В. К. Романко

Статистический анализ данных в психологии

Лаборатория ЗНАНИЙ

ПСИХОЛОГИЯ

В. К. Романко

Статистический анализ данных в психологии

Учебное пособие

Рекомендовано
Советом по психологии УМО
по классическому университетскому
образованию в качестве учебного пособия
для студентов высших учебных заведений,
обучающихся по направлению
и по специальностям психологии

4-е издание, электронное

Москва
Лаборатория знаний
2020

УДК 519.22(075.8)+159.9(075.8)
ББК 22.17я73+88я73
Р69

Романко В. К.

Р69 Статистический анализ данных в психологии : учебное пособие / В. К. Романко. — 4-е изд., электрон. — М. : Лаборатория знаний, 2020. — 315 с. — Систем. требования: Adobe Reader XI ; экран 10".— Загл. с титул. экрана. — Текст : электронный.

ISBN 978-5-00101-802-5

В учебном пособии описываются основные математические методы, предлагаемые математической теорией и широко применяемые на практике в современных психолого-педагогических исследованиях.

Излагаются основные понятия теории вероятностей и описываются конкретные математические методы обработки данных. В приложении даются общие рекомендации по использованию статистических пакетов программ.

Изложение ведется практически без строгих математических доказательств, но с подробными обсуждениями, объяснениями и иллюстрациями. Для конкретных методов статистического анализа разъясняются их сущность и границы применимости. Приведено большое количество задач для самостоятельной работы.

Для студентов и преподавателей вузов.

УДК 519.22(075.8)+159.9(075.8)
ББК 22.17я73+88я73

Деривативное издание на основе печатного аналога: Статистический анализ данных в психологии : учебное пособие / В. К. Романко. — М. : БИНОМ. Лаборатория знаний, 2009. — 312 с. : ил. — ISBN 978-5-94774-849-9.

ISBN 978-5-00101-802-5

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Эта книга является учебным пособием для студентов-психологов по курсу «Математические методы в психологии». В ней описываются основные математические методы, которые предлагаются математической теорией и которые широко применяются на практике в современных психолого-педагогических исследованиях.

В предлагаемом учебном пособии автор поставил цель доступно познакомить читателей, не имеющих специального математического образования, с основами современного статистического анализа данных. Эта цель определяет характер изложения материала. Изложение ведется практически без строгих математических доказательств, но с подробными обсуждениями, объяснениями и иллюстрациями. Ограничиваясь простыми математическими средствами, мы стараемся по возможности не только описывать конкретные методы статистического анализа, но и разъяснять их сущность и границы применимости.

Имеющиеся в настоящее время руководства по статистическому анализу данных либо требуют достаточно хорошей математической подготовки, либо предназначены для технических приложений, либо не отражают современный уровень статистических методов обработки данных, либо являются для широкого круга читателей малопонятными из-за отсутствия точного описания излагаемых статистических моделей. Эти обстоятельства послужили одной из причин, побудивших автора написать эту книгу.

Книга написана на основе курса лекций «Математические методы в психологии», которые в течение ряда лет автор читал студентам Московского городского психолого-педагогического университета (МГППУ). Необходимость дать студентам сравнительно простое учебное пособие по прикладной статистике было второй причиной появления этой книги.

Поскольку книга адресована психологам и педагогам, то этим определяются не только отбор материала и характер его изложения, но и подбор иллюстрированных примеров. Эти примеры берутся из психологии и педагогики. Однако они носят учебный (гипотетический) характер и не являются результатами каких-либо реальных психолого-педагогических исследований. Приведено большое количество примеров для самостоятельной работы.

Для приведенных в учебном пособии примеров характерен небольшой объем исходных данных. Это сделано не только дня облегчения расчетов вручную, но и потому, что на практике большинство психолого-педагогических исследований проводится именно для малых объемов исходных данных. Решение простых примеров в учебном процессе позволяет наглядно ощутить, как работают на практике статистические методы и подготовить студентов к использованию компьютерных статистических программ. Решение задач вручную возможно во многих случаях лишь при наличии соответствующего набора таблиц математической статистики. Учебное пособие (см. приложение 1) снабжено достаточно полным набором таких таблиц, который включает малодоступные статистические таблицы для основных непараметрических критериев.

В главах 1—2 данного пособия излагаются основные понятия теории вероятностей. Знание психологами элементов теории вероятностей важно потому, что многие психологические признаки и связи между ними описываются вероятностными характеристиками. Далее, в главах 3—13 излагаются конкретные математические методы обработки экспериментальных данных в современной психологии.

Если психологические исследования базируются на некоторых конкретных измерениях того или другого психологического феномена, свойства, характеристики, черты и т. д., то для получения научных и практических выводов из таких измерений необходимо использовать математические методы современного статистического анализа данных. Эти методы позволяют психологу-практику обосновывать правильность используемых методик и приемов, обобщать данные эксперимента, находить зависимости между исследуемыми психологическими признаками, выявлять существенные различия между различными группами испытуемых с точки зрения исследуемого признака, строить прогнозы поведения испытуемых, избегать логических и содержательных ошибок при интерпретации полученных данных и т. д.

При этом необходимо помнить, что математические методы статистического анализа данных — лишь инструменты психологических исследований. Наиболее важным в каждом эксперименте является четкая постановка задачи, детальное планирование опыта, выбор непротиворечивых гипотез и метода исследования, содержательная интерпретация результатов обработки экспериментальных данных, полученных с помощью использования того или иного метода статистического анализа. С этой точки зрения психолог играет ведущую роль. Выработку профессионализма и хорошей интуиции

психолога стимулирует его знакомство с основами современного статистического анализа данных. Это позволяет психологу либо самостоятельно, либо совместно с привлекаемыми математиками грамотно ставить задачу психологического эксперимента, выбирать конкретный метод статистического анализа, использовать статистические пакеты программ для персональных компьютеров при реализации выбранного метода анализа, давать содержательную интерпретацию полученных математических результатов анализа данных. Противоречия в психологических и математических выводах свидетельствуют о некорректности решения задачи или некорректности интерпретации результатов.

Множество математических методов современного статистического анализа данных можно разбить на две большие группы. К первой группе относятся методы, предназначенные для анализа данных, имеющих вероятностную (случайную) природу. Они предусматривают вероятностную интерпретацию обрабатываемых данных и полученных в результате обработки статистических выводов. Для понимания таких методов необходимо знание элементов теории вероятностей. Теория вероятностей изучает математические модели, имитирующие механизмы функционирования гипотетического (неконкретного) реального явления, результат которого зависит от влияния большого числа взаимодействующих случайных (не поддающихся строгому учету и контролю) факторов. Такого рода математические модели называют вероятностными моделями. Они используются в тех ситуациях, когда имеется хотя бы принципиально мыслимая возможность приближенного многократного воспроизведения всего комплекса условий, при которых производились измерения анализируемых данных.

Методы первой группы статистического анализа данных — это методы математической статистики. Они позволяют по результатам наблюдений конкретного явления или системы (исходным статистическим данным) строить статистическую модель явления или системы, т. е. такую вероятностную модель, которая в определенном смысле наилучшим образом соответствует исходным статистическим данным.

В большинстве руководств по математической статистике излагаются лишь так называемые параметрические методы, которые предполагают, что исследуемые психологические (случайные) признаки имеют нормальный закон распределения. В данном учебном пособии кроме параметрических методов излагаются и наиболее распространенные на практике непараметрические (ранговые) методы,

т. е. методы, свободные от предположений о законе распределения признака и поэтому более универсальные.

Ко второй группе методов современного статистического анализа данных относятся методы, предназначенные для обработки данных, не имеющих вероятностной природы. Другими словами, эти методы используются в ситуации, когда изучаемое явление детерминировано (т. е. не зависит от мешающего влияния случайных факторов), или в ситуации, когда принципиально невозможно многократно повторить опыт хотя бы в приблизительно одинаковых условиях. В подобных ситуациях невозможна вероятностная интерпретация исходных статистических данных и получаемых в результате их обработки выводов. Методы анализа и моделирования данных из второй группы принято называть логико-алгебраическими, поскольку эти методы основаны на обычной логике и используют алгебраический и геометрический подходы. Ко второй группе методов современного статистического анализа данных относятся, например, методы кластерного анализа и многомерного шкалирования.

Если методы первой группы (назовем их вероятностно-статистическими) позволяют решать традиционные для статистики задачи (например, оценка неизвестных параметров, проверка гипотез, моделирование связи изучаемых признаков), то методы второй группы (логико-алгебраические) позволяют решать новые, специфические задачи, такие, как классификация объектов и признаков, сжатие информации, визуализация (наглядное представление) данных, отбор наиболее информативных показателей, включая выявление латентных факторов, моделирование сложных систем (латентно-структурный анализ) и т. д. Настоящее учебное пособие дает достаточно полное представление о современных логико-алгебраических методах. Отметим также, что логико-алгебраические методы на практике могут применяться в комплексе с вероятностно-статистическими методами.

Статистические программные пакеты для персональных компьютеров сделали методы статистического анализа данных более доступными, менее трудоемкими и наглядными.

В приложении 2 учебного пособия даются общие рекомендации по использованию статистических пакетов программ, а также подготовленные В. Т. Бордуковой и В. И. Бордуковой методические рекомендации по использованию пакета STATISTICA. Знакомство с данным учебным пособием позволяет успешно применять математические методы для обработки материалов психолого-педагогических экспериментальных измерений, начиная с постановки задач,

выбора метода ее решения, включая использование статистического пакета программ, и кончая интерпретацией результатов анализа и практическими рекомендациями.

В приведенном списке литературы можно найти более глубокое описание приведенных в учебном пособии методов анализа данных, а также некоторые другие, более специальные методы.

Первый вариант этого учебного пособия был издан в 2006 году издательским центром МГППУ при всемерной поддержке ректора МГППУ, академика РАО В. В. Рубцова и первого проректора МГППУ, профессора А. А. Марголиса. Им автор выражает свою глубокую благодарность.

В настоящем издании учебного пособия были устранены замеченные опечатки и внесены небольшие дополнения при изложении некоторых вопросов.

Автор благодарен В. Т. Бордуковой и В. И. Бордуковой за их согласие включить в учебное пособие подготовленные ими методические указания по пакету STATISTICA. Автор также благодарен М. В. Ивановой за большую техническую помощь при подготовке настоящего издания учебного пособия.

ГЛАВА 1

СЛУЧАЙНЫЕ СОБЫТИЯ И ИХ ВЕРОЯТНОСТИ

§ 1. Случайные события и операции над ними

В природе часто наблюдаются явления и события, исходы которых практически однозначно определяются заданными условиями. Такие явления и события принято называть детерминированными или закономерными. Например, при равномерном и прямолинейном движении материальной точки пройденный путь S однозначно определяется формулой $S = v \cdot t$, если задать скорость v и время t. Другой пример дает второй закон Ньютона классической механики: сила, действующая на тело, равна произведению массы тела на сообщаемое этой силой ускорение.

Но бывают такие явления и события, для которых сохранение основных условий опыта не гарантирует однозначность исхода. При подбрасывании монеты нельзя предсказать исход: упадет монета гербом вверх или нет. При измерении одной и той же физической характеристики одним и тем же прибором, в одних и тех же условиях получаем различные результаты. Результаты таких опытов не определяются заранее однозначно в силу влияния большого числа разнообразных причин, не поддающихся строгому учету и контролю. Такого рода явления и события принято называть *недетерминированными* или *случайными*.

Перейдем к более детальному описанию событий с неопределенным исходом.

Будем считать, что проводится реальный или мысленный опыт при не изменяющемся во времени действии большого числа случайных (не поддающихся строгому учету и контролю) факторов, не позволяющих сделать однозначные выводы о том, произойдет или не произойдет интересующее нас событие. При этом предполагается, что имеется принципиальная возможность многократного повторения опыта при одних и тех же условиях. Неконтролируемый (случайный) разброс в результатах наблюдений объясняется многочисленностью, сложностью и неизученностью формирующих их факторов.

Наиболее простые примеры опытов, для которых полностью выполнены описанные условия, дают азартные игры. Действительно, опыты с подбрасыванием монеты, игральной кости или с вытягива-

нием наугад карты из колоды карт можно многократно повторять в одинаковых условиях, но они не позволяют делать полностью определенные заключения о том, произойдет или не произойдет в результате данного опыта интересующее нас событие, например, появление герба, шестерки или туза пик.

Соблюдение вышеуказанных условий проведения опытов в более серьезных и сложных сферах человеческой деятельности — в психологии, медицине, экономике, технике и т. д. — требует в каждом конкретном случае специального рассмотрения.

С каждым опытом связано множество всех возможных взаимно исключающих исходов опыта. Каждый из этих исходов будем называть *элементарным событием* (или *элементарным исходом*), а множество всех таких исходов — *пространством элементарных событий* (или исходов). Элементарные события будем обозначать ω, а пространство элементарных событий — Ω. Таким образом, например, в результате опыта с конечным числом исходов ω обязательно происходит одно из элементарных событий, причем одновременно с ним не может произойти никакое другое элементарное событие.

Если пространство Ω содержит конечное или счетное число элементарных событий, то Ω называется *дискретным* пространством элементарных событий. В противном случае Ω называется *непрерывным*.

Пусть сначала Ω — дискретное пространство. Тогда элементарные события можно занумеровать. Обозначив их через $\omega_1, \omega_2, \ldots$ $\ldots, \omega_n, \ldots$, получаем, что все пространство $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$.

Рассмотрим несколько примеров опытов с дискретными пространствами Ω. В этих примерах под монетой понимается идеально симметричная монета, очень тонкая и однородная по плотности, а под игральной костью понимается геометрически правильный куб с занумерованными гранями, однородный по плотности.

Здесь уместно отметить модельный характер вышеуказанных условий проведения опытов с монетой и игральной костью. Опыты проводятся не с реальными монетами и костями, а с их абстракциями, точнее с их математическими моделями, т. е. такими абстракциями, в которых отношения между реальными элементами заменены подходящими отношениями между математическими категориями. Точность математической модели описания реального явления проверяется практикой.

Как будет ясно из дальнейшего, теория вероятностей изучает вероятностные модели реальных явлений, т. е. математические модели, имитирующие механизмы функционирования таких явлений, на которые существенно влияет большое число случайных факторов.

Примеры.

1) Подбрасывание монеты. Если обозначить через ω_1 выпадения герба *Г*, а через ω_2 выпадение решки *Р*, то $\Omega = \{\omega_1, \omega_2\}$.

2) Выбрасывание игральной кости. Тогда $\Omega = \{\omega_1 = 1, \ \omega_2 = 2, \ \omega_3 = 3, \ \omega_4 = 4, \ \omega_5 = 5, \ \omega_6 = 6\}$.

3) Двукратное бросание монеты. Имеем: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, где ω_1 — выпадение *ГГ*, ω_2 — выпадение *ГР*, ω_3 — выпадение *РГ*, ω_4 — выпадение *РР*.

4) Двукратное бросание игральной кости. Тогда Ω представляет собой множество $\{\omega_k, \ k = 1, 2, 3, \ldots, 36\}$, где каждое ω_k обозначает некоторую пару чисел (m, n), где $m = 1, 2, 3, 4, 5, 6$, $n = 1, 2, 3, 4, 5, 6$ и m означает количество выпавших очков при первом бросании, а n означает количество выпавших очков при втором бросании.

Кроме элементарных (или неразложимых) событий приходится иметь дело и с составными (или разложимыми) событиями. Составное событие в случае дискретного Ω имеем в том случае, если происходит какое-либо из элементарных событий, причем таких элементарных событий можно указать, по меньшей мере, два. В таком случае говорят, что составное событие состоит, по меньшей мере, из двух элементарных событий.

Случайным событием называется любое подмножество дискретного пространства Ω элементарных событий.

Случайные события будем обозначать буквами A, B, C, ... Приведем примеры случайных событий.

Примеры.

1) В опыте с выбрасыванием игральной кости можно говорить о событии A, состоящем в том, что выпало четное число очков, или о событии B, состоящем в том, что выпало нечетное число очков. Тогда $A = \{\omega_2, \omega_4, \omega_6\}$, $B = \{\omega_1, \omega_3, \omega_5\}$.

2) При двукратном бросании монеты можно говорить о событии A, состоящем в том, что выпал хотя бы один раз герб, или о событии B, состоящем в том, что дважды выпал герб или дважды выпала решка. Тогда $A = \{\omega_1, \omega_2, \omega_3\}$, $B = \{\omega_1, \omega_4\}$.

3) При двукратном бросании игральной кости можно говорить о событии A, состоящем в том, что сумма выпавших очков не менее десяти. В этом случае событие A содержит 6 элементарных событий, означающих соответственно выпадение пар чисел (4, 6); (5, 5); (5, 6); (6, 4); (6, 5); (6, 6).

До сих пор мы рассматривали задачи, в которых пространство Ω состояло из не более чем счетного числа элементарных событий.

Однако легко привести примеры задач, где пространство всех элементарных событий несчетно, т. е. элементарные события нельзя занумеровать натуральными числами. Например, случайный опыт с измерением температуры имеет континуум исходов, так как результатом может быть любая точка рассматриваемого отрезка $[t_1, t_2]$.

Именно с более чем счетным, т. е. с континуальным, пространством элементарных событий приходится иметь дело в случайных экспериментах по измерению физических характеристик непрерывной природы (времени, длины, веса, температуры, давления и т. п.).

Если пространство Ω содержит континуальное множество элементарных событий, то Ω называется непрерывным пространством. Принципиальное отличие непрерывного Ω от дискретного Ω заключается в том, что в общем (непрерывном) случае Ω нельзя назвать случайным событием любое подмножество Ω, подобно тому, как это сделано для дискретного Ω. Дело в том, что событие характеризуется принципиальной возможностью его наблюдения в результате случайного эксперимента. Среди же подмножеств непрерывного Ω могут быть ненаблюдаемые подмножества, т. е. такие, для которых не существует принципиальной возможности их наблюдения в результате случайного эксперимента. Их нельзя назвать случайными событиями.

Отмеченная особенность общего (непрерывного) пространства Ω требует другого определения случайного события. Такое определение дается в рамках аксиоматического построения теории вероятностей, которое впервые изложил А. Н. Колмогоров в 1933 г. Мы не будем здесь давать определение случайного события в общем случае. В каждой конкретной реальной ситуации это делается, исходя их физических, содержательных соображений.

Теперь введем операции над случайными событиями.

Суммой $A + B$ двух случайных событий A и B называется событие, которое заключается в наступлении хотя бы одного из событий A и B.

В случае дискретного Ω сумма $A + B$ состоит из всех элементарных событий, принадлежащих хотя бы одному из событий A и B.

Примеры.

1) Пусть при двукратном бросании монеты событие A означает, что выпал один герб, а событие B означает, что выпало два герба. Тогда $A + B$ — это событие, заключающееся в том, что выпали один герб или два герба.

2) Пусть из колоды карт наудачу вынимается одна карта, и пусть событие A состоит в том, что карта оказалась тузом, а событие B —

в том, что выбрана карта пиковой масти. Тогда событие $A + B$ состоит в том, что выбранная карта является тузом или картой пиковой масти.

3) Пусть при двукратном бросании игральной кости событие A состоит в том, что сумма выпавших очков не превосходит трех, а событие B — в том, что при втором бросании выпала единица. Тогда событие $A + B$ состоит в том, что выпала одна из следующих пар чисел: (1, 1); (2, 1); (3, 1); (4, 1); (5, 1); (6, 1); (1, 2).

Произведением AB двух случайных событий A и B называется событие, заключающееся в одновременном наступлении A и B.

В случае дискретного Ω произведение AB состоит из всех элементарных событий, одновременно принадлежащих и A и B.

Примеры.

1) Если из колоды карт наудачу вынимается одна карта и событие A состоит в том, что карта оказалась тузом, а событие B — в том, что выбрана карта пиковой масти, то событие AB состоит в том, что выбранная карта является тузом пиковой масти.

2) Пусть при двукратном бросании игральной кости событие A состоит в том, что сумма выпавших очков не превосходит четырех, а событие B — в том, что при втором бросании выпали два очка. Тогда событие AB состоит в том, что выпали (1, 2) или (2, 2).

Определения суммы и произведения двух случайных событий очевидным образом обобщаются на случай любого конечного числа случайных событий.

Разностью $A - B$ событий A и B называется событие, которое состоит в том, что одновременно A произошло и B не произошло.

В случае дискретного Ω разность $A - B$ состоит из всех тех элементарных событий, которые принадлежат событию A и не принадлежат событию B.

Примеры.

1) Пусть при двукратном бросании монеты событие A состоит, что герб выпал хотя бы один раз, а событие B состоит в том, что герб выпал не более одного раза. Тогда событие $A - B$ состоит в том, что выпали два герба.

2) Пусть при бросании двух игральных костей событие A состоит в том, что сумма выпавших очков не превосходит четырех, а событие B — в том, что сумма выпавших очков является четным числом. Тогда событие $A - B$ состоит в том, что выпали (1, 2) или (2, 1).

Введенные операции над событиями допускают геометрическую интерпретацию. Пусть событиями A и B являются попадания соответственно в большой и малый круг (рис. 1).

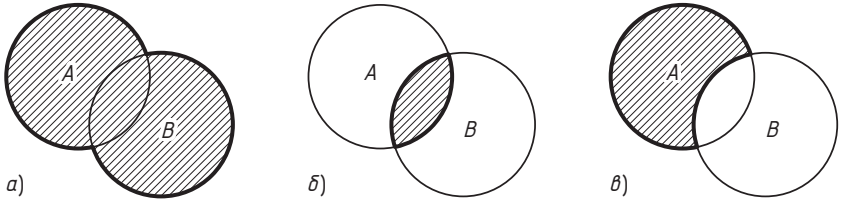


Рис. 1

Тогда событием $A + B$ является заштрихованная область на рис. 1, *а*, событием AB является заштрихованная область на рис. 1, *б*, событием $A - B$ является заштрихованная область на рис. 1, *в*.

Дадим еще некоторые определения.

Все пространство Ω элементарных событий называется *достоверным событием*, а пустое множество $\varnothing$, т. е. множество, не содержащее ни одного элементарного события, называется *невозможным событием*.

Событие $\overline{A} = \Omega - A$ называется *дополнением к A* или *противоположным событию A*. Например, если при двукратном бросании монеты событие A состоит в том, что герб выпал хотя бы один раз, то событие $\overline{A}$ означает, что герб не выпал ни разу. В событии $\overline{A}$ содержатся всевозможные элементарные события, не входящие в A, если Ω — дискретное пространство.

Два события A и B называются *равными* ($A = B$), если A происходит только в том случае, когда происходит B. События A и B в этом случае содержат одни и те же элементарные события для дискретного Ω. События A и B называются *несовместными*, если $AB = \varnothing$. Это значит, что A и B не могут произойти одновременно. В случае дискретного Ω у событий A и B нет ни одного общего элементарного события. Например, если при двукратном бросании игральной кости событие A означает, что сумма выпавших очков является нечетным числом, а событие B означает, что выпали (6,6), то A и B — несовместные события.

Несколько событий A_1, A_2, …, A_n называются *попарно несовместными*, если несовместной является каждая пара из этих событий. Например, все элементарные события являются несовместными.

Несколько несовместных событий A_1, A_2, …, A_n *образуют полную систему событий*, если их сумма является достоверным событием, т. е. $A_1 + A_2 + \ldots + A_n = \Omega$.

Множество всех элементарных событий, входящих в дискретное пространство Ω, образует полную систему событий. Еще один пример полной системы событий дают события A и B в опыте с подбрасыванием игральной кости, если под событием A понимать выпадение нечетного числа очков, а под событием B — выпадение четного числа очков.

В дальнейшем тот факт, что элементарное событие ω принадлежит пространству Ω элементарных событий, условимся обозначать следующим образом: $\omega \in \Omega$.

§ 2. Вероятности случайных событий

Пусть сначала Ω — дискретное пространство элементарных событий. Чтобы определить вероятность любого случайного события, аксиоматически вводится понятие вероятности элементарного события $\omega \in \Omega$.

Аксиома. Каждому элементарному событию $\omega \in \Omega$ ставится в соответствие некоторая числовая характеристика $\mathsf{P}(\omega) \geqslant 0$ шансов его появления, называемая вероятностью события ω, причем выполняется следующее условие нормировки:

$$\sum \mathsf{P}(\omega) = 1,$$

где знак $\sum$ означает конечную сумму вероятностей всех элементарных событий, если Ω содержит конечное число элементарных событий ω, и знак $\sum$ означает сумму ряда из вероятностей всех элементарных событий, если Ω содержит счетное число исходов ω.

Из приведенной аксиомы сразу следует, что для любого элементарного события $\omega \in \Omega$ вероятность $\mathsf{P}(\omega)$ удовлетворяет двойному неравенству $0 \leqslant \mathsf{P}(\omega) \leqslant 1$.

Вероятностью любого случайного события A называется сумма вероятностей всех элементарных событий, составляющих событие A.

Из этого определения и из аксиомы немедленно следует, что всегда $0 \leqslant \mathsf{P}(A) \leqslant 1$, $\mathsf{P}(\Omega) = 1$, $\mathsf{P}(\varnothing) = 0$ (здесь Ω — достоверное событие, $\varnothing$ — невозможное событие). Кроме того, если A и B — несовместные события, то $\mathsf{P}(A + B) = \mathsf{P}(A) + \mathsf{P}(B)$.

В общем случае из конкретных условий решаемой задачи обосновать определенное введение вероятностей элементарных событий не так-то просто. Это нетрудно сделать в том частном случае, когда

пространство Ω состоит из конечного числа N элементарных событий, причем условия задачи таковы, что вероятности каждого из этих N элементарных событий нам представляются равными. Именно такая ситуация наблюдается при подбрасывании симметричной монеты, бросании правильной игральной кости, извлечении наугад карты из хорошо перемешанной колоды игральных карт и т. д. В силу аксиомы в этом случае вероятность $\mathsf{P}(\omega) = \frac{1}{N}$ для каждого элементарного события ω. Если случайное событие A содержит $N(A)$ элементарных событий, то вероятность

$$\mathsf{P}(A) = \frac{N(A)}{N}.$$

Получили так называемое *классическое определение вероятности* события A: вероятность события A равна отношению числа благоприятных элементарных событий (т. е. элементарных событий, входящих в A) к числу всех возможных элементарных событий.

Разумеется, что при опытах с несимметричными монетами или с неправильными игральными костями было бы неправомерным использование классического определения вероятности.

Иногда на практике для приближенного определения вероятности элементарного события ω предлагается брать относительную частоту появления ω в достаточно длинном ряду случайных экспериментов в неизменных условиях. Такое частотное определение вероятностей элементарных событий имеет ряд существенных недостатков. Например, невозможно неограниченно долго поддерживать неизменность условий случайного эксперимента.

