое уравнение роста по обхвату груди $\bar{x}_y=0{,}528y+119{,}65$.

Сумма членов ряда $\bar{y}_x'$ регрессии, рассчитанных по корреляционному уравнению, должна быть равна сумме членов эмпирического ряда, т. е. $\Sigma\bar{y}_x=\Sigma\bar{y}_x'$. Если окажется, что $\Sigma\bar{y}_x'\neq\Sigma\bar{y}_x$ или $\Sigma\bar{x}_y'\neq\Sigma\bar{x}_y$ (как следствие приближенных вычислений параметров), нужно эмпирические уравнения регрессии скорректировать так, чтобы указанные равенства осуществлялись. В данном случае этому условию удовлетворяют уравнения $\bar{y}_x=0{,}289x+37{,}5$ и $\bar{x}_y=0{,}528y+119{,}1$. Рассчитанные по этим уравнениям значения $\bar{y}_x$ и $\bar{x}_y$ изображены в виде плавно идущих (сглаженных) линий регрессии Y по X и X по Y на рис. 28. Они неплохо согласуются с эмпирическими (ломаными) линиями регрессии.

Уравнение линейной регрессии можно выразить в виде отклонений членов ряда от их средних:

$$\bar{y}_x-\bar{y}=b_{yx}(x-\bar{x});\quad \bar{x}_y-\bar{x}=b_{xy}(y-\bar{y}). \qquad (187)$$

Система нормальных уравнений в этом случае будет выглядеть так:

$$an+b\sum(x_i-\bar{x})=\sum(y_i-\bar{y});$$

$$a\sum(x_i-\bar{x})+b\sum(x_i-\bar{x})^2=\sum(y_i-\bar{y})(x_i-\bar{x}).$$

Так как $\Sigma(y_i-\bar{y})=0$ и $\Sigma(x_i-\bar{x})=0$, то параметр определяют по формуле (187), а параметр a легко найти по формуле (183).

Если средние $\bar{y}$ и $\bar{x}$ перенести в правую часть уравнения (187), то получим

$$\bar{y}_x=\bar{y}+b_{yx}(x_i-\bar{x});\quad \bar{x}_y=\bar{x}+b_{xy}(y_i-\bar{y}). \qquad (188)$$

Система нормальных уравнений для определения параметров a и b будет следующая:

$$an+b\sum(x_i-\bar{x})=\sum y;$$

$$a\sum(x_i-\bar{x})+b\sum(x_i-\bar{x})^2=\sum y\,(x_i-\bar{x}).$$

Так как $\Sigma(x-\bar{x})=0$, то система уравнений оказывается такой:

$$an=\sum y;$$

$$b\sum(x_i-\bar{x})^2=\sum y\,(x_i-\bar{x}).$$

Отсюда параметры уравнения линейной регрессии, выраженной в виде отклонений члеиов ряда от их средних величин, оказываются следующими:

$$a=\frac{\sum y}{n}=\bar{y}; \qquad (189)$$

$$b=\frac{\sum y\,(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}. \qquad (190)$$

Эти формулы особенно удобны при определении параметров эмпирических уравнений рядов динамики (см. ниже).

Множественная линейная регрессия. Зависимость между несколькими переменными величинами принято выражать уравнением множественной регрессии, которая может быть *линейной* и *нелинейной*. В простейшем виде множественная линейная регрессия выражается уравнением с двумя независимыми переменными величинами (x, z):

$$y=a+bx+cz, \qquad (191)$$

где a — свободный член уравнения; b и c — параметры уравнения. Для нахождения параметров этого уравнения (по спосо-

бу наименьших квадратов) применяют следующую систему нормальных уравнений:

$$an+b\sum x+c\sum z=\Sigma y;$$

$$a\sum x+b\sum x^2+c\sum xz=\sum xy;$$

$$a\sum z+b\sum xz+c\sum z^2=\sum yz.$$

Чтобы по эмпирическим данным составить такую систему, необходимо предварительно рассчитать Σx, Σy, Σyz, Σxz, Σx^2 и Σz^2.

Пример 6. Найти эмпирическое уравнение регрессии между числом колосков y, количеством зерен z и длиной колосьев X у озимой ржи. Данные о корреляционной зависимости между этими признаками приведены в табл. 115. Объем выборки $n=10$. Предполагая линейный характер связи между этими признаками и учитывая их буквенные обозначения, возьмем за исходное уравнение регрессии уравнение вида

$$x=a+by+cz,$$

которому отвечает выше приведенная система нормальных уравнений. Необходимые суммы см. в табл. 115. Подставляем их в уравнения системы:

$$10a+165b+294c=575;$$

$$165a+2891b+5202c=9908;$$

$$294a+5202b+9456c=17816.$$

Чтобы решить эту систему относительно параметров a, b и c, разделим каждое уравнение на коэффициент при a, что дает:

$$a+16{,}5000b+29{,}4000c=57{,}5000; \quad \text{(I)}$$

$$a+17{,}5212b+31{,}5273c=60{,}0485; \quad \text{(II)}$$

$$a+17{,}6939b+32{,}1633c=60{,}5986. \quad \text{(III)}$$

Затем, вычитая первое уравнение из второго, а второе — из третьего, получим

$$1{,}0212b+2{,}1273c=2{,}5485;$$

$$0{,}1727b+0{,}6360b=0{,}5501.$$

Разделим каждое уравнение на коэффициент при b и найдем разность между полученными уравнениями:

$$\begin{array}{r} b+2{,}0831c=2{,}4956 \\ -\quad b+3{,}6827c=3{,}1853 \\ \hline -1{,}5996c=-0{,}6897 \end{array}$$

Отсюда $c=\frac{-0,6897}{-1,5996}=0,4312$. Подставляя в одно из этих уравнений вместо c его значение, находим $b+2,0831(0,4312)=2,4956$, откуда $b=2,4956-0,8982=1,5974$.

Наконец, в первое (исходное) уравнение вместо b и c подставляем их значения: $10a+165(1,5974)+294(0,4312)=575$. Отсюда $a=\frac{575-390,3438}{10}=\frac{184,6562}{10}=18,466$. В итоге

$$\bar{x}_y=18,466+1,597y+0,431z.$$

Подставляя в это уравнение задаваемые значения переменных y и z, можно определить ожидаемую величину переменной x, т. е. среднюю длину колосьев этой культуры. Так, для $y=10$ и $z=8$ $\bar{x}_y=18,466+10(1,597)+8(0,431)=37,334\approx37,9$ см; для $y=15$ и $z=14$ средняя длина колоса $\bar{x}_y=18,466+15(1,597)+14(0,431)=48,455\approx48,5$ см и т. д.

Найденное эмпирическое уравнение регрессии показывает, что при изменении длины колосьев X на 1 см число колосков Y при постоянном количестве зерен Z изменится в среднем на 1,60, а число Z при постоянной величине Y изменится в среднем на 0,43.

Ряды динамики. Выравнивание рядов. Изменение признаков во времени образует так называемые *временные ряды* или *ряды динамики*. Характерной особенностью таких рядов является то, что в качестве независимой переменной X здесь всегда выступает фактор времени, а зависимой Y — изменяющийся признак. В отличие от рядов регрессии зависимость между переменными X и Y носит односторонний характер, так как фактор времени не зависит от изменчивости признаков. Несмотря на указанные особенности, ряды динамики можно уподобить рядам регрессии и обрабатывать их одними и теми же методами.

Как и ряды регрессии, эмпирические ряды динамики несут на себе влияние не только основных, но и многочисленных второстепенных (случайных) факторов, затушевывающих ту главную тенденцию в изменчивости признаков, которая на языке статистики называется *трендом*.

Анализ рядов динамики начинается с выявления формы тренда. Для этого временной ряд изображают в виде линейного графика в системе прямоугольных координат. При этом по оси абсцисс откладывают временные точки (годы, месяцы и другие единицы времени), а по оси ординат — значения зависимой переменной Y. При наличии линейной зависимости между переменными X и Y (т. е. линейного тренда) для выравнивания ряда динамики способом наименьших квадратов наиболее подходящим является уравнение регрессии в виде отклонений чле-

нов ряда зависимой переменной Y от средней арифметической ряда независимой переменной X [см. формулу (188)].

Пример 7. Наблюдения над физическим развитием макак-резусов в первый год их жизни показали, что масса тела малышей увеличивается с возрастом по закону линейной функции. В этом легко убедиться, если результаты наблюдений над развитием макак-резусов изобразить в виде линейного графика в системе прямоугольных координат.

Соответствующие данные приведены в табл. 117.

Таблица 117

Возраст x_i, мес	Масса тела y_i, кг	$(x_i-\bar{x})$	$y(x_i-\bar{x})$	$(x_i-\bar{x})^2$	$\bar{y}_x$
1	0,53	—5,5	—2,915	30,25	0,59
2	0,71	—4,5	—3,195	20,25	0,70
3	0,79	—3,5	—2,765	12,25	0,81
4	0,98	—2,5	—2,450	6,25	0,92
5	1,06	—1,5	—1,590	2,25	1,03
6	1,13	—0,5	—0,565	0,25	1,14
7	1,25	+0,5	+0,625	0,25	1,26
8	1,43	+1,5	+2,145	2,25	1,37
9	1,51	+2,5	+3,775	6,25	1,45
10	1,59	+3,5	+5,565	12,25	1,59
11	1,65	+4,5	+7,425	20,25	1,70
12	1,77	+5,5	+9,735	30,25	1,81
$\Sigma=78$	14,40	—	15,790	143,00	14,40

Определяем среднюю арифметическую ряда независимой переменной: $\bar{x}=78/12=6,5$. Эту величину можно получить и по полусумме крайних значений ряда: $\bar{x}=(1+112)/2=6,5$. Отклонения от этой величины членов ряда зависимой переменной (с учетом знаков) помещены в третьей графе табл. 117. Остальные действия понятны из этой таблицы.

Подставляя известные величины в формулы (189) и (190), определяем параметры линейной регрессии: $a=14,40/12=1,20$ и $b=15,790/143,00=0,1104$. Отсюда эмпирическое уравнение массы тела Y по возрасту X детенышей макак-резусов

$$\bar{y}_y=1,20+0,1104(x_i-\bar{x}).$$

Подставляя вместо $(x_i-\bar{x})$ их значения, находим ожидаемые (выравнивающие) значения зависимой переменной Y:

$$\bar{y}_x=1,20+0,1104\,(-5,5)=1,20-0,61=0,59;$$

$$\bar{y}_x=1,20+0,1104\,(-4,5)=1,20-0,50=0,70 \text{ и т. д.}$$

Рассчитанные таким образом значения $\bar{y}_x$ приведены в последнем столбце табл. 117. Видно, что они хорошо согласуются с эмпирически найденными значениями этого ряда.

Таблица 11

Временные точки: годы	Временные точки: x_i	Процент отличных оценок y_i	$(x_i-\bar{x})$	$y(x_i-\bar{x})$	$(x_i-\bar{x})^2$	$\bar{y}_x$
1962	1	22	—4	—88	16	20,9
1963	2	28	—3	—84	9	22,8
1964	3	16	—2	—32	4	24,7
1965	4	28	—1	—28	1	26,6
1966	5	34	0	0	0	28,4
1967	6	22	+1	+22	1	30,3
1968	7	30	+2	+60	4	32,2
1969	8	41	+3	+123	9	34,1
1970	9	35	+4	+140	16	36,0
Сумма	45	256	—	113	60	256,0

Пример 8. На протяжении 9-летнего периода обучения процент отличных оценок, получаемых студентами на экзаменационных сессиях по курсу дарвинизма, колебался следующим образом (табл. 118).

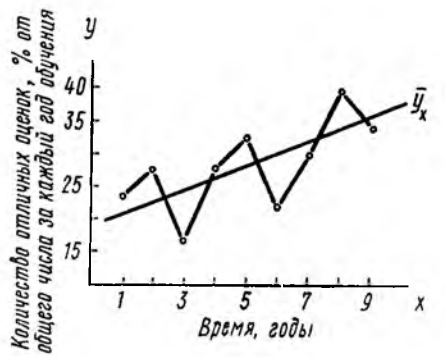


Рис. 29. Эмпирическая и вычисленная линии регрессии отличных оценок учащихся за девятилетний срок обучения

В данном случае среднюю арифметическую для независимой переменной определяем по временным точкам, обозначенным числами натурального ряда: $\bar{x}=(1+9)/2=5$. Затем, как и в предыдущем примере, берем отклонения членов ряда зависимой переменной Y от этой величины (тоже с учетом знаков!) и производим операции, показанные в табл. 118. Подставляя найденные значения в формулы (189) и (190), определяем параметры линейной регрессии: $a=256/9=28{,}44$; $b=113/60=1{,}883$. Отсюда эмпирическое уравнение ряда динамики отличных оценок знаний студентов по курсу дарвинизма за десятилетний период оказывается следующим: $\bar{y}_x=28{,}44+1{,}883(x_i-\bar{x})$

Рассчитанные по этому уравнению значения $\bar{y}_x$ зависимой переменной помещены в последнем столбце табл. 118. Они неплохо согласуются с эмпирическими членами этого ряда. Более наглядное представление об этом дает рис. 29, на котором изображены эмпирическая (ломаная) и вычисленная (плавно идущая) линии регрессии этого ряда.

Числовые характеристики рядов динамики. К числу основных обобщающих числовых характеристик рядов динамики относят *среднюю геометрическую* $\bar{x}_g$ и близкую к ней *среднюю арифметическую* $\bar{x}$ величины, о которых речь шла в гл. II. Они характеризуют среднюю скорость, с какой изменяется величина зависимой переменной за определенные периоды времени. Так, судя по данным табл. 117, средняя месячная прибавка массы тела макак-резусов за первый год их жизни определяется следующим образом [см. формулу (12)]:

$$\lg \bar{x}_g = \frac{\lg x_n - \lg x_0}{n-1} = \frac{\lg 117 - \lg 53}{12-1} = \frac{2{,}24797 - 1{,}72428}{11} =$$

$$= \frac{0{,}52369}{11} = 0{,}047608.$$

Отсюда $\bar{x}_g = 0{,}11$ мг. Эта величина получается и при вычислении средней арифметической $\bar{x}$ из месячных абсолютных прибавок массы тела макак-резусов за первый год их жизни (читателю предлагается рассчитать эту величину).

Оценкой изменчивости членов ряда динамики служит *среднее квадратическое отклонение*. Примеры такой оценки будут рассмотрены ниже. При выборе уравнений регрессии для описания рядов динамики учитывают форму тренда, которая может быть линейной (или приведена к линейной) и нелинейной. О правильности выбора уравнения регрессии обычно судят по сходству эмпирически наблюденных и вычисленных значений зависимой переменной. Более точным в решении этой задачи является метод дисперсионного анализа регрессии (см. ниже).

Корреляция рядов динамики. Нередко приходится сопоставлять динамику параллельно идущих временных рядов, связанных друг с другом некоторыми общими условиями, например выяснять связь между производством сельскохозяйственной продукции и ростом поголовья скота за определенный промежуток времени, определять влияние агротехники возделывания сельскохозяйственных культур на их урожайность и т. д. В таких случаях характеристикой связи между переменными X и Y служит *коэффициент корреляции* r_{xy} (при наличии линейного тренда).

Известно, что главное направление изменчивости, или тренд рядов динамики, как правило, затушевывается колебаниями членов ряда зависимой переменной Y. Отсюда возникает задача двоякого рода: измерение зависимости между сопоставляе-

мыми рядами, не исключая тренд, и измерение зависимости между соседними членами одного и того же ряда, исключая тренд. В первом случае показателем тесноты связи между сопоставляемыми рядами динамики служит *коэффициент корреляции* (если связь линейна), во втором — *коэффициент автокорреляции*. Эти показатели имеют разные значения, хотя и вычисляются по одним и тем же формулам [см. формулы (144), (145) и др.].

Таблица 11

Временные точки, годы	Площадь черного пара x_i, га	Собрано зерна y_i, т	$x^*=x_i-260$	$y^*=y_i-50$	x^*y^*	$(x^*)^2$	$(y^*)^2$
1	154	25	—106	—25	2650	11236	625
2	158	28	—102	—22	2244	10404	484
3	216	43	—44	—7	308	1936	49
4	280	64	+20	+14	280	400	196
5	325	55	+65	+5	325	4225	25
6	340	68	+89	+18	1440	6400	324
7	354	79	+94	+29	2726	8836	841
8	350	82	+90	+32	2880	8100	1024
Сумма	—	—	97	44	12853	51537	3568

Пример 9. В табл. 119 приведены данные об увеличении за 8 лет черного пара в одном из колхозов РСФСР и сборе зерна пшеницы с паровых полей. Вычислим коэффициент корреляции между этими рядами исходя из того, что зависимость между ними следует закону линейной регрессии. Чтобы упростить расчеты, каждый член ряда независимой переменной X уменьшим на 260, а члены ряда зависимой переменной Y — на 50. Такого рода преобразование чисел не сказывается на значении коэффициента корреляции, которое будет одним и тем же при вычислении его по значениям x_i и y_i или же по преобразованным значениям $x^*=x_i-260$ и $y^*=y_i-50$.

Применим формулу (147) и предварительно рассчитаем:

$$\frac{\sum x \sum y}{n}=\frac{97\cdot 44}{8}=533{,}5;\quad D_x=\sum x^2-\frac{(\sum x)^2}{n}=51537-\frac{97^2}{n}=$$

$$=50361;\quad D_y=\sum y^2-\frac{(\sum y)^2}{n}=3568-\frac{44^2}{8}=3326.\quad \sqrt{D_xD_y}=$$

$$=\sqrt{50361\cdot 3326}=12942{,}2.\quad \text{Отсюда}\quad r_{xy}=\frac{\sum xy-(\sum x\sum y)/n}{\sqrt{D_xD_y}}=$$

$$=\frac{12\,853-533{,}5}{12942{,}2}=\frac{12319{,}5}{12942{,}2}=0{,}952.$$

Это довольно высокий показатель, свидетельствующий о весьма сильной положительной

вязи между количеством собранного зерна пшеницы и увеличением парового клина в общей структуре посевных площадей колхоза.

Вычислим коэффициент автокорреляции как меру сопряженности между членами одного и того же ряда динамики. Для этого необходимо сдвинуть члены ряда на принятую единицу времени, в данном случае равную одному году, что позволит образовать ряды двух переменных Y и X. При этом число парных значений двойного ряда n уменьшается на единицу. Сдвиг ряда динамики на единицу времени оправдывается и тем, что влияние пара на урожай сказывается обычно через год.

Таблица 120

x	y	xy	x^2	y^2
—106	—102	10 812	11 236	10 404
—102	—44	4 488	10 404	1 936
—44	+20	—880	1 936	400
+20	+65	1 300	400	4 225
+65	+80	5 200	4 225	6 400
+80	+94	7 520	6 400	8 836
+94	+90	8 460	8 836	8 100
7	203	36 900	43 437	40 301

Необходимые данные и расчет вспомогательных величин приведены в табл. 120. В данном случае $\Sigma x \Sigma y/n = 7 \cdot 203/7 = 203$; $D_x = 43\,437 - 7^2/7 = 43\,430$; $D_y = 40\,301 - 203^2/7 = 34\,414$. Подставляем эти величины в формулу (147): $r_{xy} = \frac{36\,900 - 203}{\sqrt{43\,430 \cdot 34\,414}} = \frac{36\,697}{38660{,}1} = 0{,}949$. Это означает, что в ряду динамики парового клина между членами независимой переменной X существует высокая положительная автокорреляция.

Нетрудно заметить, что на значении коэффициента автокорреляции сказывается изменчивость членов ряда зависимой переменной: чем меньше члены ряда отклоняются от тренда, тем выше коэффициент автокорреляции, и наоборот. В этом легко убедиться на примерах рядов динамики с разной изменчивостью членов ряда. Так, рассмотренный выше ряд динамики отличных оценок знаний студентов по курсу дарвинизма (см. табл. 118) отличается сильной изменчивостью его членов и характеризуется коэффициентом автокорреляции, равным 0,171, тогда как слабоколеблющийся ряд возрастных изменений мас-

сы тела макак-резусов (см. табл. 117) характеризуется коэффициентом автокорреляции, равным 0,992 (читателю предлагается вычислить эти показатели).

IX.2. НЕЛИНЕЙНАЯ РЕГРЕССИЯ

Регрессия, выражаемая уравнением параболы второго порядка. Как уже было показано, наряду с линейными корреляциями в биологии встречаются и нелинейные корреляции между переменными величиниами. Хорошо известна, например, нелинейная зависимость между сроками лактации и удоем коров, логистическая закономерность возрастания численного состава популяции в замкнутой среде обитания и многие другие явления подобного рода. Все оии отражают те или иные биологические закономерности и могут быть описаиы соответствующими корреляционными уравнениями, формулами или выражены в виде эмпирических или теоретически построенных линий регрессии и динамики.

Нередко зависимость между переменными величинами Y и X выражается *уравнением параболы второго порядка*

$$y=a+bx+cx^2. \qquad (192)$$

Отысканию параметров a, b и c этого уравнения удовлетворяет следующая система нормальных уравнений:

$$an+b\sum x+c\sum x^2=\sum y;$$

$$a\sum x+b\sum x^2+c\sum x^3=\sum xy;$$

$$a\sum x^2+b\sum x^3+c\sum x^4=\sum yx^2.$$

Чтобы решить эту систему относительно параметров a, b и c, нужно предварительно рассчитать Σx, Σy, Σxy, Σx^2, Σyx^2, Σx^3 и Σx^4.

Пример 10. Наблюдеиия показали, что удой Y группы коров ярославской породы изменяется по срокам лактации X следующим образом (табл. 121).

Из табл. 121 видно, что значения зависимой перемеиной Y сначала возрастают, а с седьмого месяца лактации иачинают убывать. Это признак параболической зависимости между переменными Y и X. Найдем эмпирическое уравнеиие этой зависимости. Предварительно рассчитаем вспомогательные величины Σy, Σxy, Σyx^2 и др. Расчет приведен в табл. 121.

Составим систему нормальных уравнений:

$$9a + 45b + 285c = 203{,}3;$$

$$45a + 285b + 2025c = 1030{,}0;$$

$$285a + 2025b + 15\,333c = 6439{,}6.$$

Решая эту систему (описанным выше способом) относительно коэффициентов a, b и c, находим: $a = 13{,}466$; $b = 4{,}587$ и $c = -0{,}436$. Отсюда эмпирическое уравнение параболы второго порядка таково:

$$\bar{y}_x = 13{,}466 + 4{,}587x - 0{,}436x^2.$$

Таблица 121

Лактация x_i, мес.	Удой, y_i, ц	xy	x^2	yx^2	x^3	x^4	$\bar{y}_x$
1	18,2	18,2	1	18,2	1	1	17,6
2	20,1	40,2	4	80,4	8	16	20,9
3	23,4	70,2	9	210,6	27	81	23,3
4	24,6	98,4	16	393,6	64	256	24,8
5	25,6	128,0	25	640,0	125	625	25,5
6	25,9	155,4	36	932,4	216	1296	25,3
7	23,6	165,2	49	1156,4	343	2401	24,2
8	22,7	181,6	64	1452,8	512	4096	22,3
9	19,2	172,8	81	1555,2	729	6561	19,4
$\Sigma = 45$	203,3	1030,0	285	6439,6	2025	15333	203,3

Подставляя в это уравнение вместо x значения независимой переменной X, можно рассчитать ожидаемые величины удоя коров данной группы за любую лактацию:

$$\bar{y}_x = 03{,}466 + 4{,}587 - 0{,}436 = 17{,}6;$$

$$\bar{y}_x = 13{,}466 + 4{,}587 \cdot 2 - 0{,}436 \cdot 2^2 = 20{,}9;$$

$$\bar{y}_x = 13{,}466 + 4{,}587 \cdot 3 - 0{,}436 \cdot 3^2 = 23{,}3 \text{ и т. д.}$$

Эти величины приведены в последнем столбце табл. 121. Они хорошо согласуются с фактическими данными. Более наглядно это показано на рис. 30, где изображены эмпирическая и вычисленная (более плавно идущая) линии регрессии. Равенство $\Sigma y = \Sigma \bar{y}_x$ указывает на то, что расчет значений $\bar{y}_x$ произведен правильно.

Вычисление параметров параболы второго порядка значительно упрощается, если воспользоваться следующими форму-

лами, найденными путем решения системы нормальных уравнений:

$$a=\frac{1}{D}(\sum y\sum x^4-\sum x^2\sum yx^2);\quad b=\frac{\sum xy}{\sum x^2}$$

$$\text{и}\quad c=\frac{1}{D}(n\sum yx^2-\sum x^2\sum y),$$

где $D=n\Sigma x^4-(\Sigma x^2)^2$ — определитель системы; n — число членов ряда регрессии; y_i — значения зависимой переменной Y, а через x обозначены отклонения членов ряда независимой переменной от средней величины, т. е. $x=(x_i-\bar{x})$. Чтобы применить эти формулы, достаточно рассчитать следующие вспомогательные величины: Σy, Σyx, Σx^2, Σyx^2 и Σx^4.

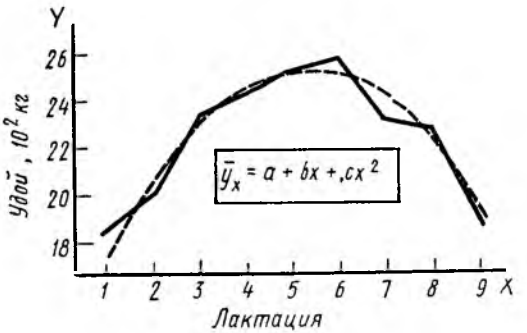


Рис. 30. Эмпирическая и вычисленная по уравнению параболы второго порядка кривые лактации

Пример 11. Воспользуемся данными о связи между удоем Y и сроками лактации X у коров ярославской породы и рассчитаем с помощью указанных формул эмпирическое уравнение этой связи. Расчет вспомогательных величин приведен в табл. 122. Среднюю ($\bar{x}$) вычисляют, как и в предыдущих случаях, по формуле $\bar{x}=\Sigma x/n=45/9=5$. Подставляя известные величины в формулы, находим: $a=\frac{203{,}3\cdot 708-1222{,}1\cdot 60}{9\cdot 708-60\cdot 60}=$ $=\frac{70610{,}4}{2772{,}0}=25{,}4727$; $b=\frac{+13{,}5}{60}=0{,}225$; $c=\frac{9\cdot 1222{,}1-203{,}3\cdot 60}{9\cdot 708-60\cdot 60}=$ $=\frac{-1199{,}1}{2772{,}0}=-0{,}43266$. Отсюда эмпирическое уравнение регрессии удоя коров по срокам лактации, выраженное в отклонениях членов независимой переменной от их средней,

$$\bar{y}_x=24{,}473+0{,}225x-0{,}4326x^2.$$

Подставляя в это уравнение вместо x конкретные значения независимой переменной X, выраженные в виде отклонений их от средней арифметической $\bar{x}$ данного ряда, можно определить ожидаемые значения зависимой переменной Y:

$$\bar{y}_x=25{,}473+0{,}225\,(-4)-0{,}4326\,(16)=25{,}473-7{,}822=17{,}6;$$

$$\bar{y}_x=25{,}473+0{,}225\,(-3)-0{,}4326\,(9)=25{,}473-4{,}568=20{,}9 \text{ и т. д.}$$

Вычисленные таким образом значения $\bar{y}_x$ зависимой переменной помещены в последнем столбце табл. 122. Как и в предыдущем случае, они хорошо согласуются с членами эмпирического ряда.

Таблица 122

Лактация X, мес	Удой Y, ц	$x=\bar{x}-X$	yx	x^2	yx^2	x^4	$\bar{y}_x$
1	18,2	—4	—72,8	16	291,2	256	17,6
2	20,1	—3	—60,3	9	180,9	81	20,9
3	23,4	—2	—46,8	4	93,6	16	23,3
4	24,6	—1	—24,6	1	24,6	1	24,8
5	25,6	0	0	0	0	0	25,5
6	25,9	+1	+25,9	1	25,9	1	25,3
7	23,6	+2	+47,2	4	94,4	16	24,2
8	22,7	+3	+68,1	9	204,3	81	22,2
9	19,2	+4	+76,8	16	307,2	256	19,4
$\Sigma=45$	203,3	—	13,5	60	1222,1	708	203,3

Расчет параметров a, b и c еще более упрощается, если каждый член зависимой переменной Y уменьшить на некоторое произвольно взятое число K, т. е. значения y_i заменить на $y_i^*=y_i-K$. При этом параметр a определяют с поправкой на ве-

Таблица 123

Лактация X	Удой Y	$x=\bar{x}-X$	$y^*=Y-K$	x^2	y^*x	y^*x^2	x^4
1	18,2	—4	—1,8	16	+7,2	—28,8	256
2	20,1	—3	0,1	9	—0,3	0,9	81
3	23,4	—2	3,4	4	—6,8	13,6	16
4	24,6	—1	4,6	1	—4,6	4,6	1
5	25,6	0	5,6	0	0	0	0
6	25,9	+1	5,9	1	+5,9	5,9	1
7	23,6	+2	3,6	4	+7,2	14,4	16
8	22,7	+3	2,7	9	+8,1	24,3	81
9	19,2	+4	—0,8	16	—3,2	—12,8	256
Сумма	—	—	23,3	60	13,5	22,1	708

личину K, которую прибавляют к a^*, т. е. $a=a^*+K$. При вычислении параметров b и c поправки не нужны. Этот прием значительно облегчает вычисление вспомогательных величин, особенно в тех случаях, когда зависимая переменная представлена рядом многозначных чисел.

Пример 12. Как и в предыдущих случаях, воспользуемся фактическими данными о корреляционной зависимости между удоем Y и сроками лактации X у коров ярославской породы и найдем эмпирическое уравнение регрессии Y по X. Предварительно каждый член ряда зависимой переменной Y уменьшим на величину $K=20$, т. е. вместо значений y_i примем $y_i^*=y_i-20$. Исходя из преобразованных таким образом членов ряда зависимой переменной, рассчитаем вспомогательные величины (табл. 123).

Подставляя известные суммы в формулы, находим: $a=a'+K=\frac{23,3\cdot 708-22,1\cdot 60}{9\cdot 708-60\cdot 60}+20=\frac{15170,4}{2772,0}+20=5,4727+20=25,4727$; $b=\frac{13,5}{60}=0,225$; $c=\frac{9\cdot 22,1-23,3\cdot 60}{9\cdot 708-60\cdot 60}=\frac{-1199,1}{2772}=-0,4326$. Получились те же результаты, что и выше.

Если к параметру a прибавить $\left(\frac{c\bar{x}^2}{\lambda}-\frac{b\bar{x}}{\lambda^2}\right)$, параметр b умножить на $\left(\frac{b}{\lambda}-\frac{2c\bar{x}}{\lambda^2}\right)$, а параметр c разделить на квадрат интервала λ^2 между членами ряда независимой переменной, т. е. вместо c взять c/λ^2, то получатся следующие коэффициенты уравнения параболы второго порядка (учитывая, что в данном случае $\lambda=1$): $a'=25,4727+(-0,4326\cdot 25)-(0,225\cdot 5)=25,4727-11,940=13,533$; $b'=0,225-2(-0,4326\cdot 5)=0,225-2(2,163)=4,551$; $c'=-0,43257$. Отсюда эмпирическое уравнение регрессии удоя коров по срокам лактации, в котором переменная x обозначает не отклонения членов ряда X от их средней $\bar{x}$, а непосредственные значения членов этого ряда, будет таковым:

$$\bar{y}_x=13,53+4,55x-0,4326x^2.$$

Подставляя в это уравнение значения независимой переменной X, находим выравнивающие значения зависимой переменной Y:

$$\bar{y}_{x_1}=13,53+4,55\cdot(1)-0,4326\cdot(1^2)=18,08-0,4326=17,6;$$

$$y_{x_2}=13,53+4,55\cdot(2)-0,4326\cdot(2^2)=22,63-1,73=20,9 \text{ и т. д.}$$

Остальные действия объяснены выше.

Регрессия, выражаемая уравнением параболы третьего порядка. Среди различных форм параболической зависимости между переменными величинами встречаются и такие, которые наилучшим образом описываются *уравнением параболы третьего порядка*:

$$y=a+bx+cx^2+dx^3. \qquad (193)$$

Для определения параметров этого уравнения используют следующую систему нормальных уравнений:

$$na+b\sum x+c\sum x^2+d\sum x^2=\sum y;$$
$$a\sum x+b\sum x^2+c\sum x^3+d\sum x^4=\sum xy;$$
$$a\sum x^2+b\sum x^3+c\sum x^4+d\sum x^5=\sum yx^2;$$
$$a\sum x^3+b\sum x^4+c\sum x^5+d\sum x^6=\sum yx^3.$$

Решение этой системы относительно параметров a, b, c и d приводит к следующим формулам:

$$a=\frac{1}{D_1}(\sum y\sum x^4-\sum x^2\sum yx^2);$$

$$b=\frac{1}{D_2}(\sum xy\sum x^6-\sum x^4\sum x^3y);$$

$$c=\frac{1}{D_1}(n\sum x^2y-\sum y\sum x^2);$$

$$d=\frac{1}{D_2}(\sum x^2\sum x^3y-\sum xy\sum x^4),$$

где $D_1=n\Sigma x^4-(\Sigma x^2)^2$ и $D_2=\Sigma x^2\Sigma x^6-(\Sigma x^4)^2$ — определители системы; n — число членов ряда регрессии; x_i и y_i — значения переменных, из которых независимая переменная выражается отклонениями членов ряда от их средней величины

Для нахождения параметров a, b, c и d нужно предварительно рассчитать Σy, Σyx, Σx^2, Σx^2y, Σx^3y, Σx^4 и Σx^6.

Пример 13. В отношении некоторого объекта было проведено девять испытаний, которые дали следующие результаты:

X	5	6	7	8	9	10	11	12	13
Y	78,0	76,1	73,6	72,9	70,8	69,4	69,3	69,0	69,1

В данном случае с увеличением независимой переменной X зависимая переменная Y закономерно убывает, т. е. ведет себя не так, как это имело место в отношении кривой лактации. Для нахождения выравнивающих значений этого ряда уравнение параболы второго порядка не подходит, в чем легко убедиться, проделав необходимую вычислительную работу. Применим к отысканию эмпирического уравнения регрессии этого ряда параболу третьего порядка. Предварительно, чтобы облегчить вычислительную работу, уменьшим каждый член ряда зависимой переменной y на $K=68$, т. е. заменим значения y на $y^*=y-68$.

Средняя арифметическая для членов ряда независимой переменной $\bar{x}=9$. Отклонения от этой величины и расчет необходимых вспомогательных сумм приведены в табл. 124.

Находим значения определителей: $D_1 = 9 \cdot 708 - 60^2 = 2772$ и $D_2 = 60 \cdot 9780 - 708^2 = 85\,536 - 501\,264 = 85\,536$. Переходим к вычислению параметров; при этом параметр a надо увеличить на $K = 68$. При вычислении параметров b, c и d поправки не нужны:

$$a = a' + K = \frac{36{,}2 \cdot 708 - 60 \cdot 293{,}4}{2772} + 68 = \frac{8025{,}6}{2772} + 68 =$$

$$= 2{,}895 + 68 = 70{,}895; \quad b = \frac{-69 \cdot 9780 - 708(-799{,}2)}{85\,536} = \frac{-108\,986{,}4}{85\,536} =$$

$$= -1{,}274; \quad c = \frac{9 \cdot 293{,}4 - 36{,}2 \cdot 60}{2772} = \frac{486{,}6}{2772} = 0{,}169;$$

$$d = \frac{60(-799{,}2) - (-69 \cdot 708)}{85\,536} = \frac{900{,}0}{85\,536} = 0{,}0105.$$

Таблица 12

x	y	y^*	$x = \bar{x} - X$	xy^*	x^2	x^2y^*	x^3	x^3y^*	x^4	x^6	$\bar{y}_x$
5	78,0	10,0	−4	−40,0	16	160,0	−64	−640,0	256	4096	78,0
6	76,1	8,1	−3	−24,3	9	72,9	−27	−218,7	81	729	76,0
7	73,6	5,6	−2	−11,2	4	22,4	−8	−44,8	16	64	74,0
8	72,9	4,9	−1	−4,9	1	4,9	−1	−44,9	1	1	72,2
9	70,8	2,8	0	0	0	0	0	0	0	0	70,9
10	69,4	1,4	+1	+1,4	1	1,4	+1	+1,4	1	1	69,8
11	69,3	1,3	+2	+2,6	4	5,2	+8	+10,4	16	64	69,1
12	69,0	1,0	+3	+3,0	9	9,0	+27	+27,0	81	729	68,9
13	69,1	1,1	+4	+4,4	16	17,6	+64	+70,4	256	4096	69,2
Σ	648,2	36,2	—	−69,0	60	293,4	—	−799,2	708	9780	648,2

Отсюда эмпирическое уравнение регрессии Y по X:

$$\bar{y}_x = 70{,}895 - 1{,}274x + 0{,}169x^2 + 0{,}0105x^3.$$

Подставляя в это уравнение вместо x отклонения членов ряда независимой переменной X от их средней арифметической $\bar{x}$, находим ожидаемые (выравнивающие) значения зависимой переменной $\bar{y}_{x_i}$: $\bar{y}_{x_1} = 70{,}895 - 1{,}274(-4) + 0{,}169(16) + 0{,}0105(-64) = 70{,}895 + 5{,}096 + 2{,}704 - 0{,}672 = 78{,}0$; $\bar{y}_{x_2} = 70{,}895 - 1{,}274(-3) + 0{,}169(9) + 0{,}0105(-27) = 70{,}895 + 3{,}822 + 1{,}521 - 0{,}2835 = 76{,}0$ и т. д.

Рассчитанные таким образом выравнивающие значения $\bar{y}_x$ приведены в последнем столбце табл. 124. Видно, что они неплохо согласуются с эмпирически найденными членами ряда

:ависимой переменной Y, о чем более наглядно свидетельствует рис. 31.

Как и в предыдущем случае, параметры a, b, c и d параболы третьего порядка можно корректировать, применяя для этого следующие формулы:

$$a^\circ = a + \frac{b\bar{x}}{\lambda} + \frac{c\bar{x}^2}{\lambda^2} + \frac{d\bar{x}^3}{\lambda^3}; \quad b^\circ = \frac{b}{\lambda} - \frac{2c\bar{x}}{\lambda^2} + \frac{3d\bar{x}^2}{\lambda^3};$$

$$c^\circ = \frac{c}{\lambda^2} - \frac{3d\bar{x}}{\lambda^3}; \quad d^\circ = \frac{d}{\lambda^3}.$$

Применительно к рассматриваемому примеру (учитывая, что $\lambda = 1$) это выглядит так: $a^\circ = 70{,}895 + (-1{,}274)(9) + 0{,}169(9^2) -$

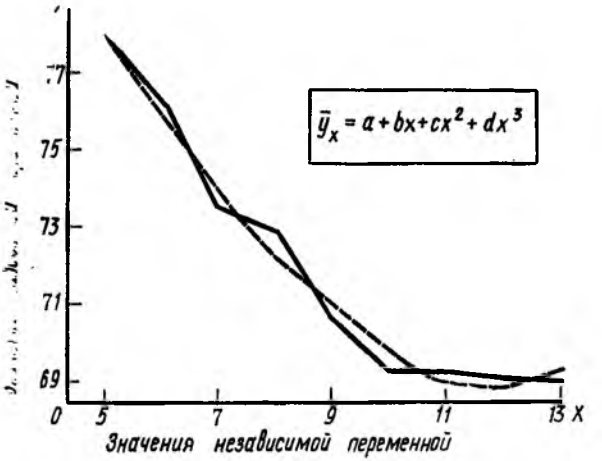


Рис. 31. Эмпирическая и вычисленная по уравнению параболы третьего порядка линии регрессии Y по X

$-0{,}0105(9^3) = 88{,}396$; $b^\circ = -1{,}274 - 2 \cdot 0{,}169(9) + 3 \cdot 0{,}0105(9^2) = -1{,}764$; $c^\circ = 0{,}169 - 3 \cdot 0{,}0105(9) = -0{,}1145$; $d^\circ = 0{,}0105$. Отсюда эмпирическое уравнение регрессии Y по X:

$$\bar{y}_x = 88{,}396 - 1{,}764x - 0{,}1145x^2 + 0{,}0105x^3.$$

Подставляя в это уравнение вместо x значения независимой переменной X, находим: $\bar{y}_{x_1} = 88{,}396 - 1{,}764(5) - 0{,}1145(5^2) + 0{,}0105(5^3) = 88{,}396 - 8{,}820 - 2{,}8625 + 1{,}3125 = 89{,}708 - 11{,}682 = 78{,}03$; $\bar{y}_{x_2} = 88{,}396 - 1{,}764(6) - 0{,}1145(6^2) + 0{,}0105(6^3) = 75{,}96 \approx 76{,}0$ и т. д.

Регрессия, выражаемая уравнением гиперболы первого порядка. В зависимости от наклона кривой регрессии к осям прямоугольных координат корреляционная зависимость между пе-

ременными величинами может быть выражена тем или иным уравнением гиперболы. В простейшем виде гиперболическая зависимость между переменными Y и X описывается *уравнением гиперболы первого порядка*:

$$\bar{y}_x = a + \frac{b}{x}. \qquad (194$$

Для определения параметров a и b этого уравнения служит следующая система нормальных уравнений:

$$an + b\sum\frac{1}{x} = \sum y;$$

$$a\sum\frac{1}{x} + b\sum\frac{1}{x_2} = \sum\frac{y}{x}.$$

Таблица 12

x, кг	y	x^2	$\frac{y}{x}$	$\frac{1}{x}$	$\frac{1}{x^2}$	$\bar{y}_x$
1,4	673	1,96	480,7	0,714	0,5102	669
2,2	489	4,84	222,3	0,454	0,2066	487
2,3	451	5,29	196,1	0,435	0,1890	473
2,6	405	6,76	155,8	0,385	0,1479	438
3,6	485	12,96	134,7	0,278	0,0772	364
4,1	330	16,81	80,5	0,244	0,0595	340
4,4	288	19,36	65,5	0,227	0,0517	329
5,8	268	33,64	46,2	0,172	0,0297	290
Сумма	3389	—	1381,8	2,909	1,2718	

Совместное решение этой системы относительно параметров a и b приводит к следующим формулам:

$$a = \frac{1}{D}\left(\sum y\sum\frac{1}{x^2} - \sum\frac{y}{x}\sum\frac{1}{x}\right);$$

$$b = \frac{1}{D}\left(n\sum\frac{y}{x} - \sum y\sum\frac{1}{x}\right),$$

где $D = n\sum\frac{1}{x^2} - \left(\sum\frac{1}{x}\right)^2$ — определитель системы; x — значения независимой; y — значения зависимой переменных величин; n — число членов ряда регрессии. Для нахождения параметров a и b по этим формулам необходимо предварительно рассчитать Σy, $\Sigma y/x$, $\Sigma 1/x$ и $\Sigma 1/x^2$.

Пример 14. Зависимость основного обмена y, выраженного в килоджоулях на 1 кг массы тела обезьян за 24 ч, характеризуется следующими величинами (табл. 125).

Если эти данные изобразить графически в системе прямоугольных координат, можно убедиться в том, что они выглядят в виде гиперболической зависимости между переменными Y и X. Необходимые суммы для вычисления параметров a и b по уравнению (194) содержатся в табл. 125. Подставляя эти данные в формулы, находим:

$$a=\frac{33{,}89\cdot 1{,}272-1381{,}8\cdot 2{,}91}{8\cdot 1{,}272-(2{,}91)^2}=169{,}7\approx 170;$$

$$b=\frac{8\cdot 1381{,}8-3389\cdot 2{,}91}{8\cdot 1{,}272-(2{,}91)^2}=698{,}1\approx 698.$$

Отсюда уравнение регрессии Y по X: $\bar{y}_x=170+698/x$.

Рассчитанные по этому уравнению ожидаемые величины основного обмена $\bar{y}_x$ приведены в последнем столбце табл. 125. Видно, что вычисленные величины неплохо согласуются с данными опыта. Более наглядное представление об этом дает рис. 32, на котором изображены эмпирическая (ломаная) и вычисленная (плавно идущая) линия регрессии Y по X.

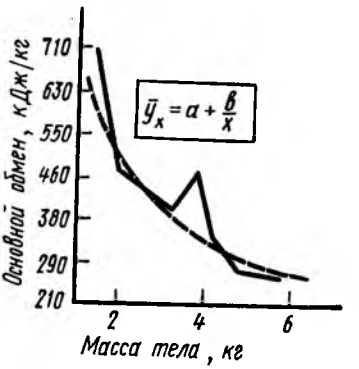


Рис. 32. Зависимость величины осиовного обмеиа обезьян от массы их тела

Регрессия, выражаемая уравнением гиперболы второго порядка. Для нахождения выравнивающих значений зависимой переменной иногда более подходящим оказывается *уравнение гиперболы второго порядка*

$$\bar{y}_x=a+\frac{b}{x^2}. \qquad (195)$$

Для определения параметров a и b этого уравнения служит следующая система нормальных уравнеиий:

$$an+b\sum\frac{1}{x^2}=\sum y;$$

$$a\sum\frac{1}{x^2}+b\sum\frac{1}{x^4}=\sum\frac{y}{x^2}.$$

Решение этой системы относительно параметров a и b приводит к следующим формулам:

$$a=\frac{1}{D}\left(\sum y\sum\frac{1}{x^4}-\sum\frac{y}{x^2}\sum\frac{1}{x^2}\right);$$

$$b=\frac{1}{D}\left(n\sum\frac{y}{x^2}-\sum y\sum\frac{1}{x^2}\right),$$

где $D=n\sum\frac{1}{x^4}-\left(\sum\frac{1}{x^2}\right)^2$.

Для определения параметров a и b необходимо предварительно рассчитать Σy, $\sum\frac{y}{x^2}$, $\sum\frac{1}{x^2}$ и $\sum\frac{1}{x^4}$.

Пример 15. В семи хозяйствах района сопоставили урожаи озимой пшеницы с себестоимостью 1 ц зерна этой культуры. Полученные результаты и их обработка приведены в табл. 126.

Таблица 126

X	$x=X/8$	y	$\frac{y}{x^2}$	$\frac{1}{x^2}$	$\frac{1}{x^4}$	$\bar{y}_x$
8	1,0	12,0	12,0	1,0000	1,0000	11,7
11	1,4	8,0	4,08	0,5102	0,2603	8,4
13	1,6	7,3	2,85	0,3906	0,1526	7,6
19	2,4	6,0	1,04	0,1736	0,0301	6,1
21	2,6	6,3	0,93	0,1479	0,0219	5,9
27	3,4	5,8	0,50	0,0865	0,0075	5,5
29	3,6	5,2	0,40	0,0772	0,0060	5,4
Сумма	—	50,6	21,80	2,3860	1,4784	50,6

В этой таблице через X обозначен урожай пшеницы (ц/га) в разных хозяйствах района, а через y — себестоимость 1 ц пшеницы (руб.). Чтобы облегчить вычисление вспомогательных величин, значения независимой переменной X сокращены на $K=8$, полученные результаты помещены во второй графе (x табл. 126. Используя суммы из табл. 126, находим значения определителей системы: $D=7\cdot1,4784-(2,386)^2=10,3488-5,6930=4,6558$; $A=50,6\cdot1,4784-21,80\cdot2,3860=22,7922$; $B=7\cdot21,80-50,6\cdot2,3860=31,8684$. Отсюда $a=A/D=22,77/4,65=4,895$; $b=B/D=31,868/4,656=6,8445$. Эмпирическое уравнение гиперболы второго порядка оказывается следующим:

$$\bar{y}_x=4,9+\frac{6,8}{x^2}.$$

Рассчитанные по этому уравнению ожидаемые значения зависимой переменной $\bar{y}_x$ приведены в последнем столбце табл. 126. Видно, что они неплохо согласуются с эмпирическими значениями признака y. Более наглядно это показано на рис. 33, где изображены эмпирическая и выровненная по уравнению гиперболы второго порядка линии регрессии.

Регрессия, выражаемая уравнением гиперболы третьего порядка. В практике встречаются случаи, когда с увеличением независимой переменной X зависимая переменная Y, быстро убывая, вскоре стабилизируется на определенном уровне, принимая более или менее постоянные значения. В таких ситуациях для выравнивания эмпирического ряда регрессии можно использовать *уравнение гиперболы третьего порядка*:

$$\bar{y}_x = a + \frac{b}{x^3} . \qquad (196)$$

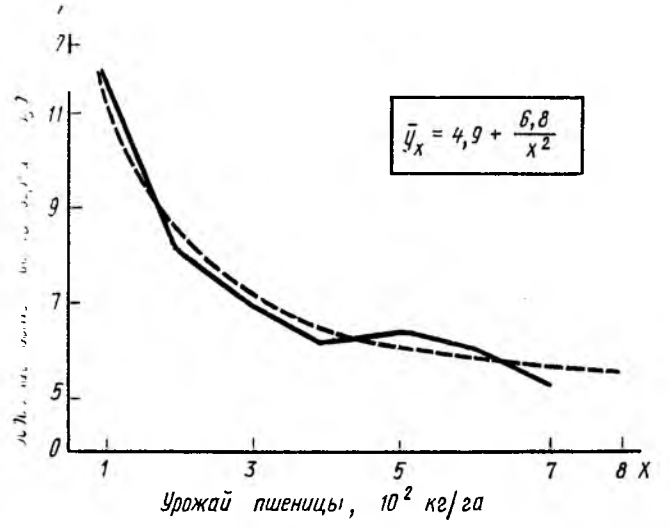


Рис. 33. Зависимость между себестоимостью пшеницы и урожаем этой культуры

Для определения параметров a и b этого уравнения применяют следующую систему нормальных уравнений:

$$an + b\sum\frac{1}{x^3} = \sum y;$$

$$a\sum\frac{1}{x^3} + b\sum\frac{1}{x^6} = \frac{y}{x^3} .$$

Решая совместно эти уравнения относительно параметров a и b, получаем следующие формулы:

$$a = \frac{1}{D}\left(\sum y\sum\frac{1}{x^6} - \sum\frac{y}{x^3}\sum\frac{1}{x^3}\right);$$

$$b = \frac{1}{D}\left(n\sum\frac{y}{x^3} - \sum y\sum\frac{1}{x^3}\right),$$

где $D=n\Sigma\frac{1}{x^6}-\left(\sum\frac{1}{x^3}\right)^2$. Отсюда следует, что для нахождения параметров a и b необходимо предварительно рассчитать Σy, $\sum\frac{y}{x^3}$, $\sum\frac{1}{x^3}$ и $\sum\frac{1}{x^6}$.

Пример 16. В табл. 127 приведены результаты восьми однотипных испытаний и их обработка по формуле (196).

Таблица 127

x	y	x^3	x^6	$\frac{y}{x^3}$	$\frac{1}{x^3}$	$\frac{1}{x^6}$	$\bar{y}_x$
1	29,0	1	1	29,000	1,0000	1,00000	28,4
2	5,9	8	64	0,738	0,1250	0,01562	5,7
3	3,4	27	729	0,126	0,0370	0,00137	3,4
4	3,8	64	4096	0,059	0,0156	0,00024	2,9
5	2,5	125	15625	0,020	0,0080	0,00006	2,7
6	2,0	216	46656	0,009	0,0046	0,00002	2,6
7	2,3	343	117649	0,007	0,0029	0,00001	2,6
8	1,9	512	262144	0,004	0,0020	0,00000	2,5
Сумма	50,8	—	—	29,963	1,1951	1,01732	50,8

Из данных табл. 127 видно, что после резкого снижения числовых значений зависимой переменной y они постепенно стабилизируются, оставаясь примерно на одном уровне. Найдем эмпирическое уравнение этой регрессии: $D=8\cdot 1{,}01732-1{,}1951^2=8{,}1386-1{,}4283=6{,}710$; $A=(50{,}8\cdot 1{,}01732-29{,}963\cdot 1{,}1951)/6{,}710=15{,}871/6{,}710=2{,}37$; $B=(8\cdot 29{,}963-50{,}8\cdot 1{,}1951)/6{,}710=178{,}993/6{,}710=26{,}676$. Эмпирическое уравнение регрессии Y по X оказывается следующим:

$$\bar{y}_x=2{,}37+\frac{26{,}676}{x^3}.$$

Испытание этого уравнения показало, что оно не удовлетворяет равенству $\Sigma y=\Sigma\bar{y}_x$; корректируя его, находим точное уравнение регрессии Y по X:

$$\bar{y}_x=\frac{26{,}4}{x^3}+2{,}40.$$

Рассчитанные по этому уравнению выравнивающие значения $\bar{y}_x$ приведены в последнем столбце табл. 127. Видно, что они неплохо согласуются с эмпирическими значениями переменной Y. Более наглядное представление об этом дает рис. 34.

Регрессия, выражаемая уравнением гиперболы первого порядка с тремя неизвестными: a, b и c. Если с увеличением независимой переменной X зависимая переменная Y быстро убывает,

достигая некоторого предела, за которым обнаруживается более или менее стабильное течение функции, то для выравнивания эмпирических значений зависимой переменной может быть использовано уравнение гиперболы следующего вида:

$$y = a + bx + \frac{c}{x}. \qquad (197)$$

Для определения параметров a, b и c этого уравнения служит следующая система нормальных уравнений:

$$an + b\sum x + c\sum \frac{1}{x} = \sum y;$$

$$a\sum x + b\sum x^2 + an = \sum xy;$$

$$a\sum \frac{1}{x} + bn + c\sum \frac{1}{x^2} = \sum \frac{y}{x}.$$

Чтобы по выборочным данным составить такую систему, необходимо предварительно рассчитать Σx, Σy, Σxy, $\sum \frac{y}{x}$, $\sum \frac{1}{x}$ и $\sum \frac{1}{x^2}$.