До сих пор рассматривались дискретные пространства Ω элементарных событий. Пусть теперь Ω — непрерывное пространство элементарных событий. Как уже разъяснялось в предыдущем параграфе, не всякое подмножество Ω можно объявить случайным событием. Не определяя случайные события, укажем только характерное свойство множества случайных событий. Множество всех случайных событий всегда содержит $\varnothing$, Ω, и если содержит события A и B, то обязано также содержать их сумму, произведение и дополнения $\overline{A}$, $\overline{B}$.

Итак, пусть выделен некоторый класс F случайных событий в рассматриваемом опыте. Все остальные подмножества Ω, не входящие в выделенный класс F, случайными событиями не являются и для них понятие вероятности не вводится. Для случайных событий из F понятие вероятности вводится аксиоматически.

Аксиома. Каждому случайному событию $A \in F$ ставится в соответствие число $\mathsf{P}(A)$, называемое вероятностью события A, причем выполнены следующие свойства:

а) $\mathsf{P}(A) \geqslant 0$ для любого события $A \in F$;

б) $\mathsf{P}(\Omega) = 1$;

в) если события A и B несовместны, то

$$\mathsf{P}(A + B) = \mathsf{P}(A) + P(B).$$

Из этой аксиомы следует, что $0 \leqslant \mathsf{P}(A) \leqslant 1$ для любого события A.

Отметим, что приведенный здесь аксиоматический способ введения вероятности событий включает в себя в качестве частного случая ранее рассмотренный способ определения вероятности событий в случае дискретного Ω.

В отличие от дискретного пространства Ω в общем (непрерывном) пространстве Ω могут существовать возможные события A (т. е. $A \neq \varnothing$), для которых $\mathsf{P}(A) = 0$. Соответственно существуют противоположные к ним недостоверные события A (т. е. $A \neq \Omega$), для которых $\mathsf{P}(A) = 1$.

В заключение § 2 приведем несколько примеров на классическое определение вероятности событий. При решении таких примеров могут использоваться основные понятия комбинаторики: размещения, перестановки, сочетания. Поэтому дадим сначала определения этих понятий и проиллюстрируем их примерами.

Пусть задано множество, содержащее n элементов. *Размещением* из n элементов по k элементов ($0 \leqslant k \leqslant n$) называется каждое его упорядоченное подмножество, состоящее из k элементов.

Из этого определения следует, что размещения из n элементов по k элементов — это всевозможные подмножества из k элементов, отличающиеся либо составом элементов, либо порядком их следования. Частный случай $k = 0$ дает лишь пустое множество.

Число всех размещений из n элементов по k элементов обозначается A_n^k. Очевидно, что $A_n^0 = 1$, так как в этом случае имеется только пустое множество. При $k > 0$

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

где символом $n!$ (читается: «эн факториал») обозначается произведение всех натуральных чисел от единицы до n. Если определить $0! = 1$, то последняя формула A_n^k дает верный результат и для случаев $k = 0$, $n = 0$.

Пример. Группа студентов изучает 6 учебных дисциплин. Сколькими способами можно составить расписание занятий для этой

группы в понедельник, если в этот день недели должно быть 3 различных занятия?

Решение. Число способов равно числу размещений из 6 элементов по 3, т. е. равно A_6^3.

Из общей формулы A_n^k, при $n=6$, $k=3$ получаем, что $A_6^3 = \frac{6!}{3!} = 120$. ■

Замечание. Здесь и в дальнейшем знак ■ означает конец примера.

Размещения из n элементов по n элементов называются *перестановками* из n элементов.

Из определения перестановок следует, что различные перестановки отличаются друг от друга только порядком следования элементов. Число всех перестановок из n элементов обозначается P_n. Из формулы A_n^k при $k=n$ получаем, что

$$P_n = n!.$$

Пример. На собрании пожелали выступить 3 человека. Сколькими способами их можно расположить в списке выступающих?

Решение. Число способов равно $P_3 = 3! = 1 \cdot 2 \cdot 3 = 6$. ■

Пусть задано множество, состоящее из n элементов. Каждое его подмножество, содержащее k элементов ($0 \leqslant k \leqslant n$), называется *сочетанием* из n элементов по k элементов.

Из этого определения следует, что различные сочетания отличаются друг от друга составом элементов. Подмножества, отличающиеся друг от друга только порядком следования элементов, не считаются различными.

Число всех сочетаний из n элементов по k элементов обозначается C_n^k и определяется следующей формулой:

$$C_n^k = \frac{n!}{(n-k)!\,k!} = \frac{n(n-1)(n-2)\cdot\ldots\cdot(n-k+1)}{k!}.$$

Пример. В группе 25 студентов. Сколькими способами можно выделить двух человек для дежурства?

Решение. Число способов равно $C_{25}^2 = 300$. ■

Числа C_n^k обладают некоторыми интересными и важными свойствами. Отметим одно такое свойство. Из формулы C_n^k легко получить равенство $C_n^k = C_n^{n-k}$.

Пользуясь этим свойством, можно упрощать вычисление чисел C_n^k в тех случаях, когда $k > \frac{n}{2}$. Например, $C_{20}^{18} = C_{20}^2 = 190$.

Рассмотрим теперь примеры на использование классического определения вероятности.

Пример 1. Монета бросается дважды. Какова вероятность того, что хотя бы один раз выпал герб?

Решение. Обозначим интересующее нас событие через A. Элементарными событиями при двукратном бросании монеты будут: *ГГ*, *ГР*, *РГ*, *РР* (здесь *Г* означает выпадение герба, а *Р* — выпадение решки). Благоприятными для события A будут первые три элементарных события. Значит, $N = 4$, $N(A) = 3$, $\mathsf{P}(A) = \dfrac{N(A)}{N} = \dfrac{3}{4}$. ■

Пример 2. Бросаются две игральные кости. Какова вероятность того, что сумма очков, выпавших на двух костях, окажется равной 8?

Решение. Пусть A — интересующее нас событие «сумма выпавших очков равна 8». При бросании двух костей число всех элементарных событий $N = 36$, поскольку каждая кость дает 6 исходов и нужно брать комбинации этих исходов. Благоприятными для A являются элементарные события (2, 6); (3, 5); (4, 4); (5, 3); (6, 2), т. е. $N(A) = 5$. Следовательно, $\mathsf{P}(A) = \dfrac{5}{36}$. ■

Пример 3. Из колоды в 36 карт наугад выбираются 3. Какова вероятность того, что это будут валет пик и два короля?

Решение. Пусть A — искомое событие «выбраны валет пик и два короля». Число всех элементарных событий $N = C_{36}^3$ равно числу сочетаний из 36 по 3, а число благоприятных элементарных событий $N(A) = C_4^2$ — числу возможных вариантов выбора двух королей из четырех. Тогда

$$\mathsf{P}(A) = \frac{C_4^2}{C_{36}^3} = \frac{1}{1190}. \blacksquare$$

Пример 4. Набирая номер телефона, студент забыл две последние цифры и помня лишь, что эти цифры различны, набрал их наудачу. Какова вероятность того, что номер набран правильно?

Решение. Пусть B — событие «цифры набраны правильно». Две последние цифры можно набрать A_{10}^2 способами (здесь A_{10}^2 — число размещений из 10 по 2), а событию B будет благоприятствовать лишь один способ. Поэтому, $N = A_{10}^2$, $N(B) = 1$, значит,

$$\mathsf{P}(B) = \frac{1}{A_{10}^2} = \frac{1}{90}. \blacksquare$$

Пример 5. Из шести букв *М*, *А*, *А*, *И*, *Н*, *Ш* наудачу выбираются одна за другой и приставляются друг к другу в порядке выбора все

шесть букв. Найти вероятность того, что при этом получится слово «*МАШИНА*».

Решение. Пусть B — искомое событие. Число N всех элементарных событий равно $P_6 = 6!$ — числу перестановок из шести букв, а число благоприятных B элементарных событий $N(B) = 2$, так как две буквы A можно расположить двумя разными способами. Следовательно,

$$\mathsf{P}(B) = \frac{2}{6!} = \frac{1}{360}. \blacksquare$$

Пример 6. В лифт на первом этаже девятиэтажного дома вошли 5 человек. Найти вероятность того, что все они выйдут на разных этажах.

Решение. Пусть B — интересующее нас событие. Число всех элементарных событий $N = 8^5$, так как для каждого человека имеется 8 вариантов выхода из лифта. Число благоприятных B элементарных событий $N(B) = A_8^5$ равно числу размещений из восьми по пять. Поэтому,

$$\mathsf{P}(B) = \frac{A_8^5}{8^5} = \frac{105}{512}. \blacksquare$$

§ 3. Основные правила действий с вероятностями случайных событий

1. Вероятность противоположного события

Пусть A — некоторое случайное событие и $\overline{A}$ — противоположное ему событие. Так как $A + \overline{A} = \Omega$, $A \cdot \overline{A} = \varnothing$, то из аксиом § 2 следует, что

$$\mathsf{P}(\overline{A}) = 1 - \mathsf{P}(A).$$

Вероятность противоположного события $\overline{A}$ равна единице минус вероятность события A.

Пример 1. Монета бросается трижды. Найти вероятность того, что герб выпадет, по крайней мере, один раз.

Решение. Пусть A — интересующее нас событие. Тогда $\overline{A}$ — событие, состоящее в том, что при трехкратном бросании монеты герб ни разу не выпал. Число всех элементарных событий $N = 2^3 = 8$ и только одно элементарное событие благоприятствует $\overline{A}$: $N(\overline{A}) = 1$. Тогда $\mathsf{P}(\overline{A}) = \frac{1}{8}$ и, значит, $\mathsf{P}(A) = 1 - \mathsf{P}(\overline{A}) = 1 - \frac{1}{8} = \frac{7}{8}$. $\blacksquare$

Пример 2. Среди 20 студентов группы, успешно сдавших экзамены, пятнадцать студентов сдали экзамены без троек. Случайным образом выбираются три студента. Найти вероятность того, что среди них хотя бы один сдал экзамены без троек.

Решение. Пусть A — искомое событие. Тогда $\overline{A}$ — событие, заключающееся в том, что все три студента сдали экзамены с тройками. Число всех элементарных событий $N = C_{20}^3$, а число элементарных событий, благоприятствующих $\overline{A}$, равно $N(\overline{A}) = C_5^3$. Тогда

$$\mathsf{P}(\overline{A}) = \frac{C_5^3}{C_{20}^3} \quad \text{и} \quad \mathsf{P}(\overline{A}) = 1 - \mathsf{P}(\overline{A}) = 1 - \frac{C_5^3}{C_{20}^3} = \frac{113}{114}. \blacksquare$$

2. Вероятность суммы двух событий

Пусть A и B — два случайных события. Из аксиом следует, что $\mathsf{P}(A + B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(AB)$.

Вероятность суммы двух событий равна сумме вероятностей событий без вероятности их произведения.

Пример 1. Из колоды в 36 карт наудачу вынимается одна. Какова вероятность того, что будет вынут туз или карта красной масти?

Решение. Пусть событие A — «выбран туз» и событие B — «выбрана карта красной масти». Тогда

$$\mathsf{P}(A) = \frac{4}{36} = \frac{1}{9}, \quad \mathsf{P}(B) = \frac{18}{36} = \frac{1}{2}, \quad \mathsf{P}(AB) = \frac{1}{18}.$$

Значит, искомая вероятность

$$\mathsf{P}(A + B) = \frac{1}{9} + \frac{1}{2} - \frac{1}{18} = \frac{5}{9}. \blacksquare$$

В частности, когда события A и B несовместны, $\mathsf{P}(AB) = 0$ и формула вероятности суммы двух событий принимает вид: $\mathsf{P}(A + B) = \mathsf{P}(A) + \mathsf{P}(B)$.

Пример 2. От студенческой группы, которая состоит из 15 девушек и 5 юношей, на конференцию выбираются наугад два студента. Какова вероятность того, что среди выбранных хотя бы один юноша?

Решение. Пусть событие A состоит в том, что среди выбранных двух человек хотя бы один юноша. Если событие A произойдет, то обязательно произойдет одно из двух несовместных событий: B — «выбраны юноша и девушка», C — «выбраны два юноши». Поэтому $A = B + C$. Находим, что

$$\mathsf{P}(B) = \frac{5 \cdot 15}{C_{20}^2}, \quad \mathsf{P}(C) = \frac{C_5^2}{C_{20}^2}.$$

Тогда

$$\mathsf{P}(A) = \mathsf{P}(B) + \mathsf{P}(C) = \frac{17}{38}. \blacksquare$$

Формула вероятности суммы двух несовместных событий обобщается на случай любого конечного числа несовместных событий следующим образом.

Если A_1, A_2, …, A_n — попарно несовместные события, то вероятность их суммы равна сумме вероятностей этих событий, т. е. $\mathsf{P}(A_1 + A_2 + \ldots + A_n) = \mathsf{P}(A_1) + \mathsf{P}(A_2) + \ldots + \mathsf{P}(A)_n$.

3. Условные вероятности. Независимость событий

Рассмотрим ситуацию, когда заранее поставленное условие или осуществление некоторого события дают дополнительную информацию о структуре пространства Ω, позволяющую исключить из числа возможных часть элементарных событий.

Пример 1. Выбрасывается один раз игральная кость. Пусть событие A — выпадение шестерки.

По классическому определению $\mathsf{P}(A) = \frac{1}{6}$. Если же дополнительно стало известно, что произошло событие B — число выпавших очков четное, то вероятность A изменится. Так как событие B состоит из трех элементарных событий и событие A вместе с B выступает в одном случае, то новая вероятность A равна $\frac{1}{3}$. Эту вероятность естественно назвать условной вероятностью события A при условии, что произошло событие B, и обозначить $\mathsf{P}\{A \mid B\}$.

Перейдем к общему определению условной вероятности. Пусть заданы два случайных события A и B, причем $\mathsf{P}(B) > 0$. Условная вероятность события A при условии, что произошло событие B, определяется формулой

$$\mathsf{P}\{A \mid B\} = \frac{\mathsf{P}(AB)}{\mathsf{P}(B)}.$$

Условная вероятность события A при условии, что произошло событие B, равна отношению вероятности произведения событий AB к вероятности события B.

Пример 2. В ящике 10 одинаковых на ощупь шаров с номерами от 1 до 10. Наудачу выбирается один шар. Найти вероятность события A — «номер выбранного шара делится на 5» при условии, что номер выбранного шара — число четное.

Решение. Пусть событие B — номер выбранного шара — четное число. Тогда $B = \{2, 4, 6, 8, 10\}$. Значит,

$$\mathsf{P}(B) = \frac{5}{10}, \quad \mathsf{P}(AB) = \frac{1}{10}, \quad \mathsf{P}\{A \mid B\} = \frac{1}{10} : \frac{5}{10} = \frac{1}{5}.$$

Заметим, что по классическому определению

$$\mathsf{P}(A) = \frac{2}{10} = \frac{1}{5}. \blacksquare$$

Пример 2 дает пример события A, независимого от события B, так как $\mathsf{P}\{A \mid B\} = \mathsf{P}(A)$. Дадим определение независимости двух событий.

Два события A и B называются *независимыми*, если $\mathsf{P}(AB) = \mathsf{P}(A) \cdot \mathsf{P}(B)$, т. е. если вероятность произведения двух событий равна произведению вероятностей этих событий.

Пример 3. Из колоды в 36 карт наудачу вынимается одна. Пусть событие A — появился туз и событие B — появилась карта черной масти. Являются ли независимыми события A и B?

Решение. Имеем:

$$\mathsf{P}(A) = \frac{4}{36} = \frac{1}{9}, \quad \mathsf{P}(B) = \frac{18}{36} = \frac{1}{2}, \quad \mathsf{P}(AB) = \frac{2}{36} = \frac{1}{18}.$$

Итак,

$$\mathsf{P}(AB) = \mathsf{P}(A) \cdot \mathsf{P}(B) = \frac{1}{18},$$

т. е. A и B — независимые. $\blacksquare$

Формула из определения независимости событий A и B является формулой вероятности произведения двух независимых событий. Из формулы условной вероятности получается формула вероятности произведения двух произвольных событий A и B:

$$\mathsf{P}(AB) = \mathsf{P}\{A \mid B\} \cdot \mathsf{P}(B).$$

Определение независимости двух событий можно распространить на систему более чем двух событий.

События $A_1, A_2, \ldots, A_m$ $(m > 2)$ называются *независимыми в совокупности*, если наряду с их попарной независимостью являются независимыми каждое из этих событий и произведение любого числа остальных событий.

Поясним разницу между попарной независимостью и независимостью в совокупности на примере трех событий A, B, C. Для попарной независимости A, B, C необходимо, чтобы все пары —

A и B, A и C, B и C — были независимы. Для независимости в совокупности помимо этого необходима независимость еще трех пар событий: A и BC, B и AC, C и AB. Таким образом, требование независимости в совокупности является более сильным, чем требование попарной независимости. Можно привести примеры событий, которые являются попарно независимыми и не являются независимыми в совокупности.

Формула вероятности произведения двух независимых событий обобщается на случай любого конечного числа независимых в совокупности событий. Именно, если события A_1, A_2, …, A_m являются независимыми в совокупности, то $\mathsf{P}(A_1 \cdot A_2 \cdot \ldots \cdot A_m) = \mathsf{P}(A_1) \times \times \mathsf{P}(A_2) \cdot \ldots \cdot \mathsf{P}(A_m)$.

Понятие независимости является весьма важным в теории вероятностей. Однако при решении реальных задач не всегда удается проверить независимость событий с помощью введенного выше определения. В таких случаях оправдано правило считать независимыми события, не связанные причинно. Следует отметить, что, как видно на примерах, определение независимости случайных событий значительно шире понятия реальной (физической) независимости в смысле принадлежности к причинно не связанным явлениям. Так как теория вероятностей изучает некоторые модели реальных явлений, то ясно, что независимость событий никак не противоречит принципу всеобщей связи всех реальных явлений.

4. Формула полной вероятности

Иногда возникает ситуация, когда прямое вычисление вероятности интересующего нас случайного события A невозможно или трудно, в то время как вполне доступно вычисление условных вероятностей A при некоторых известных условиях, при которых может произойти A. В такой ситуации для вычисления вероятности $\mathsf{P}(A)$ используется формула полной вероятности.

Пусть A — некоторое случайное событие, H_1, H_2, …, H_n — полная система попарно несовместных событий, называемых гипотезами, и пусть событие A может произойти лишь при появлении одной из гипотез H_1, H_2, …, H_n. Тогда имеет место *формула полной вероятности*:

$$\mathsf{P}(A) = \sum_{i=1}^{n} \mathsf{P}(H_i) \cdot \mathsf{P}(A \mid H_i).$$

Поскольку событие A может произойти лишь с одним из событий H_1, H_2, …, H_n, то $A = AH_1 + AH_2 + \ldots + AH_n$. Отсюда следует, что

формула полной вероятности обобщает формулу условной вероятности на случай нескольких гипотез.

Пример. Была проведена одна и та же контрольная работа в пяти параллельных группах студентов 1 курса. Распределение по группам количества студентов, а также работ, выполненных на «отлично», следующее: 1-я группа: 25 студентов и 5 работ; 2-я: 24 и 6; 3-я: 27 и 9, 4-я: 21 и 7; 5-я: 28 и 7. Найти вероятность того, что наудачу выбранная работа из наудачу выбранной группы окажется выполненной на «отлично».

Решение. Пусть A — событие, состоящее в том, что взятая работа выполнена на «отлично», а H_1, H_2, H_3, H_4, H_5 — события, состоящие в том, что работа выполнена студентом соответственно первой, второй, третьей, четвертой, пятой группы. По условиям имеем:

$$\mathsf{P}(H_1) = \mathsf{P}(H_2) = \mathsf{P}(H_3) = \mathsf{P}(H_4) = \mathsf{P}(H_5) = \frac{1}{5},$$

$$\mathsf{P}(A \mid H_1) = \frac{5}{25} = \frac{1}{5}, \quad \mathsf{P}(A \mid H_2) = \frac{6}{24} = \frac{1}{4}, \quad \mathsf{P}(A \mid H_3) = \frac{9}{27} = \frac{1}{3},$$

$$\mathsf{P}(A \mid H_4) = \frac{7}{21} = \frac{1}{3}, \quad \mathsf{P}(A \mid H_5) = \frac{7}{28} = \frac{1}{4}.$$

По формуле полной вероятности находим искомую вероятность

$$\mathsf{P}(A) = \frac{1}{5}\cdot\frac{1}{5} + \frac{1}{5}\cdot\frac{1}{4} + \frac{1}{5}\cdot\frac{1}{3} + \frac{1}{5}\cdot\frac{1}{3} + \frac{1}{5}\cdot\frac{1}{4} =$$

$$= \frac{1}{5}\left(\frac{1}{5} + \frac{1}{2} + \frac{2}{3}\right) = \frac{1}{5}\cdot\frac{31}{30} = \frac{31}{150}. \blacksquare$$

5. Формула Байеса

Пусть, как и в предыдущем пункте, A — произвольное событие и H_1, H_2, …, H_n — полная система попарно несовместных гипотез такая, что событие A может наступить лишь тогда, когда появится одно из событий H_1, H_2, …, H_n.

Для использования формулы полной вероятности необходимо знать доопытные (априорные) условные вероятности A при выполнении каждой из гипотез H_1, H_2, …, H_n, а также вероятности всех гипотез. Может оказаться, что перед опытом некоторые из этих вероятностей определены неточно. Опыт проводится с целью их уточнения. Основываясь на результатах опыта, заменяют доопытные (априорные) вероятности послеопытными (апостериорными). Это делается с помощью формул Байеса. Пусть $\mathsf{P}(A) > 0$. Тогда при сфор-

мулированных в начале этого пункта условиях справедливы *формулы Байеса*:

$$\mathsf{P}(H_k \mid A) = \frac{\mathsf{P}(H_k) \cdot \mathsf{P}(A \mid H_k)}{\mathsf{P}(A)},$$

$k = 1, 2, \ldots, n$, где $\mathsf{P}(A)$ находится по формуле полной вероятности.

Формулы Байеса позволяют найти условные вероятности гипотез при условии, что произошло событие A, через вероятности гипотез и условные вероятности $\mathsf{P}(A \mid H_i)$, вычисленные до того, как событие A произошло.

Пример 1. Три группы студентов некоторого курса сдавали экзамен по математике. В первой группе 30% студентов курса, во второй группе 35% студентов курса и в третьей группе 35% студентов курса. В первой группе 30% студентов сдали экзамен на «отлично», во второй группе 40% студентов сдали на «отлично» и в третьей группе тоже 40% студентов сдали на «отлично».

Наудачу выбранный студент оказался отличником. Найти вероятность того, что это студент из первой группы.

Решение. Пусть A — событие, состоящее в том, наудачу выбранный студент является отличником, и H_1, H_2, H_3 — события, состоящие в том, что студент соответственно из первой, второй и третьей группы. Из условий имеем: $\mathsf{P}(H_1) = 0{,}3$, $\mathsf{P}(H_2) = \mathsf{P}(H_3) = 0{,}35$, $\mathsf{P}(A \mid H_1) = 0{,}3$, $\mathsf{P}(A \mid H_2) = \mathsf{P}(A \mid H_3) = 0{,}4$.

По формуле полной вероятности $\mathsf{P}(A) = 0{,}3 \cdot 0{,}3 + 2 \cdot 0{,}35 \cdot 0{,}4 = 0{,}37$.

По формуле Байеса получаем искомый ответ

$$\mathsf{P}(H_1 \mid A) = \frac{0{,}3 \cdot 0{,}3}{0{,}37} = \frac{9}{37}. \blacksquare$$

Пример 2. Предположим, что 5% всех мужчин и 0,25% всех женщин — дальтоники. Наугад выбранное лицо оказалось дальтоником. Какова вероятность того, что это мужчина? (Считать, что количество мужчин и женщин одинаково.)

Решение. Пусть Д — событие, состоящее в том, что наугад выбранное лицо оказалось дальтоником. Так как по условиям $\mathsf{P}(\text{М}) = \mathsf{P}(\text{Ж}) = \frac{1}{2}$ и $\mathsf{P}(\text{Д} \mid \text{М}) = 0{,}05$, $\mathsf{P}(\text{Д} \mid \text{Ж}) = 0{,}0025$, то по формуле полной вероятности

$$\mathsf{P}(\text{Д}) = \frac{1}{2} \cdot 0{,}05 + \frac{1}{2} \cdot 0{,}0025 = \frac{1}{2} \cdot 0{,}0525.$$

По формуле Байеса искомая вероятность

$$\mathsf{P}(\text{М} \mid \text{Д}) = \left(\frac{1}{2} \cdot 0{,}05\right) : \left(\frac{1}{2} \cdot 0{,}0525\right) = \frac{20}{21}. \blacksquare$$

6. Формула Бернулли

Пусть при одинаковых условиях проводится n испытаний, причем вероятность наступления события A в каждом из них не зависит от исходов других испытаний. Такие испытания называются *независимыми*.

Последовательность независимых испытаний, в каждом из которых с постоянной вероятностью p может произойти или с постоянной вероятностью $q = 1 - p$ не произойти интересующее нас событие A, называются *испытаниями Бернулли*. Испытания Бернулли являются математической моделью серии опытов, повторяющихся в одинаковых условиях, с двумя исходами. Примерами таких опытов могут служить подбрасывание монеты или игральной кости.

Практически событие A может появится в n испытаниях Бернулли любое число k раз ($0 \leqslant k \leqslant n$) в разных комбинациях, чередуясь с противоположным событием $\overline{A}$. Из независимости испытаний следует независимость в совокупности и группы событий, представляющей собой произвольную комбинацию событий A и $\overline{A}$, одно из которых обязательно произойдет в каждом испытании.

Нас интересует вероятность появления k раз события A в n испытаниях Бернулли. Обозначим эту вероятность через $P_n(k)$. Вероятность $P_n(k)$ находится с помощью *формулы Я. Бернулли*:

$$P_n(k) = C_n^k p^k q^{n-k},$$

где C_n^k — число сочетаний из n по k, $p + q = 1$.

Пример 1. Игральная кость подбрасывается 4 раза. Какова вероятность того, что шестерка появится дважды?

Решение. Пусть A — событие «появилась шестерка». Тогда $p = \frac{1}{6}$, $q = \frac{5}{6}$, $n = 4$, $k = 2$ и, значит по формуле Бернулли искомая вероятность

$$P_4(2) = C_4^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^2 = \frac{25}{216}. \blacksquare$$

Пример 2. Монета подбрасывается 5 раз. Какова вероятность того, что герб появится не менее двух раз?

Решение. Пусть A — интересующее нас событие. Тогда $\overline{A}$ — событие, заключающееся в том, что при пяти бросаниях монеты герб появился не более одного раза. Так как для монеты $p = q = \frac{1}{2}$, то

$$\mathsf{P}(\overline{A}) = C_5^0 \cdot \left(\frac{1}{2}\right)^5 + C_5^1 \cdot \left(\frac{1}{2}\right)^5 = \frac{6}{2^5} = \frac{3}{16}.$$

Следовательно,

$$\mathsf{P}(A) = 1 - \mathsf{P}(\overline{A}) = \frac{13}{16}. \blacksquare$$

Пример 3. Если вероятности выигрыша и проигрыша в одной партии в теннис одинаковы и равны $\frac{1}{2}$, то что более вероятно: выиграть три партии из четырех или пять из восьми?