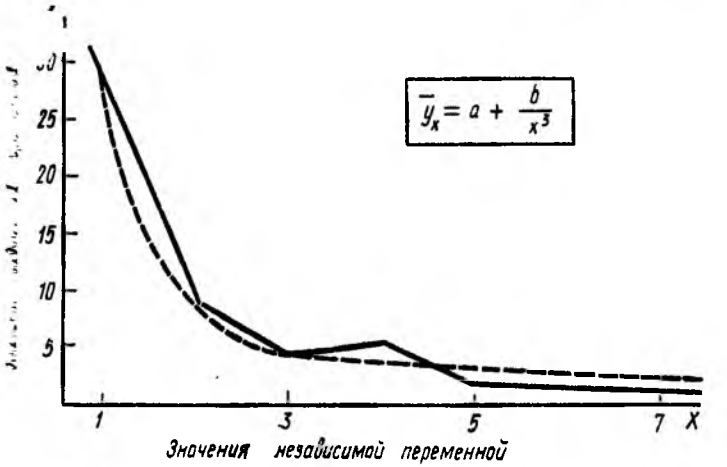


Рис. 34. Эмпирическая и вычисленная по уравнению гиперболы третьего порядка линии регрессии Y по X

Пример 17. Как показали многочисленные наблюдения, с увеличением числа независимых испытаний (n) величина ошибки среднего результата $s_{\bar{x}}$ закономерно уменьшается:

Число испытаний	5	10	15	20	25	30	35	40	45
Ошибка средней	6,2	2,9	1,6	1,9	1,1	0,9	1,2	0,9	0,9

Эта связь между переменными n и $s_{\bar{x}}$ имеет гиперболический характер и может быть описана с помощью уравнения (197). Найдем эмпирическое уравнение этой связи. Чтобы облегчить вычислительную работу, обозначим переменные n и $s_{\bar{x}}$ соответственно через X и Y и сократим значения независимой переменной X на 5. Тогда аргумент X выразится рядом натуральных чисел 1, 2, 3, ... Расчет вспомогательных величин приведен в табл. 128.

Таблица 128

Преобразованные значения аргумента x	Величина ошибки y	xy	$\frac{y}{x}$	x^2	$\frac{1}{x}$	$\frac{1}{x^2}$	$\bar{y}_x$
1	6,2	6,2	6,200	1	1,000	1,0000	6,2
2	2,9	5,8	1,450	4	0,500	0,2500	2,9
3	1,6	4,8	0,533	9	0,333	0,1111	1,9
4	1,9	7,6	0,475	16	0,250	0,0625	1,4
5	1,1	5,5	0,220	25	0,200	0,0400	1,2
6	0,9	5,4	0,150	36	0,167	0,0278	1,1
7	1,2	8,4	0,171	49	0,143	0,0204	1,0
8	0,9	7,2	0,113	64	0,125	0,0156	1,0
9	0,9	8,1	0,100	81	0,111	0,0123	0,9
45	17,6	59,0	0,412	285	2,829	1,5397	17,6
$\Sigma=45$							

Составляем систему нормальных уравнений:

$$9a+45b+2{,}83c=17{,}6;$$

$$45a+285b+9c=59{,}0;$$

$$2{,}83a+9b+1{,}54c=9{,}412.$$

Решая эту систему относительно параметров a, b и c (как описано выше), находим: $a=-0{,}571$; $b=0{,}0871$; $c=6{,}6496$. Отсюда эмпирическое уравнение регрессии Y по X:

$$\bar{y}_x=-0{,}571+0{,}0871x+\frac{6{,}6496}{x}.$$

Подставляя в это уравнение вместо x преобразованные значения аргумента, находим:

$$\bar{y}_x=-0{,}571+0{,}0871\cdot 1+6{,}6496/1=6{,}2;$$

$$\bar{y}_x=-0{,}571+0{,}0871\cdot 2+6{,}6496/2=2{,}9 \text{ и т. д.}$$

Рассчитанные таким образом значения $\bar{y}_x$ зависимой переменной y приведены в последнем столбце табл. 128. Они хорошо со-

ласуются с эмпирическими значениями функции, что более наглядно иллюстрирует рис. 35. Заметим, что в других подобных случаях для выравнивания рядов регрессии более подходящими могут быть уравнения гиперболы второго и третьего порядков с тремя неизвестными, т. е.

$$\bar{y}_x = a + bx + c/x^2; \quad \bar{y}_x = a + bx + c/x^3 \text{ и т. д.}$$

Регрессия, выражаемая уравнением показательного типа. В тех случаях, когда основная тенденция эмпирического ряда регрессии следует или оказывается близкой закону геометрической

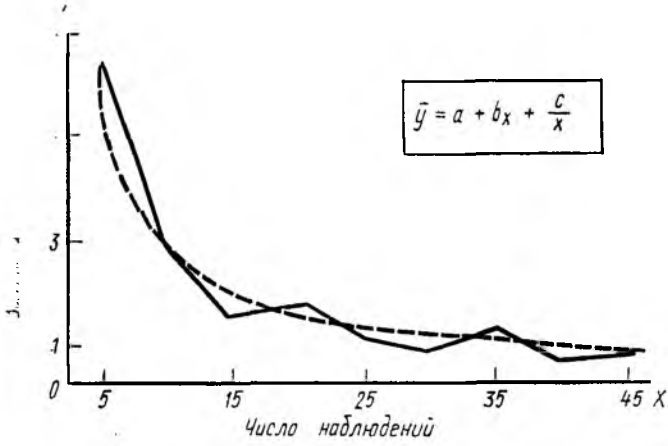


Рис. 35. Зависимость между числом наблюдений n и величиной ошибки $s_{\bar{x}}$ среднего результата $\bar{x}$

прогрессии, его удается описать уравнением *экспоненциального, или показательного, типа:*

$$y = ab^x \text{ или } y = ae^{xb}. \qquad (198)$$

Использование уравнений такого вида связано с их логарифмированием, что позволяет трансформировать их в уравнение прямой линии. Так, в данном случае

$$\lg y = \lg a + x \lg b. \qquad (199)$$

Логарифмическое преобразование исходного уравнения регрессии не только облегчает вычисление параметров a и b, но и служит своего рода контролем того, насколько правильно выбрано применяемое уравнение. В частности, условием правильного выбора уравнения показательного типа служит требование, чтобы точки x и $\lg y$ в системе прямоугольных координат находились на одной прямой.

Для определения параметров уравнения (199) служит следующая система нормальных уравнений:

$$n \lg a + \lg b \sum x = \sum \lg y;$$

$$\lg a \sum x + \lg b \sum x^2 = \sum (x \lg y).$$

Совместное решение этой системы приводит к следующим формулам:

$$\lg a = \frac{1}{D}\left[\sum \lg y \sum x^2 - \sum (x \lg y) \sum x\right];$$

$$\lg b = \frac{1}{D}\left[n \sum (x \lg y) - \sum x \sum \lg y\right],$$

где $D = n\Sigma x^2 - (\Sigma x)^2$; n — число членов ряда; y — значения членов ряда зависимой переменной Y; x — значения членов ряда независимой переменной X, которые обычно выражают, как и в предыдущем случае, числами натурального ряда.

Таблица 12

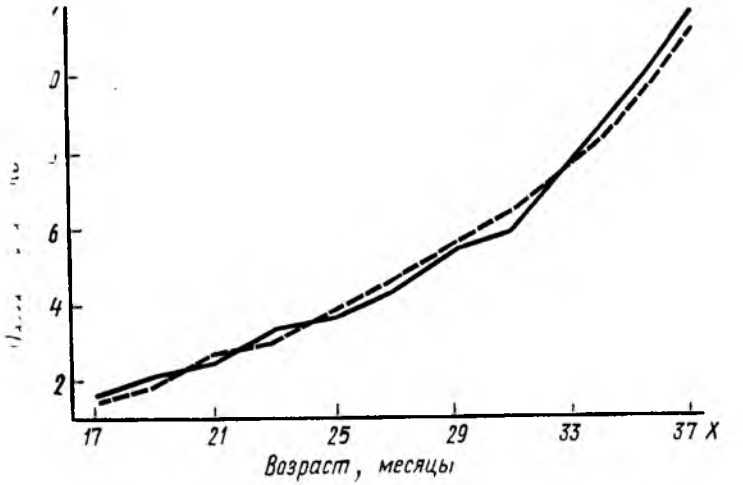
Возраст животных		Масса тела, y кг	x^2	$\lg y$	$x \lg y$	$\bar{y}_x$
фактический X	выраженный числами натурального ряда x					
20	1	4,6	1	0,66276	0,66276	4,3
26	2	4,5	4	0,65321	1,30642	5,0
32	3	6,4	9	0,80618	2,41854	5,7
38	4	6,1	16	0,78533	3,14132	6,6
42	5	7,5	25	0,87506	4,37530	7,6
48	6	8,0	36	0,90399	5,42394	8,7
52	7	11	49	1,04139	7,28973	10,2
Сумма	28	48,1	140	5,72792	2461801	48,1

Из системы уравнений и приведенных формул следует, что для отыскания параметров a и b нужно предварительно найти Σx, Σx^2, $\Sigma \lg y$ и $\Sigma (x \lg y)$.

Пример 18. Наблюдения за развитием самцов павианов-гамадрилов в период полового созревания показали, что масса их тела изменяется с возрастом следующим образом:

Возраст X, мес	20	26	32	38	42	48	52
Масса тела Y, кг	4,6	4,5	6,4	6,1	7,5	8,0	11,0

Графический аиализ этих даиных показывает, что оии изме-яются по закону экспоненциальиой функции. Об этом свидетель-твует тот факт, что значения иезависимой перемеиной X и лога-ифмированные значения зависимой переменной Y располагают-я в системе прямоугольных координат на одной прямой. Найдем мпирическое уравнение, описывающее эту закономерность. Предварительно рассчитаем вспомогательные величины табл. 129).



ис. 36. Эмпирическая и вычислеииая по уравиению показательиой ункции линии регрессии возрастных изменеиий массы тела самцов гамадрилов в период полового созревания

Подставляя известные величиины в формулы, находим

$$\lg a = \frac{5{,}72792 \cdot 140 - 24{,}61801 \cdot 28}{7 \cdot 140 - 28 \cdot 28} = \frac{112{,}60452}{196} = 0{,}57451;$$

$$\lg b = \frac{7 \cdot 24{,}61801 - 28 \cdot 5{,}72792}{7 \cdot 140 - 28 \cdot 28} = \frac{11{,}94431}{196} = 0{,}060940.$$

тсюда эмпирическое уравнение регрессии Y по X:

$$\lg \bar{y}_x = 0{,}061x + 0{,}575.$$

Подставляя в это уравнение вместо x значения независимой еременной, выраженные числами натурального ряда, находим $g\,\bar{y}_x$, а затем по таблице логарифмов определяем значения $\bar{y}_x$. Рассчитаиные таким образом значения $\bar{y}_x$ приведены в послед-ем столбце табл. 129. Они неплохо согласуются с помещенными той же таблице эмпирическими членами ряда. Более наглядное редставление об этом дает рис. 36, иа котором изображены эм-

пирическая и выровненная по уравнению регрессия лнин массы тела Y по возрасту X самцов павианов-гамадрилов.

Регрессия, выражаемая уравнением степенно́го типа. Зависимость между переменными величинами Y и X иногда хорошо описывается уравнением *степенно́го типа*

$$\bar{y}_x = ax^b, \tag{200}$$

которое в результате логарифмирования превращается в уравнение прямой линии:

$$\lg \bar{y}_x = \lg a + b \lg x. \tag{201}$$

Условием правильного применения этого уравнения служит требование, чтобы точки $\lg y$ и $\lg x$ в системе прямоугольных координат находились на одной прямой. Эта особенность отличает уравнение степенной функцин от уравнения регрессни показательного типа, когда в системе координат на одной прямой оказываются точки $\lg y$ и x.

Для определения параметров a и b уравнения степенного типа служит следующая система нормальных уравнений

$$n \lg a + b \sum \lg x = \sum \lg y;$$

$$\lg a \sum \lg x + b \sum (\lg x)^2 = \sum (\lg x \lg y).$$

Из решения этой системы получаются формулы

$$\lg a = \frac{1}{D}\left[\sum \lg y \sum (\lg x)^2 - \sum (\lg x \lg y) \sum \lg x\right] \quad \text{и}$$

$$b = \frac{1}{D}\left[n \sum (\lg x \lg y) - \sum \lg x \sum \lg y\right],$$

где $D = n\Sigma (\lg x)^2 - (\Sigma \lg x)^2$ — определитель системы; n — число членов ряда регрессии; x и y — значения членов ряда независимой и зависимой переменных величин. Из этих формул следует, что для нахождения параметров a и b нужно предварительно рассчитать $\Sigma \lg y$, $\Sigma (\lg y, \lg x)$, $\Sigma \lg x$ и $\Sigma (\lg x)^2$.

Пример 19. Наблюдения за развитием самок павианов-гамадрилов показали, что между массой их тела и длиной туловнща существует положительная связь. Соответствующие данные и их обработка приведены в табл. 130.

Если эти данные изобразить в виде графика, как показано на рис. 37, можно убедиться в том, что между этими признаками имеет место нелинейная связь. Прологарифмировав значения переменных x и y и выразив полученные величины в процентах от их общей суммы для каждого ряда в отдельности, легко построить линейный график, показывающий, что точки $\lg y$ и $\lg x$ располагаются на одной прямой (рис. 38). Следовательно, зависи-

Длина туловища x, см	Масса тела y, кг	$\lg x$	$\lg y$	$\lg x \lg y$	$(\lg x)^2$	$\bar{y}_x$
17	1,5	1,23045	0,17609	0,21667	1,51401	1,4
19	2,0	1,27875	0,30103	0,38494	1,63520	1,9
21	2,3	1,32222	0,36173	0,47829	1,74826	2,4
23	3,3	1,36173	0,51851	0,70607	1,85431	3,1
25	3,6	1,39794	0,55630	0,77767	1,95424	3,8
27	4,4	1,43136	0,64345	0,92101	2,04879	4,6
29	5,5	1,46240	0,74036	1,08270	2,13861	5,6
31	6,0	1,49136	0,77815	1,16050	2,22415	6,7
33	7,8	1,51851	0,89209	1,35465	2,30587	7,8
35	9,6	1,54407	0,98227	1,51669	2,38415	9,1
37	12,0	1,56820	1,07918	1,69237	2,45925	11,6
Сумма	58,0	15,60699	7,02916	10,29156	22,26684	58,0

...ность между этими признаками можно выразить уравнением ...тепенного типа. Найдем эмпирическое уравнение этой зависимости. Вспомогательные величины содержатся в табл. 130. Подставляя эти величины в формулы, находим

$$\lg a = \frac{7,02916 \cdot 22,26684 - 10,29 \cdot 56 \cdot 15,60699}{11 \cdot 22,26684 - (15,60699)^2} = \frac{-4,1031}{1,3571} = -3,023;$$

$$b = \frac{11 \cdot 10,29156 - 15,60699 \cdot 7,02916}{1,3571} = \frac{3,50314}{1,3571} = 2,5813.$$

...тсюда эмпирическое уравнение массы тела y по длине туловища x самок гамадрилов имеет следующий вид:

$$\lg \bar{y}_x = 2,58 \lg x - 3,023.$$

...ис. 37. Зависимость между длиной ...уловища и массой тела у самок павианов-гамадрилов

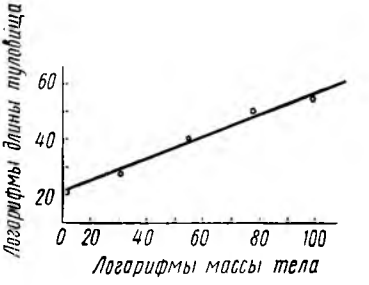


Рис. 38. График регрессии массы тела по длине туловища у самок павианов-гамадрилов

Подставляя в это уравнение вместо x конкретные данные о длине туловища самок гамадрилов, можно определить возможную массу их тела, соответствующую известной длине туловища. Например, длине туловища самки гамадрила, равной 28 см, должна соответствовать следующая масса ее тела: $\lg \bar{y}_x = 2{,}58 \lg 28 - 3{,}023 = 2{,}58 \cdot 1{,}44716 - 3{,}023 = 0{,}711$, или $\bar{y}_x = 5{,}0$ кг. Рассчитанные таким образом ожидаемые величины массы тела $\bar{y}_x$ самок гамадрилов, соответствующие размерам туловища этих животных, приведены в последней графе табл. 130. Видно, что они неплохо согласуются с эмпирически найденными значениями этого ряда.

Пример 20. По данным А. Д. Слонима и О. П. Щербаковой (1949), величина основного обмена обезьян, выраженная в килоджоулях на 1 кг массы тела за 24 ч, изменяется с возрастом следующим образом (табл. 131).

Как и в предыдущем случае, зависимость между этими признаками можно описать уравнением степенного типа. Основанием для этого служит тот факт, что точки $\lg x$ и $\lg y$ располагаются в системе прямоугольных координат на одной прямой (читатель может это проверить). Найдем эмпирическое уравнение регрессии величины основного обмена y по возрасту обезьян x. Предварительно рассчитываем вспомогательные величины. Рас-

Таблица 131

Возраст обезьян x, мес	Основной обмен y	$\lg x$	$\lg y$	$\lg x \lg y$	$(\lg x)^2$	$\bar{y}_x$
2	1045	0,30103	3,01912	0,90884	0,09062	991
4	673	0,60206	2,82802	1,70263	0,36248	703
8	489	0,90309	2,68931	2,42869	0,81557	499
9	451	0,95424	2,65418	2,53272	0,91057	471
10	405	1,00000	2,60746	2,60746	1,00000	447
16	485	1,20412	2,68574	3,23396	1,44990	355
19	293	1,27875	2,46687	3,15451	1,63520	326
20	318	1,30103	2,50243	3,25573	1,69268	318
22	288	1,34242	2,45939	3,30154	1,80209	303
28	268	1,44716	2,42813	3,51390	2,09427	269
Сумма	4715	1033390	26,34065	26,63998	11,85338	

чет приведен в табл. 131. Подставляя найденные величины в формулы, определяем параметры:

$$\lg a = \frac{76{,}34065 \cdot 11{,}85338 - 26{,}63998 \cdot 10{,}33390}{10 \cdot 11{,}85338 - (10{,}33390)^2} = 3{,}145;$$

$$b = \frac{10 \cdot 26{,}63998 - 10{,}33390 \cdot 26{,}34065}{10 \cdot 11{,}85338 - (10{,}33390)^2} = -0{,}494.$$

Отсюда эмпирическое уравнение регрессии основного обмена Y по возрасту X обезьян:

$$\lg \bar{y}_x = 3{,}145 - 0{,}494 \lg x.$$

Рассчитанные по этому уравнению ожидаемые величины основного обмена в зависимости от возраста обезьян приведены в последней графе табл. 131. Видно, что они хорошо согласуются между собой, что более наглядно показано на рис. 39, где изображены эмпирическая и выровненная линии регрессии Y по X.

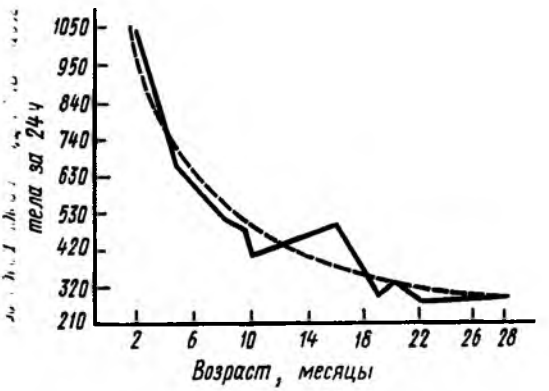


Рис. 39. Эмпирическая и вычисленная по уравнению степенной функции кривые возрастных изменений основного обмена у обезьян

Регрессия, выражаемая уравнением логистической кривой. Значительный интерес для биолога представляет *логистическая зависимость* между переменными величинами. Зависимость такого рода встречается во многих случаях, например при изменении состава популяции в замкнутой среде обитания, когда начальное число особей сначала быстро возрастает, затем темп роста популяции также быстро снижается и популяция переходит в состояние динамического равновесия. Графически эта закономерность изображается в виде S-образной кривой, которая описывается уравнением Ферхюльста:

$$y = \frac{N}{1 + 10^{a+bt}} + c, \qquad (202)$$

где y — учитываемый признак; t — время, прошедшее от начальной, или базисной (c), величины признака, с которой начато его измерение, до предельной в данных условиях величины N, которой он достиг за время t; a и b — параметры уравнения, определяющие характер логистической кривой.

Путем логарифмического преобразования это уравнение приобретает следующее выражение:

$$\lg\left(\frac{N}{y-c}-1\right)=a+bt.$$

Обозначив $\left(\frac{N}{y-c}-1\right)$ через z, получаем уравнение линейной регрессии:

$$\lg z=a+bt. \qquad (203$$

Таблица 13.

t	y	t^2	$\frac{N}{y}$	$\frac{N}{y}-1=z$	$\lg z$		$t\lg z$
0	5	—	—	—	—	—	—
1	20	1	19,50000	18,50000	+1,2672	+1,2672	+1,2672
2	100	4	3,90000	2,90000	+0,4624	+0,4624	+0,9248
3	300	9	1,30000	0,30000	1,4771	—0,5229	—1,5687
4	350	16	1,11430	0,11430	1,0581	—0,9419	—3,7676
5	380	25	1,02630	0,02630	2,4200	—1,5800	—7,9000
6	385	36	1,01299	0,01299	2,1136	—1,8864	—11,3184
7	389	—	—	—	—	—	—
Σ 21	—	91	—	—	—	—3,2016	—22,3627

Определению параметров a и b этого уравнения удовлетворяет следующая система нормальных уравнений:

$$an+b\sum t=\sum \lg z;$$

$$a\sum t+b\sum t^2=\sum(t\lg z).$$

Решая эту систему относительно параметров a и b, получаем следующие формулы:

$$a=\frac{1}{D}\left[\sum \lg z\sum t^2-\sum t\sum(t\lg z)\right];$$

$$b=\frac{1}{D}\left[n\sum(t\lg z)-\sum t\sum \lg z\right],$$

где $D=n\Sigma t^2-(\Sigma t)^2$.

Из этих формул следует, что для получения эмпирического уравнения логистической зависимости между переменными t и y необходимо предварительно рассчитать Σt, Σt^2, $\Sigma \lg z$ и $\Sigma(t\lg z)$. Затем, определив параметры a и b, найти для каждого значения t (в пределах учитываемого промежутка времени) величины $\lg z$ и z, что и приведет к нахождению ожидаемых значений $\bar{y}_t$.

Пример 21. На протяжении семи суток проводили наблюдения над ростом численности особей инфузории туфельки в замкнутой среде обитания (аквариуме). Результаты опыта и их обработка приведены в табл. 132.

Наблюдения начаты с помещения в аквариум пяти особей туфельки. К концу первых суток их численность увеличилась в 4 раза, затем в 5 раз и т. д.

Принимая $c=0$ и $N=390$, составляем систему нормальных уравнений:

$$6a+21b=-3{,}2016;$$

$$21a+91b=-22{,}3627.$$

Решая эту систему, находим

$$a=\frac{-3{,}2016\cdot 91-21\,(-22{,}3627)}{6\cdot 91-21\cdot 21}=\frac{178{,}271}{105}=1{,}6978\approx 1{,}698;$$

$$b=\frac{6-22{,}3627-21\,(-3{,}2016)}{105}=\frac{-134{,}1762-67{,}2336}{105}=-0{,}6375.$$

Отсюда эмпирическое уравнение, выражающее закономерность роста численности особей популяции туфельки в замкнутой среде обитания, оказывается следующим:

$$\bar{y}_t=\frac{390}{1+10^{1{,}698-0{,}6375t}}.$$

Это уравнение позволяет определять ожидаемую численность особей туфельки $\bar{y}_t$ в любой момент ее развития в замкнутой среде обитания. Для простоты следует логарифмировать уравнение прямой $z=a+bt$, затем с помощью таблицы логарифмов найти значения $z=\frac{N}{y}-1$, а также $z+1$. Делением N на $z+1$ получаются искомые величины $\bar{y}_z$. В отношении рассматриваемого примера эти вычисления выглядят так (табл. 133).

Таблица 133

t	$\lg z=a+bt$	$\lg z$	$z=\frac{N}{y}-1$	$z+1$	$\bar{y}_t$
0	+1,698	+1,698	49,900	50,900	8
1	+1,061	+1,061	11,510	12,510	31
2	+0,424	+0,424	2,650	3,650	107
3	−0,213	1,787	0,612	1,612	242
4	−0,850	1,150	0,142	1,142	342
5	−1,488	2,512	0,032	1,032	378
6	−2,124	3,876	0,0075	1,0075	387

При сопоставлении эмпирически полученных величин y_t с вычисленными $\bar{y}_t$ видно, что они неплохо согласуются между собой. Более наглядно об этом свидетельствует рис. 40, на котором на фоне ломаной эмпирической кривой изображена плавно идущая кривая ожидаемых значений $\bar{y}_t$ переменной Y[1].

IX.3. ОЦЕНКА ДОСТОВЕРНОСТИ ПОКАЗАТЕЛЕЙ РЕГРЕССИИ

Выборочные показатели регрессии являются *оценками соответствующих генеральных параметров* и, как величины случайные, сопровождаются статистическими ошибками. Ошибку выборочного коэффициента регрессии Y по X определяют по формуле

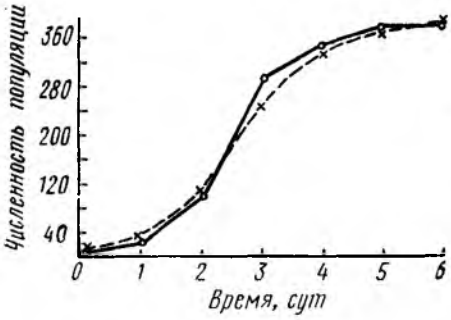


Рис. 40. Возрастание численности популяции инфузории туфельки в замкнутой среде обитания

$$s_{b_{yx}}=\sqrt{\frac{(1-r^2)\sum(y_i-\bar{y})^2}{(n-2)\sum(x_i-\bar{x})^2}}, \qquad (204)$$

а ошибку коэффициента регрессии X по Y — соответственно

$$s_{b_{xy}}=\sqrt{\frac{(1-r^2)\sum(x_i-\bar{x})^2}{(n-2)\sum(y_i-\bar{y})^2}}. \qquad (204а)$$

Достоверность выборочных коэффициентов регрессии оценивают с помощью t-критерия Стьюдента. Нулевую гипотезу отвергают на принятом уровне значимости (α) с числом степеней свободы $k=n-2$, если $t_ф \geqslant t_{st}$.

Пример 22. В гл. VIII было установлено, что между массой тела новорожденных гамадрилов y_i и массой тела их матерей x существует положительная связь ($r_{xy}=0{,}564$). Уравнение регрессии, описывающее эту связь, следующее:

$$\bar{y}_x=0{,}42+0{,}024x.$$

Определим ошибку коэффициента регрессии $b_{yx}=0{,}024$ и оценим достоверность этого показателя. Необходимые данные содержатся в табл. 96: $n=20$; $\Sigma y=14{,}06$; $\Sigma x=237{,}4$; $\Sigma y^2=9{,}9596$ и $\Sigma x^2=2861{,}60$. Рассчитаем девиаты:

$$D_y=\sum(y_i-\bar{y})^2=\sum y_i^2-\frac{(\Sigma y_i)^2}{n}=9{,}9596-\frac{(14{,}06)^2}{20}=0{,}0754;$$

[1] Более подробно этот способ описан в книге Н. А. Плохинского «Биометрия» (1970). Там же приведены способы выравнивания других видов нелинейной регрессии, в частности регрессии периодического типа.

$$D_x=\sum(x_l-\bar{x})^2=\sum x_l^2-\frac{(\sum x_l)^2}{n}=2861{,}60-\frac{(237{,}4)^2}{20}=43{,}66;$$

$$s_{b_{xy}}=\sqrt{\frac{1-0{,}564^2\cdot 0{,}0754}{(20-2)\,43{,}662}}=\sqrt{\frac{0{,}05142}{785{,}916}}=\sqrt{0{,}000066}=0{,}008.$$

Критерий $t_\Phi=0{,}024/0{,}008=3{,}0$. Эта величина превосходит критическую точку $t_{st}=2{,}88$, для $k=20-2$ и $\alpha=1\%$, что дает основание для непринятия нулевой гипотезы.

Если исходные данные сгруппированы в вариационные ряды, а их частоты распределяются по ячейкам корреляционной таблицы, ошибку коэффициентов регрессии определяют с учетом классовых интервалов по следующим формулам:

$$s_{b_{yx}}=\frac{s_y\lambda_y}{s_x\lambda_x}\sqrt{\frac{1-r^2}{n-2}};\qquad s_{b_{xy}}=\frac{s_x\lambda_x}{s_y\lambda_y}\sqrt{\frac{1-r^2}{n-2}}. \qquad (205)$$

Пример 23. Выше были найдены характеристики корреляционной зависимости между массой тела Y и годовым удоем X коров горбатовской породы $r_{xy}=0{,}523$; $b_{yx}=0{,}0564$, а также $s_y=3{,}272$; $s_x=2{,}843$, классовые интервалы: $\lambda_y=14$ и $\lambda_y=152$. Вычислим ошибку коэффициента регрессии удоя коров по живой массе их тела:

$$s_{b_{yx}}=\frac{3{,}272\cdot 14}{2{,}843\cdot 152}\sqrt{\frac{1-(0{,}523)^2}{100-2}}=0{,}106\cdot 0{,}086=0{,}0091.$$

Критерий $t_\Phi=\frac{0{,}0564}{0{,}0091}=6{,}19>t_{st}=3{,}29$ для $k=100-2=98$ и $\alpha=0{,}1\%$. Нулевую гипотезу отвергают на высоком уровне значимости ($P<0{,}001$).

Эмпирические уравнения регрессии также сопровождаются ошибками. Последние, обозначаемые символами s_{yx} и s_{xy}, могут быть рассчитаны по формулам

$$s_{yx}=s_y\sqrt{1-r_{xy}^2}\,\frac{n-1}{n-2}; \qquad (206)$$

$$s_{xy}=s_x\sqrt{1-r_{xy}^2}\,\frac{n-1}{n-2}; \qquad (206а)$$

или $$s_{yx}=\sqrt{\frac{\sum(y_l-\bar{y}_x)^2}{n-2}}\quad\text{и}\quad s_{xy}=\sqrt{\frac{\sum(x_l-\bar{x}_y)^2}{n-2}}, \qquad (207)$$

где $\bar{y}_x$ и $\bar{x}_y$ — частные средние переменных Y и X; $n-2$ — число степеней свободы.

Значения s_{xy} и s_{yx} также называют *частными, парциальными* или *остаточными средними квадратическими отклонениями.* Они

описывают величину изменчивости отдельных наблюдений по отношению к линии регрессии, т. е. частным средним $\bar{y}_x$ (или $\bar{x}_y$), составляющим эту линию. Величина s_{yx} или s_{xy} позволяет судить, насколько можно ошибиться, пытаясь найти значение признака y или x как значение частной средней, полученной по уравнению регрессии.

Пример 24. Зависимость между массой тела y и возрастом детенышей макак-резусов в первый год их жизни описывается уравнением линейной регрессии $\bar{y}_x = 1{,}20 + 0{,}1104\ (x - \bar{x})$. Определим ошибку для этого уравнения. Необходимые данные содержатся в табл. 117. Расчет величины $\Sigma(y_i - \bar{y}_x)^2$ приведен в табл. 134.

Таблица 13-

Возраст детенышей x, мес	Масса тела детенышей, кг		$(y_i - \bar{y}_x)$	$(y_i - \bar{y}_x)^2$
	фактическая y	вычисленная y_x		
1	0,53	0,59	0,06	0,0036
2	0,71	0,70	0,01	0,0001
3	0,79	0,81	0,02	0,0004
4	0,98	0,92	0,06	0,0036
5	1,06	1,03	0,03	0,0009
6	1,13	1,14	0,01	0,0001
7	1,25	1,26	0,01	0,0001
8	1,43	1,37	0,06	0,0036
9	1,51	1,48	0,03	0,0009
10	1,59	1,59	0,00	0,0000
11	1,65	1,70	0,05	0,0025
12	1,77	1,81	0,04	0,0016

Подставляя известные величины в формулу (207), находим

$$s_{yx} = \sqrt{\frac{0{,}0174}{12-2}} = \sqrt{0{,}00174} = 0{,}042.$$

Если исходные данные сгруппированы в виде корреляционной таблицы, ошибку уравнения регрессии вычисляют с учетом частот ряда распределения, т. е. вместо $\Sigma(x_i - \bar{x}_y)^2$ нужно брать $\Sigma f_y(x_i - \bar{x}_y)^2$, а вместо $\Sigma(y_i - \bar{y}_x)^2$ следует находить $\Sigma f_x(y_i - \bar{y}_x)^2$.

Случайная вариация отдельных частных средних $\bar{y}_x$, принадлежащих линии регрессии, зависит от величины остаточной вариации признака Y, т. е. от s_{yx}, объема выборки n, по которой оценивали регрессионную связь, и от того, насколько далеко от средней $\bar{x}$ отстоит значение x, для которого по уравнению регрессии $\bar{y}_x = a + bx$ была найдена величина $\bar{y}_x$. Квадратическая ошибка частной средней может быть получена по формуле

$$s_{\bar{y}_x} = s_{yx}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{ns_x^2}},$$

а доверительный интервал может быть задан выражением

$$\bar{y}_x \pm ts_{\bar{y}_x},$$

где t зависит от числа степеней свободы $k=n-2$ и от принятого уровня значимости α.

Иногда практический интерес может представлять построение доверительного интервала для отдельных наблюдений, например если требуется очертить зону, включающую в себя определенный

Таблица 135

Длина тела x_i	Обхват груди y_x	$(x_i-\bar{x})/s_x$	$\bar{y}_x - ts_{y_x}$	$\bar{y}_x + ts_{y_x}$
148,5	80,5	—2,95	73,2	87,8
150,5	81,1	—2,58	73,8	88,4
152,5	81,6	—2,22	74,3	88,9
154,5	82,1	—1,85	74,8	89,4
156,5	82,7	—1,48	75,4	90,0
158,5	83,3	—1,12	76,0	90,6
160,5	83,9	—0,76	76,6	91,2
162,5	84,5	—0,38	77,2	91,8
164,5	85,0	—0,02	77,7	92,3
166,5	85,6	0,35	78,3	92,9
168,5	86,2	0,71	78,9	93,5
170,5	86,7	1,08	79,4	94,0
172,5	87,4	1,45	80,1	94,7
174,5	87,9	1,81	80,6	95,2
176,5	88,6	2,18	81,3	95,9

роцент всех эмпирических наблюдений, располагающихся возле инии регрессии. В этом случае может быть использована форму-ла квадратической ошибки отдельного наблюдения

$$s_{y_x} = s_{yx}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{ns_x^2}},$$

доверительный интервал будет иметь границы

$$\bar{y}_x \pm ts_{y_x},$$

де t имеет $k=n-2$. Следует заметить, что границы доверительного интервала для разных значений x будут расширяться в той мере, в какой эти значения будут отличаться от среднего уровня x.

Пример 25. По данным примера с нахождением уравнения регрессии длины тела X и обхвата груди Y найти доверительный интервал для индивидуальных значений, включающих 95% всех наблюдений. Воспользуемся для этого регрессией Y по X. Исход-

ными данными для примера являются значения $\bar{x}=164{,}6$; $s_y=4{,}04$; $s_x=5{,}46$; $r_{xy}=0{,}391$ и $n=727$.

Сначала найдем остаточную дисперсию признака Y. Она равна $s_{y_x}=4{,}04\sqrt{1-0{,}391^2}(726/725)=3{,}72$. Затем требуется найти нормированные отклонения для середин классовых интервалов по признаку X. Эти значения приведены в табл. 135. После этого для каждого классового интервала находим квадратическую ошибку s_{y_x}, на основании которой легко рассчитываются границы доверительного интервала.

Например, для первого класса эти границы определяют следующим образом. Квадратическая ошибка отдельного наблюдения $s_{y_x}=3{,}72\sqrt{1+1/727+(-2{,}95)^2/727}=3{,}74$. Нижняя граница доверительного интервала равна $80{,}5-1{,}96\cdot 3{,}74=73{,}2$, а верхняя его граница равна $80{,}5+1{,}96\cdot 3{,}74=87{,}8$. Границы доверительного интервала для отдельных наблюдений в соответствии с линией регрессии обхвата груди Y по длине тела X приведены на рис. 41. Эти границы можно считать своего рода нормативом формы тела. Если некоторое наблюдение окажется на графике расположенным ниже нижней границы, то для такого индивида можно констатировать нетипичное сочетание значений двух признаков, когда обхват груди по отношению к длине тела характеризуется весьма слабым развитием.

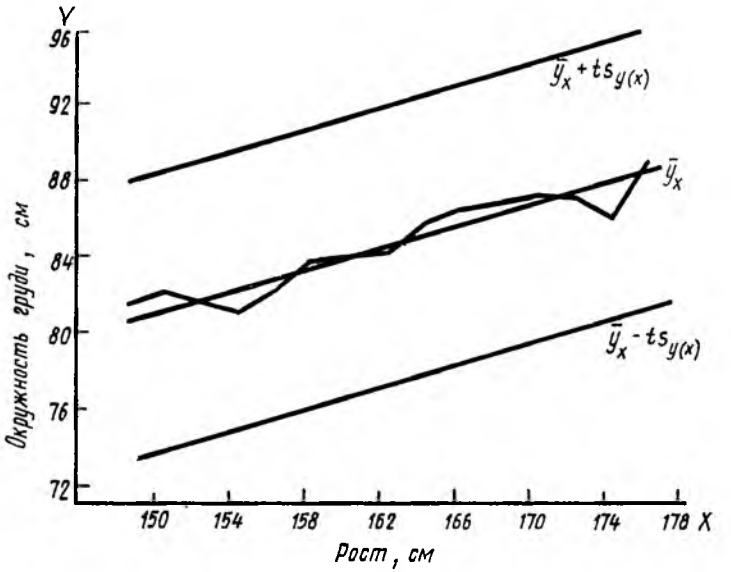


Рис. 41. Доверительная зона регрессии окружности груди Y по росту X мужчин, включающая 95% всех наблюдений

Следует также отметить еще одно обстоятельство, проявившееся в данном примере. В общем случае доверительные границы отдельных наблюдений должны расширяться по мере удаления значений признака X от центральной точки. Однако этот эффект будет выражен тем больше, чем меньше окажется объем выборки. В данном случае этот объем достаточно велик и границы доверительного интервала обнаруживают весьма слабую криволинейность. Так, для первого классового интервала ширина границ 87,8—73,2=14,6, тогда как для центрального класса она составляет 92,3—77,7=14,6. Это объясняется тем, что объем выборки достаточно велик, что обусловливает небольшую величину значений радикала $\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{ns_x^2}}$, учитывающего степень отклонения точек x_i от центра $\bar{x}$.

Решая такую задачу, следует иметь в виду, что использование уравнений линейной регрессии допустимо лишь в тех случаях, когда исходные данные распределяются нормально или же их распределение не очень сильно отклоняется от нормальной кривой. Если же генеральная совокупность, из которой извлечена выборка, распределяется по другому закону, применять уравнение линейной регрессии к нормативным оценкам варьирующих объектов нельзя. В таких случаях более подходящими будут непараметрические оценки, в частности перцентильные, о которых шла речь выше.

IX.4. ВЫБОР УРАВНЕНИЙ РЕГРЕССИИ

Важной задачей в области регрессионного анализа является выбор уравнения, которое бы наилучшим образом описывало исследуемую закономерность. Обычно эту задачу решают следующим образом. Эмпирический ряд регрессии или динамики, для которого подыскивают наилучшее корреляционное уравнение, изображают в виде точечного графика в системе прямоугольных координат. Если эмпирические точки располагаются на одной прямой или могут быть аппроксимированы прямой линией, зависимость между переменными величинами описывают уравнением линейной регрессии. Труднее выбрать наилучшее уравнение регрессии при наличии нелинейной связи между переменными величинами. В таких случаях подходящее уравнение подбирают на основании сравнения эмпирического графика с известными образцами кривых. Немаловажное значение при выборе уравнения регрессии имеют личный опыт и профессиональные знания исследователя. Иногда форма связи между переменными Y и X сама по себе подсказывает выбор наилучшего уравнения регрессии. Примером может служить лактационная кривая или кривая, от-

ражающая закономерность возрастания численного состава популяции в замкнутой среде обитания и т. п. явления.

Графический анализ не гарантирует от возможных ошибок, особенно в тех случаях, когда главное направление регрессии или динамики (их тренд) снльно затушевывается колебаниями членов ряда. Поэтому наряду с графическим анализом применяют и аналитические способы проверки правильности выбора корреляционных уравнений.

Одним из них является применение принципов *дисперсионного анализа.* Неадекватность линии регрессии, найденной теоретическим способом по отношению к эмпирической линии, может быть описана суммой квадратов отклонений $D_R^2 = \sum \left(\bar{y}_{x_l} - \hat{y}_{x_l}\right)^2$. Случайная вариация наблюдений y_{ij} по отношению к условным средним $\bar{y}_{x_l}$, найденным эмпирически, определит девиату $D_e^2 = \sum_{l=1}^{a} \sum_{j=1}^{f_x} (y_{lj} - \bar{y}_{x_l})^2$. Здесь для различения эмпирических и теоретических условных средних использованы обозначения: $\bar{y}_{x_l}$, для первой и $\hat{y}_{x_l}$ для второй. По каждой из девиат может быть найдена соответствующая дисперсия. Так, дисперсию неадекватности регрессионной модели находят по формуле

$$s_R^2 = \frac{1}{a-m} D_R^2,$$

где a — число классовых интервалов, для которых находили эмпирические средние $\bar{y}_{x_l}$; m — число параметров, определяемых в регрессионной модели. Это количество для прямолинейной регрессии равно двум, для квадратической параболы — трем и т. д. Остаточная дисперсия может быть получена по формуле

$$s_e^2 = \frac{1}{n-a} D_e^2.$$

Сравнение двух дисперсий осуществляют по F-критерию Фишера ($F = s_R^2/s_e^2$) с числами степеней свободы $k_1 = a-m$ и $k_2 = n-a$.

Характерно, что для случая проверки неадекватности прямолинейной регрессии величина F-критерия сводится к выражению

$$F = \frac{h_{yx}^2 - r_{yx}^2}{1 - h_{yx}^2} \cdot \frac{n-a}{a-2},$$

которое полностью совпадает с уже известным критерием прямолинейности корреляционных связей.

Однако при работе с непрямолинейными уравнениями регрессии эта формула оказывается неприменимой, и следует поль-

зоваться непосредственным определением девиат неадекватности и остаточной случайной изменчивости. Это осуществляется следующим образом.

Пусть по признаку X существует группировка (вариационный ряд), включающая a классов, в каждом из которых содержится f_{x_l} наблюдений. Пусть теперь в каждом классе признака X по значениям признака Y определена условная средняя $\bar{y}_{x_l} = \frac{1}{f_{x_l}} \sum_{j=1}^{f_{x_l}} y_{ij}$. Здесь суммирование производят по всему количеству наблюдений, попадающих в i-й класс. Совокупность условных средних $\bar{y}_{x_l}$ в зависимости от значений x_i (центральных значений классового интервала) составит эмпирическую линию регрессии Y по X. Пусть теперь найдена теоретическая линия регрессии $\hat{y}_x = f(x)$ и определены теоретические значения условных средних для каждого интервала по признаку X.

В соответствии с принципами дисперсионного анализа необходимо найти две девиаты: D_R^2 — неадекватности регрессии и D_e^2 — остаточную. Для проведения вычислений удобнее воспользоваться простыми формулами

$$D_e^2 = \sum_{l=1}^{a} \left(\sum_{j=1}^{f_{x_l}} y_{lj}^2 - f_{x_l} \bar{y}_{x_l}^2 \right);$$

$$D_R^2 = \sum_{l=1}^{a} f_{x_l} (\bar{y}_{x_l} - \hat{y}_{x_l})^2.$$

Согласно одиой из них, для каждого интервала находится разность двух значений условных средних $\bar{y}_{x_l}$ и $\hat{y}_{x_l}$, , которую возводят в квадрат и домножают на число наблюдений f_{x_l}. Суммирование производят по всем a классовым интервалам. Согласно второй формуле, для каждого класса вычисляют сумму квадратов значений признака $y \left(\sum_j y_{ij}^2 \right)$, свойственных тем наблюдениям, которые оказались в этом классе. Затем для каждого интервала находят величину $\sum_j y_{ij}^2 - f_{x_l} \bar{y}_{x_l}^2$, после чего все эти a значений суммируют. Делением на числа степеней свободы $k_R = a - m$ и $k_e = n - a$ получают искомые дисперсии, сопоставляемые в рамках F-критерия Фишера. Если F_Φ превысит табличное значение F_{st}, найденное для k_R и k_e, а также выбранного уровня значимости α, то необходимо будет признать, что предположение об адекватности построенной теоретической регрессии следует отклонить и попытаться испытать иную модель. В противном слу-

чае ($F_\Phi < F_{st}$) можно считать, что неадекватность проверяемой линии регрессин оказалась недоказанной, поэтому ее можно использовать.

Дисперсионный анализ применим не только к оценке рядов регрессни, но и рядов динамики, которые рассматривают как одностороние регрессии. Например, на рис. 29 показано, что на протяжении 9-летнего периода оценки знаний студентов по курсу дарвинизма варьировали. Возникает вопрос: правильно ли выбрано уравнение линейной регрессии для описания этого ряда? Читателю предлагается ответить на этот вопрос самостоятельно. Нужные данные содержатся в табл. 118.

В заключение этой главы необходимо отметить: уравнения регрессии позволяют прогнозировать возможные значения зависимой переменной на основании известных величин аргумента. При этом, однако, не следует экстраполировать регрессию за пределы проведенных опытов, так как она может менять свое направление. Область применения уравнений регрессии лучше ограничить теми данными, на которых получены эмпирические уравнения. Это предостережет исследователя от возможных ошибок.

Не следует также при отыскании эмпирических уравнений регрессии и динамики допускать значительные сокращения приближенных чисел. Округлять числа следует лишь после того, как закончены расчеты вспомогательных величин.

ГЛАВА X

ВОПРОСЫ ПЛАНИРОВАНИЯ ИССЛЕДОВАНИЙ

Классические работы Р. Фишера открыли новую страницу в истории биометрии: они показали, что планирование экспериментов и обработка их результатов — это две тесно связанные между собой задачи статистического анализа. Это открытие легло в основу разработки теории планирования экспериментов, которая в настоящее время находит применение не только при проведении сельскохозяйственных опытов, на базе которых она возникла, но и в различных областях биологии, медицины, антропологии, в сфере других научно-практических дисциплин, включая и социально-экономические исследования.

Планирование экспериментов, как уже отмечалось в предисловии к этой книге, стало самостоятельным разделом биометрии, которому посвящена огромная литература. В начальном курсе биометрии невозможно осветить все аспекты теории экспериментов. Здесь будут рассмотрены лишь некоторые общие положения, относящиеся к этой сложной и многогранной проблеме.

Термин «эксперимент» (от лат. experimentum — опыт) означает искусственно организуемый комплекс условий, в которых испытывают воздействие того или иного фактора или одновременно нескольких факторов на результативный признак. В земледелии это полевые опыты; в животноводстве — опыты по кормлению животных, по уходу за ними; в педагогике — опыты по проверке новых методов обучения и воспитания учащихся; в фармакологии — испытание эффективности новых лечебных препаратов; в медицине — проверка разных способов лечения больных и т. д.

Исследовательская работа не только сводится к экспериментам; ее проводят и вне их на основе непосредственных наблюдений. Так что выражение «планирование исследований» оказывается более емким, а следовательно, и более подходящим, чем введенный Р. Фишером (1930) термин «планирование экспериментов». Конечно, и термин «эксперимент» можно применять в более широком смысле, понимая под ним любые испытания, проводимые исследователем в отношении изучаемого объекта. При всем разнообразии методов исследовательской работы задача планирования сводится к тому, чтобы при возможно минимальных объемах наблюдений получать достаточно полную информацию об изучаемых объектах.

С варьированием результатов наблюдений связана *повторность вариантов опыта,* позволяющая повысить точность оценок генеральных параметров, надежность выводов, которые делает исследователь на основании выборочных показателей. Под повторностью в полевом опыте понимают число одноименных делянок для каждого варианта опыта. В лабораторных условиях повторность может выражаться числом одинаковых проб серий одновременных испытаний, измерений и т. п. повторений одного и того же варианта опыта. Очевидно, чем шире диапазон варьирования признака, тем больше должна быть и повторность опыта, и, наоборот, при слабом варьировании учитываемого признака число вариантов опыта, т. е. их повторность, уменьшается. В такой же зависимости от размаха варьирования признаков находится и организация планирования минимально допустимого числа испытаний.

X.1. ПРИБЛИЖЕННЫЕ ОЦЕНКИ ОСНОВНЫХ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Прежде чем наметить необходимый объем выборки, надо определить *среднюю величину* и *ее ошибку* для варьирующего признака — характеристики, которые позволяют использовать показатель точности выборочной средней при решении этой задачи. Приближенное значение средней арифметической $\bar{x}$ можно опре-

делить по полусумме лимитов:

$$\bar{x} = \frac{x_{min} + x_{max}}{2}, \qquad (208)$$

а среднее квадратическое отклонение s_x — по разности лимитов, отнесенной к коэффициенту K, который устанавливают в зависимости от объема выборки (n) с помощью табл. 136 (по Н. А. Плохинскому, 1970), т. е. по формуле

$$s_x = \frac{x_{max} - x_{min}}{K}. \qquad (209)$$

Пример 1. Зная лимиты $x_{min} = 9{,}0$ мг% и $x_{max} = 14{,}7$ мг% кальция в сыворотке крови обследованной группы обезьян ($n = 100$), можно определить основные характеристики для этой выборки:

$$\bar{x} = \frac{9{,}0 + 14{,}7}{2} = 11{,}85 \text{ мг \%} \quad \text{и} \quad s_x = \frac{14{,}7 - 9{,}0}{5} = 1{,}14.$$

Эти величины близки к фактически найденным: $\bar{x} = 11{,}94$ мг% и $s_x = 1{,}26$.

Таблица 136

n	2—5	6—15	16—49	50—200	201—1000	>1000
K	2	3	4	5	6	7

Величину ошибки средней $s_{\bar{x}}$ можно определить по следующей приближенной формуле:

$$s_{\bar{x}} = \frac{x_{max} - x_{min}}{K\sqrt{n}}. \qquad (210)$$

Так, в данном случае $s_{\bar{x}} = \frac{14{,}7 - 9{,}0}{5\sqrt{100}} = 0{,}114$. Эта же величина получается и при использовании основной формулы $s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{1{,}14}{100} = 0{,}114$. Отсюда показатель точности Cs выборочной средней $\bar{x}$ $Cs = 100\frac{s_{\bar{x}}}{\bar{x}} = 100\frac{0{,}114}{11{,}85} = 0{,}96$. Это очень высокая точность. Намечаемый таким образом объем выборки можно считать вполне достаточным для получения надежных оценок генеральных параметров (при условии , что совокупность, из которой взята выборка, распределяется по нормальному закону).

X.2. ОПРЕДЕЛЕНИЕ НЕОБХОДИМОГО ОБЪЕМА ВЫБОРКИ

Элементарная логика и практический опыт подсказывают, что неразумно стремиться к неоправданно большому числу испытаний, если убедительный результат можно получить при минимально допустимом объеме выборки. Необходимая численность выборки n, отвечающая точности, с какой намечено получить средний результат, зависит от величины ошибки выборочной средней и определяется по формуле

$$n = \frac{t^2 s_x^2}{\Delta^2} \quad \text{или} \tag{211}$$

$$n = \frac{t^2}{\Delta^2 / s_x^2} = \left(\frac{t}{K}\right)^2, \tag{212}$$

где t — нормированное отклонение, с которым связан тот или иной уровень значимости (α); s_x^2 — выборочная дисперсия; $\Delta = ts_{\bar{x}}$ — величина, определяющая границы доверительного интервала (здесь $s_{\bar{x}} = \sqrt{\frac{s_x^2}{n}}$ — ошибка выборочной средней); $K = \Delta / s_x$.

Пример 2. Случайная выборка девяти вариант характеризуется средней $\bar{x} = 12{,}1 \pm 0{,}68$. Точность выборочной средней оказалась недостаточно высокой: $Cs = 100 \frac{0{,}68}{12{,}1} = 5{,}62 = 6$. Какое число испытаний n нужно провести, чтобы ошибку средней уменьшить вдвое? В данном случае $s_x = s_{\bar{x}} \sqrt{n} = 0{,}68 \sqrt{9} = 2{,}04$. Примем $t = 1{,}96 \approx 2$, что соответствует 5%-ному уровню значимости. Предварительно определим $\Delta = 2 \frac{0{,}68}{2} = 0{,}68$; $K = \frac{0{,}68}{2{,}04} = 0{,}33$. Подставляем найденные величины в формулу (212): $n = (2/0{,}33)^2 = 6^2 = 36$.