Решение. По формуле Бернулли вероятность выигрыша трех партий из четырех

$$P_4(3) = C_4^3\left(\frac{1}{2}\right)^4 = \frac{1}{4},$$

а вероятность выигрыша пяти партий из восьми

$$P_8(5) = C_8^5\left(\frac{1}{2}\right)^8 = \frac{7}{32}.$$

Следовательно, $P_4(3) > P_8(5)$. ■

Замечание. При больших значениях n вычисления вероятностей по формуле Бернулли становятся затруднительными. В таких случаях используются асимптотические приближения формулы Бернулли, которые дают теорема Пуассона и теорема Муавра—Лапласа.

В испытаниях Бернулли в каждом из независимых испытаний может появиться лишь одно из двух несовместных событий A и $\overline{A}$. Обобщением схемы Бернулли с двумя несовместными исходами A и $\overline{A}$ является полиномиальная схема с N несовместными исходами A_1, A_2, …, A_N с заданными вероятностями p_1, p_2, …, p_N: $p_1 + p_2 + \ldots + p_N = 1$. Примером полиномиальной схемы служит n бросаний игральной кости. Здесь $N = 6$ и $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$. Имеется обобщение формулы Бернулли на случай полиномиальной схемы, позволяющее найти вероятность того, что в n независимых испытаниях событие A_1 появится k_1 раз, событие A_2 появится k_2 раз, …, событие A_N появится k_N раз, где k_1, k_2, …, k_N — неотрицательные целые числа, удовлетворяющие условию $k_1 + k_2 + \ldots + k_N = n$.

7. Понятие о цепях Маркова

Обобщением понятия испытаний Бернулли является понятие цепи Маркова*).

Пусть проводится последовательность испытаний, в каждом из которых может появиться только одно из двух несовместных событий A_1 и A_2, образующих полную группу событий ($A_1 + A_2 = \Omega$). Обозначим через $p_{ij}(n)$ условную вероятность того, что в n-м испытании наступит событие A_j ($j = 1, 2$) при условии, что в $(n-1)$-м испытании наступило событие A_i ($i = 1, 2$).

*) А. Марков (1856—1922) — русский математик.

Последовательность таких испытаний называют *цепью Маркова*, если условная вероятность $p_{ij}(n)$ не зависит ни от результатов предшествующих испытаний, ни от номера испытания. Поэтому для цепи Маркова вместо $p_{ij}(n)$ пишут p_{ij}. Таким образом, в отличие от независимых испытаний в теории цепей Маркова допускается зависимость исхода любого испытания от исхода предыдущего испытания, но только от него.

Обычно события A_1 и A_2 называют *состояниями* и рассматривается некоторая система, которая в каждый момент времени может находиться в одном из двух состояний. В результате испытания система либо переходит из одного состояния в другое, либо остается в том же состоянии. Тогда $p_{ij}(n)$ означает условную вероятность того, что в n-м испытании система будет находиться j-м состоянии ($j = 1,\ 2$) при условии, что в $(n-1)$-м испытании система находилась в i-м состоянии ($i = 1,\ 2$).

Используя введенную терминологию, цепью Маркова называют последовательность испытаний, в которых условная вероятность $p_{ij}(n)$ перехода системы из i-го состояния в j-е состояние ($i,\ j = 1,\ 2$) не зависит ни от результатов предшествующих испытаний, ни от номера испытания. Условную вероятность p_{ij} называют переходной вероятностью ($i,\ j = 1,\ 2$).

Приведем примеры цепей Маркова.

1. Пусть для учащегося в конце каждого учебного года отмечаются два состояния: «успевающий» и «неуспевающий». Если считать вероятность принятия учащимся одного из состояний в конце года зависящей лишь от его состояния в конце предыдущего года, то получаем пример цепи Маркова.

2. Пусть для семьи в конце каждого года отмечаются два состояния: «доход семьи достаточный» и «доход семьи недостаточный». Если считать вероятность принятия семьей одного из состояний в конце года зависящей лишь от ее состояния в конце предыдущего года, то опять получаем пример цепи Маркова.

Из переходных вероятностей p_{ij} ($i,\ j = 1,\ 2$) строится так называемая *матрица перехода*

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}.$$

Так как p_{11} и p_{12} — вероятности перехода из первого состояния в первое состояние и во второе соответственно, то $p_{11} + p_{12} = 1$. Аналогично получаем, что $p_{21} + p_{22} = 1$.

Зная матрицу перехода P, можно найти матрицу перехода $P(n)$ за n шагов по формуле $P(n) = P^n$.

Матрица P является матрицей перехода за один шаг, т. е. $P = P(1)$. Например, если

$$P = \begin{pmatrix} 0{,}4 & 0{,}6 \\ 0{,}3 & 0{,}7 \end{pmatrix},$$

то матрица перехода за два шага

$$P_2 = P^2 = \begin{pmatrix} 0{,}4 & 0{,}6 \\ 0{,}3 & 0{,}7 \end{pmatrix}\begin{pmatrix} 0{,}4 & 0{,}6 \\ 0{,}3 & 0{,}7 \end{pmatrix} = \begin{pmatrix} 0{,}34 & 0{,}66 \\ 0{,}33 & 0{,}67 \end{pmatrix}.$$

Понятие цепи Маркова обобщается на тот случай, когда вместо двух возможных состояний A_1 и A_2 рассматриваются $k \geqslant 2$ несовместных состояний A_1, A_2, …, A_k, образующих полную систему состояний.

Задачи к главе 1

1. Брошены две игральные кости. Найти вероятность того, что: а) сумма выпавших очков нечетная, б) произведение выпавших очков больше 15.

2. Из колоды в 36 карт взяты наугад три карты. Найти вероятность того, что это будут король пик и два туза.

3. В коробке 20 одинаковых конфет, причем 15 из них изготовлены на фабрике № 1, остальные — на фабрике № 2. Наугад берут три конфеты. Найти вероятность того, что все эти конфеты изготовлены: а) на фабрике № 1, б) на фабрике № 2.

4. В конверте 7 мужских и 3 женские фотографии, которые неразличимы на ощупь. Из конверта вынимают наугад 4 фотографии. Найти вероятность того, что из них: а) две женские фотографии, б) есть хотя бы одна женская фотография.

5. На полке в случайном порядке расставлено 9 различных книг. Найти вероятность того, что определенные 4 книги стоят рядом в определенном порядке.

6. Найти вероятность того, что дни рождения 12 человек придутся на различные месяцы года.

7. В лифт на первом этаже девятиэтажного здания входят восемь человек. Найти вероятность того, что на одном из этажей из лифта не выйдет ни один человек.

8. Случайно выбранная группа из пяти человек классифицируется по восьми профессиям. Найти вероятность того, что все люди разных профессий.

9. В студенческой группе из 25 человек учится студентка Иванова. По списку группы наугад выбираются пять студентов. Найти вероятность того, что среди них есть Иванова.

10. Пусть каждую неделю в городе происходит 7 автомобильных аварий. Предполагая, что все возможные распределения аварий равновероятны, найти вероятность того, что ежедневно происходит одна авария.

11. Из перетасованной колоды в 36 карт наугад вынимаются шесть карт. Найти вероятность того, что эти карты различных значений.

12. Имеется коллектив туристов, в котором 15 человек из Москвы, 13 — из Петербурга и 7 — из Самары. Из них отбирают группу из 10 человек. Найти вероятность того, что среди отобранных будут 6 москвичей.

13. В группе из 25 студентов десять имеют спортивные разряды. Найти вероятность того, что выбранные наудачу 3 студента являются разрядниками.

14. Имеется 6 карточек, на каждой из которых написана одна буква из набора $\{A, A, A, H, H, C\}$. Ребенок случайным образом прикладывает одну к другой все 6 карточек. Найти вероятность того, что получится слово «*АНАНАС*».

15. В детском саду работают воспитателями 6 женщин с высшим образованием и 4 женщины без высшего образования. По списку воспитателей наудачу отобраны 7 женщин. Найти вероятность того, что среди них окажутся три женщины с высшим образованием.

16. В записанном телефонном номере стерлись три последние цифры. Предполагая, что все комбинации стершихся цифр равновероятны, найти вероятности событий:

а) $A = \{$стерлись различные цифры$\}$,

б) $B = \{$стерлись одинаковые цифры$\}$.

17. Трижды подбрасывается монета. Найти вероятность того, что герб выпадет один раз, если известно, что число выпавших гербов нечетно.

18. Рассматриваются семьи с двумя детьми. Найти вероятность того, что оба ребенка — мальчики, если известно, что в семье есть мальчик. (Рождения мальчика и девочки считать равновероятными.)

19. Брошены две игральные кости. Найти вероятность того, что выпали две пятерки, если известно, что сумма выпавших очков делится на пять.

20. Среди 25 экзаменационных билетов 5 «хороших». Два студента по очереди берут по одному билету. Найти вероятность того,

что: а) оба студента взяли «хорошие» билеты, б) оба студента взяли не «хорошие» билеты, в) первый студент взял «хороший» билет, а второй студент взял не «хороший» билет.

21. Из перетасованной колоды в 36 карт наудачу вынута одна карта. Является ли событие A (появился король) независимым от события B (вынута карта черной масти)?

22. Вероятность сдачи студентом первого экзамена равна 0,9, второго экзамена — 0,9 и третьего экзамена — 0,8. Найти вероятность того, что студент сдаст: а) только первый экзамен, б) все три экзамена, в) хотя бы один экзамен, г) только один экзамен, д) по крайней мере, два экзамена.

23. Из колоды в 36 карт наудачу вынимается одна. Найти вероятность того, что будет вынута дама или бубна.

24. Группа из 25 студентов писала 2 контрольные работы по математике. Первую контрольную работу на «отлично» написали 8 студентов, а вторую контрольную работу на «отлично» написали 6 студентов. По списку группы наудачу выбирается один студент. Найти вероятность того, что выбранный студент написал на «отлично» хотя бы одну контрольную работу.

25. В группе 25 студентов, из которых 4 студента не сдали зачет по физике, 3 студента не сдали зачет по английскому языку и два студента не сдали оба зачета. По списку студентов группы наудачу выбирается один студент. Найти вероятность того, что выбранный студент не сдал хотя бы один зачет.

26. Бросается игральная кость. Найти вероятность того, что выпадет пятерка или нечетное число.

27. Бросаются две игральные кости. Найти вероятность того, что выпадут две пятерки или что сумма выпавших очков делится на пять.

28. Бросаются две игральные кости. Найти вероятность того, что выпадут две единицы или что сумма выпавших очков четная.

29. Теннисные мячи выпускаются двумя фабриками спортивного инвентаря. Известно, что объем продукции второй фабрики в 3 раза превосходит объем продукции первой, а доля брака у первой фабрики 0,2%, у второй — 1%. Из ящика, в котором вперемешку лежат мячи, изготовленные двумя фабриками за одинаковый период времени, наугад вынимается один мяч. Найти вероятность того, что: а) выбранный мяч является бракованным, б) выбранный мяч изготовленный первой фабрикой, если он оказался бракованным.

30. Из 50 конфет 18 изготовлены кондитерской фабрикой № 1, 20 — фабрикой № 2, остальные — фабрикой №3. Фабрики № 1 и № 3 дают продукцию высшего сорта с вероятностью 0,9, а фабрика

№ 2 — с вероятностью 0,6. Найти вероятность того, что а) взятая наудачу конфета будет высшего сорта, б) взятая наудачу конфета изготовлена фабрикой № 1, если она оказалась высшего сорта.

31. Была проведена одна и та же контрольная по математике одновременно в трех параллельных классах А, Б, В. В классе А, где 30 учеников, 8 работ получили оценку «отлично», в классе Б, где 28 учеников, 6 работ получили «отлично» и в классе В, где 27 учеников, 9 работ получили «отлично». Из стопки работ, в которой находятся случайным образом перемешанные все работы учеников трех классов, наудачу берется одна работа. Найти вероятность того, что: а) взятая работа выполнена на «отлично», б) взятая работа принадлежит ученику класса А, если она выполнена на «отлично».

32. Из полного набора 28 костей домино наугад одна за другой берутся две кости. Найти вероятность того, что вторую кость можно приставить к первой.

33. Вероятности рождения мальчика и девочки одинаковы и равны 0,5. Вероятность того, что в семье один ребенок, равна 0,7, вероятность того, что в семье два ребенка, равна 0,6 и вероятность того, что в семье три ребенка, равна 0,3. Наугад выбирается семья. Найти вероятность того, что: а) в семье есть мальчики и нет девочек, б) в семье только один ребенок при условии, что в семье нет девочек.

34. В библиотеке имеются книги только по психологии и математике. Вероятность того, что читатель возьмет книгу по психологии, равна 0,8, а вероятность того, что читатель возьмет книгу по математике, равна 0,2. Определить вероятность того, что 5 читателей подряд возьмут четыре книги по психологии и лишь одну книгу по математике, если каждый читатель берет ровно одну книгу.

35. В семье шесть детей. Считая вероятности рождения мальчика и девочки равными 0,5, найти вероятность того, что в данной семье: а) три мальчика, б) мальчиков не менее двух, но и не более четырех.

36. Монета бросается пять раз. Найти вероятность того, что выпали три герба.

37. Вероятность выигрыша на каждый из лотерейных билетов равна 0,02. Найти вероятность хотя бы одного выигрыша для: а) трех билетов, б) пяти билетов.

38. В партии арбузов 90% спелых, остальные недоспелые. Наугад отобраны три арбуза. Найти вероятность того, что среди них: а) не менее двух спелых, б) хотя бы один неспелый.

39. Два равносильных шахматиста играют в шахматы. Что вероятнее: выиграть три партии из шести или две партии из четырех (ничьих не бывает)?

ГЛАВА 2

СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

§ 1. Понятие случайной величины. Функции распределения случайных величин

Примеры случайных величин уже ранее встречались. Например, случайной величиной является число выпавших очков при бросании игральной кости. Бросание монеты тоже дает пример случайной величины, если, скажем, выпадению герба приписать значение «единица», а выпадению решки — значение «ноль». В первом примере случайная величина принимает одно из шести возможных значений в зависимости от случая, а во втором примере в зависимости от случая случайная величина принимает одно из двух возможных значений. Заранее предсказать числовой результат опыта с бросанием игральной кости или монеты невозможно.

Под *случайной величиной*, связанной с некоторым опытом, понимается величина, числовые значения которой при повторении опыта предсказать невозможно. Если пространство Ω элементарных событий является дискретным, то понятие случайной величины можно уточнить. В этом случае случайная величина — это функция, определенная на множестве Ω, со значениями из множества вещественных чисел $\mathbb{R}$. Таким образом, при проведении опыта случайная величина принимает одно из возможных числовых значений, но заранее не известно какое. Множество всех возможных значений случайной величины может быть хорошо известно, но конкретное значение случайной величины нельзя предугадать до опыта.

Случайная величина будет обозначаться одной из букв греческого алфавита — ξ, η, ζ и т. д. или буквой латинского алфавита — X, Y, Z и т. д.

Также примерами случайных величин могут служить:

а) число новорожденных в некотором городе в течение суток,

б) количество дорожно-транспортных происшествий в некотором городе в течение определенной недели,

в) число телефонных звонков, поступивших в определенную квартиру в течение дня,

г) рост или вес случайно выбранного человека из группы людей.

Для характеристики поведения случайной величины используется так называемая функция распределения вероятностей.

Функцией распределения вероятностей $F_\xi(x)$ случайной величины ξ называют функцию, которая ставит в соответствие любому заданному значению x величину вероятности события $\{\xi < x\}$, т. е.

$$F_\xi(x) = \mathsf{P}\{\xi < x\}, \quad x \in (-\infty, +\infty).$$

В дальнейшем вместо $F_\xi(x)$ иногда будем писать $F(x)$ и называть ее просто «функцией распределения».

Из определения $F(x)$ непосредственно вытекают следующие ее свойства:

1) $0 \leqslant F(x) \leqslant 1$ для любого x,
2) $F(x)$ — неубывающая функция x,
3) $\lim\limits_{x\to-\infty} F(x) = 0$, $\lim\limits_{x\to+\infty} F(x) = 1$,
4) $F(x)$ — функция, непрерывная слева в каждой точке x,
5) $\mathsf{P}\{a < \xi < b\} = F(b) - F(a)$ для любых значений a и b.

Наиболее распространенные на практике случайные величины подразделяются на два типа: дискретные и непрерывные.

Случайная величина ξ называется *дискретной случайной величиной*, если множество ее значений является конечным или счетным. В последнем случае множество значений не должно содержать предельных точек. Примерами дискретных случайных величин могут быть: число отличников в случайно выбранной группе учащихся, число браков в городе в определенном месяце, число баллов при тестировании ребенка и т. д.

Для полной характеристики дискретной случайной величины ξ необходимо знать множество всех ее значений $x_1, x_2, \ldots, x_n, \ldots$ и вероятности всех этих значений $p_1 = \mathsf{P}\{\xi = x_1\}$, $p_2 = \mathsf{P}\{\xi = x_2\}$, $\ldots$, $p_n = \mathsf{P}\{\xi = x_n\}$, $\ldots$, причем все $p_i > 0$ и $\sum\limits_{i=1}^{\infty} p_i = 1$. Если это так, то говорят, что задан закон распределения (вероятностей) дискретной случайной величины ξ. Этот закон удобно записывать в виде следующей таблицы:

x_1	x_2	$\ldots$	x_n	$\ldots$
p_1	p_2	$\ldots$	p_n	$\ldots$

В первой строке таблицы записываются все значения случайной величины, а во второй строке под ними — соответствующие вероятности.

Пример 1. Пусть случайная величина ξ — число выпавших очков при подбрасывании игральной кости. Найти закон распределения ξ.

Решение. Случайная величина ξ принимает значения $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$, $x_5 = 5$, $x_6 = 6$ с вероятностями $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$. Поэтому закон распределения ξ задается таблицей:

1	2	3	4	5	6
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

■

Пример 2. Найти закон распределения случайной величины ξ — числа появлений герба при двукратном бросании монеты.

Решение. Величина ξ принимает три значения: $x_1 = 0$, $x_2 = 1$, $x_3 = 2$. Вероятности этих значений находятся по формуле Бернулли: $p_1 = \frac{1}{4}$, $p_2 = \frac{1}{2}$, $p_3 = \frac{1}{4}$. Закон распределения ξ задается таблицей:

0	1	2
$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

■

Заметим, что областью определения дискретной случайной величины ξ является дискретное пространство Ω элементарных событий.

По закону распределения дискретной случайной величины ξ всегда можно найти функцию распределения $F(x)$ для ξ. Она имеет вид

$$F(x) = \sum_{x_i < x} \mathrm{P}\{\xi = x_i\},$$

где неравенство $x_i < x$ под знаком суммы означает суммирование по всем тем значениям x_i величины ξ, которые меньше x. График

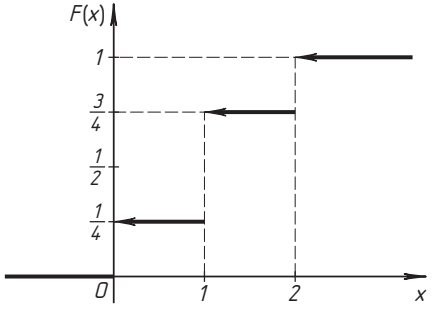


Рис. 2

функции распределения $F(x)$ дискретной величины ξ представляет собой разрывную ступенчатую линию. В интервалах между значениями ξ функция $F(x)$ постоянна, а в точках, соответствующих принимаемым значениям x_i величины ξ, $F(x)$ имеет скачок, равный вероятности этого значения. Для случайной величины ξ примера 2 график $F(x)$ имеет вид, указанный на рис. 2.

На рис. 2 острие стрелки указывает выколотую точку. Отметим, что по функции распределения дискретной случайной величины легко восстановить ее закон распределения.

Приведем два часто встречающиеся на практике законы распределения дискретных случайных величин.

Биномиальный закон распределения. В n испытаниях Бернулли число появлений интересующего нас события является случайной величиной ξ с биномиальным законом распределения. Возможными значениями ξ являются все целые неотрицательные числа от нуля до n. Вероятности этих значений задаются формулами Бернулли:

$$\mathsf{P}\{\xi = k\} = C_n^k p^k q^{n-k}, \quad k = 0, 1, \ldots, n,$$

где p — вероятность появления интересующего нас события в одном испытании, $q = 1 - p$.

Многократное бросание монеты или игральной кости дает примеры случайных величин с биномиальным законом распределения.

Числа n и p называются параметрами биномиального распределения.

Распределение Пуассона с параметром $\lambda > 0$. Дискретная случайная величина ξ имеет распределение Пуассона с параметром $\lambda > 0$, если возможными ее значениями являются все целые неотрицательные числа, а вероятности этих значений задаются формулами

$$\mathsf{P}\{\xi = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2 \ldots .$$

Имеются таблицы значений $\mathsf{P}\{\xi = k\} = P_k(\lambda)$ (см. табл. I).

Распределение Пуассона имеет, например, число заболевших студентов или число выполненных тестовых заданий за определенный период времени.

Заметим, что значения дискретной случайной величины ξ могут быть измерены в любой из шкал: количественной, порядковой или номинальной. Количественная шкала — это числовая шкала определенного физического смысла. Порядковая шкала упорядочивает

значения ξ по степени их важности, а номинальная шкала разбивает объекты на однородные по анализируемому свойству классы. Например, если для качества жилищных условий предусмотреть четыре возможные категории: «плохое», «удовлетворительное», «хорошее», «очень хорошее», то по этому свойству можно упорядочить обследование жилищных условий группы студентов. Примером измерения в номинальной шкале служит разбиение студентов группы на градации с точки зрения социальной принадлежности семьи или профессии главы семьи. Если нет оговорки, то в дальнейшем всегда значения дискретной случайной величины ξ предполагаются измеренными в количественной шкале.

Случайная величина ξ называется *непрерывной*, если существует такая функция $p(x) \geqslant 0$, что при любом $x \in \mathbb{R}$ функция распределения ξ имеет вид

$$F(x) = \int\limits_{-\infty}^{x} p(t)\, dt.$$

Функция $p(x)$ называется *плотностью распределения вероятностей* случайной величины ξ. Далее будем всегда предполагать, что $p(x)$ непрерывна для всех $x \in (-\infty, +\infty)$ за исключением, быть может, конечного числа точек разрыва первого рода. Функция $p(x)$ обладает следующими свойствами:

1) $\int\limits_{-\infty}^{+\infty} p(t)\, dt = 1$,

2) $F'(x) = p(x)$ в точках непрерывности $p(x)$.

Из определения непрерывной случайной величины ξ сразу получаем следующие важные свойства вероятностей ее значений:

1) вероятность каждого отдельного значения непрерывной случайной величины ξ равна нулю, т. е. $\mathsf{P}\{\xi = c\} = 0$ для каждого отдельного значения c;

2) $\mathsf{P}\{a \leqslant \xi < b\} = \mathsf{P}\{a \leqslant \xi \leqslant b\} = \mathsf{P}\{a < \xi \leqslant b\} = \mathsf{P}\{a < \xi < b\} = \int\limits_{a}^{b} p(t)\, dt$;

3) вероятность попадания значений ξ в различные промежутки числовой оси является положительной, если плотность $p(x) \neq 0$ на этих промежутках.

Множество значений непрерывной случайной величины не является счетным и сплошь заполняет некоторый промежуток числовой оси или всю числовую ось.

Таким образом, для непрерывной случайной величины ξ нет смысла рассматривать вероятность принятия отдельного значения c,

но имеет смысл рассматривать вероятность попадания ее значений в любой промежуток, пусть даже сколь угодно малый. График функции распределения $F(x)$ непрерывной случайной величины ξ представляет собой непрерывную линию в отличие от разрывной линии для $F(x)$ в случае дискретной ξ. Пример графика функции распределения приведен на рис. 3.

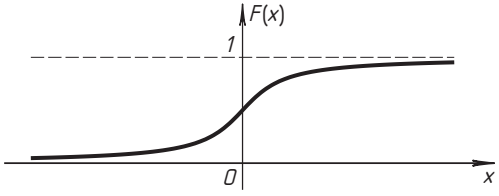


Рис. 3

Измерения непрерывных по своей физической природе величин (роста, веса, температуры, давления и т. п.) дают примеры непрерывных случайных величин.

Значения непрерывной случайной величины измеряются всегда количественной шкалой (см. § 2 главы 3).

Приведем наиболее известные типы распределений непрерывных случайных величин.

Нормальное распределение случайной величины является наиболее распространенным и наиболее важным. Оно используется для описания тех непрерывных случайных величин, значения которых формируются под воздействием суммы большого количества независимых случайных факторов, среди которых нет сильно выделяющихся. Распределение показателей, получаемых в эмпирических психодиагностических исследованиях при большом числе наблюдений, как правило, приближается к нормальному распределению. Говорят, что непрерывная случайная величина ξ имеет *нормальное распределение вероятностей с параметрами a и* σ^2 (краткое обозначение: $\xi \sim N(a, \sigma^2)$), если ее плотность распределения имеет вид:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad x \in (-\infty, +\infty).$$

Как будет установлено в следующем параграфе, параметры a и σ^2 нормального распределения имеют ясный вероятностный смысл.

Заметим, что при $x = a$ функция $\varphi(x)$ имеет максимум, равный $\frac{1}{\sqrt{2\pi}\cdot\sigma}$, и что график $\varphi(x)$ симметричен относительно прямой $x = a$.

Кроме того, $\varphi(x) > 0$ для всех x, $y = 0$ является горизонтальной асимптотой графика $\varphi(x)$, а $x = a \pm \sigma$ являются точками перегиба графика $\varphi(x)$. На основании сказанного можно построить график плотности $\varphi(x)$, который приведен на рис. 4.

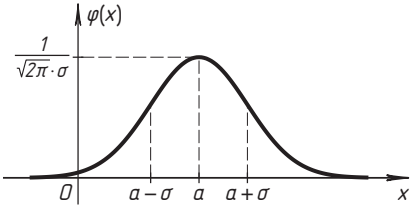


Рис. 4

График $\varphi(x)$ имеет колоколообразную форму, являющейся отличительной чертой нормального распределения. Иногда полученную кривую называют *нормальной кривой* или *кривой Гаусса*.

При изменении параметра a форма графика $\varphi(x)$ не изменяется, но график сдвигается влево или вправо (см. рис. 5, *а*). При изменении параметра σ^2 изменяется форма графика $\varphi(x)$. А именно: при увеличении σ^2 график $\varphi(x)$ сжимается к оси OX и растягивается (см. рис. 5, *б*), а при уменьшении σ^2 график $\varphi(x)$ стягивается к прямой $x = a$ (см. рис. 5, *в*). Таким образом, параметр σ^2 характеризует степень сжатия или растяжения графика плотности.

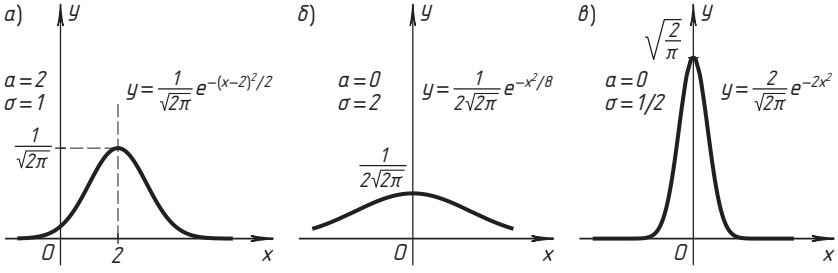


Рис. 5

В психологических исследованиях нормальное распределение используется в первую очередь при разработке тестов интеллекта и способностей. Например, при использовании шкалы Стэнфорд—Бине умственного развития рассматривался нормальный закон общего интеллекта с параметрами $a = 100$ и $\sigma = 16$. Случайные ошибки измерения физических величин также имеют нормальное распределение.

Вся площадь между осью Ox и графиком плотности $\varphi(x)$ равна единице. Площадь криволинейной трапеции, ограниченной на отрезке $[\alpha, \beta]$ графиком $\varphi(x)$, задает вероятность попадания значений нормально распределенной случайной величины ξ на отрезок $[\alpha, \beta]$:

$$\mathsf{P}\{\alpha \leqslant \xi \leqslant \beta\} = \int_{\alpha}^{\beta} \varphi(x)\, dx.$$

Известно, что $\mathsf{P}\{\alpha - \sigma < \xi < \alpha + \sigma\} \approx 0{,}68$; $\mathsf{P}\{\alpha - 2\sigma < \xi < \alpha + 2\sigma\} \approx 0{,}95$ и $\mathsf{P}\{\alpha - 3\sigma < \xi < \alpha + 3\sigma\} \approx 0{,}99$.

Отсюда следует так называемый «закон 3σ»: для нормально распределенной случайной величины ξ вероятность попадания ее значений в интервал $(\alpha - 3\sigma, \alpha + 3\sigma)$ приближенно равна 0,99.

Особую роль играет нормальное распределение с параметрами $\alpha = 0$, $\sigma = 1$, т. е. распределение $N(0, 1)$, которое называется *стандартным нормальным распределением*. Имеются таблицы значений плотности случайной величины $\xi \sim N(0, 1)$:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

(см. табл. II).

Для вычисления вероятности попадания значений $\xi \sim N(0, 1)$ в заданный интервал (α, β) используется функция

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{x^2}{2}}\, dx,$$

для значений которой имеются таблицы (см. табл. III). $\Phi(x)$ — нечетная функция, и поэтому в таблицах приведены ее значения лишь для некоторых $x > 0$. Если $\xi \sim N(0, 1)$, то вероятность $\mathsf{P}\{a < \xi < \beta\} = \Phi(\beta) - \Phi(\alpha)$.

Заметим, что если $\xi \sim N(\alpha, \sigma^2)$, то $\eta = \frac{\xi - \alpha}{\sigma} \sim N(0, 1)$. Это позволяет вычислять вероятности попадания значений величины $\xi \sim N(\alpha, \sigma^2)$ в интервал (α, β) с помощью таблиц величины $\xi \sim N(0, 1)$, поскольку вероятность $\mathsf{P}\{\alpha < \xi < \beta\} = \mathsf{P}\left\{\frac{\alpha - a}{\beta} < \frac{\xi - a}{\sigma} < \frac{\beta - a}{\sigma}\right\} = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right)$.

Пример. Пусть $\xi \sim N(6, 9)$. Найти вероятность того, что значения ξ попадут в интервал (5, 14).

Решение. Имеем

$$\mathrm{P}\{5<\xi<14\}=\Phi\left(\frac{14-6}{3}\right)-\Phi\left(\frac{5-6}{3}\right)=\Phi\left(\frac{8}{3}\right)-\Phi\left(-\frac{1}{3}\right)=\Phi\left(\frac{8}{3}\right)+\Phi\left(\frac{1}{3}\right)\approx$$

$$\approx \Phi(2{,}7)+\Phi(0{,}3)=0{,}4965+0{,}1179=0{,}6144. \blacksquare$$

Следует отметить, что теоретические исследования, относящиеся к нормальному закону распределения случайной величины, являются наиболее полными по сравнению с исследованиями других законов распределений. Это делает закон нормального распределения весьма удобным в применении, что будет видно в дальнейшем. Наряду с нормально распределенными случайными величинами на практике используется также логарифмически-нормально распределенные случайные величины. Случайная величина ξ называется *логарифмически-нормально распределенной*, если $\ln \xi$ подчиняется нормальному закону распределения.

Например, такими случайными величинами будут заработная плата работника, доход семьи, долговечность жизни изделия.

Рассмотрим несколько других типов распределений.

Равномерное распределение на отрезке $[a, b]$ задается плотностью распределения

$$p(x)=\begin{cases}\frac{1}{b-a}, & x\in[a, b],\\ 0, & x\notin[a, b].\end{cases}$$

Числа a и b называют параметрами равномерного распределения. Соответственно функция $F(x)$ равномерного распределения имеет вид

$$F(x)=\begin{cases}0, & x\leqslant a,\\ \frac{x-a}{b-a}, & x\in[a, b],\\ 1, & x\geqslant b.\end{cases}$$

Графики функций $p(x)$ и $F(x)$ приведены на рис. 6.

Например, случайные ошибки округления при проведении числовых расчетов являются равномерно распределенными.

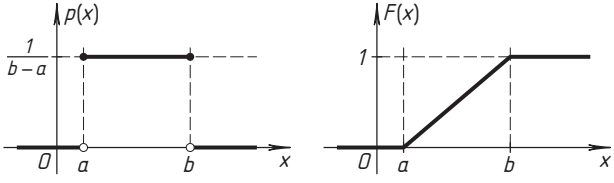


Рис. 6

Показательное распределение с параметром $\lambda > 0$ имеет плотность распределения

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0, \\ 0, & x < 0. \end{cases}$$

Например, показательное распределение имеют интервалы времени между выполнением двух разных вопросов теста, между вызовами скорой помощи, между обращениями клиентов и т. д.

Распределение Парето с параметрами $\alpha > 0$ *и* $c_0 \neq 0$ имеет функцию распределения

$$F(x) = \mathrm{P}\{\xi < x\} = 1 - \left(\frac{c_0}{x}\right)^{\alpha}, \quad x \in (c_0, +\infty).$$

Плотность $p(x)$ распределения Парето задает кривую типа гиперболы, выходящей из точки $\left(c_0, \frac{\alpha}{c_0}\right)$.

Этому распределению, например, подчиняется годовой доход тех людей, для которых он превосходит некоторый порог $c_0 > 0$.

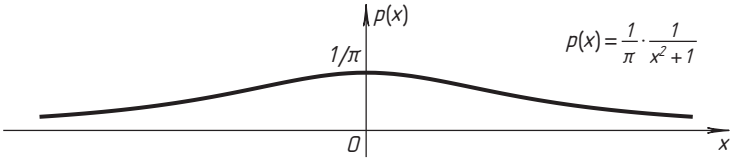


Рис. 7

Распределение Коши с параметрами $a > 0$ *и* b имеет плотность распределения

$$p(x) = \frac{1}{\pi} \cdot \frac{a}{a^2 + (x-b)^2}, \quad x \in (-\infty, +\infty).$$

Например, при $a = 1$, $b = 0$ график $p(x)$ имеет вид, указанный на рис. 7.

§ 2. Многомерные случайные величины. Функции случайных величин

В § 1 были приведены примеры случайных опытов, результат каждого из которых описывается одним числом. Однако нетрудно привести примеры случайных опытов, когда каждый результат опыта одновременно описывается не одним числом, а парой чисел или, в общем случае, набором из нескольких чисел.

Примеры. 1) При одновременном подбрасывании двух игральных костей результат записывается парой двух чисел, выпавших на этих костях. Если, скажем, выпали две шестерки, то результат опыта имеет вид (6, 6).

2) При случайном выборе человека из группы могут одновременно измеряться рост, вес и температура человека. Тогда результат такого случайного опыта записывается в виде системы трех чисел.

3) При случайном выборе семьи могут одновременно регистрироваться количество детей, среднедушевой доход, расходы на питание, расходы на промышленные товары, количество комнат в квартире. Тогда результат такого случайного эксперимента записывается в виде системы пяти чисел.

Описанные случайные опыты дают примеры многомерных случайных величин (или случайных векторов) в отличие от случайных величин из §1, которые можно назвать одномерными. Таким образом, если пространство Ω элементарных событий является дискретным и если каждому $\omega \in \Omega$ поставлено в соответствие $r \geqslant 2$ чисел, то говорят, что задана *r-мерная случайная величина* $(\xi_1, \xi_2, \ldots, \xi_r)$. Аналогичным образом определяется r-мерная случайная величина в случае непрерывного пространства Ω.

В дальнейшем для простоты изложения ограничимся рассмотрением двумерной случайной величины (ξ_1, ξ_2). Для нее, как и для любой случайной величины, заранее предсказать результат опыта невозможно, даже если известно множество значений этой случайной величины.

Для двумерной случайной величины (ξ_1, ξ_2) вводится следующее понятие функции распределения.

Функцией распределения (или двумерной функцией распределения) называется функция двух переменных $F(x_1, x_2) = \mathsf{P}\{\xi_1 < x_1, \xi_2 < x_2\}$, где $\mathsf{P}\{\xi_1 < x_1, \xi_2 < x_2\}$ означает вероятность события $\{\xi_1 < x_1, \xi_2 < x_2\}$ при любых значениях x_1 и x_2.

Как и одномерные случайные величины ξ, двумерные случайные величины (ξ_1, ξ_2) подразделяются, в основном, на дискретные и непрерывные.

Двумерная случайная величина (ξ_1, ξ_2) называется *дискретной*, если множество ее значений конечно или счетно.

Если же существует такая непрерывная функция $p(x_1, x_2) \geqslant 0$, называемая *плотностью* распределения, что для любого прямоугольника $\Pi = \{|x_1| \leqslant a, |x_2| \leqslant b\}$ вероятность $\mathsf{P}\{(\xi_1, \xi_2) \in \Pi\} = \iint\limits_{\Pi} p(x_1, x_2) \times$

$\times dx_1\, dx_2$, то двумерная случайная величина (ξ_1, ξ_2) называется *непрерывной*.

Если множество значений дискретной случайной величины (ξ_1, ξ_2) имеет вид $\{(x_i, y_j),\ i=1,\ 2,\ 3,\ \ldots,\ j=1,\ 2,\ 3,\ \ldots\}$, и вероятность $\mathrm{P}\{\xi_1 = x_i,\ \xi_2 = y_j\} = p_{ij}$, то $\sum\limits_{i,j} p_{ij} = 1$. В случае же непрерывной величины (ξ_1, ξ_2) справедливы следующие свойства плотности $p(x_1, x_2)$:

$$\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} p(x_1, x_2)\, dx_1\, dx_2 = 1, \qquad \frac{\partial^2 F(x_1, x_2)}{\partial x_1\, \partial x_2} = p(x_1, x_2),$$

$$F(x_1, x_2) = \int\limits_{-\infty}^{x_1}\int\limits_{-\infty}^{x_2} p(t_1, t_2)\, dt_1\, dt_2.$$

Закон распределения дискретной двумерной случайной величины (ξ_1, ξ_2) является заданным, если заданы все ее значения и вероятность каждого значения.

По заданному закону распределения дискретной величины (ξ_1, ξ_2) всегда можно однозначно найти законы распределения каждой из величин ξ_1 и ξ_2 в отдельности. Обратное в общем случае сделать нельзя.

Пример 1. Пусть закон распределения дискретной двумерной случайной величины (ξ_1, ξ_2) задан следующей таблицей

ξ_1 \ ξ_2	-1	0	1
2	0,05	0,3	0,35
3	0,15	0,12	0,03

Найти законы распределения каждой из величин ξ_1 и ξ_2 в отдельности.

Решение. Вероятность $\mathrm{P}\{\xi_1 = 2\} = 0{,}05 + 0{,}3 + 0{,}35 = 0{,}7$, а вероятность $\mathrm{P}\{\xi_1 = 3\} = 0{,}15 + 0{,}12 + 0{,}03 = 0{,}3$. Аналогичным образом находим, что $\mathrm{P}\{\xi_2 = -1\} = 0{,}2$, $\mathrm{P}\{\xi_2 = 0\} = 0{,}42$, $\mathrm{P}\{\xi_2 = 1\} = 0{,}38$.

Следовательно законы распределения ξ_1 и ξ_2 задаются соответственно следующими таблицами:

ξ_1	2	3
P	0,7	0,3

ξ_2	-1	0	1
P	0,2	0,42	0,38

■

Две дискретные случайные величины ξ_1 и ξ_2 называются *независимыми*, если для любых значений x_i величины ξ и любых значений y_j величины ξ_2 вероятность $p_{ij} = \mathrm{P}\{\xi_1 = x_i,\ \xi_2 = y_j\} = \mathrm{P}\{\xi_1 = x_i\} \cdot \mathrm{P}\{\xi_2 = y_j\}$.

Случайные величины ξ_1 и ξ_2 из примера 1 не являются независимыми, так как например, вероятность $\mathsf{P}\{\xi_1 = 2, \xi_2 = -1\} = 0{,}05 \neq$ $\neq \mathsf{P}\{\xi_1 = 2\} \cdot \mathsf{P}\{\xi_2 = -1\} = 0{,}14$.

Пример 2. Закон распределения двумерной случайной величины (ξ_1, ξ_2) задан следующей таблицей

ξ_1 \ ξ_2	0	1
0	1/4	1/4
6	1/4	1/4

Доказать, что величины ξ_1 и ξ_2 являются независимыми.

Решение. Законы распределения ξ_1 и ξ_2 в отдельности задаются соответственно следующими таблицами:

ξ_1	0	1
P	1/2	1/2

ξ_2	0	1
P	1/2	1/2

Тогда определение независимости ξ_1 и ξ_2 выполняется, что легко проверить. Например, $\mathsf{P}\{(\xi_1 = 0, \xi_2 = 0\} = \mathsf{P}\{\xi_1 = 0\} \cdot \mathsf{P}\{\xi_2 = 0\} = \frac{1}{4}$. Аналогично проверяются и остальные равенства. ■

Зная плотность $p(x_1, x_2)$ распределения непрерывной двумерной случайной величины (ξ_1, ξ_2), можно однозначно установить плотность распределения каждой из величин ξ_1 и ξ_2 в отдельности. Например, для ξ_1 плотность распределения $p(x_1) = \int\limits_{-\infty}^{\infty} p(x_1, x_2)\, dx_2$.

Среди непрерывных двумерных случайных величин наиболее важными являются величины, распределенные по нормальному закону.

Пусть $p(x_1)$ — плотность распределения непрерывной случайной величины ξ_1 и $p(x_2)$ — плотность распределения непрерывной случайной величины ξ_2. Тогда случайные величины ξ_1 и ξ_2 называются *независимыми*, если плотность распределения двумерной случайной величины $p(x_1, x_2) = p(x_1) \cdot p(x_2)$.

Таким образом, если известны одномерные распределения независимых случайных величин ξ_1 и ξ_2, то всегда можно найти совместное распределение этих величин. Например, если ξ_1 и ξ_2 — независимые случайные величины, причем $\xi_1 \sim N(a_1, \sigma_1^2)$, $\xi_2 \sim N(a_2, \sigma_2^2)$, то плотность распределения двумерной (ξ_1, ξ_2) имеет вид:

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1 \cdot \sigma_2} \cdot e^{-\frac{(x_1 - a_1)^2}{2\sigma_1^2} - \frac{(x_2 - a_2)^2}{2\sigma_2^2}}.$$

Рассмотрим теперь понятие функции от случайных величин. Пусть сначала ξ — одномерная случайная величина и пусть функция $f(x)$ — заданная непрерывная функция на множестве всех вещественных чисел. Тогда $\eta = f(\xi)$ является случайной величиной. Зная закон распределения величины ξ, можно найти закон распределения новой случайной величины $\eta = f(\xi)$. Рассмотрим примеры.

Пример 3. Пусть закон распределения задан следующей таблицей:

ξ	-1	0	1
P	1/3	1/3	1/3

Найти законы распределения случайных величин $\eta_1 = 2\xi + 3$ и $\eta_2 = \xi^2$.

Решение. Закон распределения η_1 задается таблицей

η_1	1	3	5
P	1/3	1/3	1/3

Закон распределения величины η_2 имеет следующий вид:

η_2	0	1
P	1/3	2/3

■

Пример 4. Доказать, что если $\xi \sim N(a, \sigma^2)$ и $\eta = \alpha\xi + \beta$, где $\alpha > 0$ и β — заданные числа, то $\eta \sim N(\alpha a + \beta, \alpha^2\sigma^2)$.

Решение. Функция распределения случайной величины η имеет вид $F_\eta(x) = \mathrm{P}\{\eta < x\} = \mathrm{P}\{\alpha\xi + \beta < x\} = \mathrm{P}\left\{\xi < \dfrac{x-\beta}{\alpha}\right\} = F_\xi\left(\dfrac{x-\beta}{\alpha}\right)$, где $F_\xi(x)$ — функция распределения величины ξ. Если $p_\eta(x)$ — плотность распределения величины η, то отсюда

$$p_\eta(x) = F'_\eta(x) = \frac{1}{\alpha} F'_\xi\left(\frac{x-\beta}{\alpha}\right) = \frac{1}{\alpha} p_\xi\left(\frac{x-\beta}{\alpha}\right) =$$

$$= \frac{1}{\alpha} \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot e^{\frac{-\left(\frac{x-\beta}{\alpha}-a\right)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\cdot\sigma_1} e^{\frac{-(x-a_1)^2}{2\sigma_1^2}},$$

где $a_1 = \alpha a + \beta$, $\sigma_1 = \alpha\sigma$.

Следовательно, $\eta \sim N(a_1, \sigma_1^2)$. ■

Если (ξ_1, ξ_2) — двумерная случайная величина и $f(x_1, x_2)$ — непрерывная функция двух переменных x_1 и x_2, заданная на всей

плоскости с декартовыми прямоугольными координатами (x_1, x_2), то $\eta = f(\xi_1, \xi_2)$ также является случайной величиной. Например, можно говорить о случайных величинах $\xi_1 \pm \xi_2$, $\xi_1 \cdot \xi_2$, $\frac{\xi_1}{\xi_2}$, $\xi_1^2 + \xi_2^2$ и т. д. Если известен закон распределения двумерной величины (ξ_1, ξ_2), то можно находить закон распределения величины $\eta = f(\xi_1, \xi_2)$. Например, если ξ_1 и ξ_2 — независимые непрерывные случайные величины, причем $p_1(x)$ плотность распределения ξ_1 и $p_2(x)$ — плотность распределения ξ_2, то плотность $p_\eta(x)$ распределения величины $\eta = \xi_1 + \xi_2$ находится по формуле

$$p_\eta(x) = \int\limits_{-\infty}^{\infty} p_1(t)\, p_2(x-t)\, dt.$$

Рассмотрим еще пример.

Пример 5. Заданы законы распределения независимых случайных величин ξ_1 и ξ_2:

ξ_1	0	1
P	1/2	1/2

ξ_2	0	1
P	1/2	1/2

Найти законы распределения величин $\xi_1 + \xi_2$ и $\xi_1 \cdot \xi_2$.

Решение. Величина $\xi_1 + \xi_2$ принимает значения 0, 1, 2, а величина $\xi_1 \cdot \xi_2$ принимает значения 0, 1. Найдем вероятности этих значений. Поскольку ξ_1 и ξ_2 — независимы, то:

$$\mathrm{P}\{\xi_1 + \xi_2 = 0\} = \mathrm{P}\{\xi_1 = 0,\, \xi_2 = 0\} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},$$

$$\mathrm{P}\{\xi_1 + \xi_2 = 2\} = \mathrm{P}\{\xi_1 = 1,\, \xi_2 = 1\} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},$$

$$\mathrm{P}\{\xi_1 + \xi_2 = 1\} = \mathrm{P}\{\xi_1 = 0,\, \xi_2 = 1\} + \mathrm{P}\{\xi_1 = 1,\, \xi_2 = 0\} = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2},$$

$$\mathrm{P}\{\xi_1 \cdot \xi_2 = 0\} = \mathrm{P}\{\xi_1 = 0,\, \xi_2 = 0\} + \mathrm{P}\{\xi_1 = 0,\, \xi_2 = 1\} + \mathrm{P}\{\xi_1 = 1,\, \xi_2 = 0\} = 3 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4},$$

$$\mathrm{P}\{\xi_1 \cdot \xi_2 = 1\} = \mathrm{P}\{\xi_1 = 1,\, \xi_2 = 1\} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Получаем следующие законы распределения $\xi_1 + \xi_2$ и $\xi_1 \cdot \xi_2$:

$\xi_1 + \xi_2$	0	1	2
P	1/4	1/2	1/4

$\xi_1 \cdot \xi_2$	0	1
P	3/4	1/4

■

Особую роль играют некоторые специальные функции от нескольких независимых и одинаково нормально распределенных случайных величин. Дело в том, что их распределения широко применяются при анализе статистических данных, о чем будет идти речь в дальнейшем. Наиболее важными среди этих распределений являются: χ^2-распределение («хи-квадрат»-распределение), t-распределение (или распределение Стьюдента) и F-распределение (или распределение Фишера).

Пусть случайные величины $\xi_1, \xi_2, \ldots, \xi_n$ являются независимыми и каждая из них имеет стандартное нормальное распределение $N(0, 1)$. Тогда случайная величина $\chi^2(n) = \xi_1^2 + \xi_2^2 + \ldots + \xi_n^2$ называется *случайной величиной «хи-квадрат» с n степенями свободы*, а ее распределение называется *«хи-квадрат»-распределением с n степенями свободы*.

Для случайной величины $\chi^2(n)$ известна плотность распределения и составлены разнообразные таблицы. Ясно, что $\chi^2(n)$ для любого значения параметра n принимает лишь неотрицательные значения (см. табл. VI).

Пусть случайные величины $\xi_0, \xi_1, \xi_2, \ldots, \xi_n$ — независимы и каждая из них имеет стандартное нормальное распределение $N(0, 1)$. Случайная величина

$$t(n) = \frac{\xi_0}{\sqrt{\frac{1}{n}(\xi_1^2 + \xi_2^2 + \ldots + \xi_n^2)}}$$

называется *t-величиной с n степенями свободы* (или случайной величиной с *распределением Стьюдента*).

Для t-величины известна плотность распределения $p_n(x)$. Она является симметричной функцией относительно прямой $x=0$ при каждом значении параметра n. Имеются таблицы для случайных величин $t(n)$ (см. табл. VII).

Наконец, пусть $\xi_1, \xi_2, \ldots, \xi_m; \eta_1, \eta_2, \ldots, \eta_n$ — независимые случайные величины, каждая из которых распределена по стандартному нормальному закону $N(0, 1)$. Тогда случайная величина

$$F(m, n) = \frac{\frac{1}{m}(\xi_1^2 + \xi_2^2 + \ldots + \xi_m^2)}{\frac{1}{n}(\eta_1^2 + \eta_2^2 + \ldots + \eta_n^2)}$$

называется *F(m, n)-величиной с m и n степенями свободы*, а ее распределение называется *F-распределением с m и n степенями свободы*

(или *распределением Фишера*). Ясно, что $F(m, n)$-величина принимает лишь неотрицательные значения при любых значениях параметров m и n.

Известна плотность F-распределения и имеются таблицы для $F(m, n)$-величины (см. табл. VIII).

§ 3. Числовые характеристики случайных величин

Функция распределения случайной величины ξ дает исчерпывающие сведения о законе распределения ξ. Однако на практике для случайной величины ξ зачастую достаточно задать лишь несколько числовых характеристик ее распределения. Они в сжатой форме должны выражать наиболее важные особенности распределения ξ, такие, например, как центр группирования значений ξ, разброс значений ξ вокруг центра группирования и т. п.

К основным числовым характеристикам случайной величины ξ относятся: математическое ожидание, дисперсия, среднее квадратическое отклонение (или стандартное отклонение), моменты, центральные моменты, квантили.

Сразу отметим, что существуют случайные величины, для которых может не быть некоторых из числовых характеристик.

1. Математическое ожидание

В качестве основной числовой характеристики центра группирования значений случайной величины ξ используется математическое ожидание ξ (или среднее значение ξ). Его обозначают через $\mathsf{M}\xi$ или $\mathsf{E}\xi$.

Пусть сначала ξ — дискретная случайная величина со значениями $x_1, x_2, \ldots, x_n, \ldots$, имеющих соответственно вероятности $p_1, p_2, \ldots, p_n, \ldots$. Тогда

$$\mathsf{M}\xi = \sum_n x_n \cdot p_n.$$

Если число возможных значений ξ конечно, то $\mathsf{M}\xi$ всегда существует. Если же число возможных значений ξ счетно, то для существования $\mathsf{M}\xi$ необходимо требовать, чтобы этот ряд сходился абсолютно.

Пример 1. Найти математическое ожидание случайной величины ξ — числа выпавших очков при бросании игральной кости.

Решение. Случайная величина ξ — дискретная, принимающая значения 1, 2, 3, 4, 5, 6. Вероятность каждого значения равна $\frac{1}{6}$. Тогда $\mathsf{M}\xi = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = \frac{7}{2}$. ■

Пример 2. Найти математическое ожидание случайной величины ξ — суммы выпавших очков при бросании двух игральных костей.

Решение. Случайная величина ξ — дискретная, принимающая значения 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. Найдем вероятности этих значений. Имеем: $\mathsf{P}\{\xi=2\} = \mathsf{P}\{\xi=12\} = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}$, $\mathsf{P}\{\xi=3\} = \mathsf{P}\{\xi=11\} = 2\cdot\frac{1}{6}\cdot\frac{1}{6} = \frac{2}{36}$, $\mathsf{P}\{\xi=4\} = \mathsf{P}\{\xi=10\} = 3\cdot\frac{1}{6}\cdot\frac{1}{6} = \frac{3}{36}$, $\mathsf{P}\{\xi=5\} = \mathsf{P}\{\xi=9\} = 4\cdot\frac{1}{6}\cdot\frac{1}{6} = \frac{4}{36}$, $\mathsf{P}\{\xi=6\} = \mathsf{P}\{\xi=8\} = 5\cdot\frac{1}{6}\cdot\frac{1}{6} = \frac{5}{36}$, $\mathsf{P}\{\xi=7\} = \frac{6}{36}$.