Чтобы уменьшить ошибку репрезентативности вдвое, нужно объем выборки увеличить в четыре раза ($9 \cdot 4 = 36$). Обобщая эти данные, можно сделать вывод: для уменьшения ошибки выборочной средней в K раз нужно увеличить объем выборки в K^2 раз.

При определении необходимого объема выборки для получения статистически достоверной разности между средними $(\bar{x}_1 - \bar{x}_2) = d$ применяют формулу

$$n_2 = \left(\frac{t}{\Delta}\right)^2 \left(\frac{s_1^2}{a} + s_2^2\right). \tag{213}$$

Здесь $\Delta = ts_d$, где s_d — заданная величина ошибки для разности сравниваемых средних; s_1^2 и s_2^2 — дисперсии для сравниваемых

выборок, причем s_1^2 — дисперсия для большей выборки; $a = n_1/n_2$ — отношение объема большей выборки к объему меньшей выборки. При $n_1 = n_2$ формула (213) принимает следующий вид:

$$n = \left(\frac{t}{\Delta}\right)^2 (s_1^2 + s_2^2). \qquad (214)$$

Пример 3. Изучали влияние лечебного препарата на массу тела лабораторных мышей (см. пример 2 гл. V). Были получены следующие результаты. Характеристики опытной группы ($n_1 = 9$):

$$\bar{x}_1 = 74{,}1 \text{ г}; \quad s_1^2 = \frac{302{,}89}{9-1} = 37{,}86;$$

контрольной группы ($n_2 = 11$):

$$\bar{x}_2 = 68{,}8 \text{ г}; \quad s_2^2 = \frac{443{,}64}{11-1} = 44{,}36.$$

Разность между $\bar{x}_1$ и $\bar{x}_2$, равная $5{,}3 \pm 2{,}89$, оказалась статистически недостовериой. Определим число наблюдений n, которое необходимо провести при уменьшении ошибки разности вдвое, т. е. $s_d = 2{,}87/2 = 1{,}445$. Примем $t = 2$. Имеем $a = 11/9 = 1{,}222$ и $\Delta = 2 \cdot 1{,}445 = 2{,}89$. Отсюда $n = \left(\frac{2}{2{,}89}\right)^2 \cdot \left(\frac{44{,}364}{1{,}222} + 37{,}861\right) =$ $= 35{,}52 = 36$.

При альтернативной группировке данных, когда численность выборочных групп выражают в долях единицы, планируемый объем наблюдений определяют по формуле

$$n = \frac{t^2 p (1-p)}{\Delta^2}, \qquad (215)$$

где p — доля вариант, обладающих данным признаком; $\Delta = ts_p$.

Если доли выражают в процентах от общего числа наблюдений, формула (215) принимает следующий вид:

$$n = \frac{t^2 p (100-p)}{\Delta^2}. \qquad (216)$$

Пример 4. По предварительным данным, число гельминтоносителей среди лиц, проживающих в N-м населенном пункте, равно 8%. Определить необходимое число наблюдений, при котором величина максимальной ошибки Δ не превысит 4% для уровня значимости, равного 0,05, и соответственно $t = 2$.

Подставляя известные значения в формулу (216), находим $n = \frac{2^2 \cdot 8 (100-8)}{4^2} = \frac{2944}{16} = 184.$

Объем сведений по биометрии, рассматриваемый в данном учебном пособии, касается главным образом классической ситуации, когда анализируют отдельный признак или несколько признаков, каждый из которых рассматривают отдельно от других. Вместе с тем в последних главах, где описаны методы корреляции и регрессии, по сути дела, вскрываются возможности *биометрического анализа одновременно двух переменных.* Дальнейшее развитие теории корреляции позволило разработать так называемые *методы многомерной статистики*, которые для биолога могут считаться составляющими особый раздел биометрии — *многомерной биометрии,* рассматривающей способы анализа изменчивости не одного отдельного признака, а целых их комплексов.

В рамках небольшого послесловия можно дать лишь краткий вводный обзор многомерных методов и отослать читателя к существующей специальной литературе, часть которой написана достаточно доступно для знакомящихся с этим предметом впервые.

Среди признаков, имеющих различную форму варьирования, многомерные методы лучше разработаны для *количественных переменных*, тогда как приемы анализа качественных показателей интеисивно разрабатывают лишь в течение последних двух десятилетий. Поэтому основное изложение будет касаться первых из них.

Когда исследователь имеет в своем распоряжении набор из m признаков, то в качестве характеристик, описывающих расположение некоторой анализируемой совокупности наблюдений на осях измерений этих признаков, можно получить m средних арифметических величин $\bar{x}_i$, которые записывают в виде строки

$$[\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m],$$

называемой *вектором средних* и являющейся многомерным аналогом средней арифметической величины. Иными словами, в многомерной статистике вектор средних играет такую же роль, как и средняя арифметическая величина в одномерной биометрии.

Для любой пары признаков в качестве показателя тесноты их статистической взаимозависимости может быть найдено значение коэффициента корреляции r_{ij} или ковариация cov_{ij}. Если будут найдены все характеристики, которые можно получить для набо-

ра m признаков, то их можно записать в виде таблиц

$$\begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1m} \\ r_{12} & 1 & r_{23} & \dots & r_{2m} \\ r_{13} & r_{23} & 1 & \dots & r_{3m} \\ r_{1m} & r_{2m} & r_{3m} & \dots & 1 \end{bmatrix}$$

а также

$$\begin{bmatrix} s_1^2 & cov_{12} & cov_{13} & \dots & cov_{1m} \\ cov_{12} & s_2^2 & cov_{23} & \dots & cov_{2m} \\ cov_{13} & cov_{23} & s_3^2 & \dots & cov_{3m} \\ \dots & \dots & \dots & \dots & \dots \\ cov_{1m} & cov_{2m} & cov_{3m} & \dots & s^2_m \end{bmatrix}$$

первая из которых называется *корреляционной матрицей*, а вторая — *ковариационной матрицей* и включает в себя также значения m дисперсий признаков s_i^2, располагающиеся на диагонали таблицы. Эти две матрицы описывают закономерности изменчивости и коррелированности признаков, которые можно обнаружить для их набора. Ковариационная матрица является многомерным аналогом дисперсии признака и играет в многомерной статистике такую же роль, что и дисперсия признака в обычной биометрии.

В соответствии с двумя элементами описания изменчивости набора признаков (вектором средних и ковариационной матрицей) многомерные статистические методы грубо можно разделить на три крупных класса. 1. Приемы, которые позволяют решать задачи, аналогичные рассматриваемым в одномерной биометрии. 2. Методы анализа внутригрупповой изменчивости, когда структуру ковариационной или корреляционной матрицы исследуют с целью выявления и интерпретации закономерностей соотносительной изменчивости и коррелированности признаков. 3. Методы анализа межгрупповой изменчивости, когда сопоставляют векторы средних, найденные для нескольких выборок.

Следует заметить, что обычные алгебраические символы в применении к многомерной статистике становятся малоконструктивными и не позволяют строить по формулам, выписанным в этих символах, вычислительные алгоритмы. Поэтому основным математическим аппаратом многомерной биометрии является *матричная алгебра*, которая позволяет записывать формулы в очень компактном виде и получать по ним алгоритмы вычислений. Для использования тех многомерных статистических методов, которые могут быть интересны для биолога, достаточно ознакомления лишь с элементарными сведениями из теории матриц, которые почти всегда приводят как приложение к книгам по многомерной статистике.

Второе обстоятельство, которое следует учитывать, касается *вычислительных аспектов* многомерной статистики. Действия над векторами и матрицами в основном несложны, хотя и трудоемки. Отдельные матричные вычисления (нахождение определителей, обратных матриц, собственных чисел и векторов) часто описываются в книгах по многомерной статистике [4, 11, 17] [1], где даны рекомендации либо по ручному счету с применением калькуляторов, либо по составлению программ для ЭВМ. Матричные операции, как правило, входят в программное математическое обеспечение современных компьютеров.

Многомерные методы — аналоги одномерных. Среди таких приемов анализа многомерных данных наибольшее значение имеют *проверки статистических гипотез* по отношению к векторам средних и ковариационным матрицам, которые получены по двум или нескольким выборкам, извлеченным из двух или нескольких генеральных совокупностей. Так, при двух выборках проверку достоверности различий векторов средних осуществляют при помощи так называемого T^2*-критерия Хотеллинга*, похожего по конструкции на свой одномерный аналог — t-критерий Стьюдента. При наличии нескольких выборок, в которых найдены векторы средних, их однородность проверяют с применением многомерного аналога дисперсионного анализа. Для межвыборочной изменчивости определяют межгрупповую ковариационную матрицу, которую сопоставляют с такой же внутригрупповой матрицей в конструкции специального критерия, например критерия Уилкса. Это аналогично сравнению двух дисперсий (межгрупповой и внутригрупповой), аналогами которых являются эти ковариационные матрицы.

При использовании других многомерных методов очень распространенным является *прием нахождения некоторого нового признака* y на базе набора исходных переменных x_i в виде линейной конструкции

$$y = c_1x_1 + c_2x_2 + \ldots + c_mx_m. \qquad (*)$$

Коэффициенты c_i вычисляют таким образом, чтобы обеспечить признаку y определенные желаемые свойства, которых не имеют признаки x_i. Например, для разделения двух генеральных совокупностей (двух подвидов животных, больных и здоровых людей и т. д.) по комплексу признаков x коэффициенты c_i должны быть найдены так, чтобы y имел минимальную трансгрессию своих величин в этих генеральных совокупностях. Значения коэффициентов c_i у разных признаков позволяют интерпретировать смысл, который имеет новый признак y, т. е. описать те комплек-

[1] Цифрами в квадратных скобках указаны литературные источники в списке рекомендуемой литературы к послесловию редактора.

сы значений переменных x, которые свойственны его большим и малым величинам.

Методы анализа внутригрупповой изменчивости. Приемы многомерного анализа данных, относящиеся к этому разделу, направлены на выявление *закономерностей внутригрупповой вариации и коррелированности больших наборов переменных x.*

Наиболее близок к традиционно используемым методам парной корреляции и регрессии раздел, включающий в себя множественную корреляцию и регрессию, который кратко рассмотрен в настоящем пособии. Уравнение множественной регрессии можно рассматривать как линейную конструкцию типа (*), позволяющую находить на базе большого набора исходных признаков x такую новую переменную, которая была бы максимально скоррелирована с $(m+1)$-м признаком x_{m+1}. Эта корреляция называется *множественной.* По значениям коэффициентов c_i, которые в данном случае являются *коэффициентами множественной регрессии,* можно из всего набора m признаков x выделить только n из них, которые обнаруживают наибольшие значения этих коэффициентов. По уменьшенному набору признаков x можно построить новое уравнение регрессии, основанное на меньшем числе переменных, которое будет более компактным.

Дальнейшим развитием методов множественной регрессии и корреляции является *анализ канонических корреляций и величин.* Здесь весь набор исходных признаков x делят в соответствии с качественным содержанием задачи на две части, включающие n и $m—n$ признаков. Необходимо выявить закономерности признаков, входящих в разные их наборы. Для каждого из них определяется новый признак:

$$y=c_1x_1+c_2x_2+\ldots+c_nx_n,$$

$$y_n'=c_{n+1}x_{n+1}+c_{n+2}x_{n+2}+\ldots+c_mx_m.$$

Переменные y_1 и $\overline{y}_1'$ должны быть скоррелированы между собой максимально тесно. Смысл переменных y может быть истолкован по значениям коэффициентов c_i. Таким образом, можно считать, что y_1 и y_1' описывают наиболее важную закономерность коррелированности, которая проявляется в статистических связях признаков $x_1, x_2, \ldots, x_n$ и $x_{n+1}, x_{n+2}, \ldots, x_m$. Вместе с тем эта закономерность может оказаться не единственной, которую следует рассматривать. Тогда можно выделить другие новые переменные: y_2 и y_2', y_3 и y_3' и т. д. Новые признаки y называют *каноническими переменными,* а коэффициенты корреляции между ними — *каноническими корреляциями.*

Способ анализа корреляций большого набора признаков x может быть иным, когда невозможно или нежелательно разделять его на части, а следует рассматривать как единое целое.

Наилучшим путем анализа здесь является применение *компонентного или факторного анализа.* Согласно целям каждого из них, по корреляционной матрице признаков x находят новые линейные переменные y, которые обычно бывают не скоррелированными друг с другом (возможно выделение и связанных переменных y) и описывают определенные закономерности вариации и коррелированности исходных признаков. Эти новые переменные называют в зависимости от используемого метода *главными компонентами* или *факторами.* По значениям коэффициентов c_i у разных признаков x можно интерпретировать смысл этих переменных.

В тех случаях, когда интерпретация оказывается затруднительной, можно трансформировать эти коэффициенты с помощью специальных приемов, что часто облегчает истолкование выделенных закономерностей коррелированности признаков x.

Весьма важным является то обстоятельство, что величина каждой главной компоненты может быть получена у любого объекта исследования (экземпляра, особи, индивида и т. д.). При этом число главных компонент или факторов, суммарно описывающих весьма значительную часть информации о закономерностях вариации и коррелированности признаков, бывает гораздо меньшим, чем количество этих исходных переменных. Таким образом, применение компонентного или факторного анализа позволяет значительно уменьшить количество анализируемых переменных. Кроме того, главные компоненты являются комплексными интегративными показателями, каждый из которых зависит от многих признаков, что также весьма ценно.

Эти методы используют весьма широко, и им посвящена значительная литература. Среди наиболее простых изложений можно отметить [2, 3, 4, 5, 10, 13, 14]. Существуют и более сложные, но и более подробные описания этих методов [11, 12, 17].

Методы анализа межгрупповой изменчивости. При анализе межгрупповой изменчивости признаков решают обычно две задачи: *дискриминации* и *классификации.* В первом случае имеются две или большее число совокупностей, из которых извлечены выборки. По ним требуется получить так называемое *решающее правило,* которое позволяет на основании набора признаков x правильно отнести взятое наугад наблюдение (экземпляр, особь, индивид и т. д.) к одной из этих двух совокупностей, причем возможность ошибиться должна быть минимальной. Способы построения таких решающих правил рассматривают дискриминантным анализом. При этом на основе информации о генеральных совокупностях, полученной по выборкам, находят новый признак y, который отличается минимально возможной в данной ситуации трансгрессией своих распределений в двух совокупностях. Этот новый признак называют *дискриминантной функцией.* Величина трансгрессии, измеренная тем или иным способом, может послу-

жить основой для оценки вероятности ошибки при неправильном отнесении некоторого наблюдения.

Вопросы, связанные с вычислением и применением дискриминантных функций, относительно доступно изложены в [2, 3, 4, 15].

Задача классификации наблюдений заключается в выявлении естественного, объективно существующего порядка, присутствующего в наборе выборок, которые относятся к различным генеральным совокупностям, причем их взаимоотношения априорно обычно неясны. При решении подобных вопросов используют *методы кластерного анализа*, которые также называют *методами распознавания образов* или *числовой таксономией*.

Кластерный анализ включает в себя осуществление двух этапов обработки материала. Первый из них заключается в получении представления о взаимной близости расположения центров сравниваемых выборок по значениям комплекса признаков. Для этой цели используют различные методы. Так, для измерительных количественных признаков и многих качественных показателей по любой паре анализируемых выборок может быть найдено значение таксономического расстояния. Его величина зависит от степени сходства этих выборок по значениям признаков. Чем меньше оказываются различия векторов средних, тем меньше будет величина таксономического расстояния.

Существуют различные конструкции таксономических расстояний, среди которых одной из лучших является *расстояние Махаланобиса*, выгодно отличающееся от других учетом внутригрупповых закономерностей коррелированности признаков. Хороший обзор различных конструкций таксономических расстояний дан в [19]; об этом же можно прочесть в [4]. Для качественных признаков, имеющих альтернативную форму варьирования, могут быть найдены в качестве мер сходства выборок так называемые *коэффициенты подобия*. Здесь по всем признакам подсчитываются количества совпадающих или несовпадающих вариантов, которые затем определенным образом нормируются.

Совокупность мер сходства между всеми парами выборок может быть записана в табличном виде так называемой матрицы расстояний или коэффициентов подобия. Первая из них может быть изображена в виде

$$\begin{bmatrix} 0 & D_{12} & D_{13} & \dots & D_{1k} \\ D_{12} & 0 & D_{23} & \dots & D_{2k} \\ D_{13} & D_{23} & 0 & \dots & D_{3k} \\ \dots & \dots & \dots & \dots & \dots \\ D_{1k} & D_{2k} & D_{3k} & \dots & 0 \end{bmatrix}$$

где D_{ij} — расстояние между i-й и j-й выборками; k — число выборок.

Эти таблицы являются исходными для выделения групп выборок, включающих в себя только те из них, у которых взаимные расстояния относительно невелики. Напротив, сходство выборок из разных таких групп должно быть небольшим, а расстояния — значительными. Подобные группы относительно сходных выборок называют *кластерами* (образами, таксонами), а процесс их выделения — *кластеризацией*.

Существуют различные *методы кластеризации*. Так, согласно так называемым *агломеративным иерархическим процедурам*, которые наиболее часто используют в биологических исследованиях, процесс выделения кластеров осуществляется пошаговым образом. На первом шаге в матрице находят минимальную величину расстояния между некоторыми единицами, которые объединяют и в дальнейшем рассматривают как кластер. После нахождения расстояний этого кластера с остальными единицами отыскивают новую минимальную величину D_{ij}, так что образуется новый кластер. Такой процесс последовательного укрупнения таксонов продолжают до получения некоторой их структуры. Методы кластерного анализа описаны в [3, 4, 7].

Существует также *метод многомерного анализа межвыборочной изменчивости*, который позволяет одновременно решать как задачи дискриминантного анализа, так и проблемы классификации. Этот метод называют *каноническим анализом* (множественным дискриминантным анализом). В соответствии с ним рассматривают межгрупповые и внутригрупповые корреляционные матрицы и дисперсии. В результате находят новые линейные признаки так, чтобы каждый из них разделял анализируемые выборки с достижением минимальной трансгрессии, т. е. был *дискриминантной функцией*. Любая из них может считаться описывающей некоторую закономерность межгрупповой вариации, конкретный смысл которой истолковывают при рассмотрении коэффициентов c_i у разных признаков x. Наиболее важные из этих дискриминантных функций при попарном рассмотрении позволяют получить плоскости, расположение на которых центров выборок наглядно представляет их взаимоотношения. По этим графикам возможно выделение кластеров. О каноническом анализе читатель может прочесть в [1, 4, 20].

Использование вычислительной техники при проведении биометрических расчетов. В данном учебном пособии приведены главным образом алгоритмы, ориентированные преимущественно на ручные вычисления при помощи простейших электронных калькуляторов. Вместе с тем к настоящему времени существуют вполне доступные программируемые калькуляторы отечественного производства БЗ-34, МК-54, МК-56, МК-61, МК-52, к которым разработаны и опубликованы [6, 8, 16] значительные библиотеки программ; среди них программы автоматического проведения биометрических расчетов. Несмотря на невысокое быстродейст-

вие этих калькуляторов, их применение позволяет в несколько раз ускорить проведение биометрических вычислений, а также исключить многие возможные ошибки.

Гораздо бóльшие возможности открывает использование ЭВМ, особенно персональных. С принципами их работы и применения читатель может познакомиться по соответствующей литературе [9, 18]. Следует лишь помнить о том, что при написании программ вычисления биометрических характеристик необходимо ориентироваться на применение формул и алгоритмов, в которых фигурируют суммы анализируемых показателей: Σx, Σx^2, Σx^3, Σx^4, $\Sigma x_1 x_2$ и т. д. Получение этих величин весьма просто программируется; на их основе могут быть определены средние величины, коэффициенты асимметрии, эксцесса, корреляции и т. д. Программирование обработки вариационных рядов целесообразно главным образом для получения кривых распределения, сглаживающих эмпирическую картину.

При вводе данных в ЭВМ полезно предусмотреть программное выявление в иих ошибок ввода. Для этой цели можно использовать, например, простейшую проверку для каждого наблюдения выполнения неравенства $x_{\min} \leqslant x \leqslant x_{\max}$. Предельные значения могут быть найдены предварительно и введены в ЭВМ до начала ввода всего массива данных.

Основной трудностью обработки биометрических массовых данных на ЭВМ является их точный ввод, исключающий весьма вероятные ошибки. Поэтому целесообразнее оказывается обработка не отдельных признаков, которая позволяет получить лишь небольшой набор характеристик (среднюю, дисперсию, их ошибки, коэффициенты асимметрии и эксцесса), а одновременный обсчет сразу всех исследуемых признаков. Это позволяет вычислять кроме перечисленных одномерных показателей для каждой переменной также и значения коэффициентов корреляции для всех попарных сочетаний признаков, параметров уравнений регрессии. Для этого следует последовательно вводить в ЭВМ не отдельные значения одного признака у разных единиц наблюдения, а целые наборы признаков для каждой такой единицы.

Таблица I. Значения интеграла вероятностей для разных значений t

t	Сотые доли t									
	0	1	2	3	4	5	6	7	8	9
0,0	0000	0080	0160	0239	0319	0399	0478	0558	0638	0717
0,1	0797	0876	0955	1034	1114	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2961	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6679	6729	6778
1,0	6827	6875	6923	6970	7017	7063	7109	7154	7199	7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7995	8030
1,3	8064	8098	8182	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8788	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9089
1,7	9108	9127	9146	9164	9182	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9425	9439	9451	9464	9476	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9586	9596	9608	9615	9625	9634
2,1	9643	9652	9660	9668	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9755	9762	9768	9774	9780
2,3	9786	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9840	9845	9849	9853	9857	9861	9866	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9909	9912	9915	9917	9920	9922	9924	9926	9929
2,7	9931	9933	9935	9937	9939	9940	9942	9944	9946	9947
2,8	9949	9950	9952	9953	9955	9956	9956	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972
3,0	9973	9974	9975	9976	9976	9977	9978	9979	9979	9980
3,1	9981	9981	9982	9983	9983	9984	9984	9985	9985	9986
3,2	9986	9987	9987	9988	9988	9988	9989	9989	9990	9990
3,3	9990	9991	9991	9991	9992	9992	9992	9992	9993	9993
3,4	9993	9993	9994	9994	9994	9994	9995	9995	9995	9995
3,5	9995	9995	9996	9996	9996	9996	9996	9996	9997	9997

Примечание. Значения вероятности даны числами после запятой.

Таблица II. Значения функции $f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$

(ординаты нормальной кривой)

t	Сотые доли t									
	0	1	2	3	4	5	6	7	8	9
0,0	3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0356	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0041	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001
4,0	0001	0001	0001	0001	0001	0001	0001	0001	0001	0001

Примечание. Значения вероятности даны числами после запятой,

Таблица III. Значения вероятности $p_n(m) = \frac{a^m}{m} e^{-a}$

m \ a	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	9048	8187	7408	6703	6065	5488	4966	4493	4066	3679
1	0905	1637	2222	2681	3033	3293	3476	3595	3659	3679
2	0045	0164	0333	0536	0758	0988	1217	1438	1647	1839
3	0002	0011	0033	0072	0126	0198	0284	0383	0494	0613
4	0000	0001	0003	0007	0016	0030	0050	0077	0111	0153
5				0001	0002	0004	0007	0012	0020	0031
6							0001	0002	0003	0005
7										0001

m \ a	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	3329	3012	2725	2466	2231	2019	1827	1653	1496	1353
1	3662	3614	3543	3452	3347	3230	3106	2975	2842	2707
2	2014	2169	2303	2417	2510	2584	2640	2675	2700	2707
3	0738	0867	0998	1128	1255	1378	1496	1607	1710	1804
4	0203	0260	0324	0395	0471	0551	0636	0723	0812	0902
5	0045	0063	0084	0111	0141	0176	0216	0260	0309	0361
6	0008	0013	0018	0026	0035	0047	0061	0078	0098	0120
7	0001	0002	0003	0005	0008	0011	0015	0020	0027	0034
8			0001	0001	0001	0002	0003	0005	0006	0009
9							0001	0001	0001	0002

m \ a	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	1225	1108	1003	0903	0821	0743	0672	0608	0550	0498
1	2572	2438	2306	2177	2052	1931	1815	1703	1596	1494
2	2700	2681	2652	2613	2565	2510	2450	2384	2314	2240
3	1890	1964	2083	2090	2138	2176	2205	2225	2234	2240
4	0992	1087	1169	1254	1336	1414	1488	1557	1622	1680
5	0417	0476	0538	0602	0668	0735	0804	0872	0941	1008
6	0146	0175	0206	0241	0278	0319	0362	0407	0456	0504
7	0044	0055	0068	0083	0099	0118	0140	0163	0188	0216
8	0012	0015	0020	0025	0031	0039	0047	0057	0068	0081
9	0003	0004	0005	0007	0009	0011	0014	0018	0022	0027
10	0001	0001	0001	0002	0002	0003	0004	0005	0006	0008

m \ a	3,5	4,0	4,5	5,0	6,0	7,0	8,0	9,0	10,0	11,0
0	0302	0183	0111	0067	0025	0009	0003	0001	0000	0000
1	1057	0733	0500	0337	0149	0064	0027	0011	0005	0002
2	1850	1465	1125	0842	0446	0223	0107	0050	0023	0010
3	2153	1954	1687	1404	0892	0521	0286	0150	0076	0037
4	1888	1954	1898	1755	1339	0912	0573	0337	0189	0102
5	1327	1563	1708	1755	1606	1277	0916	0607	0378	0224
6	0771	1042	1281	1462	1606	1490	1221	0911	0631	0411
7	0386	0595	0824	1044	1377	1490	1396	1171	0901	0646
8	0169	0298	0463	0653	1033	1304	1396	1318	1126	0888
9	0066	0132	0232	0363	0688	1014	1241	1318	1251	1085
10	0023	0053	0104	0181	0413	0710	0993	1186	1251	1294
11	0007	0019	0043	0082	0225	0452	0722	0970	1137	1194
12	0002	0006	0016	0034	0113	0264	0481	0728	0948	1094
13	0001	0002	0006	0013	0052	0142	0296	0504	0729	0926
14	0000	0001	0002	0005	0022	0071	0169	0324	0521	0728

Примечание. Значения вероятности даны числами после запятой.