Тогда $\mathsf{M}\xi = 2\cdot\frac{1}{36} + 3\cdot\frac{2}{36} + 4\cdot\frac{3}{36} + 5\cdot\frac{4}{36} + 6\cdot\frac{5}{36} + 7\cdot\frac{6}{36} + 8\cdot\frac{5}{36} + 9\cdot\frac{4}{36} + 10\cdot\frac{3}{36} + 11\cdot\frac{2}{36} + 12\cdot\frac{1}{36} = 7$. ■

Можно доказать, что при биномальном распределении дискретной случайной величины ξ с параметрами n и p математическое ожидание $\mathsf{M}\xi = np$, а при распределении Пуассона с параметром λ дискретная случайная величина ξ имеет $\mathsf{M}\xi = \lambda$.

Пусть теперь ξ — непрерывная случайная величина с плотностью распределения $p(x)$. Тогда

$$\mathsf{M}\xi = \int\limits_{-\infty}^{+\infty} x\, p(x)\, dx,$$

если интеграл абсолютно сходится.

Пример 3. Найти $\mathsf{M}\xi$, если ξ — непрерывная случайная величина, равномерно распределенная на отрезке $[a, b]$.

Решение. Как известно, для равномерно распределенной ξ на $[a, b]$ плотность распределения

$$p(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & x \notin [a, b]. \end{cases}$$

Тогда $\mathsf{M}\xi = \int\limits_{-\infty}^{+\infty} xp(x)\, dx = \int\limits_{a}^{b} \frac{x}{b-a}\, dx = \frac{1}{b-a}\cdot\frac{x^2}{2}\Big|_a^b = \frac{1}{b-a}\cdot\frac{b^2-a^2}{2} = \frac{1}{2}(a+b)$. ■

Можно доказать, что для нормально распределенной случайной величины ξ с параметрами a и σ^2, т. е. для $\xi \sim N(a, \sigma^2)$, математическое ожидание $\mathsf{M}\xi = a$ и что $\mathsf{M}\xi = \lambda$ для случайной величины ξ, имеющей показательное распределение с параметром $\lambda > 0$.

Перечислим основные свойства математического ожидания.

1) Математическое ожидание постоянной равно этой постоянной, т. е. $\mathsf{M}c = c$, если c — постоянная.

2) Постоянный множитель можно выносить за знак математического ожидания, т. е.

$$\mathsf{M}(c\xi) = c\mathsf{M}\xi,$$

если c — постоянная и ξ — случайная величина.

3) Математическое ожидание суммы случайных величин ξ_1 и ξ_2 равно сумме их математических ожиданий, т. е.

$$\mathsf{M}(\xi_1 + \xi_2) = \mathsf{M}\xi_1 + \mathsf{M}\xi_2.$$

4) Если случайные величины ξ_1 и ξ_2 независимы, то математическое ожидание произведения ξ_1 и ξ_2 равно произведению их математических ожиданий, т. е.

$$\mathsf{M}(\xi_1 \cdot \xi_2) = \mathsf{M}\xi_1 \cdot \mathsf{M}\xi_2.$$

Приведенные свойства помогают при вычислении математических ожиданий. Например, если случайная величина $\xi \sim N(a, \sigma^2)$ и случайная величина $\eta = \alpha\xi + \beta$, где α и β — заданные числа, то на основании приведенных свойств и того, что $\mathsf{M}\xi = a$, сразу можно записать, что $\mathsf{M}\eta = \alpha a + \beta$.

В заключение этого пункта отметим, что далеко не для всех случайных величин существуют математические ожидания. Например, случайная величина, распределенная по закону Коши (см. с. 45), не имеет математического ожидания.

2. Дисперсия, стандартное отклонение и коэффициент вариации

В качестве основной числовой характеристики степени рассеяния значений случайной величины ξ вокруг своего центра группирования $\mathsf{M}\xi$ используется дисперсия $\mathsf{D}\xi$ случайной величины ξ.

Пусть случайная величина ξ имеет конечное математическое ожидание $\mathsf{M}\xi$. *Дисперсией* $\mathsf{D}\xi$ случайной величины ξ называется математическое ожидание квадрата отклонения ξ от своего математического ожидания $\mathsf{M}\xi$, т. е.

$$\mathsf{D}\xi = \mathsf{M}(\xi - \mathsf{M}\xi)^2.$$

Преобразуем эту формулу, воспользовавшись свойствами математического ожидания. Имеем $\mathsf{M}(\xi - \mathsf{M}\xi)^2 = \mathsf{M}[\xi^2 - 2\xi \cdot \mathsf{M}\xi + (\mathsf{M}\xi)^2] = \mathsf{M}\xi^2 - 2\mathsf{M}\xi \cdot \mathsf{M}\xi + (\mathsf{M}\xi)^2 = \mathsf{M}\xi^2 - (\mathsf{M}\xi)^2$.

Таким образом, получена удобная для вычислений формула дисперсии:

$$\mathsf{D}\xi = \mathsf{M}\xi^2 - (\mathsf{M}\xi)^2.$$

Остается заметить, что для дискретной случайной величины ξ со значениями $x_1, x_2, \ldots, x_n, \ldots$ и соответствующими им вероятностями $p_1, p_2, \ldots, p_n, \ldots$ первое слагаемое $\mathsf{M}\xi^2 = \sum\limits_n x_n^2 \cdot p_n$, а для непрерывной случайной величины ξ с плотностью распределения $p(x)$ имеем $\mathsf{M}\xi^2 = \int\limits_{-\infty}^{\infty} x^2 p(x)\, dx$.

Разумеется здесь предполагается, что ряд, задающий $\mathsf{M}\xi^2$ для дискретной ξ, сходится и что сходится несобственный интеграл, задающий $\mathsf{M}\xi^2$ для непрерывной ξ.

Пример 4. Найти дисперсию случайной величины ξ — числа выпавших очков при бросании игральной кости.

Решение. В примере 1 было найдено, что $\mathsf{M}\xi = \frac{7}{2}$. Тогда $\mathsf{D}\xi = \mathsf{M}(\xi - \mathsf{M}\xi)^2 = \frac{1}{6}\left[\left(1 - \frac{7}{2}\right)^2 + \left(2 - \frac{7}{2}\right)^2 + \left(3 - \frac{7}{2}\right)^2 + \left(4 - \frac{7}{2}\right)^2 + \left(5 - \frac{7}{2}\right)^2 + \left(6 - \frac{7}{2}\right)^2\right] = \frac{1}{24}[25 + 9 + 1 + 1 + 9 + 25] = \frac{70}{24} = \frac{35}{12}$. ■

Пример 5. Найти дисперсию равномерно распределенной случайной величины ξ на $[a, b]$.

Решение. В примере 3 было найдено, что $\mathsf{M}\xi = \frac{a+b}{2}$. Так как $\mathsf{M}\xi^2 = \int\limits_{-\infty}^{\infty} x^2 p(x)\, dx = \frac{1}{b-a}\int\limits_a^b x^2\, dx = \frac{1}{b-a} \cdot \frac{x^3}{3}\Big|_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + ab + a^2}{3}$, то $\mathsf{D}\xi = \mathsf{M}\xi^2 - (\mathsf{M}\xi)^2 = \frac{b^2 + ab + a^2}{3} - \frac{1}{4}(a+b)^2 = \frac{(b-a)^2}{12}$. ■

Можно показать, что для случайной величины $\xi \sim N(a, \sigma^2)$ дисперсия $\mathsf{D}\xi = \sigma^2$.

Кроме того, для случайной величины ξ, распределенной по биномальному закону с параметрами n и p, $\mathsf{D}\xi = np(1-p)$, а для случайной величины ξ, распределенной по закону Пуассона с параметром λ, $\mathsf{D}\xi = \lambda$.

Заметим также, что распределенная по закону Коши случайная величина не имеет дисперсии.

Приведем свойства дисперсии случайной величины ξ.

1) Для каждой случайной величины ξ дисперсия $\mathsf{D}\xi \geqslant 0$.

2) Если c — постоянная, то ее дисперсия

$$\mathsf{D}c = 0.$$

3) Постоянный множитель c можно выносить за знак дисперсии, предварительно возведя его в квадрат, т. е.

$$\mathsf{D}(c\xi) = c^2\mathsf{D}\xi.$$

4) Дисперсия суммы двух независимых случайных величин ξ_1 и ξ_2 равна сумме дисперсий этих величин, т. е.

$$\mathsf{D}(\xi_1 + \xi_2) = \mathsf{D}\xi_1 + \mathsf{D}\xi_2.$$

Эти свойства в ряде случаев упрощают вычисление дисперсии. Например, если $\xi \sim N(a, \sigma^2)$ и $\eta_1 = \alpha\xi + \beta$, $\eta_2 = \frac{\xi - a}{\sigma}$, то $\mathsf{D}\eta_1 = a^2\sigma^2$ и $\mathsf{D}\eta_2 = 1$.

Наряду с дисперсией для характеристики степени рассеяния значений случайной величины ξ вокруг своего центра в ряде случаев бывает удобным использование среднего квадратического (или стандартного) отклонения величины ξ и коэффициента вариации ξ.

Стандартным отклонением ξ называют $\sqrt{\mathsf{D}\xi}$ и применяют его в тех случаях, когда необходимо разброс ξ измерить в тех же единицах, что и значение самой случайной величины ξ.

Коэффициентом вариации V случайной величины ξ называется отношение (в процентах) стандартного отклонения $\sqrt{\mathsf{D}\xi}$ к его математическому ожиданию $\mathsf{M}\xi$, если $\mathsf{M}\xi \neq 0$, т. е. $V = \frac{\sqrt{\mathsf{D}\xi}}{\mathsf{M}\xi} \cdot 100\%$.

Ясно, что V — величина безразмерная. Она применяется в тех случаях, когда нужно сравнить степени рассеяния двух случайных величин, значения которых измеряются разными единицами.

3. Другие числовые характеристики

Наряду с рассмотренными выше числовыми характеристиками случайной величины ξ часто используются так называемые начальные моменты порядка k и центральные моменты порядка k.

Начальным моментом порядка k случайной величины ξ называется математическое ожидание величины ξ^k, т. е. число $\mathsf{M}\xi^k$.

Для дискретной ξ

$$\mathsf{M}\xi^k = \sum_n x_n^k \cdot p_n,$$

для непрерывной ξ

$$\mathsf{M}\xi^k = \int\limits_{-\infty}^{\infty} x^k\, p(x)\, dx,$$

при условии, что ряд для дискретной величины ξ и несобственный интеграл для непрерывной величины ξ абсолютно сходятся.

Начальным моментом первого порядка случайной величины ξ является ее математическое ожидание $\mathsf{M}\xi$. *Центральным моментом порядка k* случайной величины ξ называется математическое ожидание величины $(\xi - \mathsf{M}\xi)^k$, т. е. число $\mathsf{M}(\xi - \mathsf{M}\xi)^k$.

Для дискретной ξ

$$\mathsf{M}(\xi - \mathsf{M}\xi)^k = \sum (x_n - \mathsf{M}\xi)^k \cdot p_n,$$

для непрерывной ξ

$$\mathsf{M}(\xi - \mathsf{M}\xi)^k = \int\limits_{-\infty}^{+\infty} (x - \mathsf{M}\xi)^k\, p(x)\, dx,$$

при условии, что ряд для дискретной величины ξ и несобственный интеграл для непрерывной величины ξ абсолютно сходятся.

Очевидно, что центральный момент первого порядка всегда равен 0, а центральный момент второго порядка представляет собой $D\xi$. Из центральных моментов более высокого порядка чаще других используются центральные моменты третьего и четвертого порядков. С их помощью определяют так называемые коэффициент асимметрии и коэффициент эксцесса.

Для случайной величины ξ *коэффициент асимметрии* $A = \dfrac{\mathsf{M}(\xi - \mathsf{M}\xi)^3}{(\mathsf{D}\xi)^{3/2}}$,

а *коэффициент эксцесса* $E = \dfrac{\mathsf{M}(\xi - \mathsf{M}\xi)^4}{(\mathsf{D}\xi)^2} - 3$.

Для справедливости этих формул требуется, чтобы $\mathsf{D}\xi \neq 0$. Формула для A дает количественную безразмерную характеристику скошенности распределения ξ. Для всех симметричных относительно $\mathsf{M}\xi$ распределений ξ всегда $A = 0$. Например, это так в случае $\xi \sim N(a, \sigma^2)$.

Если же длинная часть кривой плотности распределения непрерывной ξ расположена справа от ее вершины, то $A > 0$, а если длинная часть кривой плотности распределения ξ расположена влево от ее вершины то $A < 0$.

Формула для E дает количественную безразмерную характеристику «островершинности» распределения величины ξ.

$E = 0$ для нормально распределенной ξ. Как правило, распределения непрерывных случайных величин ξ с более высокой и более острой вершиной кривой плотности по сравнению с нормальной кривой имеют $E > 0$, а с менее острой — $E < 0$.

Кроме начальных моментов и центральных моментов, для характеристики случайной величины ξ используются понятия моды, медианы, квантилей.

Модой непрерывной случайной величины ξ называется такое ее значение x, при котором плотность $p(x)$ распределения ξ имеет максимум. *Модой дискретной случайной величины* ξ называется ее значение, имеющее наибольшую вероятность. Распределения случайных величин могут быть одномодальными, двумодальными и т. д. Для одномодальных ξ мода является характеристикой центра группирования значений величины ξ. Мода является как бы наиболее типичным (наиболее часто принимаемым) значением случайной величины.

Медианой непрерывной случайной величины ξ называется такое ее значение x_μ, что вероятность $\mathsf{P}\{\xi < x_\mu\} = \mathsf{P}\{\xi > x_\mu\} = 0{,}5$.

Для дискретной случайной величины ξ со значениями x_1, x_2 , …, …, x_n *медиана* определяется следующим образом. Из значений x_1, x_2, …, x_n стоится так называемый вариационный ряд $x^{(1)}$, $x^{(2)}$, …, $x^{(n)}$, где $x^{(1)} \leqslant x^{(2)} \leqslant \ldots \leqslant x^{(n)}$. По определению, медиана $x_\mu = x^{(k+1)}$, если число n — нечетное: $n = 2k + 1$, и медиана $x_\mu = \frac{1}{2}(x_k + x_{k+1})$, если число n — четное: $n = 2k$.

Из этого определения следует, что медиана x_μ может быть одним из значений величины ξ, а может и не быть значением ξ. Например, для величины ξ со значениями 1, 2, 3, 4, 5 медиана $x_\mu = 3$, а для величины ξ со значениями 1, 2, 3, 4 медиана $x_\mu = \frac{1}{2}(2 + 3) = 2{,}5$.

В случае симметричной плотности распределения непрерывной случайной величины ξ всегда математическое ожидание, мода и медиана совпадают между собой. Для асимметричных распределений это не так.

Пусть непрерывная случайная величина ξ обладает непрерывной функцией распределения $F(x)$. Тогда q-квантилью величины ξ называют такое ее значение x_q, что $F(x_q) = q$.

Очевидно, чем больше значение q $(0 < q < 1)$, тем больше будет и величина q-квантили x_q. В частном случае при $q = 0{,}5$ квантиль x_q является медианой.

Синонимом слова «квантиль» является слово «процентиль». Заметим, что для некоторых значений $0 < q < 1$ уравнение $F(x) = q$ может иметь неединственное решение, если $F(x)$ не является строго монотонно возрастающей функцией.

§ 4. Корреляционно-регрессионный анализ зависимости двух случайных величин

Пусть X и Y — две случайные величины, значения которых измерены в количественных шкалах. Нас интересует зависимость Y от X. Будем называть X *аргументом* (фактором), входной случайной величиной, а Y — *откликом*, выходной, результирующей или зависимой случайной величиной.

Для случайных величин X и Y ранее были введены понятия независимости и функциональной зависимости вида $Y = f(X)$. Однако, гораздо чаще наблюдается стохастическая (или вероятностная) зависимость случайных величин, заключающаяся в том, что каждому фиксированному значению одной из них соответствует не одно, а множество значений другой величины, причем заранее не известно, какое из этих значений примет другая величина. Это значит, что при изменении одной величины, например X, меняется функция распределения другой величины Y.

Примерами стохастически зависимых случайных величин могут быть: величина урожая зерна с некоторого участка земли и количество на этот участок удобрений, возраст мужа и возраст жены, рост отца и рост его детей, уровень квалификации рабочего и его выработка за смену, успеваемость студента в вузе и его успеваемость по математике в школе и т. д. Например, величина урожая зерна не является функцией от количества внесенных удобрений, так как на величину урожая влияют такие случайные факторы, как количество осадков, температура воздуха и многие другие неконтролируемые случайные факторы.

Таким образом, стохастическая зависимость Y от X объясняется влиянием на Y не только X, но и некоторых других неучтенных случайных факторов. Частным случаем стохастической зависимости X и Y является их корреляционная зависимость.

Если имеется стохастическая зависимость величины Y от величины X, то рассматривают, прежде всего, изменения центра условного распределения величины Y при изменении значений величины X. Иными словами рассматривается условное математическое

ожидание $\mathsf{M}[Y \mid X = x]$ величины Y при условии, что величина X равна фиксированному значению x. Если условное математическое ожидание $\mathsf{M}[Y \mid X = x]$ изменяется при изменении значений x, то стохастическую зависимость Y от X называют *корреляционной* зависимостью. Если же условное математическое ожидание $\mathsf{M}[Y \mid X = x]$ остается постоянным при изменении значений x, то говорят, что случайная величина Y *не коррелирует* с величиной X.

В случае корреляционной зависимости Y от X условное математическое ожидание $\mathsf{M}[Y \mid X = x]$ называется *функцией регрессии* величины Y по X или просто — *регрессией Y по X*. Корреляционную зависимость Y от X также часто называют *регрессионной зависимостью Y* от X.

Корреляционная зависимость Y от X, например, имеет место, когда на функциональную зависимость $f(X)$ случайной величины Y от X накладываются неконтролируемые случайные возмущения ε, так что получается $Y = f(X) + \varepsilon$, причем остаточная случайная величина ε такова, что ее условное математическое ожидание $\mathsf{M}[\varepsilon \mid X = x] = 0$ при всех значениях x. Для такой модели зависимости Y от X получаем, что функция регрессии $\mathsf{M}[Y \mid X = x] = f(x)$, так как $\mathsf{M}[Y \mid X = x] = f(x) + \mathsf{M}[\varepsilon \mid X = x] = f(x)$.

В такой модели зависимости Y от X фактор X может быть как случайным, так и не случайным. Однако зависимая величина Y всегда является случайной величиной в силу обязательного присутствия случайного остатка ε. Случайный остаток ε отражает влияние на величину Y факторов, неучтенных зависимостью $f(X)$, и включает в себя также случайную погрешность в измерении значений результирующей величины Y. Случайные величины ε называют *случайными ошибками* или *регрессионными остатками*.

Если X и Y — две произвольные случайные величины, то имеет смысл говорить как о корреляционной зависимости Y от X, так и о корреляционной зависимости X от Y. В том случае, когда функция регрессии Y по X представляет собой непрерывную функцию $f(x)$, график функции $y = f(x)$ на плоскости с декартовыми прямоугольными координатами (x, y) называется *линией регрессии Y* по X. Она характеризует форму корреляционной зависимости Y от X. Когда Y не коррелирует с X, то линией регрессии Y по X является прямая $y = \mathsf{M}Y$, где $\mathsf{M}Y$ — математическое ожидание Y. Эта прямая параллельна оси x.

Если задан закон распределения двумерной случайной величины (X, Y), то нетрудно исследовать коррелируемость X и Y. Приведем пример.

Пример 1. Проверить, коррелирует ли величина Y с величиной X, если задана следующая таблица распределения двумерной случайной величины (X, Y):

X \ Y	$y_1 = -1$	$y_2 = 0$	$y_3 = 1$
$x_1 = 2$	0,1	0,25	0,15
$x_2 = 3$	0,3	0,1	0,1

Решение. Имеем: $\mathsf{P}\{X=2\}=0,1+0,25+0,15=0,5$; $\mathsf{P}\{X=3\}=0,3+0,1+0,1=0,5$.

Найдем теперь условные математические ожидания $\mathsf{M}[Y \mid X=2]$ и $\mathsf{M}[Y \mid X=3]$, используя формулу условной вероятности (см. с. 24)

$$\mathsf{P}\{Y=y_i \mid X=x_j\} = \frac{\mathsf{P}\{Y=y_i,\ X=x_j\}}{\mathsf{P}\{x=x_j\}}.$$

Тогда

$$\mathsf{M}[Y \mid X=x_j] = \sum_{i=1}^{3} y_i \cdot \mathsf{P}\{Y=y_i \mid X=x_j\} =$$

$$= \sum_{i=1}^{3} y_i \frac{\mathsf{P}\{Y=y_i,\ X=x_j\}}{\mathsf{P}\{X=x_j\}} = \frac{1}{\mathsf{P}\{X=x_j\}} \sum_{i=1}^{3} y_i \cdot \mathsf{P}\{Y=y_i,\ X=x_j\}.$$

Подставляя последовательно $x_j = 2$ и $x_j = 3$, получаем:

$$\mathsf{M}[Y \mid X=2] = \frac{1}{0,5}[(-1)\cdot 0,1 + 0\cdot 0,25 + 1\cdot 0,15] = 0,1;$$

$$\mathsf{M}[Y \mid X=3] = \frac{1}{0,5}[(-1)\cdot 0,3 + 0\cdot 0,1 + 1\cdot 0,1] = -0,4.$$

Поскольку $\mathsf{M}(Y \mid X=2) \neq \mathsf{M}(Y \mid X=3)$, то Y коррелирует с X. ■

Имеются показатели степени (тесноты) корреляционной зависимости двух случайных величин X и Y. Выбор таких показателей существенно зависит от того, является ли функция регрессии линейной или нелинейной.

Пусть сначала функция регрессии Y по X является линейной, т. е. $\mathsf{M}[Y \mid X=x] = \alpha x + \beta$, где $\alpha \neq 0$ и β — некоторые числа. В случае линейной функции регрессии Y по X степень корреляционной зависимости Y и X характеризует так называемый *коэффициент корреляции Пирсона* $r(Y, X)$, если он существует. По определению,

$$r(Y, X) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\mathsf{D}X \cdot \mathsf{D}Y}}.$$

Здесь $\operatorname{cov}(X, Y)$ — число, называемое *ковариацией* X и Y и задаваемое формулой

$$\operatorname{cov}(X, Y) = \mathsf{M}[(X - \mathsf{M}X)(Y - \mathsf{M}Y)],$$

а $\mathsf{D}X \neq 0$ — дисперсия X, $\mathsf{D}Y \neq 0$ — дисперсия Y. Если воспользоваться свойствами математического ожидания, то формулу ковариации X и Y можно упростить следующим образом:

$$\begin{aligned}\operatorname{cov}(X, Y) &= \mathsf{M}[(X - \mathsf{M}X)(Y - \mathsf{M}Y)] = \\ &= \mathsf{M}[XY - Y \cdot \mathsf{M}X - X \cdot \mathsf{M}Y + \mathsf{M}X \cdot \mathsf{M}Y] = \\ &= \mathsf{M}(X \cdot Y) - \mathsf{M}Y \cdot \mathsf{M}X - \mathsf{M}X \cdot \mathsf{M}Y + \mathsf{M}X \cdot \mathsf{M}Y = \mathsf{M}(X \cdot Y) - \mathsf{M}X \cdot \mathsf{M}Y.\end{aligned}$$

Таким образом, наиболее удобное для практики выражение коэффициента корреляции Пирсона имеет вид:

$$r(Y, X) = \frac{\mathsf{M}(XY) - \mathsf{M}X \cdot \mathsf{M}Y}{\sqrt{\mathsf{D}X \cdot \mathsf{D}Y}}.$$

Заметим, что с помощью $\operatorname{cov}(X, Y)$ выражается дисперсия суммы произвольных случайных величин X и Y:

$$\mathsf{D}(X + Y) = \mathsf{D}X + \mathsf{D}Y + 2\operatorname{cov}(X, Y).$$

Перечислим *свойства* коэффициента корреляции Пирсона.

1) $r(Y, X)$ — безразмерная величина и $r(X, X) = 1$.

2) Коэффициент корреляции симметричен по отношению к X и Y, т. е. $r(Y, X) = r(X, Y)$. Поэтому в дальнейшем коэффициент корреляции величин X и Y будет обозначаться просто r.

3) Модуль коэффициента корреляции r не превосходит 1, т. е. $|r| \leqslant 1$.

4) Модуль коэффициента корреляции r равен единице тогда и только тогда, когда случайные величины X и Y связаны линейно, т. е. существуют такие числа $a \neq 0$ и b, что $Y = aX + b$.

5) Если случайные величины X и Y независимы, то коэффициент корреляции $r = 0$.

Обратное утверждение не всегда верно, т. е. из того, что коэффициент корреляции $r = 0$, в общем случае не следует независимость случайных величин X и Y. Приведем пример. Пусть величина X принимает значения ± 1, ± 2 с вероятностями $p = \frac{1}{4}$ и пусть величина $Y = X^2$. Тогда Y принимает значения 1 и 4 с вероятностями $p = \frac{1}{2}$ и $\mathsf{M}X = \mathsf{M}(XY) = 0$. Следовательно, $r = 0$, но X и Y не являются независимыми, а связаны функционально.

6) Если $r = 0$ и двумерная случайная величина (X, Y) распределена нормально, то X и Y независимы. Следовательно, для величин X, Y, имеющих совместное двумерное нормальное распределение, условие $r = 0$ равносильно их независимости.

7) При $r > 0$ значение величин X и Y одновременно возрастают, при $r < 0$ при возрастании значений одной величины значения другой величины убывают.

Итак, в случае линейной функции регрессии при $r = 0$ случайные величины X и Y не коррелируют, а при $r \neq 0$ они зависимы и коррелируют. Коэффициент корреляции r дает численную меру корреляционной зависимости в случае линейной функции регрессии. Заметим, что r не существует, если не существует хотя бы одно из чисел $\operatorname{cov}(X, Y)$, $\mathsf{D}X$, $\mathsf{D}Y$.

Пример 2. Найти коэффициент корреляции по заданной в примере 1 таблице распределения двумерной случайной величины (X, Y).

Решение. Из заданной в примере 1 таблицы распределения (X, Y) получаем следующие таблицы распределения для каждой из величин X, Y, XY в отдельности:

X	2	3
P	0,5	0,5

Y	-1	0	1
P	0,4	0,35	0,25

XY	-3	-2	0	2	3
P	0,2	0,2	0,35	0,125	0,125

Из полученных таблиц распределения находим, что: $\mathsf{M}X = 2 \cdot 0{,}5 + 3 \cdot 0{,}5 = 2{,}5$, $\mathsf{M}Y = (-1) \cdot 0{,}4 + 0 \cdot 0{,}35 + 1 \cdot 0{,}25 = -0{,}15$, $\mathsf{M}(XY) = (-3) \times 0{,}2 + (-2) \cdot 0{,}2 + 0 \cdot 0{,}35 + 2 \cdot 0{,}125 + 3 \cdot 0{,}125 = -0{,}375$, $\mathsf{D}X = \mathsf{M}X^2 - (\mathsf{M}X)^2 = [4 \cdot 0{,}5 + 9 \cdot 0{,}5] - (2{,}5)^2 = 0{,}25$, $\mathsf{D}Y = \mathsf{M}Y^2 - (\mathsf{M}Y)^2 = [1 \cdot 0{,}65 + 0 \cdot 0{,}35] - (-0{,}15)^2 = 0{,}6275$.