Таблица IV. Случайные числа

3393	6270	4228	6069	9407	1865	8549	3217	2351	8410
9108	2330	2157	7416	0398	6173	1703	8132	9065	6717
7891	3590	2502	5945	3402	0491	4328	2365	6175	7695
9085	6307	6910	9174	1753	1797	9229	3422	9861	8357
2638	2908	6368	0398	5495	3283	0031	5955	6544	3883
1313	8338	0623	8600	4950	5414	7131	0134	7241	0651
3897	4202	3814	3505	1599	1649	2784	1994	5775	1406
4380	9543	1640	2850	8415	9120	8062	2421	6161	4634
1618	6309	7909	0874	0401	4301	4517	9197	3350	0434
4858	4676	7363	9141	6133	0549	1972	3461	7116	1496
5354	9142	0847	5393	5416	6505	7156	5634	9703	6221
0905	6986	9396	3975	9255	0537	2479	4589	0562	5345
1420	0470	8679	2328	3939	1292	0406	5428	3789	2882
3218	9080	6604	1813	8209	7039	2086	3369	4437	3798
9697	8431	4387	0622	6893	8788	2320	9358	5904	9539
0912	4964	0502	9683	4636	2861	2876	1273	7870	2030
4636	7072	4868	0601	3894	7182	8417	2367	7032	1003
2515	4734	9878	6761	5636	2949	3979	8650	3430	0635
5964	0412	5012	2369	6461	0678	3693	2928	3740	8047
7848	1523	7904	1521	1455	7089	8094	9872	0898	7174
5192	2571	3643	0707	3434	6818	5729	8615	4298	4129
8438	8325	9886	1805	0226	2310	3675	5058	2515	2388
8106	6349	0319	5436	6838	2460	6433	0644	7428	8556
9158	8263	6504	2562	1160	1526	1816	9690	1215	9590
6061	3525	4048	0382	4224	7148	8259	6526	5340	4064

Таблица V. Критические точки t-критерия Стьюдента при различных уровнях значимости α

Число степеней свободы k	α, %			Число степеней свободы k	α, %		
	5	1	0,1		5	1	0,1
1	12,71	63,66	64,60	18	2,10	2,88	3,92
2	4,30	9,92	31,60	19	2,09	2,86	3,88
3	3,18	5,84	12,92	20	2,09	2,85	3,85
4	2,78	4,60	8,61	21	2,08	2,83	3,82
5	2,57	4,03	6,87	22	2,07	2,82	3,79
6	2,45	3,71	5,96	23	2,07	2,81	3,77
7	2,37	3,50	5,41	24	2,06	2,80	3,75
8	2,31	3,36	5,04	25	2,06	2,79	3,73
9	2,26	3,25	4,78	26	2,06	2,78	3,71
10	2,23	3,17	4,59	27	2,05	2,77	3,69
11	2,20	3,11	4,44	28	2,05	2,76	3,67
12	2,18	3,05	4,32	29	2,05	2,76	3,66
13	2,16	3,01	4,22	30	2,04	2,75	3,65
14	2,14	2,98	4,14	40	2,02	2,70	3,55
15	2,13	2,95	4,07	60	2,00	2,66	3,46
16	2,12	2,92	4,02	120	1,98	2,62	3,37
17	2,11	2,90	3,97	∞	1,96	2,58	3,29
P	0,05	0,01	0,001	—	0,05	0,01	0,001

Таблица VI. Значения F-критерия Фишера при уровнях значимости α=5% (верхняя строка) и α=1% (нижняя строка)

k_2	k_1 — степени свободы для большей дисперсии							
	1	2	3	4	5	6	7	8
1	161	200	216	225	230	234	237	239
	4052	4999	5403	5625	5764	5859	5928	5982
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37
	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85
	34,12	30,82	29,16	28,71	28,42	27,91	27,67	27,49
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	6,00
	21,20	18,00	16,89	15,98	15,52	15,21	14,98	14,80
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82
	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15
	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73
	12,25	9,55	8,47	7,85	7,46	7,19	6,99	6,84
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44
	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03

Продолжение табл. V

k_2	k_1 — степени свободы для большей дисперсии							
	1	2	3	4	5	6	7	8
9	5,12 10,56	4,26 8,02	3,86 6,99	3,63 6,42	3,48 6,06	3,37 5,80	3,29 5,61	3,23 5,47
10	[illegible] 10,04	4,10 7,56	3,71 6,55	3,48 5,99	3,33 5,64	3,22 5,39	3,14 5,20	3,07 5,06
11	4,84 9,65	3,98 7,21	3,59 6,22	3,36 5,67	3,20 5,32	3,09 5,07	3,01 4,89	2,95 4,74
12	4,75 9,33	3,80 6,93	3,49 5,95	3,26 5,41	3,11 5,06	3,00 4,82	2,91 4,64	2,85 4,50
13	4,67 9,07	3,80 6,70	3,41 5,74	3,18 5,21	3,02 4,86	2,92 4,62	2,83 4,44	2,77 4,30

Продолжение табл. V

k_2	k_1 — степени свободы для большей дисперсии						
	9	10	12	15	20	30	∞
1	241 6022	242 6056	244 6106	246 6157	248 6209	250 6261	254 6366
2	19,38 99,39	19,40 99,40	19,41 99,42	19,43 99,43	19,45 99,45	19,46 99,47	19,50 99,50
3	8,81 27,35	8,79 27,23	8,74 27,05	8,70 26,87	8,66 26,69	8,62 26,50	8,53 26,13
4	5,94 14,66	5,94 14,55	5,91 14,37	5,86 14,20	5,80 14,02	5,75 13,84	5,63 13,46
5	4,77 10,16	4,74 10,05	4,68 9,89	4,62 9,72	4,56 9,55	4,50 9,38	4,36 9,02
6	4,10 17,98	4,06 7,87	4,00 7,72	3,94 7,56	3,87 7,40	3,81 7,23	3,67 6,88
7	3,68 6,72	3,64 6,62	3,57 6,47	3,51 6,31	3,44 6,16	3,38 5,99	3,23 5,65
8	3,39 5,91	3,35 5,81	3,28 5,67	3,22 5,52	3,15 5,36	3,08 5,20	2,93 4,86
9	3,18 5,35	3,14 5,26	3,07 5,11	3,01 4,96	2,94 4,81	2,86 4,65	2,71 4,31
10	3,02 4,94	2,98 4,85	2,91 4,71	2,85 4,56	2,77 4,41	2,70 4,25	2,54 3,91
11	2,90 4,63	2,85 4,54	2,79 4,40	2,72 4,25	2,65 4,10	2,57 3,94	2,40 3,60
12	2,80 4,39	2,75 4,30	2,69 4,16	2,62 4,01	2,54 3,86	2,47 3,70	2,30 3,36
13	2,71 4,19	2,67 4,10	2,60 3,96	2,53 3,82	2,46 3,66	2,38 3,51	2,21 3,16

k_2	k_1 — степени свободы для большей дисперсии							
	1	2	3	4	5	6	7	8
14	4,60 8,86	3,74 6,51	3,34 5 56	3,11 5,04	2,96 4,69	2,85 4,46	2,76 4,28	2,70 4,14
15	4,54 8,68	3,68 6,36	3,29 5,42	3,06 4,89	2,90 4,56	2,79 4,32	2,71 4,14	2,64 4,00
16	4,49 8,53	3,63 6,23	3,24 5,29	3,01 4,77	2,85 4,44	2,74 4,20	2,66 4,03	2,59 3,89
17	4,45 8,40	3,59 6,11	3,20 5,18	2,96 4,67	2,81 4,34	2,70 3,93	2,61 3,79	2,55 3,68
18	4,41 8,29	3,55 6,01	3,16 5,09	2,93 4,58	2,77 4,25	2,66 4,01	2,58 3,84	2,51 3,71
19	4,38 8,18	3,52 5,93	3,13 5,01	2,90 4,50	2,74 4,17	2,63 3,94	2,54 3,77	2,48 3,61
20	4,35 8,10	3,49 5,85	3,10 4,94	2,87 4,43	2,71 4,10	2,60 3,87	2,51 3,70	2,45 3,56
21	4,32 8,02	3,47 5,78	3,07 4,87	2,84 4,37	2,68 4,04	2,57 3,81	2,49 3,64	2,42 3,51
22	4,30 7,95	3,44 5,72	3,05 4,82	2,82 4,31	2,66 3,99	2,55 3,76	2,46 3,59	2,40 3,45
23	4,28 7,88	3,42 5,66	3,03 4,76	2,80 4,26	2,64 3,94	2,53 3,71	2,44 3,54	2,37 3,41
24	4,26 7,82	3,40 5,61	3,01 4,72	2,78 4,22	2,62 3,90	2,51 3,67	2,42 3,50	2.36 3,36
25	4,24 7,77	3,39 5,57	2,99 4,68	2,76 4,18	2,60 3,85	2,49 3,63	2,40 3,46	2,34 3,32
26	4,22 7,72	3,37 5,53	2,98 4,64	2,74 4,14	2,59 3,82	2,47 3,59	2,39 3,42	2,32 3,29
27	4,21 7,68	3,35 5,49	2,96 4,60	2,73 4,11	2,57 3,78	2,46 3,56	2,37 3,39	2,31 3,26
28	4,20 7,64	3,34 5,45	2,95 4,57	2,71 4,07	2,56 3,75	2,45 3,53	2,36 3,36	2,29 3,23

k_2	k_1 — степени свободы для большей дисперсии						
	9	10	12	15	20	30	∞
14	2,65 4,03	2,60 3,94	2,53 3,80	2,46 3,66	2,39 3,51	2,31 3,35	2,13 3,00
15	2,59 3,89	2,54 3,80	2,48 3,67	2,40 3,52	2,33 3,37	2,25 3,21	2,07 2,87
16	2,54 3,78	2,49 3,69	2,42 3,55	2,35 3,41	2,28 3,26	2,19 3,10	2,01 2,75
17	2,49 3,68	2,45 3,59	2,38 3,46	2,31 3,31	2,23 3,16	2,15 3,00	1,96 2,65
18	2,42 3,60	2,38 3,51	2,31 3,37	2,23 3,23	2,16 3,08	2,07 2,92	1,88 2,57
19	2,42 3,52	2,38 3,43	2,31 3,30	2,23 3,15	2,16 3,00	2,07 2,84	1,88 2,49
20	2,39 3,46	2,35 3,37	2,28 3,23	2,20 3,09	2,12 2,94	2,04 2,78	1,84 2,42
21	2,37 3,40	2,32 3,31	2,25 3,17	2,18 3,03	2,10 2,88	2,01 2,72	1,81 2,36
22	2,34 3,35	2,30 3,26	2,23 3,12	2,15 2,98	2,07 2,83	1,98 2,67	1,78 2,31
23	2,32 3,30	2,27 3,21	2,20 3,07	2,13 2,93	2,05 2,78	1,96 2,62	1,76 2,26
24	2,30 3,26	2,25 3,17	2,18 3,03	2,11 2,89	2,03 2,74	1,94 2,58	1,73 2,21
25	2,28 3,22	2,24 3,13	2,16 2,99	2,09 2,85	2,01 2,70	1,92 2,54	1,71 2,17
26	2,27 3,18	2,22 3,09	2,15 2,96	2,07 2,81	1,99 2,66	1,90 2,50	1,69 2,13
27	2,25 3,15	2,20 3,06	2,13 2,93	2,06 2,78	1,97 2,63	1,88 2,47	1,67 2,10
28	2,24 3,12	2,19 3,03	2,12 2,90	2,04 2,75	1,96 2,60	1,87 2,44	1,65 2,06

k_2	k_1 — степени свободы для большей дисперсии							
	1	2	3	4	5	6	7	8
29	4,18 7,60	3,33 5,42	2,93 4,54	2,70 4,04	2,55 3,73	2,43 3,50	2,35 3,33	2,28 3,20
30	4,17 7,56	3,32 5,39	2,92 4,51	2,69 4,02	2,53 3,70	2,42 3,47	2,33 3,30	2,27 3,17
32	4,15 7,50	3,29 5,34	2,90 4,46	2,67 3,97	2,51 3,65	2,40 3,43	2,31 3,25	2,24 3,13
34	4,13 7,44	3,28 5,29	2,88 4,42	2,65 3,93	2,49 3,61	2,38 3,39	2,29 3,22	2,23 3,09
36	4,11 7,40	3,26 5,25	2,87 4,38	2,63 3,89	2,48 3,57	2,36 3,35	2,28 3,18	2,21 3,05
38	4,10 7,35	3,24 5,21	2,85 4,34	2,62 3,86	2,46 3,54	2,35 3,32	2,26 3,15	2,19 3,02
40	4,08 7,31	3,23 5,18	2,84 4,31	2,61 3,85	2,45 3,51	2,34 3,29	2,25 3,12	2,18 2,99
42	4,07 7,28	3,22 5,15	2,83 4,29	2,59 3,80	2,44 3,49	2,32 3,27	2,24 3,10	2,17 2,97
44	4,06 7,25	3,21 5,12	2,82 4,26	2,58 3,78	2,43 3,47	2,31 3,24	2,23 3,08	2,16 2,95
46	4,05 7,22	3,20 5,10	2,81 4,24	2,57 3,76	2,42 3,44	2,30 3,22	2,22 3,06	2,15 2,93
48	4,04 7,20	3,19 5,08	2,80 4,22	2,57 3,74	2,41 3,43	2,30 3,20	2,21 3,04	2,14 2,91
50	4,03 7,17	3,18 5,06	2,79 4,20	2,56 3,72	2,40 3,41	2,29 3,19	2,20 3,02	2,13 2,89
60	4,00 7,08	3,15 4,98	2,76 4,13	2,53 3,63	2,37 3,34	2,25 3,12	2,17 2,95	2,10 2,82
70	3,98 7,01	3,13 4,92	2,74 4,08	2,50 3,60	2,35 3,29	2,23 3,07	2,14 2,91	2,07 2,78
80	3,96 6,96	3,11 4,88	2,72 4,04	2,49 3,56	2,33 3,26	2,21 3,04	2,13 2,87	2,06 2,74
100	3,94 6,90	3,09 4,82	2,70 3,98	2,46 3,51	2,31 3,21	2,19 2,99	2,10 2,82	2,03 2,69
150	3,90 6,81	3,06 4,75	2,66 3,92	2,43 3,45	2,27 3,14	2,16 2,92	2,07 2,76	2,00 2,63
200	3,89 6,76	3,04 4,71	2,65 3,88	2,42 3,41	2,26 3,11	2,14 2,89	2,06 2,73	1,98 2,60
∞	3,64 6,63	3,00 4,61	2,60 3,78	2,37 3,32	2,21 3,02	2,10 2,80	2,01 2,64	1,94 2,51

k_2	k_1 — степени свободы для большей дисперсии								
	9	10	11	12	14	16	20	30	∞
29	2,22 3,09	2,18 3,00	2,14 2,93	2,10 2,87	2,05 2,77	2,01 2,69	1,94 2,57	1,83 2,41	1,64 2,03
30	2,21 3,00	2,16 2,17	2,13 2,90	2,09 2,84	2,04 2,74	1,99 2,66	1,93 2,55	1,84 2,38	1,62 2,01
32	2,19 3,02	2,14 2,93	2,10 2,86	2,07 2,80	2,01 2,70	1,97 2,62	1,91 2,50	1,82 2,34	1,39 1,96
34	2,17 2,98	2,12 2,89	2,08 2,28	2,05 2,76	1,99 2,66	1,95 2,58	1,89 2,46	1,80 2,30	1,57 1,91
36	2,15 2,95	2,11 2,86	2,07 2,76	2,03 2,72	1,98 2,62	1,93 2,54	1,87 2,43	1,78 2,26	1,55 1,87
38	2,14 2,95	2,09 2,82	2,05 2,75	2,02 2,69	1,96 2,59	1,92 2,51	1,85 2,40	1,76 2,23	1,53 1,84
40	2,12 2,89	2,08 2,80	2,04 2,73	2,00 2,66	1,95 2,56	1,90 2,48	1,84 2,37	1,74 2,20	1,51 1,80
42	2,11 2,86	2,06 2,78	2,03 2,70	1,99 2,64	1,93 2,54	1,89 2,46	1,83 2,34	1,73 2,18	1,49 1,78
44	2,10 2,84	2,05 2,75	2,01 2,68	1,98 2,62	1,92 2,52	1,88 2,44	1,81 2,32	1,72 2,15	1,48 1,75
46	2,09 2,82	2,04 2,73	2,00 2,66	1,97 2,60	1,91 2,50	1,87 2,42	1,80 2,30	1,71 2,13	1,46 1,73
48	2,08 2,80	2,03 2,72	1,99 2,64	1,96 2,58	1,90 2,48	1,86 2,40	1,79 2,28	1,70 2,12	1,45 1,70
50	2,07 2,79	2,03 2,70	1,99 2,63	1,95 2,56	1,89 2,46	1,85 2,38	1,78 2,26	1,69 2,10	1,44 1,68
60	2,04 2,72	1,99 2,63	1,95 2,56	1,92 2,50	1,86 2,39	1,82 2,31	1,75 2,20	1,65 2,03	1,39 1,60
70	2,02 2,67	1,97 2,59	1,93 2,51	1,89 2,45	1,84 2,35	1,79 2,27	1,72 2,15	1,62 1,98	1,35 1,53
80	2,00 2,64	1,95 2,55	1,91 2,48	1,88 2,42	1,82 2,31	1,77 2,23	1,70 2,12	1,60 1,94	1,32 1,49
100	1,97 2,59	1,93 2,50	1,89 2,43	1,85 2,37	1,79 2,26	1,75 2,19	1,68 2,06	1,57 1,89	1,28 1,43
150	1,94 2,53	1,89 2,44	1,85 2,37	1,82 2,31	1,76 2,20	1,71 2,12	1,64 2,00	1,53 1,83	1,22 1,33
200	1,93 2,50	1,88 2,41	1,84 2,34	1,80 2,27	1,74 2,17	1,69 2,09	1,62 1,97	1,52 1,79	1,19 1,28
∞	1,88 2,41	1,83 2,32	1,79 2,25	1,76 2,18	1,69 2,08	1,64 2,00	1,57 1,88	1,46 1,70	1,00 1,00

Таблица VII. χ^2-Распределение. Критические (процентные) точки для разных значений вероятности P и чисел степеней свободы k

k	α, %					P, %				
	5	2,5	1	0,5	0,1	99,9	99,5	99,0	97,5	95,0
1	3,84	5,02	6,64	7,88	10,83	—	—	—	—	—
2	5,99	7,38	9,21	10,60	13,82	—	0,01	0,02	0,05	0,10
3	7,82	9,35	11,34	12,84	16,27	0,02	0,07	0,12	0,22	0,35
4	9,49	11,14	13,28	14,86	18,47	0,09	0,21	0,30	0,48	0,71
5	11,07	12,83	15,09	16,75	20,52	0,21	0,41	0,55	0,83	1,14
6	12,59	14,15	16,81	18,55	22,46	0,38	0,68	0,87	1,24	1,64
7	14,07	16,01	18,48	20,28	24,32	0,60	0,99	1,24	1,69	2,17
8	15,51	17,54	20,09	21,96	26,12	0,86	1,44	1,65	2,18	2,73
9	16,92	19,02	21,67	23,59	27,88	1,15	1,54	2,09	2,70	3,32
10	18,31	20,48	23,21	25,19	29,59	1,48	2,65	2,56	3,25	3,94
11	19,68	21,92	24,72	26,76	31,26	1,83	2,31	3,05	3,82	4,58
12	21,03	23,34	26,22	28,30	32,91	2,21	3,47	3,57	4,40	5,23
13	22,36	24,74	27,69	29,82	34,53	2,62	3,56	4,11	5,01	5,89
14	23,68	26,12	29,14	31,32	36,12	3,04	4,57	4,66	5,63	6,57
15	25,00	27,49	30,58	32,80	37,70	3,48	4,11	5,23	6,26	7,26
16	26,30	28,84	32,00	34,27	39,25	3,94	5,24	5,81	6,91	7,96
17	27,59	30,19	33,41	35,72	40,79	4,42	5,80	6,41	7,56	8,67
18	28,87	31,53	34,80	37,16	42,31	4,91	6,56	7,02	8,23	9,39
19	30,14	32,85	36,19	38,18	43,82	5,41	6,45	7,63	8,91	10,12
20	31,41	34,17	37,57	40,00	45,32	5,92	7,43	8,27	9,59	10,85
21	32,67	35,48	38,93	41,40	46,80	6,45	8,43	8,90	10,28	11,59
22	33,92	36,78	40,29	42,80	48,27	6,98	8,35	9,54	10,98	12,34
23	35,17	38,08	41,64	44,18	49,73	7,53	9,06	10,20	11,69	13,09
24	36,42	39,36	42,98	45,56	51,18	8,09	9,69	10,86	12,40	13,85
25	37,65	40,65	44,31	46,93	52,62	8,65	10,02	11,52	13,12	14,61
26	38,88	41,92	45,64	48,29	54,05	9,22	11,06	12,20	13,84	15,38
27	40,11	43,19	46,96	49,64	55,48	9,80	11,81	12,88	14,57	16,15
28	41,34	44,46	48,28	50,99	56,89	10,39	12,46	13,56	15,31	16,93
29	42,56	45,72	49,59	52,34	58,30	10,99	13,12	14,25	16,05	17,71
30	43,77	46,98	50,89	53,67	59,70	11,59	13,79	14,95	16,79	18,49
31	44,93	48,23	52,19	55,00	61,10	12,20	14,46	15,66	17,54	19,28
32	46,19	49,48	53,49	56,33	62,49	12,81	15,13	16,36	18,29	20,07
33	47,40	50,72	54,78	57,65	63,87	13,43	15,82	17,07	19,05	20,88
34	48,60	51,97	56,06	58,96	65,25	14,06	16,50	17,79	19,81	21,66
35	49,80	53,20	57,34	60,28	66,62	14,69	17,19	18,51	20,57	22,46
36	51,00	54,44	58,62	61,58	67,98	15,32	17,89	19,23	21,34	23,27
37	52,19	55,67	59,89	62,88	69,35	15,94	18,59	19,96	22,11	24,08
38	53,38	56,90	61,18	64,18	70,70	16,61	19,29	20,69	22,88	24,88
39	54,57	58,12	62,43	65,48	72,06	17,26	20,00	21,43	23,65	25,70
40	55,76	59,34	63,69	66,77	73,40	17,92	20,71	22,16	24,43	26,51
41	56,94	60,56	64,95	68,05	74,74	18,58	21,42	22,91	25,22	27,33
42	58,12	61,78	66,21	69,34	76,08	19,24	22,14	23,65	26,00	28,14
43	59,30	62,99	67,46	70,62	77,42	19,91	22,86	24,40	26,78	28,97
44	60,48	64,20	68,71	71,89	78,75	20,58	23,58	25,15	27,58	29,79
45	61,66	65,41	69,96	73,17	80,08	21,25	24,31	25,90	28,37	30,61
46	62,83	66,62	71,20	74,44	81,40	21,93	25,04	26,66	29,16	31,44
47	64,00	67,82	72,44	75,70	82,72	22,61	25,78	27,42	29,96	32,27
48	65,17	69,02	73,68	76,97	84,04	23,30	26,51	28,18	30,76	33,10
49	66,34	70,22	74,92	78,23	85,35	23,98	27,25	28,94	31,56	33,93
50	67,51	71,42	76,15	79,49	86,66	24,67	27,99	29,71	32,36	34,76
51	68,67	72,62	77,39	80,75	87,97	25,37	28,74	30,48	53,16	35,60

Продолжение табл. VI

k	α, %					P, %				
	5	2,5	1	0,5	0,1	99,9	99,5	99,0	97,5	95,0
52	69,83	73,81	78,62	82,00	89,27	26,06	29,48	31,25	33,97	36,44
53	70,99	75,00	79,84	83,25	90,57	26,76	30,23	32,02	34,78	37,28
54	72,15	76,19	81,07	84,50	91,87	27,47	30,98	32,79	35,59	38,12
55	73,31	77,38	82,29	85,75	93,17	28,17	31,74	33,57	36,40	38,96
56	74,47	78,57	83,51	86,99	94,46	28,88	32,49	34,35	37,21	39,80
57	75,62	79,75	84,73	88,24	95,75	29,59	33,25	35,13	38,03	40,65
58	76,78	80,94	85,95	89,48	97,04	30,30	34,01	35,91	38,84	41,49
59	77,93	82,12	87,17	90,72	98,32	31,02	34,77	36,70	39,66	42,34
60	79,08	83,30	88,38	91,95	99,61	31,74	35,54	37,48	40,48	43,19
61	80,23	84,48	89,59	93,19	100,89	32,46	36,30	38,27	41,30	44,04
62	81,38	85,65	90,80	94,42	102,17	33,18	37,07	39,06	42,13	44,89
63	82,53	86,83	92,01	95,65	103,44	33,91	37,84	39,86	42,95	45,74
64	83,68	88,00	93,22	96,88	104,72	34,63	38,61	40,65	43,78	46,60
65	84,82	89,18	94,42	98,11	105,99	35,36	39,38	41,44	44,60	47,45
66	85,97	90,35	95,63	99,33	107,26	36,09	40,16	42,24	45,43	48,30
67	87,11	91,52	96,83	100,6	108,53	36,83	40,94	43,04	46,26	49,16
68	88,25	92,69	98,03	101,8	109,79	37,56	41,71	43,84	47,09	50,02
69	89,39	93,86	99,23	103,0	111,06	38,30	42,49	44,64	47,92	50,88
70	90,53	95,02	100,4	104,2	112,32	39,04	43,28	45,44	48,76	51,74
71	91,67	96,19	101,6	105,4	113,58	39,78	44,06	46,25	49,59	52,60
72	92,81	97,35	102,8	106,6	114,84	40,52	44,84	47,05	50,43	53,46
73	93,94	98,52	104,0	107,8	116,09	41,26	45,63	47,86	51,26	54,32
74	95,08	99,68	105,2	109,1	117,35	42,01	46,42	48,67	52,10	55,19
75	96,22	100,8	106,4	110,3	118,60	42,76	47,21	49,48	52,94	56,05
76	97,35	102,00	107,58	111,50	119,85	43,51	48,00	50,29	53,78	56,92
77	98,48	103,16	108,77	112,70	121,10	44,26	48,79	51,10	54,62	57,79
78	99,62	104,32	109,96	113,91	122,35	45,01	49,58	51,91	55,47	58,65
79	100,75	105,47	111,14	115,12	123,59	45,76	50,38	52,72	56,31	59,52
80	101,88	106,63	112,33	116,32	124,84	46,52	51,17	53,54	57,15	60,39
81	103,01	107,78	113,51	117,52	126,08	47,28	51,97	54,36	58,00	61,26
82	104,14	108,94	114,70	118,73	127,32	48,04	52,77	55,17	58,84	62,13
83	105,27	110,09	115,88	119,93	128,56	48,80	53,57	55,99	59,69	63,00
84	106,40	111,24	117,06	121,13	129,80	49,56	54,37	56,81	60,54	63,88
85	107,52	112,39	118,24	122,32	131,04	50,32	55,17	57,63	61,39	64,75
86	108,65	113,54	119,41	123,52	132,28	51,08	55,97	58,46	62,24	65,62
87	109,77	114,69	120,59	124,72	133,51	51,85	56,78	59,28	63,09	66,50
88	110,90	115,84	121,77	125,91	134,74	52,62	57,58	60,10	63,94	67,37
89	112,02	117,00	122,94	127,11	135,98	53,39	58,39	60,93	64,79	68,25
90	113,14	118,14	124,12	128,30	137,21	54,16	59,20	61,75	65,65	69,13
91	114,27	119,28	125,29	129,49	138,44	54,93	60,00	62,58	66,50	70,00
92	115,39	120,43	126,46	130,68	139,67	55,70	60,82	63,41	67,36	70,88
93	116,51	121,57	127,63	131,87	140,89	56,47	61,62	64,24	68,21	71,76
94	117,63	122,72	128,80	133,06	142,12	57,25	62,44	65,07	69,07	72,64
95	118,75	123,86	129,97	134,25	143,34	58,02	63,25	65,90	69,92	73,52
96	119,87	125,00	131,14	135,43	144,57	58,80	64,06	66,73	70,78	74,40
97	120,99	126,14	132,31	136,62	145,79	59,58	64,88	67,56	71,64	75,28
98	122,11	127,29	133,48	137,80	147,01	60,36	65,69	68,40	72,50	76,16
99	123,22	128,42	134,64	138,99	148,23	61,14	66,51	69,23	73,36	77,05
100	124,34	129,56	135,81	140,17	149,45	61,92	67,33	70,06	74,22	77,93
P	0,05	0,025	0,01	0,005	0,001	0,999	0,995	0,990	0,975	0,950