Следовательно, $r = \dfrac{\mathsf{M}(XY) - \mathsf{M}X \cdot \mathsf{M}Y}{\sqrt{\mathsf{D}X \cdot \mathsf{D}Y}} = \dfrac{-0{,}375 + 2{,}5 \cdot 0{,}15}{\sqrt{0{,}25 \cdot 0{,}6275}} = 0$. ■

Сделаем еще два замечания относительно коэффициента корреляции r. Если двумерная случайная величина (X, Y) подчиняется нормальному закону распределения, то функция регрессии $\psi(x)$ величины Y по X всегда является линейной, причем

$$\psi(x) = \mathsf{M}Y - r \cdot \sqrt{\frac{\mathsf{D}Y}{\mathsf{D}X}}(x - \mathsf{M}X),$$

а условная дисперсия $\mathsf{D}[Y \mid X = x] = (1 - r^2)\mathsf{D}Y$.

Число r^2 называется *коэффициентом детерминации* Y по X. Он показывает, какая часть изменения величины Y может быть объяснена только изменением величины X в случае модели зависимости

Y от X вида $Y = \alpha X + \beta + \varepsilon$, где ε — остаточная случайная величина, $\alpha \neq 0$ и β — некоторые числа.

Пусть теперь функция регрессии Y по X является нелинейной. Тогда коэффициент корреляции r теряет свой смысл как характеристика степени (тесноты) корреляционной зависимости Y от X. В случае нелинейной функции регрессии степень корреляционной зависимости Y от X описывается так называемым *корреляционным отношением* $\rho(Y/X)$ величины Y по X, которое при $\mathsf{D}Y > 0$ задается формулой

$$\rho(Y/X) = \sqrt{\frac{\mathsf{D}[M(Y \mid X = x)]}{\mathsf{D}Y}},$$

где $\mathsf{D}[\mathsf{M}(Y \mid X = x)]$ обозначает дисперсию регрессии Y по X. Аналогично вводится и корреляционное отношение $\rho(X/Y)$ величины X по Y. Основные свойства корреляционного отношения $\rho(Y/X)$ следующие:

1. В общем случае $\rho(Y/X)$ несимметрично по отношению к Y и X, т. е. $\rho(Y/X) \neq \rho(X/Y)$. Например, возможно $\rho(Y/X) = 0$ и $\rho(X/Y) = 1$.

2. $0 \leqslant \rho(Y/X) \leqslant 1$.

3. $\rho(Y/X) = 0$ тогда и только тогда, когда нет корреляционной зависимости Y от X.

4. $\rho(Y/X) = 1$ тогда и только тогда, когда случайная величина Y является некоторой функцией величины X, т. е. $Y = f(X)$ для некоторой функции f.

5. Если функция регрессии Y по X является линейной, то корреляционное отношение Y по X совпадает с модулем коэффициента корреляции, т. е. $\rho(Y/X) = |r|$.

6. Если функция регрессии Y по X является нелинейной, то всегда корреляционное отношение Y по X больше модуля коэффициента корреляции, т. е. $\rho(Y/X) > |r|$.

Число $\rho^2(Y/X)$ называется коэффициентом детерминации величины Y по X. В случае модели зависимости Y от X вида $Y = f(X) + \varepsilon$, где ε — остаточная случайная величина, коэффициент детерминации дает численную характеристику той доли общего изменения Y, которая объясняется изменением функции регрессии $f(x)$.

Чаще всего нелинейная функция регрессии $f(x)$ может задаваться одной из следующих формул: $f(x) = ax^2 + bx + c$, $f(x) = a \cdot e^{bx}$, $f(x) = \frac{a}{x} + b$, $f(x) = ax^3 + bx^2 + cx + d$, где $a \neq 0$.

Заметим, что задача исследования нелинейной функции регрессии в общем достаточно трудная и там, где это возможно, желательно свести ее к задаче о линейной функции регрессии. Пусть, например, из каких-либо соображений известно, что функция регрессии Y по X имеет вид $f(x) = ab^x$, где числа $a > 0$ и $b > 0$. Тогда $\ln f(x) = \ln a + x \cdot \ln b$, т. е. $\ln f(x)$ уже линейно зависит от x.

На практике подбор функции регрессии проводится с помощью статистических методов, о чем будет речь впереди.

Пример 3. Используя данные примеров 1 и 2, найти корреляционное отношение Y по X.

Решение. Заметим сначала, что условное математическое ожидание $\mathsf{M}(Y \mid X)$ так же как и условная дисперсия $\mathsf{D}(Y \mid X)$, является случайной величиной, поскольку X — случайная величина. Вероятности значений $\mathsf{M}(Y \mid X)$ и $\mathsf{D}(Y \mid X)$ будут такими же, как вероятности значений величины X. Поэтому имеем:

$\mathsf{M}(Y \mid X)$	0,1	−0,4
P	0,5	0,5

Следовательно,

$$\mathsf{D}[\mathsf{M}(Y \mid X = x)] = \mathsf{M}[\mathsf{M}(Y \mid X) - \mathsf{M}Y]^2 = \\ = (0{,}1 + 0{,}15)^2 \cdot 0{,}5 + (-0{,}4 + 0{,}15)^2 \cdot 0{,}5 = 0{,}0625.$$

Так как $\mathsf{D}Y = 0{,}6275$ (см. пример 2), то $\rho(Y/X) = \sqrt{\frac{0{,}0625}{0{,}6275}} \approx 0{,}31$.

Напомним, что в примере 2 получен $r = 0$. Таким образом, $|r| \neq \rho(Y/X)$. Это объясняется тем, что функция регрессии, заданная вышеприведенной таблицей, не является линейной. ■

§ 5. Закон больших чисел. Центральная предельная теорема

Известно, что результаты испытаний, которые проведены в относительно одинаковых условиях, могут сильно отличаться друг от друга. Однако оказывается, что средние значения результатов испытаний при большом числе испытаний обладают хорошей устойчивостью. Этот факт обосновывается с помощью группы теорем, называемой законом больших чисел.

Сформулируем простейший вариант закона больших чисел.

Теорема 1 (закон больших чисел). *Пусть* $\xi_1, \xi_2, \ldots, \xi_n$ *— независимые и одинаково распределенные случайные величины, имеющие общее конечное математическое ожидание* $\mathsf{M}\xi_1 = \mathsf{M}\xi_2 = \ldots = \mathsf{M}\xi_n = a$. *Тогда для любого* $\varepsilon > 0$ *при* $n \to \infty$ *вероятность*

$$\mathsf{P}\left(\left|\frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n} - a\right| < \varepsilon\right) \to 1.$$

Смысл закона больших чисел в том, что среднее арифметическое достаточно большого числа независимых случайных величин с вероятностью, близкой к единице, принимает значение, близкое к числу a. Таким образом, среднее арифметическое достаточно большого числа рассматриваемых случайных величин теряет характер случайной величины, и с вероятностью, близкой к единице, можно сказать, что среднее арифметическое проявляет свойство устойчивости, заключающееся в концентрации среднего арифметического вокруг числа a, причем независимо от типов законов распределения случайных величин $\xi_1, \xi_2, \ldots, \xi_n$.

Теорема 1 имеет важные применения.

Пусть, например, проводятся n испытаний Бернулли и пусть p — неизвестная вероятность интересующего нас события A. Если m — число тех испытаний, в которых произошло событие A, то из теоремы 1 следует, что при достаточно большом n относительная частота $\frac{m}{n}$ наступления события A практически достоверно приближается к вероятности p события A.

На основании этого часто на практике за неизвестные вероятности p при большом числе испытаний Бернулли принимают относительную частоту $\frac{m}{n}$ появления события A.

Эту относительную частоту называют статистической вероятностью события A. Она дает приближенное значение неизвестной вероятности p. Недостатком такого определения p является его неоднозначность. Тем не менее, метод приближенного определения неизвестной вероятности события с помощью большого числа испытаний Бернулли является важным для практики. Другие важные применения закон больших чисел имеет в математической статистике.

Еще одним важным утверждением для теории вероятностей и математической статистики является центральная предельная теорема. Различные ее формы отличаются друг от друга лишь степенью общности и видом условий.

Пусть $\xi_1, \xi_2, \ldots, \xi_n$ — независимые случайные величины, которые одинаково распределены и имеют конечные математические ожидания $\mathsf{M}\xi_1 = \mathsf{M}\xi_2 = \ldots = \mathsf{M}\xi_n = a$ и дисперсии $\mathsf{D}\xi_1 = \mathsf{D}\xi_2 = \ldots = \mathsf{D}\xi_n = \sigma^2 > 0$. Рассмотрим случайную величину

$$S_n = \frac{\frac{1}{n}(\xi_1 + \xi_2 + \ldots + \xi_n) - a}{\sigma/\sqrt{n}}.$$

Она является так называемой нормированной случайной величиной, так как для нее $\mathsf{M}S_n = 0$ и $\mathsf{D}S_n = 1$.

Обозначим через $F_n(x)$ функцию распределения нормированной случайной величины S_n, а через $F(x)$ — функцию стандартного нормального распределения, т. е.

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}}\, dt.$$

Теорема 2 (центральная предельная теорема). *Для любого* $x \in (-\infty, +\infty)$ *при* $n \to \infty$ *имеет место сходимость* $F_n(x) \to F(x)$.

Другими словами, центральная предельная теорема утверждает, что функции распределения нормированных сумм S_n случайных величин асимптотически (при $n \to \infty$) стремятся к функции стандартного нормального распределения для любого значения x. Иногда говорят коротко, что нормированные суммы S_n являются асимптотически (при $n \to \infty$) нормальными. Очень важно, что центральная предельная теорема справедлива независимо от типа распределения случайных величин $\xi_1, \xi_2, \ldots, \xi_n$.

Центральная предельная теорема не только еще раз подчеркивает широкое использование нормального распределения случайных величин. Она имеет большое практическое значение для теории вероятностей и математической статистики. Например, опираясь на центральную предельную теорему, можно получить следующие полезные для практики факты.

1) Случайная величина $\xi(n, p)$, распределенная по биномиальному закону с параметрами (n, p), асимптотически (при $n \to \infty$) имеет нормальное распределение с параметрами $(np, np \cdot (1-p))$. Этот результат называется теоремой Муавра—Лапласа.

2) Случайная величина $\xi(\lambda)$, распределенная по закону Пуассона с параметром λ, асимптотически (при $\lambda \to \infty$) имеет нормальное распределение с параметрами (λ, λ).

3) Распределение случайной величины $\chi^2(n)$ асимптотически (при $n \to \infty$) нормально с параметрами $(n, 2n)$.

4) Распределение случайной величины $t(n)$ асимптотически (при $n \to \infty$) нормально с параметрами (0, 1).

Заметим в заключение, что скорость сходимости распределения суммы S_n к нормальному закону существенно зависит от типа распределения исходных слагаемых $\xi_1, \xi_2, \ldots, \xi_n$.

§ 6. Понятие о случайных процессах

Теория случайных процессов изучает законы изменения случайных величин, зависящих от времени.

Пусть Ω — множество элементарных событий ω и t — время. *Случайным* или *стохастическим процессом* называется функция двух аргументов

$$\xi(t) = f(\omega, t), \quad \omega \in \Omega.$$

Для каждого значения t функция $\xi(t)$ является случайной величиной, а для каждого $\omega \in \Omega$ функция $\xi(t)$ является вполне определенной числовой функцией аргумента t. При фиксированном ω функция $\xi(t)$ называется *реализацией* (или *траекторией*) случайного процесса. Для различных $\omega \in \Omega$ эти функции в общем случае различны.

Если $t = 0, 1, 2, 3, \ldots$, то $\xi(t)$ называют *случайным процессом с дискретным временем*. Если же $t \in [0, T]$, то $\xi(t)$ называют *случайным процессом с непрерывным временем*.

Например, число $\xi(t)$ просмотренных программ передач от начала работы телевизора до момента t является случайным процессом с непрерывным временем, а температура $\xi(t)$ заболевшего человека, измеренная в определенные моменты времени $t_1, t_2, t_3, \ldots$, является случайным процессом с дискретным временем.

Функция распределения

$$F_t(x) = \mathsf{P}[\xi(t) < x]$$

задает распределение значений $\xi(t)$ в момент времени t. Так как знание $F_t(x)$ при всех t не позволяет сказать о зависимости значений $\xi(t)$ в какие-либо фиксированные моменты времени, то для полного описания процесса $\xi(t)$ необходимо указать всевозможные совместные распределения случайных величин $\xi(t)$ для всевозможных наборов $0 < t_1 < t_2 < \ldots < t_n$, $n = 1, 2, 3, \ldots$. Это сложная задача. Однако на практике для случайного процесса $\xi(t)$ чаще всего достаточно

знать три характеристики: математическое ожидание $M\xi(t)$, дисперсию $D\xi(t)$ или среднее квадратическое отклонение $\sigma[\xi(t)] = \sqrt{D\xi(t)}$ и автоковариационную функцию $\operatorname{cov}[\xi(t'), \xi(t'')]$ или автокорреляционную функцию

$$r[\xi(t'), \xi(t'')] = \frac{\operatorname{cov}[\xi(t'), \xi(t'')]}{\sigma[\xi(t')] \cdot \sigma[\xi(t'')]},$$

где t' и t'' пробегают всевозможные значения времени t. Отметим, что автоковариационная и автокорреляционная функции процесса $\xi(t)$ необходимы для описания зависимости между случайными величинами, отвечающим разным значения времени t.

Введенные характеристики обладают свойствами, аналогичными свойствам соответствующих характеристик случайной величины. В частности, $M\xi(t)$ дает некоторую среднюю функцию, вокруг которой группируются всевозможные реализации данного процесса, а $D\xi(t)$ и $\sigma[\xi(t)]$ характеризуют степень рассеяния всевозможных реализаций $\xi(t)$ относительно $M\xi(t)$. Функция $r[\xi(t'), \xi(t'')]$ симметрична относительно ξ' и ξ'', и всегда $|r[\xi(t'), \xi(t'')]| \leqslant 1$ для любого процесса $\xi(t)$.

Опишем основные типы случайных процессов $\xi(t)$ с непрерывным временем.

Наиболее простой природой обладают так называемые однородные процессы с независимыми приращениями. Случайный процесс $\xi(t)$ с непрерывным временем $t \in [0, T]$ называется *однородным процессом с независимыми приращениями*, если выполнены три условия:

1) при любых $0 \leqslant t_1 < t_2 < \ldots < t_n \leqslant T$, $n = 1, 2, 3, \ldots$ случайные величины $\xi(0)$, $\xi(t_1) - \xi(0)$, $\xi(t_2) - \xi(t_1)$, $\ldots$, $\xi(t_n) - \xi(t_{n-1})$ независимы,

2) при любых $t_1 < t_2$ и τ приращения $\xi(t_1) - \xi(t_2)$, $\xi(t_1 + \tau) - \xi(t_2 + \tau)$ одинаково распределены,

3) $\xi(0) = 0$.

Первое условие означает, что в процессе отсутствует последействие, т. е. имеется взаимная независимость появления числа событий в непересекающихся промежутках. Второе условие означает однородность процесса, т. е. вероятность числа появления событий в каждом временном интервале зависит лишь от длины этого интервала и не зависит от начального момента $t = 0$.

Два наиболее распространенных класса однородных процессов с независимыми приращениями — это так называемые пуассоновские и винеровские процессы. Пуассоновским процессом называется однородный процесс с независимыми приращениями, для которого

приращение $\xi(t) - \xi(0)$ распределено по закону Пуассона с параметром λt:

$$\mathsf{P}\{\xi(t) = k\} = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad \lambda > 0, \quad k = 0, 1, 2\ldots.$$

Реализации пуассоновского процесса являются скачкообразными, причем число скачков процесса на $[0, t]$ равно значению $\xi(t)$. Заметим, что в данном процессе $\mathsf{M}\xi(t) = \mathsf{D}\xi(t) = \lambda t$. Пуассоновский процессом описываются, например, число телефонных звонков на телефонную станцию, число вызовов скорой помощи или аварийной службы в каком-нибудь временном промежутке.

Простейшим обобщение пуассоновского процесса является так называемый процесс размножения и гибели. Он является моделью изменения объема совокупности за счет размножения или гибели ее членов. Сначала эту модель стали изучать при исследовании численности популяции распространения эпидемий.

Винеровским процессом называется однородный процесс с независимыми приращениям, для которого приращения $\xi(t+\tau) - \xi(\tau)$ распределены нормально с параметрами $(\mu t, \sigma\sqrt{t})$, где $\sigma > 0$. Из этого определения следует, что все совместные распределения $\xi(t_1)$, ..., $\xi(t_n)$ также будут нормальными. Таким образом, для винеровского процесса функция распределения

$$F_t(x) = \frac{1}{\sigma\sqrt{2\pi t}} \int_{-\infty}^{x} e^{-\frac{(u-at)^2}{2\sigma^2 t}}\, du.$$

Винеровский процесс с параметрами называют $(0, \sqrt{t})$ стандартным винеровским процессом. Реализации винеровского процесса $\xi(t)$ с вероятностью единица могут быть выбраны непрерывными.

Винеровские процессы часто называют также процессами броуновского движения, поскольку они описывают закономерности, возникающие в совокупностях большого числа движущихся и взаимодействующих малых частиц.

Все рассматривавшиеся до сих пор типы случайных процессов с непрерывным временем являются процессами без последействия. Их еще называют процессами марковского типа. Они характерны тем, что прошлые состояния системы не влияют на вероятности ее будущих состояний, если известно настоящее состояние. Однако нетрудно привести примеры процессов, когда прошлые состояния системы оказывают весьма сильное влияние на вероятности ее будущих состояний. Например, успеваемость учащегося по данному предмету в данном году зависит не только от его успеваемости в

прошлом году, но и от его успеваемости в предыдущие годы. Поэтому был выделен класс случайных процессов с последействием, называемых стационарными процессами.

Случайный процесс $\xi(t)$ называется *стационарным процессом*, если для произвольных n моментов времени $t_1, t_2, \ldots, t_n$ и любого числа τ функции распределения вероятностей системы величин $\xi(t_1), \xi(t_2), \ldots, \xi(t_n)$ и системы величин $\xi(t_1+\tau), \xi(t_2+\tau), \ldots, \xi(t_n+\tau)$ совпадают, т. е. не зависят от сдвига всех моментов времени на одну и ту же величину τ.

Для стационарного процесса математическое ожидание и дисперсия являются постоянными. Примером стационарного процесса является нормальный процесс, т. е. процесс, для которого при любом $n = 1, 2, 3, \ldots$ система величин $\{\xi(t_1), \xi(t_2) \ldots \xi(t_n)\}$ распределена по нормальному закону. Стационарным является и так называемый однородный марковский процесс, т. е. марковский процесс, для которого функция распределения $F(t, x, \tau, y)$ зависит лишь от x, y, и $\tau - t$. Однако на практике установить стационарность процесса весьма непросто. Поэтому чаще всего вместо стационарных процессов рассматривают так называемые стационарные в широком смысле случайные процессы.

Случайный процесс $\xi(t)$ называется *стационарным в широком смысле*, если его математическое ожидание и дисперсия не зависят от времени t, а коэффициент корреляции между $\xi(t)$ и $\xi(t+\tau)$ является функцией только сдвига τ. Всякий стационарный процесс является стационарным в широком смысле, но обратное утверждение неверно.

Понятно, что теория стационарных в широком смысле процессов не в состоянии полностью заменить теорию стационарных процессов. Однако для многих вопросов теория стационарных в широком смысле процессов дает удовлетворенный ответ. Эту теорию иногда называют корреляционной теорией. Она существенно использует корреляционную функцию $r(\tau)$, т. е. коэффициент корреляции между $\xi(t)$ и $\xi(t+\tau)$.

Например, случайный процесс $\xi(t) = \sin(t+\xi)$, где $t \geqslant 0$ и случайная величина ξ распределена равномерно на $[0, 2\pi]$, является стационарным в широком смысле, а случайный процесс $\xi(t) = \xi \cdot \sin t$, где $t \geqslant 0$ и случайная величина $\xi \sim N(2, 4)$, не является стационарным в широком смысле.

До сих пор рассматривались случайные процессы $\xi(t)$ с непрерывным временем, когда случайные явления происходят в любой момент времени. Однако можно рассматривать и такие случайные

процессы $\xi(t)$, изменения которых происходят лишь в некоторые фиксированные моменты времени t, например, при $t = 0, 1, 2, 3, \ldots$. Такие случайные процессы называют еще *случайными последовательностями*.

Наиболее изученными среди процессов с дискретным временем являются марковские процессы с дискретным временем. Марковские цепи — их частный случай. Для марковских случайных процессов условная вероятность состояния системы в любой последующий момент времени t_n зависит лишь от состояния системы в предыдущий момент времени t_{n-1} и не зависит от поведения системы в предыдущее время $t < t_{n-1}$. Для марковских процессов множество возможных состояний может быть конечным или счетным.

Задачи к главе 2

1. Подбрасывают монету три раза. Случайной величиной ξ является число выпадений герба. Построить для ξ функцию распределения и ее график.

2. Дважды бросается игральная кость. Случайной величиной ξ является сумма выпавших очков. Найти закон распределения ξ и построит функцию распределения.

3. Дискретная случайная величина ξ задана следующим законом распределения:

ξ	-1	0	1	2
p	0,1	0,2	0,3	0,4

Построить функцию распределения ξ и ее график.

4. Дискретная случайная величина ξ задана следующим законом распределения:

ξ	1	2	3	4
p	0,2	0,1	0,3	0,4

Построить функцию распределения ξ и найти вероятность того, что $\xi > 2$.

5. Из группы в 10 студентов, среди которых имеются 4 отличника, случайным образом выбраны 3 студента. Найти закон распределения случайного числа ξ отличников, содержавшихся в выборке, и вероятность того, что $\xi \leqslant 1$.

6. Непрерывная случайная величина ξ имеет функцию распределения

$$F(x)=\begin{cases}0, & x<0,\\ x^2, & 0\leqslant x\leqslant 1,\\ 1, & x>1.\end{cases}$$

Найти: а) плотность распределения ξ, б) вероятность попадания ξ на $[0, 1/2]$.

7. Плотность распределения $p(x)$ непрерывной случайной величины ξ равна $\frac{1}{2}\cos x$ на $[-\pi/2, \pi/2]$ и $p(x)=0$ вне $[-\pi/2, \pi/2]$. Найти: а) функцию распределения ξ, б) вероятность попадания ξ на $[0, \pi/2]$.

8. Плотность распределения непрерывной случайной величины ξ задана равенством

$$p(x)=\frac{c}{1+x^2}$$

для всех $x\in(-\infty, +\infty)$. Найти: а) постоянную c, б) функцию распределения ξ, в) вероятность попадания ξ в интервал $(-\infty, 1)$.

9. Непрерывная случайная величина ξ имеет нормальное распределение с параметрами $a=1$, $\sigma=2$. Найти вероятность попадания ξ: а) на отрезок $[1, 3]$, б) в промежуток $(-\infty, 3]$.

10. Случайная величина распределена равномерно на отрезке $[1, 5]$. Найти плотность распределения ξ и вероятность: а) $\mathsf{P}\{2<\xi<4\}$, б) $\mathsf{P}\{\xi<3\}$.

11. Задан закон распределения дискретной случайной величины ξ:

ξ	-1	1	2
p	1/3	1/3	1/3

Найти закон распределения случайной величины: а) $\eta=2\xi+3$, б) $\eta=2\xi^2$, в) $\eta=\xi^3$.

12. Заданы законы распределения двух независимых случайных величин ξ_1 и ξ_2:

ξ_1	1	2	3
p	1/3	1/3	1/3

ξ_2	-1	1
p	1/2	1/2

Найти закон распределения случайной величины: а) $\eta=\xi_1+\xi_2$, б) $\eta=\xi_1\cdot\xi_2$, в) $\eta=\xi_1-\xi_2^2$.

13. Зная функцию распределения $F(x)$ непрерывной случайной величины ξ, найти функцию распределения случайной величины: а) $\eta = 2\xi + 4$, б) $\eta = \xi^3$, в) $\eta = \xi^5 - 1$, г) $\eta = e^{\xi}$.

14. Найти математическое ожидание и дисперсию случайной величины ξ, если: а) $\mathsf{P}\{\xi = 0\} = 1/2$, $\mathsf{P}\{\xi = 1\} = \mathsf{P}\{\xi = 2\} = 1/4$, б) $\mathsf{P}\{\xi = -1\} = \mathsf{P}\{\xi = 1\} = 1/4$, $\mathsf{P}\{\xi = 0\} = 1/2$, в) $\mathsf{P}\{\xi = 1\} = 1/2$, $\mathsf{P}\{\xi = 2\} = 1/3$, $\mathsf{P}\{\xi = 1\} = 1/6$.

15. Монета подбрасывается три раза. Найти математическое ожидание и дисперсию числа выпавших гербов.

16. Найти математическое ожидание и дисперсию непрерывной случайной величины ξ, если ее плотность распределения $p(x) = 2x$ на отрезке [0, 1] и $p(x) \equiv 0$ вне [0, 1].

17. Найти математическое ожидание и дисперсию непрерывной случайной величины ξ, если ее плотность распределения $p(x) = 2(x - 1)$ на отрезке [1,2] и $p(x) \equiv 0$ вне [1, 2].

18. Найти математическое ожидание и дисперсию непрерывной случайной величины ξ, если ее плотность распределения $p(x) = \frac{1}{2}\sin x$ на отрезке $[0, \pi]$ и $p(x) \equiv 0$ вне $[0, \pi]$.

19. Пусть ξ_1 и ξ_2 — независимые случайные величины, причем $\mathsf{M}\xi_1 = 1$, $\mathsf{M}\xi_2 = 5$, $\mathsf{D}\xi_1 = \mathsf{D}\xi_2 = 1$. Найти математическое ожидание и дисперсию случайной величины: а) $\eta = \xi_1 + \xi_2$, б) $\eta = \xi_1 - \xi_2$, в) $\eta = \xi_1 + 3\xi_2$, г) $\eta = 2\xi_1 - 3\xi_2$.

20. Дискретная случайная величина ξ имеет только два возможных значения: $x_1 = 1$, $x_2 = 2$. Найти вероятность каждого из этих значений, если $\mathsf{M}\xi = 1{,}4$.

21. Дискретная случайная величина ξ имеет только три возможных значения: $x_1 = 1$, $x_2 = 2$, $x_3 = 3$. Найти вероятность каждого из этих значений, если $\mathsf{M}\xi = 1{,}4$, $\mathsf{D}\xi = 0{,}14$.