Таблица VIII. Значения $\varphi = 2 \arcsin \sqrt{P}$

P	0	1	2	3	4	5	6	7	8	9
0,0	0,000	0,020	0,028	0,035	0,040	0,045	0,049	0,053	0,057	0,060
0,1	0,063	0,066	0,069	0,072	0,075	0,077	0,080	0,082	0,085	0,087
0,2	0,089	0,092	0,094	0,096	0,098	0,100	0,102	0,104	0,106	0,108
0,3	0,110	0,111	0,113	0,115	0,117	0,118	0,120	0,122	0,123	0,125
0,4	0,127	0,128	0,130	0,131	0,133	0,134	0,136	0,137	0,139	0,140
0,5	0,142	0,143	0,144	0,146	0,147	0,148	0,150	0,151	0,153	0,154
0,6	0,155	0,156	0,158	0,159	0,160	0,161	0,163	0,164	0,165	0,166
0,7	0,168	0,169	0,170	0,171	0,172	0,173	0,175	0,176	0,177	0,178
0,8	0,179	0,180	0,182	0,183	0,184	0,185	0,186	0,187	0,188	0,189
0,9	0,190	0,191	0,192	0,193	0,194	0,195	0,196	0,197	0,198	0,199
1	0,200	0,210	0,220	0,229	0,237	0,246	0,254	0,262	0,269	0,277
2	0,284	0,291	0,298	0,304	0,311	0,318	0,324	0,330	0,336	0,342
3	0,348	0,354	0,360	0,363	0,371	0,376	0,382	0,387	0,392	0,398
4	0,403	0,408	0,413	0,418	0,423	0,428	0,432	0,437	0,442	0,448
5	0,451	0,456	0,460	0,465	0,469	0,473	0,478	0,482	0,486	0,491
6	0,495	0,499	0,503	0,507	0,512	0,516	0,520	0,524	0,528	0,532
7	0,535	0,539	0,543	0,546	0,551	0,555	0,559	0,562	0,566	0,570
8	0,574	0,577	0,581	0,584	0,588	0,592	0,595	0,599	0,602	0,606
9	0,609	0,613	0,616	0,620	0,623	0,627	0,630	0,633	0,637	0,640
10	0,644	0,647	0,650	0,653	0,657	0,660	0,663	0,666	0,670	0,673
11	0,676	0,679	0,682	0,686	0,689	0,692	0,695	0,698	0,701	0,704
12	0,707	0,711	0,714	0,717	0,720	0,723	0,726	0,729	0,732	0,735
13	0,738	0,741	0,744	0,747	0,750	0,752	0,755	0,758	0,761	0,764
14	0,767	0,770	0,773	0,776	0,778	0,781	0,784	0,787	0,790	0,793
15	0,795	0,798	0,801	0,804	0,807	0,809	0,812	0,815	0,818	0,820
16	0,823	0,826	0,828	0,831	0,834	0,837	0,839	0,842	0,845	0,847
17	0,850	0,853	0,855	0,858	0,861	0,863	0,866	0,868	0,871	0,874
18	0,876	0,879	0,881	0,884	0,887	0,889	0,892	0,894	0,897	0,900
19	0,902	0,905	0,907	0,910	0,912	0,915	0,917	0,920	0,922	0,925
20	0,927	0,930	0,932	0,935	0,937	0,940	0,942	0,945	0,947	0,950
21	0,952	0,955	0,957	0,959	0,962	0,964	0,967	0,969	0,972	0,974
22	0,976	0,979	0,981	0,984	0,986	0,988	0,991	0,993	0,996	0,998
23	1,000	1,003	1,005	1,007	1,010	1,012	1,015	1,017	1,019	1,022
24	1,024	1,026	1,029	1,031	1,033	1,036	1,038	1,040	1,043	1,045
25	1,047	1,050	1,052	1,054	1,056	1,059	1,061	1,063	1,066	1,068
26	1,070	1,072	1,075	1,077	1,079	1,082	1,084	1,086	1,088	1,091
27	1,093	1,095	1,097	1,100	1,102	1,104	1,106	1,109	1,111	1,113
28	1,115	1,117	1,120	1,122	1,124	1,126	1,129	1,131	1,133	1,135
29	1,137	1,140	1,142	1,144	1,146	1,148	1,151	1,153	1,155	1,157
30	1,159	1,161	1,164	1,166	1,168	1,170	1,172	1,174	1,177	1,179
31	1,182	1,183	1,185	1,187	1,190	1,192	1,184	1,196	1,198	1,200
32	1,203	1,205	1,207	1,209	1,211	1,213	1,215	1,217	1,220	1,222
33	1,224	1,226	1,228	1,230	1,232	1,234	1,237	1,239	1,241	1,243
34	1,245	1,247	1,249	1,251	1,254	1,256	1,258	1,260	1,262	1,264
35	1,266	1,268	1,270	1,272	1,274	1,277	1,279	1,281	1,283	1,285
36	1,287	1,289	1,291	1,293	1,295	1,297	1,299	1,302	1,304	1,306
37	1,308	1,310	1,312	1,314	1,316	1,318	1,320	1,322	1,324	1,326
38	1,328	1,330	1,333	1,335	1,337	1,339	1,341	1,343	1,345	1,347
39	1,349	1,351	1,353	1,355	1,357	1,359	1,361	1,363	1,365	1,367
40	1,369	1,371	1,374	1,376	1,378	1,380	1,382	1,384	1,386	1,388

P	0	1	2	3	4	5	6	7	8	9
41	1,390	1,392	1,394	1,396	1,398	1,400	1,402	1,404	1,406	1,408
42	1,410	1,412	1,414	1,416	1,418	1,420	1,422	1,424	1,426	1,428
43	1,430	1,432	1,434	1,436	1,438	1,440	1,442	1,444	1,446	1,448
44	1,451	1,453	1,455	1,457	1,459	1,461	1,463	1,465	1,467	1,469
45	1,471	1,473	1,475	1,477	1,479	1,481	1,483	1,485	1,487	1,489
46	1,491	1,493	1,495	1,497	1,499	1,501	1,503	1,505	1,507	1,509
47	1,511	1,513	1,515	1,517	1,519	1,521	1,523	1,525	1,527	1,529
48	1,531	1,533	1,535	1,537	1,539	1,541	1,543	1,545	1,547	1,549
49	1,551	1,553	1,555	1,557	1,559	1,561	1,563	1,565	1,567	1,569
50	1,571	1,573	1,575	1,577	1,579	1,581	1,583	1,585	1,587	1,589
51	1,591	1,593	1,595	1,597	1,599	1,601	1,603	1,605	1,607	1,609
52	1,611	1,613	1,615	1,617	1,619	1,621	1,623	1,625	1,627	1,629
53	1,631	1,633	1,635	1,637	1,639	1,641	1,643	1,645	1,647	1,649
54	1,651	1,653	1,655	1,657	1,659	1,661	1,663	1,665	1,667	1,669
55	1,671	1,673	1,675	1,677	1,679	1,681	1,683	1,685	1,687	1,689
56	1,691	1,693	1,695	1,697	1,699	1,701	1,703	1,705	1,707	1,709
57	1,711	1,713	1,715	1,717	1,719	1,721	1,723	1,725	1,727	1,729
58	1,731	1,734	1,736	1,738	1,740	1,742	1,744	1,746	1,748	1,750
59	1,752	1,754	1,756	1,758	1,760	1,762	1,764	1,766	1,768	1,770
60	1,772	1,774	1,776	1,778	1,780	1,782	1,784	1,786	1,789	1,791
61	1,793	1,795	1,797	1,799	1,801	1,803	1,805	1,807	1,809	1,811
62	1,813	1,815	1,817	1,819	1,821	1,823	1,826	1,828	1,830	1,832
63	1,834	1,836	1,838	1,840	1,842	1,844	1,846	1,848	1,850	1,853
64	1,855	1,857	1,859	1,861	1,863	1,865	1,867	1,869	1,871	1,873
65	1,875	1,878	1,880	1,882	1,884	1,886	1,888	1,890	1,892	1,894
66	1,897	1,899	1,901	1,903	1,905	1,907	1,909	1,911	1,913	1,916
67	1,918	1,920	1,922	1,924	1,926	1,928	1,930	1,933	1,935	1,937
68	1,939	1,941	1,943	1,946	1,948	1,950	1,952	1,954	1,956	1,958
69	1,961	1,963	1,965	1,967	1,969	1,971	1,974	1,976	1,978	1,980
70	1,982	1,984	1,987	1,989	1,991	1,993	1,995	1,998	2,000	2,002
71	2,004	2,006	2,009	2,011	2,013	2,015	2,018	2,020	2,022	2,024
72	2,026	2,029	2,031	2,033	2,035	2,038	2,040	2,042	2,044	2,047
73	2,049	2,051	2,053	2,056	2,058	2,060	2,062	2,065	2,067	2,069
74	2,071	2,074	2,076	2,078	2,081	2,083	2,085	2,087	2,090	2,092
75	2,094	2,097	2,099	2,101	2,104	2,106	2,108	2,111	2,113	2,115
76	2,118	2,120	2,122	2,125	2,127	2,129	2,132	2,134	2,136	2,139
77	2,141	2,144	2,146	2,148	2,151	2,153	2,156	2,158	2,160	2,163
78	2,165	2,168	2,170	2,172	2,175	2,177	2,180	2,182	2,185	2,187
79	2,190	2,192	2,194	2,197	2,199	2,202	2,204	2,207	2,209	2,212
80	2,214	2,217	2,219	2,222	2,224	2,224	2,227	2,229	2,231	2,237
81	2,240	2,242	2,245	2,247	2,250	2,252	2,255	2,258	2,260	2,263
82	2,265	2,268	2,271	2,273	2,276	2,278	2,281	2,284	2,286	2,289
83	2,292	2,294	2,297	2,300	2,302	2,305	2,308	2,310	2,313	2,316
84	2,319	2,321	2,324	2,327	2,330	2,332	2,335	2,338	2,341	2,343
85	2,346	2,349	2,352	2,355	2,357	2,360	2,363	2,366	2,369	2,372
86	2,375	2,377	2,380	2,383	2,386	2,389	2,392	2,395	2,398	2,402
87	2,404	2,407	2,410	2,413	2,416	2,419	2,422	2,425	2,428	2,431
88	2,434	2,437	2,440	2,443	2,447	2,450	2,453	2,456	2,459	2,462
89	2,465	2,469	2,472	2,475	2,478	2,482	2,485	2,488	2,491	2,495
90	2,498	2,501	2,505	2,508	2,512	2,515	2,518	2,522	2,525	2,529

P	0	1	2	3	4	5	6	7	8	9
91	2,532	2,536	2,539	2,543	2,546	2,550	2,554	2,557	2,561	2,564
92	2,568	2,572	2,575	2,579	2,583	2,587	2,591	2,594	2,598	2,602
93	2,606	2,610	2,614	2,618	2,622	2,626	2,630	2,634	2,638	2,642
94	2,647	2,651	2,655	2,659	2,664	2,668	2,673	2,677	2,681	2,686
95	2,691	2,695	2,700	2,706	2,709	2,714	2,719	2,724	2,729	2,734
96	2,739	2,744	2,749	2,754	2,760	2,765	2,771	2,776	2,782	2,788
97	2,793	2,799	2,805	2,811	2,818	2,824	2,830	2,837	2,844	2,851
98	2,858	2,865	2,872	2,880	2,888	2,896	2,904	2,913	2,922	2,931
99	2,941	2,942	2,943	2,944	2,945	2,946	2,948	2,949	2,950	2,951
99,1	2,952	2,953	2,954	2,955	2,956	2,957	2,958	2,959	2,960	2,961
99,2	2,963	2,964	2,965	2,966	2,967	2,968	2,969	2,971	2,972	2,973
99,3	2,974	2,975	2,976	2,978	2,979	2,980	2,981	2,983	2,984	2,985
99,4	2,987	2,988	2,989	2,990	2,992	2,993	2,995	2,996	2,997	2,999
99,5	3,000	3,002	3,003	3,004	3,006	3,007	3,009	3,010	3,012	3,013
99,6	3,015	3,017	3,018	3,020	3,022	3,023	3,025	3,027	3,028	3,030
99,7	3,032	3,034	3,036	3,038	3,040	3,041	3,044	3,046	3,048	3,050
99,8	3,052	3,054	3,057	3,059	3,062	3,064	3,067	3,069	3,072	3,075
99,9	3,078	3,082	3,085	3,089	3,093	3,097	3,101	3,107	3,113	3,122
100	3,142									

Таблица IX. Значения функции $\psi\left(\frac{R}{n+1}\right)$

$\frac{R}{n+1}$	0	1	2	3	4	5	6	7	8	9
0,00	$-\infty$	−3,09	−2,88	−2,75	−2,65	−2,58	−2,51	−2,46	−2,41	−2,37
0,01	−2,53	−2,29	−2,26	−2,23	−2,20	−2,17	−2,14	−2,12	−2,10	−2,07
0,02	−2,05	−2,03	−2,01	−2,00	−1,98	−1,96	−1,94	−1,93	−1,91	−1,90
0,03	−1,88	−1,87	−1,85	−1,84	−1,83	−1,81	−1,80	−1,79	−1,77	−1,76
0,04	−1,75	−1,74	−1,73	−1,72	−1,71	−1,70	−1,68	−1,67	−1,66	−1,65
0,05	−1,64	−1,64	−1,63	−1,62	−1,61	−1,60	−1,59	−1,58	−1,57	−1,57
0,06	−1,55	−1,55	−1,54	−1,53	−1,52	−1,51	−1,51	−1,50	−1,49	−1,48
0,07	−1,48	−1,47	−1,46	−1,45	−1,45	−1,44	−1,43	−1,43	−1,42	−1,41
0,08	−1,41	−1,40	−1,39	−1,39	−1,38	−1,37	−1,37	−1,36	−1,35	−1,35
0,09	−1,34	−1,33	−1,33	−1,32	−1,32	−1,31	−1,30	−1,30	−1,29	−1,29
0,10	−1,28	−1,28	−1,27	−1,26	−1,26	−1,25	−1,25	−1,24	−1,24	−1,23
0,11	−1,23	−1,22	−1,22	−1,21	−1,21	−1,20	−1,20	−1,19	−1,19	−1,18
0,12	−1,18	−1,17	−1,17	−1,16	−1,16	−1,15	−1,15	−1,14	−1,14	−1,13
0,13	−1,13	−1,12	−1,12	−1,11	−1,11	−1,10	−1,10	−1,09	−1,09	−1,09
0,14	−1,08	−1,08	−1,07	−1,07	−1,06	−1,06	−1,05	−1,05	−1,05	−1,04
0,15	−1,04	−1,03	−1,03	−1,02	−1,02	−1,02	−1,01	−1,01	−1,01	−1,00
0,16	−0,99	−0,99	−0,99	−0,98	−0,98	−0,97	−0,97	−0,97	−0,96	−0,96
0,17	−0,95	−0,95	−0,95	−0,94	−0,94	−0,93	−0,93	−0,93	−0,92	−0,92
0,18	−0,92	−0,91	−0,91	−0,90	−0,90	−0,90	−0,89	−0,89	−0,89	−0,88
0,19	−0,88	−0,87	−0,87	−0,87	−0,86	−0,86	−0,86	−0,85	−0,85	−0,85
0,20	−0,84	−0,84	−0,83	−0,83	−0,83	−0,82	−0,82	−0,82	−0,81	−0,81
0,21	−0,81	−0,80	−0,80	−0,80	−0,79	−0,79	−0,79	−0,78	−0,78	−0,78

$\frac{R}{n+1}$	0	1	2	3	4	5	6	7	8	9
0,22	—0,77	—0,77	—0,77	—0,76	—0,76	—0,76	—0,75	—0,75	—0,75	—0,74
0,23	—0,74	—0,74	—0,73	—0,73	—0,73	—0,72	—0,72	—0,72	—0,71	—0,71
0,24	—0,71	—0,70	—0,70	—0,70	—0,69	—0,69	—0,69	—0,68	—0,68	—0,68
0,25	—0,67	—0,67	—0,67	—0,67	—0,66	—0,66	—0,66	—0,65	—0,65	—0,65
0,26	—0,64	—0,64	—0,64	—0,63	—0,63	—0,63	—0,63	—0,62	—0,62	—0,62
0,27	—0,61	—0,61	—0,61	—0,60	—0,60	—0,60	—0,60	—0,59	—0,59	—0,59
0,28	—0,58	—0,58	—0,58	—0,57	—0,57	—0,57	—0,57	—0,56	—0,56	—0,56
0,29	—0,55	—0,55	—0,55	—0,54	—0,54	—0,54	—0,54	—0,53	—0,53	—0,53
0,30	—0,53	—0,52	—0,52	—0,52	—0,51	—0,51	—0,51	—0,50	—0,50	—0,50
0,31	—0,50	—0,49	—0,49	—0,49	—0,48	—0,48	—0,48	—0,47	—0,47	—0,47
0,32	—0,47	—0,46	—0,46	—0,46	—0,46	—0,45	—0,45	—0,45	—0,45	—0,44
0,33	—0,44	—0,44	—0,43	—0,43	—0,43	—0,43	—0,43	—0,42	—0,42	—0,42
0,34	—0,41	—0,41	—0,41	—0,40	—0,40	—0,40	—0,40	—0,39	—0,39	—0,39
0,35	—0,39	—0,38	—0,38	—0,38	—0,37	—0,37	—0,37	—0,37	—0,36	—0,36
0,36	—0,36	—0,36	—0,35	—0,35	—0,35	—0,35	—0,34	—0,34	—0,34	—0,33
0,37	—0,33	—0,33	—0,33	—0,32	—0,32	—0,32	—0,32	—0,31	—0,31	—0,31
0,38	—0,31	—0,30	—0,30	—0,30	—0,30	—0,29	—0,29	—0,29	—0,28	—0,28
0,39	—0,28	—0,28	—0,27	—0,27	—0,27	—0,27	—0,26	—0,26	—0,26	—0,26
0,40	—0,25	—0,25	—0,25	—0,25	—0,24	—0,24	—0,24	—0,24	—0,23	—0,23
0,41	—0,23	—0,23	—0,22	—0,22	—0,22	—0,21	—0,21	—0,21	—0,21	—0,20
0,42	—0,20	—0,20	—0,20	—0,19	—0,19	—0,19	—0,19	—0,18	—0,18	—0,18
0,43	—0,18	—0,17	—0,17	—0,17	—0,17	—0,16	—0,16	—0,16	—0,16	—0,15
0,44	—0,15	—0,15	—0,15	—0,14	—0,14	—0,14	—0,14	—0,13	—0,13	—0,13
0,45	—0,13	—0,12	—0,12	—0,12	—0,12	—0,11	—0,11	—0,11	—0,11	—0,10
0,46	—0,10	—0,10	—0,10	—0,09	—0,09	—0,09	—0,09	—0,08	—0,08	—0,08
0,47	—0,08	—0,07	—0,07	—0,07	—0,07	—0,06	—0,06	—0,06	—0,06	—0,05
0,48	—0,05	—0,05	—0,05	—0,04	—0,04	—0,04	—0,04	—0,03	—0,03	—0,03
0,49	—0,03	—0,02	—0,02	—0,02	—0,02	—0,01	—0,01	—0,01	—0,01	—0,00
0,50	+0,00	+0,00	+0,01	+0,01	+0,01	+0,01	+0,02	+0,02	+0,02	+0,02
0,51	0,03	0,03	0,03	0,03	0,04	0,04	0,04	0,04	0,05	0,05
0,52	0,05	0,05	0,06	0,06	0,06	0,06	0,07	0,07	0,07	0,07
0,53	0,08	0,08	0,08	0,08	0,09	0,09	0,09	0,09	0,10	0,10
0,54	0,10	0,10	0,11	0,11	0,11	0,11	0,12	0,12	0,12	0,12
0,55	0,13	0,13	0,13	0,13	0,14	0,14	0,14	0,14	0,15	0,15
0,56	0,15	0,15	0,16	0,16	0,16	0,16	0,17	0,17	0,17	0,17
0,57	0,18	0,18	0,18	0,18	0,19	0,19	0,19	0,19	0,20	0,20
0,58	0,20	0,20	0,21	0,21	0,21	0,21	0,22	0,22	0,22	0,23
0,59	0,23	0,23	0,23	0,24	0,24	0,24	0,24	0,25	0,25	0,25
0,60	0,25	0,26	0,26	0,26	0,26	0,27	0,27	0,27	0,27	0,28
0,61	0,28	0,28	0,28	0,29	0,29	0,29	0,30	0,30	0,30	0,30
0,62	0,31	0,31	0,31	0,31	0,32	0,32	0,32	0,32	0,33	0,33
0,63	0,33	0,33	0,34	0,34	0,34	0,35	0,35	0,35	0,35	0,36
0,64	0,36	0,36	0,36	0,37	0,37	0,37	0,37	0,38	0,38	0,38
0,65	0,39	0,39	0,39	0,39	0,40	0,40	0,40	0,40	0,41	0,41
0,66	0,41	0,42	0,42	0,42	0,42	0,43	0,43	0,43	0,43	0,44
0,67	0,44	0,44	0,45	0,45	0,45	0,45	0,46	0,46	0,46	0,46
0,68	0,47	0,47	0,47	0,48	0,48	0,48	0,48	0,49	0,49	0,49
0,69	0,50	0,50	0,50	0,50	0,51	0,51	0,51	0,52	0,52	0,52
0,70	0,52	0,53	0,53	0,53	0,54	0,54	0,54	0,54	0,55	0,55
0,71	0,55	0,56	0,56	0,56	0,57	0,57	0,57	0,57	0,58	0,58

Продолжение табл. IX

$\frac{R}{n+1}$	0	1	2	3	4	5	6	7	8	9
0,72	0,58	0,59	0,59	0,59	0,59	0,60	0,60	0,60	0,61	0,61
0,73	0,61	0,62	0,62	0,62	0,63	0,63	0,63	0,63	0,64	0,64
0,74	0,64	0,65	0,65	0,65	0,66	0,66	0,66	0,67	0,67	0,67
0,75	0,67	0,68	0,68	0,68	0,69	0,69	0,69	0,70	0,70	0,70
0,76	0,71	0,71	0,71	0,72	0,72	0,72	0,73	0,73	0,73	0,74
0,77	0,74	0,74	0,75	0,75	0,75	0,76	0,76	0,76	0,77	0,77
0,78	0,77	0,78	0,78	0,78	0,79	0,79	0,79	0,80	0,80	0,80
0,79	0,81	0,81	0,81	0,82	0,82	0,82	0,83	0,83	0,83	0,84
0,80	0,84	0,85	0,85	0,85	0,86	0,86	0,86	0,87	0,87	0,87
0,81	0,88	0,88	0,89	0,89	0,89	0,90	0,90	0,90	0,91	0,91
0,82	0,92	0,92	0,92	0,93	0,93	0,93	0,94	0,94	0,95	0,95
0,83	0,95	0,96	0,96	0,97	0,97	0,97	0,98	0,98	0,99	0,99
0,84	0,99	1,00	1,00	1,01	1,01	1,02	1,02	1,02	1,03	1,03
0,85	1,04	1,04	1,05	1,05	1,05	1,06	1,06	1,07	1,07	1,08
0,86	1,08	1,09	1,09	1,09	1,10	1,10	1,11	1,11	1,12	1,12
0,87	1,13	1,13	1,14	1,14	1,15	1,15	1,16	1,16	1,17	1,17
0,88	1,18	1,18	1,19	1,19	1,20	1,20	1,21	1,21	1,22	1,22
0,89	1,23	1,23	1,24	1,24	1,25	1,25	1,26	1,26	1,27	1,28
0,90	1,28	1,29	1,29	1,30	1,30	1,31	1,32	1,32	1,33	1,33
0,91	1,34	1,35	1,35	1,36	1,37	1,38	1,37	1,39	1,39	1,40
0,92	1,41	1,41	1,42	1,43	1,43	1,44	1,45	1,45	1,46	1,47
0,93	1,48	1,48	1,49	1,50	1,51	1,51	1,52	1,53	1,54	1,55
0,94	1,55	1,56	1,57	1,58	1,59	1,60	1,61	1,62	1,63	1,64
0,95	1,64	1,65	1,66	1,67	1,68	1,70	1,71	1,72	1,73	1,74
0,96	1,75	1,76	1,77	1,79	1,80	1,81	1,83	1,84	1,85	1,87
0,97	1,88	1,90	1,91	1,93	1,94	1,96	1,98	2,00	2,01	2,03
0,98	2,05	2,07	2,10	2,12	2,14	2,17	2,20	2,23	2,26	2,29
0,99	2,33	2,37	2,41	2,46	2,51	2,58	2,65	2,75	2,88	3,09

Таблица X. Критические значения X-критерия Ван-дер-Вардена

n	$n_1-n_2=0$ или 1		$n_1-n_2=2$ или 3		$n_1-n_2=4$ или 5	
	Уровни значимости α, %		Уровни значимости α, %		Уровни значимости α, %	
	5	1	5	1	5	1
8	2,40	—	2,30	—	—	—
9	2,48	—	2,40	—	—	—
10	2,60	3,20	2,49	3,10	2,30	—
11	2,72	3,40	2,58	3,40	2,40	—
12	2,86	3,60	2,79	3,58	2,68	3,40
13	2,96	3,71	2,91	3,64	2,78	3,50
14	3,11	3,94	3,06	3,88	3,00	3,76
15	3,24	4,07	3,19	4,05	3,06	3,88