22. Заданы законы распределения двух случайных величин ξ и η:

η	3	4
p	0,6	0,4

ξ	1	2
p	1/2	1/2

Найти коэффициент корреляции величин ξ и η.

23. Задан закон распределения двумерной случайной величины (ξ, η):

ξ \ η	0	1	2
−1	0	0,2	0,3
1	0,3	0,2	0

Найти: а) закон распределения каждой из случайных величин ξ и η в отдельности, б) распределение величины ξ при условии, что $\eta = 0$, в) распределение величины η при условии, что $\xi = 1$, г) коэффициент корреляции величин ξ и η.

24. Задан закон распределения двумерной случайной величины (ξ, η):

ξ \ η	0	1	2
−1	0	0,2	0
0	0,1	0,1	0,2
1	0	0,2	0,1

Найти: а) закон распределения каждой из случайных величин ξ и η в отдельности, б) распределение величины ξ при условии, что $\eta = 1$, в) распределение величины η при условии, что $\xi = 1$, г) коэффициент корреляции величин ξ и η.

25. Найти коэффициент корреляции случайных величин ξ и η, если задан закон распределения двумерной случайной величины (ξ и η):

ξ \ η	2	3
−1	0,2	0,4
1	0,1	0,3

26. Пусть ξ_1, ξ_2, ξ_3 — независимые случайные величины, для которых $\mathsf{D}\xi_1 = \mathsf{D}\xi_2 = \mathsf{D}\xi_3 = \sigma^2$. Найти коэффициент корреляции для величин $\xi_1 + \xi_2$ и $\xi_1 + \xi_3$.

27. Пусть двумерная непрерывная случайная величина (ξ, η) имеет плотность распределения $p(x, y) = \frac{1}{2}\sin(x + y)$, когда точка (x, y) принадлежит прямоугольнику $\Pi = \{(x, y)\colon 0 \leqslant x \leqslant \pi/2, 0 \leqslant y \leqslant \pi/2\}$, и $p(x, y) \equiv 0$, когда точка (x, y) не принадлежит Π. Найти: а) плотность распределения каждой из случайных величин ξ и η в отдельности, б) $\mathsf{M}\xi$ и $\mathsf{D}\xi$.

ГЛАВА 3

ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И СЛУЧАЙНАЯ ВЫБОРКА, СТАТИСТИЧЕСКАЯ МОДЕЛЬ

В этой главе описаны основополагающие понятия математической статистики — генеральная совокупность, случайная выборка, статистическая модель, основные шкалы измерений психологических признаков и методы описательной статистики.

§ 1. Основные понятия

Под *генеральной совокупностью* объектов понимают множество всех мысленно возможных объектов интересующего нас типа, для которых в заданных условиях изучается один или несколько признаков или свойств. Такие объекты называются *элементами генеральной совокупности*. Элементы одной генеральной совокупности должны обладать некоторым общим свойством, качеством, типичностью, целостностью.

В психолого-педагогических исследованиях для рассматриваемой генеральной совокупности регистрируют наблюдения одной или нескольких психологических характеристик. Так, например, можно изучать генеральную совокупность всех школьников Москвы с точки зрения их общего интеллекта или изучать распределение роста и веса совокупности всех студентов какого-нибудь университета и т. д.

Если генеральная совокупность состоит из N элементов, то N называется *объемом* этой совокупности. Практически объем N всегда является некоторым числом. Предполагается, что тот признак или свойство, наблюдения которого регистрируются для элементов генеральной совокупности, представляет собой некоторую дискретную или непрерывную случайную величину ξ, свойства которой описываются теорией вероятностей.

Чаще всего интересующая нас генеральная совокупность либо слишком велика, либо ее элементы недоступны, либо имеются временны́е, организационные, финансовые или какие-либо другие причины, не позволяющие изучить признак или свойство сразу для всех ее элементов. Тогда прибегают к изучению свойства ξ для какой-то части этой совокупности. Эта выбранная для полного исследования признака ξ группа $n < N$ элементов совокупности называется *выбор-*

кой объема n. Выборка должна быть как бы уменьшенной копией генеральной совокупности, т. е. правильно, без искажений представлять всю генеральную совокупность. Такую выборку называют *репрезентативной* (*представительной*). Репрезентативность выборки может быть обеспечена чисто случайным выбором. Однако следует помнить, что и репрезентативность, и точность случайного выбора носят вероятностный характер, о чем будет речь дальше.

Выбор одного элемента совокупности называют *простым случайным выбором*, если все элементы совокупности имеют равные вероятности быть выбранными. Если есть совокупность конечного объема N, то это означает, что для каждого элемента вероятность выбора равна $1/N$.

Выбор n элементов из генеральной совокупности N элементов называют *простым случайным выбором*, если все наборы из n элементов имеют одинаковую вероятность быть выбранными. В этом случае говорят о случайной выборке объема n. Случайную выборку объема n можно получить, извлекая из генеральной совокупности по одному элементу последовательно и чисто случайно. Однако следует понимать, что только для совокупности достаточно больших объемов N извлечение одного или нескольких элементов мало меняет вероятности выбора. Следовательно, на практике случайности выборки можно добиться лишь приближенно.

В настоящее время достаточно надежными методами простого случайного отбора являются: а) метод, использующий соответствующую программу персонального компьютера; б) метод, использующий таблицу случайных чисел. В последнем случае элементы генеральной совокупности соответствующим образом перенумеровываются и из таблицы случайных чисел, открытой на произвольной странице, из взятой наугад строки выписываются номера элементов, попадающих в выборку.

Значение простого случайного выбора в том, что с помощью набора зарегистрированных значений исследуемого признака ξ для случайной выборки n элементов совокупности можно дать вероятностную оценку характеристикам признака ξ во всей генеральной совокупности. Этим обстоятельством случайная выборка принципиально отличается от субъективно организованного отбора, от преднамеренного, неслучайного отбора, от отбора, построенного лишь на части генеральной совокупности, и т. д. При отборах такого рода невозможно никакое объективное использование результатов выборочных наблюдений для оценки характеристик распределения признака ξ во всей генеральной совокупности. Полученная оценка

будет смещенной по отношению к истинному значению оцениваемой характеристики. При простом случайном отборе точность оценивания тем лучше, чем больше объем выборки. При заданной точности оценивания можно определить необходимый объем выборки. Простой случайный отбор производится из целиком однородной генеральной совокупности.

Однако на практике встречаются разнородные генеральные совокупности, которые целесообразно разбить на отдельные непересекающиеся группы, каждая из которых более однородная, чем вся совокупность. Такие группы называются *слоями* или *стратами*. Например, при аттестации знаний выпускников средних школ Московской области целесообразно разделить школы на городские и сельские. В таких случаях применяется *типологический* (или *стратифицированный*) *случайный отбор*. Он состоит в том, что всю совокупность объектов разбивают на однородные группы и затем к каждой группе применяется простой случайный отбор. Обычно на практике берут объемы выборок из разных групп либо одинаковыми, либо пропорциональными с одним и тем же коэффициентом пропорциональности объемам групп, если только ничего другого не требует используемый статистический метод. Из других способов случайного отбора упомянем лишь о *многоступенчатом случайном отборе*, когда проводится простой случайный отбор в несколько этапов (или ступеней). Пусть, например, изучается распределение роста выпускников школ Московской области. Тогда целесообразно провести сначала случайный отбор школ отдельно среди городских и сельских школ, а затем провести простой случайный отбор отдельно юношей и отдельно девушек из отобранных на первом этапе школ.

Пусть с помощью некоторого случайного отбора выделена группа элементов из генеральной совокупности. Предположим, что наблюдения над случайной величиной ξ можно повторять независимо и в неизменных условиях. Тогда для выделенной группы элементов получаем независимые реализации $x_1, x_2, \ldots, x_n$ случайной величины ξ. С точки зрения теории вероятностей $x_1, x_2, \ldots, x_n$ будут независимыми и одинаково распределенными случайными величинами. Их дальнейшее изучение существенно основано на этом обстоятельстве.

В математической статистике все результаты, описывающие поведение всей генеральной совокупности, получаются с помощью исследования репрезентативной случайной выборки из этой совокупности. Это означает, что все суждения о генеральной совокупности носят вероятностный характер. Всякое такого рода утверждение

может быть верным лишь с некоторой вероятностью, и с некоторой вероятностью оно может оказываться неверным.

Изучение свойств генеральной совокупности на базе одних только свойств случайной выборки из нее без всякой априорной информации о совокупности практически невозможно. Приходится делать определенные допущения о генеральной совокупности, накладывать некоторые условия на нее. Точнее говоря, приходится накладывать ограничения на закон распределения исследуемого случайного признака ξ в генеральной совокупности. Например, исходя из предварительных сведений, можно требовать выполнения нормального закона распределения признака ξ в совокупности с неизвестными параметрами. Аналогичная ситуация имеет место и при исследовании с помощью случайной выборки зависимости двух случайных признаков в генеральной совокупности. Такое исследование может иметь только вероятностный характер и возможно лишь при некоторых априорных допущениях о форме зависимости признаков во всей совокупности. Например, можно считать, что с точностью до случайного остатка два признака в генеральной совокупности связаны линейно с неизвестными параметрами, которые необходимо оценить по данным случайной выборки.

Таким образом, при исследовании поведения случайных признаков ξ в генеральной совокупности приходится накладывать на них определенные математические ограничения. С точки зрения теории вероятностей выбор таких ограничений означает выбор некоторой вероятностной модели, имитирующей поведение случайных признаков. Если выбрана такая вероятностная модель для генеральной совокупности, надежность и точность которой необходимо проверить лишь на основании ограниченного количества данных для репрезентативной случайной выборки из реальной генеральной совокупности, то говорят, что *задана статистическая модель*.

Методы математической статистики позволяют выбирать такую статистическую модель, которая в определенном смысле наилучшим образом соответствует имеющимся статистическим данным, характеризующим реальное поведение конкретной исследуемой системы.

Заметим, что любая статистическая модель, так же как и любая математическая модель вообще, является лишь некоторым приближением к исследуемой реальной действительности. Она имитирует реальную действительность, так как основана на некоторых гипотетических допущениях и дает лишь упрощенное представление реальной действительности.

Статистическая модель реальных явлений, процессов или объектов должна быть описана с помощью задаваемых гипотез и должна быть изучена средствами математики. Результаты такого изучения должны быть интерпретированы (истолкованы) для рассматриваемых реальных явлений, процессов или объектов. Анализ статистических моделей позволяет выделить наиболее существенные черты рассматриваемых явлений, процессов или объектов, находить для них ранее не обнаруженные закономерности и предсказывать их поведение в будущем. Этот анализ проще и быстрее, чем экспериментальное изучение реальных явлений, процессов или объектов в различных условиях.

Математическая статистика изучает ряд типовых статистических моделей, поэтому при математической формализации эксперимента следует стараться сводить дело к стандартной статистической модели, для которой разработана подробная теория.

Чтобы можно было говорить о методах обработки наблюдений психологического признака ξ для случайной выборки n элементов генеральной совокупности, необходимо уметь измерять значения этого признака. С этой целью используются разные шкалы. В зависимости от выбранной шкалы для измерения признака выбирается метод обработки полученных наблюдений.

§ 2. Измерение психологических признаков

Исходным материалом для всякого статистического исследования в психологии и педагогике являются результаты опытов, которые проводятся по специальным измерительным процедурам. Результаты опытов для исследуемых психологических признаков могут получать количественные выражения. Чтобы правильно истолковать результаты таких опытов, необходимо понять, как измеряются значения психологических признаков. Если в естественных науках и в технике существуют стандартные единицы измерения, например метр, минута, градус, вольт и т. д., то в психологии признаки в общем случае не имеют собственных единиц измерения. Поэтому значения психологических признаков определяются при помощи специальных измерительных шкал.

Согласно С. С. Стивенсу (1951), в зависимости от природы психологического признака его значения могут измеряться при помощи одной из следующих четырех шкал измерения: номинальной шкалы, порядковой шкалы, интервальной шкалы и шкалы отношений.

Наиболее простой является *номинальная шкала*. Ее еще называют номинативной шкалой или шкалой наименований. Процедура

измерения в номинальной шкале состоит в классификации (группировке) объектов (индивидов) таким образом, что объекты одного класса (группы) однородны (одинаковы) по анализируемому признаку или свойству, тогда как объекты из разных классов (групп) различаются по анализируемому признаку или свойству.

Например, группу людей можно разбить на две группы по полу или по семейному положению, на несколько групп по цвету глаз или волос, по социальному положению, по месту проживания, по образованию, по профессии и т. д. Множество студентов университета можно разбить по факультетам, курсам или группам. Множество учеников города можно разбить по школам или классам.

Простейшим примером применения номинальной шкалы является измерение дихотомического (или двузначного) признака, т. е. признака, множество значений которого состоит лишь из двух различных значений. Например: «мальчик — девочка», «левша — нелевша», «оценка удовлетворительная — оценка неудовлетворительная», «флегматик — нефлегматик», «ответ «да» на вопрос — ответ «нет» на вопрос», «человек «экстраверт» — человек «интроверт» и т. п.

Классы, на которые номинальная шкала разбивает значения исследуемого признака, являются непересекающимися. В каждом классе можно подсчитать частоту признака, т. е. число испытуемых (объектов), попавших в данный класс и обладающих данным свойством. Например, для вычисления частоты дихотомического признака «левша — нелевша» в каждом из двух классов кодируем (присвоим число) «левшу» числом 1, а «нелевшу» кодируем числом 0. После этого отдельно подсчитываем общее количество единиц и нулей. Ясно, что вместо единицы и нуля для кодирования можно использовать и буквы, например *А* и *Б*.

Числа и буквы применяются лишь для указания различия в классах. Никакие действия над числами в таких случаях (упорядочение, сложение, вычитание, умножение, деление) не имеют никакого смысла по отношению к самим испытуемым (объектам). Например, в случае признака «левша — нелевша» нельзя сказать, что «левша» имеет большее значение, чем «нелевша». Для описания значений признака, измеренного в номинальной шкале, используются пропорции и проценты. Например, говорят, что в данной группе детей числа девочек и мальчиков относятся как 4 : 1 или что в данной группе детей 80% девочек и 20% мальчиков. Если число групп разбиения более двух, то можно указать группу с наибольшей частотой измеренного признака и группу с наименьшей частотой измеренного признака.

Более сложной является *порядковая* (или *ранговая*) *шкала*, которая для значений исследуемого признака вводит упорядоченность, направленность, степень важности или степень проявления признака. Ряд рассматриваемых индивидов или объектов порядковая шкала упорядочивает в соответствии с наблюденными для них значениями признака или свойства. Сам процесс такого упорядочения называется ранжированием. При этом каждому индивиду (объекту) присваивается некоторое, как правило, положительное число, называемое рангом, и тогда можно сравнивать между собой разные индивиды (или объекты) по величине их рангов. Если ранг объекта *А* больше ранга объекта *Б*, то *А* имеет большее значение признака, чем *Б*. Например, группу учащихся можно ранжировать в соответствии с баллами при тестировании их математических способностей, или группу спортсменов, участвующих в некотором соревновании, можно ранжировать соответственно занятым ими итоговым местам в соревновании, или группу кандидатов, претендующих на некоторую выборную должность, можно ранжировать согласно количеству поданных за них голосов.

Например, если при тестировании невербального интеллекта испытуемые *А*, *Б*, *В*, *Г*, *Д* получили соответственно баллы 64, 65, 53, 70, 52, то, если большему числу ставить больший ранг, получаем для испытуемых ряд рангов 3, 4, 2, 5, 1. Ранжирование можно проводить, ставя большему числу меньший ранг. Тогда получаем ряд рангов 3, 2, 4, 1, 5. В дальнейшем ограничимся первым способом ранжирования.

Иногда возникают ситуации, когда двум или более испытуемым (объектам) присваивают одинаковые ранги. Такие ранги называют связными (или объединенными) рангами. Пусть, например, при тестировании уровня тревожности семи испытуемых были получены баллы 24, 24, 24, 25, 30, 30, 32. Тогда каждому из первых трех испытуемых присваивается один и тот же ранг, равный

$$\frac{1+2+3}{3}=2,$$

четвертый испытуемый имеет ранг 4, пятому и шестому испытуемым присваивается ранг

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

а последний испытуемый имеет ранг 7. Получается следующий ряд рангов: 2; 2; 2; 4; 5,5; 5,5; 7.

Таким образом, в том случае, если несколько заданных величин являются равными, каждой такой величине приписывается ранг, равный среднему арифметическому тех рангов, которые эти вели-

чины получили бы, если бы они стояли по порядку друг за другом и не были бы равны.

Если в номинальной шкале были лишь два состояния признака: тождество ($x = y$) и различие ($x \neq y$), то в порядковой шкале дополнительно используются неравенства ($x > y$) и ($x < y$). Например, ($x > y$) может означать, что объект, обладающий значением x, важнее объекта, обладающего значением y, или престижнее, привлекательнее и т. д. Над рангами можно производить арифметические операции (сложение, вычитание, умножение, деление), однако их результаты не имеют реального смысла для испытуемых (или объектов), которых ранжируют. Например, нельзя сказать, что ученик, получивший по математике четверку, знает математику на единицу лучше, чем ученик, получивший по математике тройку. Точно так же не имеет смысла вычислять и сравнивать средние значения тестовых баллов, например школьных оценок.

Интервальной шкалой называют такую шкалу, в которой о двух сравниваемых индивидах (или объектах) можно сказать не только, одинаковы они или различны (как в номинальных шкалах), не только в каком из них признак выражен больше (как в порядковых шкалах), но и насколько больше этот признак выражен. Эта шкала устанавливает равные различия (равные интервалы) для индивидов (или объектов). Для интервальной шкалы выбирается единица измерения, и каждому индивиду (или объекту) присваивается число, равное количеству единиц измерения, эквивалентному количеству имеющегося в нем свойства. Равным разностям чисел соответствуют равные разности значений измеряемого признака. Измерение календарного времени или температуры дает примеры интервальной шкалы. Например, сопоставляя календарные даты двух событий, можно сказать, насколько (лет, дней, часов) одно событие произошло раньше или позже другого. В интервальной шкале задается точка отсчета — нуль шкалы. Однако нуль интервальной шкалы не означает отсутствия рассматриваемого признака или свойства в объекте. Например, нуль градусов по Цельсию не означает отсутствия температуры вообще, это температура замерзания воды. Возможны и температуры ниже нуля градусов по Цельсию.

Интервальные шкалы часто используются в психологии. Примерами измерения в интервальной шкале являются измерения различных психологических характеристик личности, социальных установок, ценностных ориентаций и др.

Однако отношение двух измерений по интервальной шкале в большинстве случаев не имеет смысла. Так, например, нельзя ска-

зать, что температура в двадцать градусов Цельсия в два раза больше температуры в десять градусов Цельсия.

Шкала отношений отличается от интервальной шкалы прежде всего тем, что начало отсчета — нуль шкалы отношений указывает на полное отсутствие измеряемого признака или свойства. По результатам двух измерений в шкале отношений можно найти отношение этих результатов и определить во сколько раз один объект превосходит другой по степени выраженности измеряемого признака. Таким образом, шкала отношений носит свое название из-за того, что отношения шкальных значений эквивалентны отношению значений исследуемого признака или свойства. Шкалами отношений являются большинство шкал, применяемых в физике и технике. В шкале отношений измеряются, например, длина, рост, вес, угол, площадь, объем, денежный доход, объем продукции, выпускаемой предприятием, и т. д. Шкала отношений редко используется в психологии, так как наличие нулевой точки, когда то или иное психологическое свойство полностью отсутствует, является проблематичным. Однако измерение в шкале отношений проводится в родственных к психологии науках, таких как психофизика, психогенетика, психофизиология.

Интервальную шкалу и шкалу отношений называют *количественными шкалами*, а признаки, измеряемые в количественных шкалах, — *количественными*.

Если измерения проведены в количественной шкале, то результаты измерений можно преобразовать в измерения по номинальной или порядковой шкалам.

До сих пор речь шла об измерении количественных одномерных психологических признаков. Разумеется, что и для количественных многомерных признаков используются количественные шкалы. В зависимости от того, в какой шкале измерены значения исследуемого признака, определяется выбор статистического метода для анализа экспериментальных данных в психологии. Наиболее разнообразны статистические методы для анализа психологических признаков, измеренных в шкале отношений.

§ 3. Первоначальная обработка наблюдений случайной выборки

Пусть в результате случайного отбора объектов из некоторой генеральной совокупности получена выборка объема n значений исследуемой количественной случайной величины ξ: $x_1, x_2, \ldots, x_n$.

Этот набор n конкретных чисел представляет собой, как правило, беспорядочную массу материала, и ему необходимо придать определенную форму и ясную структуру, чтобы извлечь из него необходимую информацию.

Прежде всего, наблюдения, составляющие выборку, располагают в порядке их возрастания, т. е. строят так называемый вариационный ряд. *Вариационный ряд* для наблюдений $x_1, x_2, \ldots, x_n$ обозначают так: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, где $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$.

Пусть, например, при опросе случайно выбранной группы школьников на уроке литературы получены оценки: 5, 3, 5, 4, 3, 4, 2, 4, 5. Тогда вариационный ряд для полученных оценок будет следующий: 2, 3, 3, 4, 4, 4, 5, 5, 5.

Имея вариационный ряд, мы получаем информацию о наименьшем значении $x_{(1)}$ выборки, о наибольшем значении $x_{(n)}$ выборки и о разности между ними, называемой выборочным размахом R: $R = x_{(n)} - x_{(1)}$.

Для нашего примера $x_{(1)} = 2$, $x_{(9)} = 5$, $R = 5 - 2 = 3$.

В том случае, когда объем наблюдений выборки небольшой, находят разные наблюдения и указывают их частоту. Полученные данные записывают в так называемую *таблицу частот дискретного вариационного ряда* следующего вида:

$x^{(1)}$	$x^{(2)}$	$\ldots$	$x^{(k)}$
n_1	n_2	$\ldots$	n_k

Здесь $x^{(1)} < x^{(2)} < \ldots < x^{(k)}$, $1 \leqslant k \leqslant n$, n_1 — частота $x^{(1)}$, n_2 — частота $x^{(2)}$, $\ldots$, n_k — частота $x^{(k)}$, причем $n_1 + n_2 + \ldots + n_k = n$.

Для вышеприведенного примера имеем таблицу следующего вида:

2	3	4	5
1	2	3	3

По этой таблице можно определить относительную частоту каждого значения.

Если объем выборки небольшой, то на этом заканчивается грубый первоначальный анализ выборки. Однако при большом объеме выборки подобный анализ не позволяет все-таки полностью понять общий смысл данных выборки. В таких случаях переходят к так называемым «группированным данным». Обычно этот переход делается так:

а) отмечают сначала наименьшее и наибольшее значения в выборке,

б) весь диапазон между наименьшим и наибольшим значениями разбивают на определенное число равных интервалов, при этом количество интервалов должно быть в пределах 7—20,

в) отмечаются крайние точки каждого из интервалов в порядке возрастания, а также их середины,

г) подсчитывается число выборочных данных, попавших в каждый из интервалов, т. е. частота по интервалам, при этом, данные, попавшие на границу двух интервалов, условливаются относить только к левому интервалу.

В зависимости от конкретного содержания задачи эту схему группирования можно видоизменять, например, можно отказаться от требования равных длин интервалов или крайние интервалы делать полубесконечными и т. д.

Пример 1. Пусть измерения роста группы школьников младшего возраста дали следующие результаты (в см): 132, 132, 133, 134, 101, 134, 135, 105, 109, 138, 138, 110, 111, 140, 115, 125, 127, 115, 116, 127, 127, 116, 117, 127, 127, 117, 128, 117, 118, 130, 119, 131, 143, 124, 124, 144, 146, 124, 125, 150, 124, 158, 125, 121, 122, 121.

Разобьем все множество результатов измерения роста на 7 интервалов: [101, 109), [109, 117), [117, 125), [125, 133), [133, 141), [141, 149), [149, 158]. Для каждого интервала найдем его середину и частоту, т. е. число наблюдений, попавших в интервал. Ясно, что сумма всех частот по интервалам равна объему выборки. Кроме того, найдем и относительные частоты (частости) в интервалах, которые получаются путем деления частоты в интервале на сумму всех частот, которая в рассматриваемом примере равна 46. Ясно, что сумма всех относительных частот равна единице. Относительная частота в интервале дает процент попадания в интервал результатов измерения из их общего числа.

Для рассматриваемого примера получаем следующую таблицу.

Интервал	Середина интервала	Частота в интервале	Относительная частота в интервале
[101,109)	105	2	0,0435
[109,117)	113	7	0,152
[117,125)	121	12	0,261
[125,133)	129	13	0,283
[133,141)	137	7	0,152
[141,149)	145	3	0,065
[149,158]	153	2	0,0435

Таблица, в которой приведены все интервалы с соответствующими частотами по интервалам для заданной выборки наблюдений, называется *таблицей частот интервального вариационного ряда*.

Распределения частот и относительных частот по интервалам можно представить не только в виде таблиц, но и графически. Графическое изображение данных интервального вариационного ряда строят в виде гистограммы частот или полигона частот.

Гистограмма частот изображается так: над каждым интервалом строится прямоугольник, основанием которого служит данный интервал, а высотой — частота в данном интервале. Как правило, для удобства рассмотрения единицы масштаба по оси абсцисс и по оси ординат выбираются разными. Кроме того, и начала отсчета по разным осям тоже могут не совпадать. Гистограмма частот для рассматриваемого примера показана на рис. 8.

Если по оси ординат откладывать не частоты в интервалах, а относительные частоты в интервалах, то подобным образом можно построить *гистограмму относительных частот*.

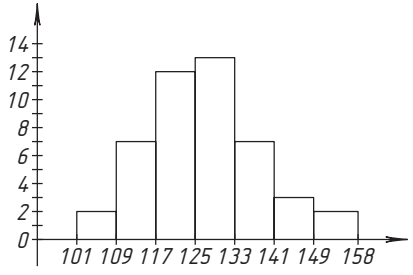


Рис. 8

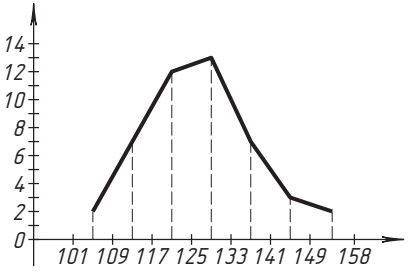


Рис. 9

Полигон частот для интервального вариационного ряда изображается так: в середине каждого интервала строится ордината, равная частоте на этом интервале, и полученные точки соединяются. Полигон частот для рассматриваемого примера показан на рис. 9.

Если же строить в середине каждого интервала ординату, равную относительной частоте на этом интервале, и соединить полученные точки, то получим *полигон относительных частот*. ■

Полигоны частот и относительных частот можно, очевидно, строить и для дискретного вариационного ряда. Однако для дискретного вариационного ряда можно также ввести понятие *выборочной функции распределения* $F_n(x)$.