Продолжение табл. 7

n	$n_1-n_2=0$ или 1		$n_1-n_2=2$ или 3		$n_1-n_2=4$ или 5	
	Уровни значимости α, %		Уровни значимости α, %		Уровни значимости α, %	
	5	1	5	1	5	1
16	3,39	4,26	3,36	4,25	3,28	4,12
17	3,49	4,44	3,44	4,37	3,36	4,23
18	3,63	4,60	3,60	4,58	3,53	4,50
19	3,73	4,77	3,69	4,71	3,61	4,62
20	3,86	4,94	3,84	4,92	3,78	4,85
21	3,96	5,10	3,92	5,05	3,85	4,96
22	4,08	5,26	4,06	5,24	4,01	5,17
23	4,18	5,40	4,15	5,36	4,08	5,27
24	4,29	5,55	4,27	5,53	4,23	5,48
25	4,39	5,68	4,36	5,65	4,30	5,58
26	4,50	5,83	4,48	5,81	4,44	5,76
27	4,59	5,95	4,56	5,92	4,51	5,85
28	4,68	6,09	4,68	6,07	4,64	6,03
29	4,78	6,22	4,76	6,19	4,72	6,13
30	4,88	6,35	4,87	6,34	4,84	6,30
31	4,97	6,47	4,95	6,44	4,91	6,39
32	5,07	6,60	5,06	6,58	5,03	6,55
33	5,15	6,71	5,13	6,69	5,10	6,64
34	5,25	6,84	5,24	6,82	5,21	6,79
35	5,33	6,95	5,31	6,92	5,28	6,88
36	5,42	7,06	5,41	7,05	5,38	7,02
37	5,50	7,17	5,48	7,15	5,45	7,11
38	5,59	7,28	5,58	7,27	5,55	7,25
39	5,67	7,39	5,65	7,37	5,62	7,33
40	5,75	7,50	5,74	7,49	5,72	7,47
41	5,83	7,62	5,81	7,60	5,79	7,56
42	5,91	7,72	5,90	7,71	5,88	7,69
43	5,99	7,82	5,97	7,81	5,95	7,77
44	6,04	7,93	6,06	7,92	6,04	7,90
45	6,14	8,02	6,12	8,01	6,10	7,98
46	6,21	8,13	6,21	8,12	6,19	8,10
47	6,29	8,22	6,27	8,21	6,25	8,18
48	6,36	8,32	6,35	8,31	6,34	8,29
49	6,43	8,41	6,42	8,40	6,39	8,37
50	6,50	8,51	6,51	8,50	6,48	8,48
P	0,05	0,01	0,05	0,01	0,05	0,01

Таблица XI. Критические значения U-критерия Уилкоксона (Манна — Уитни) (односторонний критерий, $P=0{,}01$)

n_2 \ n_1	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	n_2
3	0	0	0	0	1	1	1	2	2	2	3	3	4	4	4	5	3
4	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10	4
5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	5
6		3	4	6	**7**	8	9	11	12	14	15	16	18	19	20	22	6
7			6	7	9	11	12	14	16	18	19	21	23	24	26	28	7
8				9	11	13	15	17	20	22	24	26	28	30	32	34	8
9					14	16	19	21	23	26	28	31	33	36	38	40	9
10						19	22	24	27	30	33	36	38	41	44	47	10
11							25	28	31	34	37	41	44	47	50	53	11
12								31	35	38	42	46	49	53	56	60	12
13									39	43	**47**	51	55	59	63	67	13
14										47	51	56	60	65	69	73	14
15											56	61	66	70	75	80	15
16												66	71	76	82	87	16
17													**77**	82	88	94	17
18														88	94	100	18
19															101	107	19
20																114	20

Продолжение табл. XI (двусторонний критерий, $P=0{,}01$)

n_1 \ n_2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	n_2
5	0	0	0																5
6	0	0	1	2															6
7	0	0	1	3	4														7
8	0	1	2	4	6	7													8
9	0	1	3	5	**7**	9	11												9
10	0	2	4	6	9	11	13	16											10
11	0	2	5	7	10	13	16	19	21										11
12	1	3	6	9	12	15	18	21	24	28									12
13	1	4	7	10	13	17	20	24	27	31	34								13
14	1	4	7	11	15	18	22	26	30	34	38	42							14
15	2	5	8	12	16	20	25	29	33	37	42	46	51						15
16	2	5	9	13	18	22	27	31	36	41	46	50	55	60					16
17	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70				17
18	2	6	11	16	21	26	31	37	42	47	53	59	64	70	75	77	81		18
19	3	7	12	17	22	28	34	39	45	51	57	63	69	75	81	87	93		19
20	3	8	13	18	**24**	30	36	42	48	54	60	67	73	79	86	92	99	105	20
21	3	8	14	19	25	32	38	44	51	58	64	71	78	84	91	98	105	112	21
22	4	9	14	21	27	34	40	47	54	61	68	75	82	89	97	104	111	118	22
23	4	9	15	22	29	36	43	50	57	64	72	79	87	94	102	109	117	125	23
24	4	10	16	23	30	37	45	52	60	68	76	83	91	99	107	115	123	131	24
25	5	10	17	24	32	39	47	55	63	71	79	88	96	104	113	121	129	138	25

Таблица XII. Критические значения z-критерия знаков при разных уровня. значимости α и объеме выборки n

n	α, %		n	α, %		n	α, %		n	α, %	
	5	1		5	1		5	1		5	1
6	6	—	30	21	23	54	35	37	78	49	51
7	7	—	31	22	24	55	36	38	79	49	52
8	8	8	32	23	24	56	36	39	80	50	52
9	8	9	33	23	25	57	37	39	81	50	53
10	9	10	34	24	25	58	37	40	82	51	54
11	10	11	35	24	26	59	38	40	83	51	54
12	10	11	36	25	27	60	39	41	84	52	55
13	11	12	37	25	27	61	39	41	85	53	55
14	12	13	38	26	28	62	40	42	86	53	56
15	12	13	39	27	28	63	40	43	87	54	56
16	13	14	40	27	29	64	41	43	88	54	57
17	13	15	41	28	30	65	41	44	89	55	58
18	14	15	42	28	30	66	42	44	90	55	58
19	15	16	43	29	31	67	42	45	91	56	59
20	15	17	44	29	31	68	43	46	92	56	59
21	16	17	45	30	32	69	44	46	93	57	60
22	17	18	46	31	33	70	44	47	94	57	60
23	17	19	47	31	33	71	45	47	95	58	61
24	18	19	48	32	34	72	45	48	96	59	62
25	18	20	49	32	34	73	46	48	97	59	62
26	19	20	50	33	35	74	46	49	98	60	63
27	20	21	51	33	36	75	47	50	99	60	63
28	20	22	52	34	36	76	48	50	100	61	64
29	21	22	53	35	37	77	48	51	—	—	—
P	0,05	0,01	—	0,05	0,01	—	0,05	0,01	—	0,05	0,01

Таблица XIII. Критические значения парного Т-критерия Уилкоксона (односторонний критерий)

Число парных наблюдений n	Уровни значимости α, %		Число парных наблюдений n	Уровни значимости α, %	
	5	1		5	1
5	0	—	14	25	16
6	2	0	15	30	19
7	3	0	16	35	23
8	5	1	17	41	28
9	8	3	18	47	33
10	10	5	19	53	38
11	13	7	20	60	42
12	17	10	21	67	50
13	21	12	22	74	56
P	0,05	0,01	—	0,05	0,01

Продолжение таблицы XIII. (двусторонний критерий)

Число парных наблюдений n	Уровни значимости α, %		Число парных наблюдений n	Уровни значимости α, %	
	5	1		5	1
6	1	—	16	31	21
7	3	—	17	36	24
8	5	1	18	41	29
9	7	3	19	47	33
10	9	4	20	53	39
11	12	6	21	60	44
12	15	8	22	67	50
13	18	11	23	74	56
14	22	14	24	82	62
15	26	17	25	90	69
P	0,05	0,01	—	0,05	0,01

Примечание. Для $n>25$ критические значения T-критерия можно определить по формуле

$$T_{st}=\frac{n(n+1)}{4}-t\sqrt{\frac{n(n+1)(2n+1)}{24}},$$

где n — число парных наблюдений; t зависит от принятого уровня значимости, т. е. $t_{0,05}=1{,}96$ и $t_{0,01}=2{,}58$.

Таблица XIV. Критические значения коэффициента асимметрии As

Объем выборки n	Уровни значимости α, %		Объем выборки n	Уровни значимости α, %	
	5	1		5	1
25	0,711	1,061	250	0,251	0,360
30	0,661	0,982	300	0,230	0,329
35	0,621	0,921	350	0,213	0,305
40	0,587	0,869	400	0,200	0,285
45	0,558	0,825	450	0,188	0,269
50	0,533	0,787	500	0,179	0,255
60	0,492	0,723	550	0,171	0,243
70	0,459	0,673	600	0,163	0,233
80	0,432	0,631	650	0,157	0,224
90	0,409	0,596	700	0,151	0,215
100	0,389	0,567	750	0,146	0,208
125	0,350	0,508	800	0,142	0,202
150	0,321	0,464	850	0,138	0,196
175	0,298	0,430	900	0,134	0,190
200	0,280	0,403	950	0,130	0,185
			1000	0,127	0,180
P	0,05	0,01	—	0,05	0,01

Таблица XV. Критические значения коэффициента эксцесса Ex

Объем выборки n	Уровни значимости α, %		
	10	5	1
11	0,890	0,907	0,936
16	0,873	0,888	0,914
21	0,863	0,877	0,900
26	0,857	0,869	0,890
31	0,851	0,863	0,883
36	0,847	0,858	0,877
41	0,844	0,854	0,872
46	0,841	0,851	0,868
51	0,839	0,848	0,865
61	0,835	0,843	0,859
71	0,832	0,840	0,855
81	0,830	0,838	0,852
91	0,828	0,835	0,848
101	0,826	0,834	0,846
201	0,818	0,823	0,832
301	0,814	0,818	0,826
401	0,812	0,816	0,822
501	0,810	0,814	0,820
P	0,10	0,05	0,01

Таблица XVI. Критические значения величины нормированного отклонения при оценке сомнительных вариант с учетом объема выборки n и уровней значимости α

n	α, %		n	α, %		n	α, %	
	5	1		5	1		5	1
4	1,71	1,73	13	2,56	2,81	23	2,84	3,16
5	1,92	1,97	14	2,60	2,86	24	2,86	3,18
6	2,07	2,16	15	2,64	2,90	25	2,88	3,20
7	2,18	2,31	16	2,67	2,95	26	2,90	3,22
8	2,27	2,43	17	2,70	2,98	27	2,91	3,24
9	2,35	2,53	18	2,73	3,02	28	2,93	3,26
10	2,41	2,62	19	2,75	3,05	29	2,94	3,28
11	2,47	2,69	20	2,78	3,08	30	2,96	3,29
12	2,52	2,75	21	2,80	3,11			
P	0,05	0,01	—	0,05	0,01	—	0,05	0,01

Таблица XVII. Критические значения критерия $t_1 = \dfrac{x_2 - x_1}{x_{n-1} - x_1}$

n	Уровни значимости α, %		n	Уровни значимости α, %		n	Уровни значимости α, %	
	5	1		5	1		5	1
4	0,76	0,89	13	0,36	0,46	22	0,29	0,38
5	0,64	0,78	14	0,35	0,45	23	0,28	0,37
6	0,56	0,70	15	0,34	0,44	24	0,28	0,37
7	0,31	0,64	16	0,33	0,43	25	0,28	0,36
8	0,47	0,59	17	0,32	0,42	26	0,27	0,36
9	0,44	0,54	18	0,31	0,41	27	0,27	0,35
10	0,41	0,53	19	0,31	0,40	28	0,27	0,35
11	0,39	0,50	20	0,30	0,39	29	0,26	0,34
12	0,38	0,48	21	0,30	0,38	30	0,26	0,34
P	0,05	0,01	—	0,05	0,01	—	0,05	0,01

Таблица XVIII. Критические значения критерия $t_2 = \dfrac{x_n - x_{n-1}}{x_n - x_2}$

n	Уровни значимости α, %		n	Уровни значимости α, %		n	Уровни значимости α, %	
	5	1		5	1		5	1
4	0,96	0,99	13	0,41	0,52	22	0,32	0,41
5	0,81	0,92	14	0,40	0,50	23	0,31	0,41
6	0,69	0,80	15	0,38	0,49	24	0,31	0,40
7	0,61	0,74	16	0,37	0,47	25	0,30	0,39
8	0,55	0,68	17	0,36	0,46	26	0,30	0,39
9	0,51	0,64	18	0,35	0,45	27	0,30	0,38
10	0,48	0,60	19	0,34	0,44	28	0,29	0,38
11	0,45	0,57	20	0,33	0,43	29	0,29	0,37
12	0,43	0,54	21	0,33	0,42	30	0,28	0,37
P	0,05	0,01	—	0,05	0,01	—	0,05	0,01

Таблица XIX. Критические значения критерия χ^2_R *Фридмана*

n_s	$n_s=3$		$n_s=4$		n_s	$n_s=3$	
	α, %		α, %			α, %	
	5	1	5	1		5	1
2			6,0		9	6,22	8,67
3	6,0		7,4	9,0	10	6,20	9,60
4	6,5	8,0	7,8	9,6	11	6,54	9,46
5	6,4	8,4	7,8	9,96	12	6,17	9,50
6	7,0	9,0	7,6	10,20	13	6,0	9,38
7	7,14	8,86	7,8	10,37	14	6,14	9,00
8	6,25	9,00	7,65	10,35	15	6,40	8,93
P	0,05	0,01	0,05	0,01		0,05	0,01

Примечание. n_s — число выборок.

Таблица XX. Критические значения критерия H Краскелла — Уоллиса

n_1	n_2	n_3	Уровни значимости α, %		n_1	n_2	n_3	Уровни значимости α, %	
2	2	3	4,71	—	3	4	4	5,62	7,14
2	2	4	5,33	—	3	4	5	5,63	7,45
2	2	5	5,16	6,53	3	4	6	5,61	7,41
2	2	6	5,35	6,65	3	5	5	5,71	7,54
2	3	3	5,36	—	3	5	6	5,60	7,59
2	3	4	5,44	6,44	3	6	6	5,63	7,45
2	3	5	5,25	6,82	4	4	4	5,69	7,65
2	3	6	5,35	6,97	4	4	5	5,62	7,76
2	4	4	5,45	7,04	4	4	6	5,68	7,94
2	4	5	5,27	7,12	4	5	5	5,64	7,77
2	4	6	5,34	7,34	4	5	6	5,66	7,94
2	5	5	5,34	7,27	4	6	6	5,72	8,00
2	5	6	5,34	7,38	5	5	5	5,78	8,00
3	3	3	5,60	7,20	5	5	6	5,73	8,03
3	3	4	5,73	6,75	5	6	6	5,76	8,12
3	3	5	5,65	7,08	6	6	6	5,80	8,22
3	3	6	5,62	7,41					

Таблица XXI. Критические значения коэффициента корреляции r_{xy}

Степени свободы $k=n-2$	Уровни значимости α, %		Степени свободы $k=n-2$	Уровни значимости α, %	
	5	1		5	1
5	0,75	0,87	27	0,37	0,47
6	0,71	0,83	28	0,36	0,46
7	0,67	0,80	29	0,36	0,46
8	0,63	0,77	30	0,35	0,45
9	0,60	0,74	35	0,33	0,42
10	0,58	0,71	40	0,30	0,39
11	0,55	0,68	45	0,29	0,37
12	0,53	0,66	50	0,27	0,35
13	0,51	0,64	60	0,25	0,33
14	0,50	0,62	70	0,23	0,30
15	0,48	0,61	80	0,22	0,28
16	0,47	0,59	90	0,21	0,27
17	0,46	0,58	100	0,20	0,25
18	0,44	0,56	125	0,17	0,23
19	0,43	0,55	150	0,16	0,21
20	0,42	0,54	200	0,14	0,18
21	0,41	0,53	300	0,11	0,15
22	0,40	0,52	400	0,10	0,13
23	0,40	0,51	500	0,09	0,12
24	0,39	0,50	700	0,07	0,10
25	0,38	0,49	900	0,06	0,09
26	0,37	0,48	1000	0,06	0,09
P	0,05	0,01	—	0,05	0,01

Таблица XXII. Значения z, соответствующие значениям выборочного коэффициента корреляции r_{xy}

r_{xy}	Сотые доли коэффициента корреляции									
	0	1	2	3	4	5	6	7	8	9
0,0	0,000	0,010	0,020	0,030	0,040	0,050	0,060	0,070	0,080	0,090
0,1	0,100	0,110	0,121	0,131	0,141	0,151	0,161	0,172	0,182	0,192
0,2	0,203	0,213	0,224	0,234	0,245	0,255	0,266	0,277	0,288	0,299
0,3	0,310	0,321	0,332	0,343	0,354	0,365	0,377	0,388	0,400	0,412
0,4	0,424	0,436	0,448	0,460	0,472	0,485	0,497	0,510	0,523	0,536
0,5	0,549	0,563	0,576	0,590	0,604	0,618	0,633	0,648	0,663	0,678
0,6	0,693	0,709	0,725	0,741	0,758	0,775	0,793	0,811	0,829	0,848
0,7	0,867	0,887	0,908	0,929	0,951	0,973	0,996	1,020	1,045	1,071
0,8	1,099	1,127	1,157	1,188	1,221	1,256	1,293	1,333	1,376	1,422
0,9	1,472	1,528	1,589	1,658	1,738	1,832	1,946	2,092	2,298	2,647
0,99	2,647	2,700	2,759	2,826	2,903	2,995	3,106	3,250	3,453	3,800

Таблица XXIII. Критические значения коэффициента корреляции рангов при различных уровнях значимости α и объемах выборки n

n	α, %		n	α, %		n	α, %	
	5	1		5	1		5	1
5	0,94	—	17	0,48	0,62	29	0,37	0,48
6	0,85	—	18	0,47	0,60	30	0,36	0,47
7	0,78	0,94	19	0,46	0,58	31	0,36	0,46
8	0,72	0,88	20	0,45	0,57	32	0,36	0,45
9	0,68	0,83	21	0,44	0,56	33	0,34	0,45
10	0,64	0,79	22	0,43	0,54	34	0,34	0,44
11	0,61	0,76	23	0,42	0,53	35	0,33	0,43
12	0,58	0,73	24	0,41	0,52	36	0,33	0,43
13	0,56	0,70	25	0,40	0,51	37	0,33	0,42
14	0,54	0,68	26	0,39	0,50	38	0,32	0,41
15	0,52	0,66	27	0,38	0,49	39	0,32	0,41
16	0,50	0,64	28	0,38	0,48	40	0,31	0,40
P	0,05	0,01	—	0,05	0,01	—	0,05	0,01

Таблица XXVI. Значения величины Q, соответствующие 5%-ному уровню значимости α и числу групп (градаций), входящих в дисперсионный комплекс

Числа степени свободы k	Число градаций a										
	2	3	4	5	6	7	8	9	10	11	12
6	3,5	4,3	4,9	5,3	5,6	5,9	6,1	6,3	6,5	6,7	6,8
7	3,3	4,2	4,7	5,1	5,4	5,5	5,8	6,0	6,2	6,3	6,4
8	3,3	4,0	4,5	4,9	5,2	5,4	5,6	5,8	5,9	6,1	6,2
9	3,2	4,0	4,4	4,8	5,0	5,2	5,4	5,6	5,7	5,9	6,0
10	3,1	3,9	4,3	4,7	4,9	5,1	5,3	5,5	5,6	5,7	5,8
11	3,1	3,8	4,2	4,6	4,8	5,0	5,2	5,4	5,5	5,6	5,7
12	3,1	3,8	4,2	4,5	4,8	5,0	5,1	5,3	5,4	5,5	5,6
13	3,1	3,7	4,2	4,5	4,7	4,9	5,1	5,2	5,3	5,4	5,5
14	3,0	3,7	4,1	4,4	4,6	4,8	5,0	5,1	5,3	5,4	5,5
15	3,0	3,7	4,1	4,4	4,6	4,8	4,9	5,1	5,2	5,3	5,4
16	3,0	3,7	4,1	4,3	4,6	4,7	4,9	5,0	5,2	5,3	5,4
17	3,0	3,6	4,0	4,3	4,5	4,7	4,9	5,0	5,1	5,2	5,3
18	3,0	3,6	4,0	4,3	4,5	4,7	4,8	5,0	5,1	5,2	5,3
19	3,0	3,6	4,0	4,3	4,5	4,6	4,8	4,9	5,0	5,1	5,2
20	3,0	3,6	4,0	4,2	4,5	4,6	4,8	4,9	5,0	5,1	5,2
24	2,9	3,5	3,9	4,2	4,4	4,5	4,7	4,8	4,9	5,0	5,1
30	2,9	3,5	3,8	4,1	4,3	4,5	4,6	4,7	4,8	4,9	5,0
40	2,9	3,4	3,8	4,0	4,2	4,4	4,5	4,6	4,7	4,8	4,9
60	2,8	3,4	3,7	4,0	4,2	4,3	4,4	4,6	4,7	4,7	4,8
120	2,8	3,4	3,7	3,9	4,1	4,2	4,4	4,5	4,6	4,6	4,7

Алпатов В. В. Роковая ошибка в определении породы пчел // Природа. 1976. № 5.

Балантер Б. И., Ханин М. А., Чернавский Д. С. Введение в математическое моделирование патологических процессов. М., 1980.

Бейли Н. Математика в биологии и медицине. М., 1970.

Брукс С., Карузерс Н. Применение статистических методов в метеорологии. Л., 1963.

Ван-дер-Варден Б. П. Математическая статистика. М., 1960.

Вайну Я. Ф. Корреляция рядов динамики. М., 1977.

Глотов Н. В., Животовский Л. А., Хованов Н. В., Хромов-Борисов Н. Н. Биометрия. Л., 1982.

Гнеденко Б. В., Хинчин А. Я. Элементарное введение в теорию вероятностей. М., 1970.

Громыко Г. Л. Статистика. М., 1976.

Гублер Е. В., Генкин А. А. Применение непараметрических критериев статистики в медико-биологических исследованиях. Л., 1973.

Доспехов Б. А. Планирование полевого опыта и статистическая обработка его данных. М., 1972.

Зайцев Г. Н. Методика биометрических расчетов. Математическая статистика в экспериментальной ботанике. М., 1973.

Закс Л. Статистическое оценивание. М., 1976.

Кендэл М. Ранговые корреляции. М., 1975.

Колмогоров А. Н. Основные понятия теории вероятностей. М., 1974.

Кузьмичев Д. А., Радкевич И. А., Смирнов А. Д. Автоматизация экспериментальных исследований. М., 1983.

Куршакова Ю. С. Корреляционный и регрессионный анализ в практическом применении. Теория отбора в популяциях растений. Новосибирск, 1976.

Лысенков А. Н. Математические методы планирования многофакторных медико-биологических экспериментов. М., 1974.

Максимов Г. К., Синицын А. Н. Статистическое моделирование многомерных систем в медицине. Л., 1983.

Масальгин Н. А. Математико-статистические методы в спорте. М., 1974.

Математическая теория планирования эксперимента / С. М. Ермаков, В. З. Бродский, А. А. Жиглявский и др. М., 1983.

Меркурьева Е. К. Биометрия в селекции и генетике сельскохозяйственных животных. М., 1970.

Меркурьева Е. К., Шангин-Березовский Г. Н. Генетика с основами биометрии. М., 1983.

Митропольский А. К. Техника статистических вычислений. М., 1971.

Плохинский Н. А. Биометрия. М., 1970.

Плохинский Н. А. Алгоритмы биометрии. М., 1980.

Рокицкий П. Ф. Биологическая статистика. Минск, 1967.

Рузавин Г. И. Математизация научного знания. М., 1984.

Сепетлиев Дм. Статистические методы в научных медицинских исследованиях. М., 1968.

Славин М. Б. Системное моделирование патологических процессов. М., 1983.

Снедекор Д. У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. М., 1961.

Соколов Д. К. Математическое моделирование в медицине. М., 1974.

Терентьев П. В. Истоки биометрии. Из истории биологии. М., 1971.
Терентьев П. В., Ростова Н. С. Практикум по биометрии. Л., 1977.
Урбах В. Ю. Статистический анализ в биологических и медицинских исследованиях. М., 1975.
Фишер Р. Э. Статистические методы для исследователей. М., 1958.
Яноши Л. Теория и практика обработки результатов измерений. М., 1968.
Юл Д., Кендэл М. Теория статистики. М., 1960.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА К ПОСЛЕСЛОВИЮ РЕДАКТОРА

1. *Бартлетт М. С.* Многомерная статистика // Теоретическая и математическая биология. М., 1968.
2. *Болч Б., Хуань К. Дж.* Многомерные статистические методы для экономики. М., 1979.
3. *Девис Дж.* Статистика и анализ геологических данных. М., 1977.
4. *Дерябин В. Е.* Многомерная биометрия для антропологов. М., 1983.
5. *Дубров А. М.* Обработка статистических данных методом главных компонент. М., 1978.
6. *Дьяконов В. П.* Справочник по расчетам на микрокалькуляторах. М., 1986.
7. *Дюран Б., Оделл П.* Кластерный анализ. М., 1977.
8. *Епанченков В. А., Цветков А. Н.* Справочник по прикладным программам для микрокалькуляторов. М., 1988.
9. *Жигарев А. Н., Макарова И. В., Путинцева М. А.* Основы компьютерной грамоты. М., 1987.
10. *Жуковская В. М., Мучник И. Б.* Факторный анализ в социально-экономических исследованиях. М., 1976.
11. *Иберла К.* Факторный анализ. М., 1980.
12. *Иберског К. Г., Клован Д. И., Реймент Р. А.* Геологический факторный анализ. Л., 1980.
13. *Окунь Я.* Факторный анализ. М., 1974.
14. *Суходольский Г. В.* Основы математической статистики для психологов. Л., 1972.
15. *Урбах В. Ю.* Статистический анализ в биологических и медицинских исследованиях. М., 1975.
16. *Цветков А. И., Епанченков В. А.* Прикладные программы для микроЭВМ «Электроника БЗ-34», «Электроника МК-54», «Электроника МК-56». М., 1984.
17. *Харман Г.* Современный факторный анализ. М., 1972.
18. *Электронно*-вычислительные машины. В 8 кн. / Под ред. А. Я. Савельева. М., 1987.
19. *Constande-Westermann T. C.* Coefficients of biological distans. Netherlands, 1972.
20. *Seal H. L.* Multivariate statistical methods for biologists. London, 1964,