Пусть таблица дискретного вариационного ряда имеет вид

x_1	x_2	…	x_k
n_1	n_2	…	n_k

где $x_1 < x_2 < \ldots < x_k$, $1 \leqslant k \leqslant n$, n_1 — частота x_1, n_2 — частота x_2, … …, x_k — частота x_k, причем $n_1 + n_2 + \ldots + n_k = n$.

Выборочной функцией распределения $F_n(x)$ называется ступенчатая функция следующего вида:

$$F_n(x) = \begin{cases} 0, & x \leqslant x_1, \\ \dfrac{n_1}{n}, & x_1 < x \leqslant x_2, \\ \dfrac{n_1 + n_2}{n}, & x_2 < x \leqslant x_3, \\ \dfrac{n_1 + n_2 + n_3}{n}, & x_3 < x \leqslant x_4, \\ \quad \ldots \ldots \ldots \ldots \ldots \ldots \\ \dfrac{n_1 + n_2 + \ldots + n_{k-1}}{n}, & x_{k-1} < x \leqslant x_k, \\ 1, & x > x_k. \end{cases}$$

Выборочная функция $F_n(x)$ является постоянной на каждом интервале (x_p, x_{p+1}), а в каждой точке x_p увеличивается на величину $\dfrac{n_p}{n}$, $p = 1, 2, \ldots, k$. Кроме того, $F_n(x)$ — неубывающая функция, $0 \leqslant F_n(x) \leqslant 1$, $F_n(-\infty) = 0$, $F_n(+\infty) = 1$.

Таким образом, $F_n(x)$ обладает теми же свойствами, что и функция распределения $F(x)$ случайной величины ξ.

Иногда функцию $F_n(x)$ называют *функцией накопленных относительных частот*. Объяснение этого названия непосредственно следует из определения $F_n(x)$.

График $F_n(x)$ называют *кумулятивной кривой*.

Пример 2. Пусть таблица частот дискретного вариационного ряда имеет вид

1	2	4	6
2	1	2	1

Здесь $x_1 = 1$, $n_1 = 2$, $x_2 = 2$, $n_2 = 1$, $x_3 = 4$, $n_2 = 2$, $x_3 = 6$, $n_3 = 1$.

В примере $n = n_1 + n_2 + n_3 = 6$. Построим выборочную функцию распределения $F_6(x)$ и нарисуем ее график.

Из определения $F_n(x)$ получаем, что

$$F_6(x) = \begin{cases} 0 & \text{при } x \leqslant 1, \\ \frac{2}{6} & \text{при } 1 < x \leqslant 2, \\ \frac{3}{6} & \text{при } 2 < x \leqslant 4, \\ \frac{5}{6} & \text{при } 4 < x \leqslant 6, \\ 1 & \text{при } x > 6. \end{cases}$$

График $F_6(x)$ показан на рис. 10. На нем единицы масштаба по оси абсцисс и по оси ординат разные. График нарисован жирной линией, а стрелка на графике указывает на то, что точка на ее острие не принадлежит графику. ■

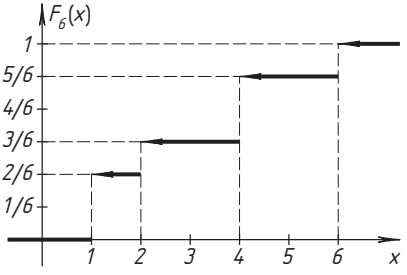


Рис. 10

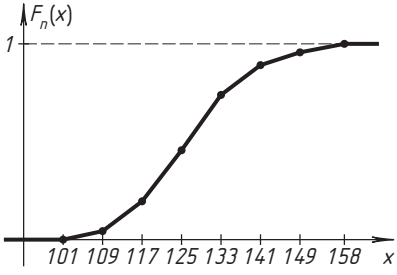


Рис. 11

Для интервального вариационного ряда также можно построить выборочную функцию распределения $F_n(x)$, если такой ряд условно заменить дискретным. В этом случае правую границу i-го интервала принимают за значение x_i, а соответствующую интервальную частоту принимают за частоту n_i значения x_i. Затем соединяют точки графика, соответствующие концам интервалов, отрезками прямой. В результате график выборочной функции распределения $F_n(x)$ для интервального вариационного ряда будет представлять собой непрерывную линию.

График $F_n(x)$ для рассмотренного в примере 1 интервального вариационного ряда показан на рис. 11. Единицы масштаба по осям взяты разные.

Построением таблиц распределения частот и относительных частот вариационного ряда и графическим изображением данных этих таблиц, завершается первоначальная обработка наблюдений случайной выборки.

§ 4. Основные выборочные характеристики и их свойства

Пусть с помощью таблиц распределения частот вариационного ряда установлена основная структура статистических данных. На следующем этапе статистического анализа необходимо указать небольшое число показателей, при помощи которых можно в сжатом виде охарактеризовать все распределение частот наблюдений случайной выборки.

Такими основными показателями являются: выборочная функция распределения, выборочные меры центра распределения наблюдений, выборочные меры рассеяния (или изменчивости) наблюдений, выборочная мера асимметрии и выборочная мера эксцесса (островершинности) наблюдений.

О выборочной функции распределения уже говорилось в § 2.

1. Выборочные меры центра распределения

К выборочным мерам центра распределения наблюдений случайной выборки принято относить выборочное среднее значение, выборочную медиану и выборочную моду.

Выборочное среднее значение — наиболее часто используемый показатель центра распределения наблюдений. Для наблюдений $x_1, x_2, \ldots, x_n$ его обычно обозначают $\overline{x}$ (или $\overline{x}(n)$) и определяют формулой

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Если воспользоваться знаком $\sum$, то $\overline{x}$ записывают в виде

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Например, если в результате опроса девяти учеников на уроке литературы получены оценки 5, 3, 5, 4, 3, 4, 2, 4, 5, то среднее значение оценок

$$\overline{x} = \frac{5+3+5+4+3+4+2+4+5}{9} = \frac{35}{9} \approx 4.$$

Для дискретного вариационного ряда (см. § 3) можно использовать такую формулу:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i \cdot x_i.$$

По этой формуле для предыдущего примера имеем

$$\overline{x} = \frac{1}{9}(2 \cdot 1 + 3 \cdot 2 + 4 \cdot 3 + 5 \cdot 3) = \frac{35}{9} \approx 4.$$

Для выборок большого объема выборочное среднее $\overline{x}$ можно находить после построения интервального вариационного ряда по формуле

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i \cdot \overline{x}_i,$$

где n — общее число наблюдений, k — число интервалов, $\overline{x_i}$ — середина i-го интервала, n_i — частота в i-м интервале.

Например, для интервального вариационного ряда из примера 1 § 3 получаем по этой формуле, что

$$\overline{x} = \frac{1}{46}(2 \cdot 105 + 7 \cdot 113 + 12 \cdot 121 + 13 \cdot 129 + 7 \cdot 137 + 3 \cdot 145 + 2 \cdot 153) =$$

$$= \frac{5830}{46} \approx 126{,}74.$$

Следует заметить, что формула $\overline{x}$ для интервального вариационного ряда рекомендуется лишь для симметричного или умеренно симметричного распределения частот по интервалам.

Отметим некоторые *свойства выборочного среднего*, применяемые на практике.

1) Если наблюдения выборки имеют вид $cx_1, cx_2, \ldots, cx_n$, где c — заданное число, то среднее значение имеет вид $c \cdot \overline{x}$, где $\overline{x}$ — среднее значение для $x_1, x_2, \ldots, x_n$.

2) Если наблюдения выборки имеют вид $x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n$, то среднее значение имеет вид $\overline{x} + \overline{y}$, где $\overline{x}$ — среднее значение для наблюдений $x_1, x_2, \ldots, x_n$, а $\overline{y}$ — среднее значение для наблюдений $y_1, y_2, \ldots, y_n$.

3) Если наблюдения выборки объема n являются объединением наблюдений двух выборок объемов n_1 и n_2 , то выборочное среднее $\overline{x}$ объединенной выборки выражается через выборочные средние $\overline{x_1}$ и $\overline{x_2}$ ее составных частей по формуле

$$n\overline{x} = n_1 \cdot \overline{x_1} + n_2 \cdot \overline{x_2}.$$

4) Сумма отклонений результатов наблюдений от среднего значения $\overline{x}$ равна нулю, т. е. $\sum_{i=1}^{n} (x_i - \overline{x}) = 0$.

5) Сумма квадратов отклонений результатов наблюдений от среднего значения $\overline{x}$ меньше, чем сумма квадратов отклонений от любого другого значения x, т. е. $\sum_{i=1}^{n}(x_i-\overline{x})^2 < \sum_{i=1}^{n}(x_i-x)^2$ при всех $x \neq \overline{x}$.

Выборочное среднее значение $\overline{x}$ является наиболее важной характеристикой наблюдений выборки. Оно является центром рассеяния данных выборки. При сравнении двух выборок значений изучаемого признака, прежде всего, сравнивают между собой средние значения для этих выборок.

Если для наблюдений выборки $x_1, x_2, \ldots, x_n$ построен дискретный вариационный ряд $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, где $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$, то *выборочной медианой* x_{med} или $x_{\text{med}}(n)$ называется такое число, которое делит вариационный ряд на две равные части. При этом x_{med} может быть наблюдением выборки, а может им не быть. Если объем n выборки нечетное число, т. е. $n = 2k+1$, то $x_{\text{med}} = x_{(k+1)}$. Если же объем n выборки четное число, т. е. $n = 2k$, то $x_{\text{med}} = \frac{1}{2}(x_{(k)} + x_{(k+1)})$. В этом случае x_{med} может не быть наблюдением выборки. Например, для выборки 1, 2, 3, 4, 5 имеем $x_{\text{med}} = 3$, а для выборки 1, 2, 3, 4, 5, 6 имеем $x_{\text{med}} = \frac{1}{2}(3+4) = 3{,}5$. Имеется также формула приближенного вычисления выборочной медианы x_{med} и для интервального вариационного ряда.

Выборочная медиана в ряде случаев имеет некоторые преимущества по сравнению с выборочным средним. Она может быть вычислена в случаях, когда вычисление выборочного среднего не возможно, например, вследствие неопределенности первого или последнего интервала группировки данных выборки в интервальном вариационном ряде.

Кроме того, выборочную медиану легко вычислять и она более устойчива по сравнению с выборочным средним при резких выбросах в наблюдениях выборки. Однако выборочная медиана имеет и свои недостатки по сравнению с выборочным средним. Например, для выборочной медианы невозможны аналоги свойств 1)—5) для выборочного среднего. Другие различия между $\overline{x}$ и x_{med} будут ясны далее.

К выборочным характеристикам центра группирования, условно относят и выборочную моду x_{mod}. *Выборочной модой* x_{mod} называется наиболее часто встречающееся значение среди наблюдений выборки.

Например, для выборки 1, 2, 5, 5, 5, 4, 4 мода $x_{\text{mod}} = 5$, а для выборки 1, 2, 5, 5, 5, 4, 4, 4 имеем две моды $x_{\text{mod}}^{(1)} = 5$, $x_{\text{mod}}^{(2)} = 4$. Таким образом, в отличие от $\overline{x}$ и x_{med} выборочная мода x_{mod} может быть не единственной. В таких случаях говорят об одномодальном, двухмодальном и т. д. распределении наблюдений выборки.

Выбор $\overline{x}$, x_{med} и x_{mod} зависит, с одной стороны, от природы данных, а с другой — от того, как этот показатель будет использоваться. Для дальнейшего теоретико-вероятностного и статистического анализа наиболее полезно $\overline{x}$.

2. Выборочные меры рассеяния

К выборочным мерам рассеяния относятся: выборочная дисперсия, выборочное среднее квадратическое отклонение, выборочный коэффициент вариации, выборочное квартильное отклонение.

Наиболее часто употребляемыми мерами рассеяния наблюдений выборки относительно центра группирования являются *выборочная дисперсия*, обозначаемая s^2 (или $s^2(n)$) и *выборочное среднее квадратическое* (или *стандартное*) *отклонение* s (или $s(n)$). Если наблюдения $x_1, x_2, \ldots, x_n$ имеют выборочное среднее $\overline{x}$, то по определению

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \qquad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} > 0.$$

Пример 1. Пусть заданы числа 3, 5, 7, 9, 11. Тогда: $\overline{x} = \frac{1}{5}(3+5+7+9+11) = 7$, $s^2 = \frac{1}{5}[(3-7)^2 + (5-7)^2 + (7-7)^2 + (9-7)^2 + (11-7)^2] = \frac{1}{5}[16+4+0+4+16] = 8$, $s = \sqrt{8} \approx 2{,}82$. ■

Для данных, сгруппированных в интервальный вариационный ряд, формула выборочной дисперсии выглядит так: $s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2$. Здесь n — общее число наблюдений, k — число интервалов, $\overline{x}$ — выборочное среднее всех данных, n_i — частота в i-м интервале, x_i — центр i-го интервала.

Выборочные дисперсия и стандартное отклонение выражаются в единицах измерения рассматриваемого количественного признака и непригодны, например, при сравнении степеней рассеяния наблюдений двух выборок различной природы. Для сравнения таких степеней рассеяния используется *выборочный коэффициент вариации*

$$\widehat{V} = \frac{s}{\overline{x}} \cdot 100\%.$$

Коэффициент $\widehat{V}$ используется также в тех случаях, когда степень рассеяния естественно описывать некоторой характеристикой в соответствии со средним значением.

Пример 2. Пусть заданы две таблицы распределений частот:

2	3	5	7
1	2	1	1

12	13	15	17
1	2	1	1

Требуется для них сравнить степени рассеяния.

Для первой таблицы имеем:

$$\overline{x}_1 = \frac{1}{5}(2\cdot 1 + 3\cdot 2 + 5\cdot 1 + 7\cdot 1) = 4, \quad s_1^2 = 3{,}2, \quad s_1 \approx 1,\,79,$$

$$\widehat{V}_1 \approx \frac{1{,}79}{4}\cdot 100\% \approx 45\%.$$

Для второй таблицы имеем:

$$\overline{x}_2 = \frac{1}{5}(12\cdot 1 + 13\cdot 2 + 15\cdot 1 + 17\cdot 1) = 14, \quad s_2^2 = 3{,}2, \quad s_2 \approx 1{,}79,$$

$$\widehat{V}_2 \approx \frac{1{,}79}{14}\cdot 100\% \approx 13\%.$$

Таким образом, хотя абсолютные степени рассеяния для обеих таблиц приблизительно одинаковы, относительные степени рассеяния у них значительно отличаются. ■

Выбор показателя рассеяния диктуется используемым показателем центра распределения. Если используется выборочное среднее $\overline{x}$, то выбираются выборочные дисперсия или среднее квадратическое отклонение.

При применении выборочной медианы как меры центра распределения показателем рассеяния будет служить квартильное отклонение. Рассмотрим такие числа Q_1 и Q_3, что Q_1, x_{med} и Q_3 разбивают все множество ранжированных (т. е. расположенных в порядке возрастания) наблюдений выборки на 4 равночисленные группы. Тогда Q_1 называется *нижней выборочной квартилью*, а Q_3 — *верхней выборочной квартилью*.

Квартильное отклонение Q определяют формулой

$$Q = \frac{1}{2}(Q_3 - Q_1).$$

Положение Q_1 определяется номером, ближайшим к числу $\frac{n+1}{4}$, а положение Q_3 определяется номером, ближайшим к числу $\frac{3(n+1)}{4}$.

Пусть, например, известны результаты (в баллах) 18 тестовых испытаний общих умственных способностей учащихся: 25, 28, 39, 39, 39, 45, 50, 52, 52, 68, 69, 69, 70, 70, 70, 70, 74, 74. Требуется определить квартили Q_1, Q_3, медиану x_{med} и квартильное отклонение Q.

Имеем $x_{\text{med}} = \frac{1}{2}(52 + 68) = 60$.

Позиция Q_1 определяется числом $\frac{n+1}{4} = \frac{19}{4}$, т. е. номером 5. Значит, $Q_1 = 39$. Позиция Q_3 определяется числом $\frac{3(n+1)}{4} = \frac{3 \cdot 19}{4} =$ $= \frac{57}{4}$, т. е. номером 14. Значит $Q_3 = 70$.

Следовательно, $Q = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(70 - 39) = 15{,}5$.

Медиана x_{med} представляет собой *центральную квартиль*, которая разделяет все исходные данные на две равные части.

Иногда разбивают все исходные данные не на 4 равные части с помощью квартилей, а на 10 равных частей с помощью децилей или на 100 равных частей с помощью процентилей.

3. Выборочные коэффициенты асимметрии и эксцесса

На практике также значительный интерес представляет собой степень отклонения одномодального распределения частот данных выборки от симметричного распределения. Если имеется асимметрия (т. е. несимметричность) распределения частот, то в этом наглядно можно убедиться по гистограмме и полигону частот. В этом случае одна из их ветвей, начиная от вершины, более длинная, чем другая. При симметричном распределении частот выборочные среднее значение, медиана и мода имеют приблизительно одинаковые значения. В случае длинной правой ветви гистограммы и полигона говорят о положительной асимметрии распределения частот наблюдений, а в случае более длинной левой ветви говорят об отрицательной асимметрии наблюдений. Для положительной асимметрии среднее значение больше медианы, которая в свою очередь больше моды. Для отрицательной асимметрии среднее значение меньше медианы, которая в свою очередь меньше моды.

Существует несколько коэффициентов для расчета асимметрии наблюдений выборки. Наиболее точным из них является *выборочный коэффициент асимметрии* $\widehat{A}$ (или $\widehat{A}(n)$) следующего вида:

$$\widehat{A} = \frac{1}{ns^3} \sum_{i=1}^{n} (x_i - \overline{x})^3,$$

где x_1, x_2, …, x_n — рассматриваемые наблюдения, $\overline{x}$ — выборочное среднее значение, s — выборочное стандартное отклонение. Например, для рассмотренной в примере 2 таблицы распределений частот

2	3	5	7
1	2	1	1

было получено, что $\overline{x} = 4$, $s \approx 1{,}79$. Тогда выборочный коэффициент асимметрии

$$\widehat{A} = \frac{1}{5 \cdot (1{,}79)^3}[(2-4)^3 \cdot 1 + (3-4)^3 \cdot 2 + (5-4)^3 \cdot 1 + (7-4)^3 \cdot 1] \approx 0{,}63.$$

Существует еще одна характеристика одновершинного распределения частот наблюдений выборки, указывающая на степень концентрации наблюдений около выборочного среднего значения $\overline{x}$. Это *выборочный коэффициент эксцесса* $\widehat{E}$ (или $\widehat{E}(n)$). Он задается формулой следующего вида:

$$\widehat{E} = \frac{1}{n \cdot s^4} \sum_{i=1}^{n} (x_i - \overline{x})^4 - 3,$$

где x_1, x_2, …, x_n — наблюдения выборки, $\overline{x}$ — выборочное среднее значение, s — выборочное стандартное отклонение.

Например, для той же таблицы распределения частот, для которой выше вычислялся $\widehat{A}$, получаем по этой формуле, что

$$\widehat{E} = \frac{1}{5 \cdot (1,\,79)^4}[(2-4)^4 \cdot 1 + (3-4)^4 \cdot 2 + (5-4)^4 \cdot 1 + (7-4)^4 \cdot 1] - 3 \approx -1{,}05.$$

Для нормального распределения коэффициент $\widehat{E} = 0$. В случае $\widehat{E} > 0$ полигон частот имеет более острую вершину, чем у нормального распределения, а в случае $\widehat{E} < 0$ полигон частот имеет более плоскую вершину, чем у нормального распределения.

4. Выборочные характеристики двумерной случайной величины

Если в каждом наблюдении определяются значения не одной случайной величины, а значения двух (или нескольких) случайных величин одновременно, то получаем двумерную (или многомерную) выборку. Пусть, например, получена случайная выборка значений двух случайных величин X и Y: (x_1, y_1); (x_2, y_2); …; (x_n, y_n).

Для характеристики такой выборки вводятся понятия выборочной ковариации и выборочного коэффициента корреляции Пирсона.

Пусть $\overline{x}$ — выборочное среднее значение и s_x — выборочное стандартное отклонение наблюдений $x_1, x_2, \ldots, x_n$, а $\overline{y}$ — выборочное среднее значение и s_y — выборочное стандартное отклонение наблюдений $y_1, y_2, \ldots, y_n$. Тогда *выборочная ковариация* $\widehat{\mathrm{cov}}(X, Y)$ определяется формулой

$$\widehat{\mathrm{cov}}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}),$$

а *выборочный коэффициент корреляции* $\widehat{r}$ (или $\widehat{r}(n)$) *Пирсона* определяется формулой

$$\widehat{r} = \frac{\widehat{\mathrm{cov}}(X, Y)}{s_x \cdot s_y}.$$

Эти формулы можно упростить следующим образом:

$$\widehat{\mathrm{cov}}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \overline{x}y_i - \overline{y}x_i + \overline{x}\overline{y}) =$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\cdot\overline{y} - n\overline{x}\cdot\overline{y} + n\overline{x}\cdot\overline{y}\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\cdot\overline{y}\right),$$

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\overline{x}x_i + \overline{x}^2)} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)}.$$

Аналогично, $s_y = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)}$.

Тогда формула выборочного коэффициента корреляции Пирсона принимает вид:

$$\widehat{r} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\cdot\overline{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)}}.$$

Свойства выборочного коэффициента корреляции $\widehat{r}$ аналогичны свойствам коэффициента корреляции $r(X, Y)$, в частности, $|\widehat{r}| \leqslant 1$ для любой выборки.

Все введенные выборочные характеристики для имеющихся данных можно посчитать с помощью компьютерных программ SPSS, STATISTICA и др.

5. Свойства выборочных характеристик

Предположим, что наблюдалась некоторая случайная величина ξ для которой $F(x)$ — функция распределения (ее называют теоретической функцией распределения) и для которой существуют основные числовые характеристики (их называют теоретическими характеристиками): математическое ожидание, дисперсия, стандартное отклонение, коэффициенты асимметрии и эксцесса и т. п.

На практике точный вид функции $F(x)$ и точные значения основных теоретических числовых характеристик случайной величины ξ бывают неизвестны. Исследователь вынужден строить свои выводы о них лишь на основании ограниченного ряда выборочных характеристик, полученных для наблюдений случайной выборки из интересующей его генеральной совокупности. К основным выборочным характеристикам относятся: выборочная функция распределения $F_n(x)$, выборочная относительная частота n_i/n появления i-го возможного значения в n наблюдениях, выборочное среднее значение $\overline{x}(n)$, выборочная дисперсия $s^2(n)$, выборочный коэффициент асимметрии $\widehat{A}(n)$, выборочный коэффициент эксцесса $\widehat{E}(n)$, выборочный коэффициент корреляции $\widehat{r}(n)$.

Из закона больших чисел теории вероятностей следует, что при неограниченном увеличении объема выборки (т. е. при $n \to \infty$) с вероятностью, близкой к единице, все основные выборочные характеристики стремятся к соответствующим теоретическим характеристикам исследуемой случайной величины ξ. Этот факт позволяет использовать выборочные характеристики для приближенного описания свойств случайной величины ξ для всей генеральной совокупности.

Все выборочные характеристики являются случайными величинами, и поэтому возникает вопрос о законе распределения вероятностей каждой из них.

Из центральной предельной теоремы теории вероятностей следует, что асимптотически (при $n \to \infty$) практически независимо от типа случайной величины ξ все основные выборочные характеристики за исключением $\widehat{r}$ ведут себя как нормально распределенные случайные величины. При этом, разумеется, параметры нормального закона, т. е. математическое ожидание и дисперсия, различны для разных выборочных характеристик. Так, например, $\mathsf{M}[\overline{x}(n)] = a$ — математическому ожиданию ξ, $\mathsf{D}[\overline{x}(n)] = \sigma^2/n$, где σ^2 — дисперсия ξ, $\mathsf{M}[s^2(n)] = \left(1 - \frac{1}{n}\right)\sigma^2$, $\mathsf{M}[F_n(x)] = F(x)$, $\mathsf{M}\left[\frac{n_i}{n}\right] = p_i$ — вероятности i-го значения соответствующей дискретной случайной величины ξ и т. д.

Из центральной предельной теоремы также следует, что асимптотически (при $n \to \infty$) случайная величина $\sqrt{n}[\widehat{r}(n) - r(X, Y)]$ распределена по нормальному закону с математическим ожиданием, равным нулю. При конечных (ограниченных) объемах случайной выборки поведение основных выборочных характеристик существенно зависит от закона распределения случайной величины ξ.

Пусть, например, случайная величина $\xi \sim N(a, \sigma^2)$. Тогда оказывается (теорема Фишера), что при любом конечном объеме n выборки (а не только при $n \to \infty$) случайная величина $\overline{x}(n) \sim N\left(a, \frac{\sigma^2}{n}\right)$, т. е. распределена по нормальному закону с параметрами a и σ^2/n. Кроме того, в этом случае случайные величины $\overline{x}(n)$ и $s^2(n)$ независимы, причем

$$\mathsf{M}[s^2(n)] = \sigma^2\left(1 - \frac{1}{n}\right), \quad \mathsf{D}[s^2(n)] = \frac{2\sigma^4}{n}\left(1 - \frac{1}{n}\right).$$

Если совместное распределение пары случайных величин (X, Y) является нормальным, то асимптотически (при $n \to \infty$) распределение $\widehat{r}(n)$ тоже является нормальным с

$$\mathsf{M}[\widehat{r}(n)] = r(X, Y), \quad \mathsf{D}[\widehat{r}(n)] = \frac{(1 - r^2)^2}{n}.$$

При малых объемах n выборки или при значениях $|\widehat{r}(n)|$, близких к единице, это приближение является достаточно грубым.

Заметим в заключение, что каждый член дискретного вариационного ряда $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, построенного для наблюдений $x_1, x_2, \ldots, x_n$, называется *порядковой статистикой*. Порядковые статистики тоже относятся к основным выборочным характеристикам. Они являются выборочными аналогами квантилей случайной величины.

Задачи к главе 3

1. При обследовании 100 семей по числу детей были получены следующие результаты: 20 семей не имеют детей, 40 семей имеют по одному ребенку, 25 семей имеют по два ребенка, 10 семей имеют по три ребенка и 5 семей имеют по четыре ребенка.

а) Построить таблицу частот дискретного вариационного ряда, полигон частот, выборочную функцию распределения и ее график.

б) Найти выборочное среднее значение, выборочную дисперсию и выборочный коэффициент вариации.

2. На экзамене по математике 25 студентов группы были оценены по пятибалльной шкале следующим образом: 5, 5, 4, 4, 4, 3, 3, 2, 5, 3, 4, 4, 4, 3, 2, 5, 2, 3, 4, 5, 5, 2, 3, 3, 3.

а) Построить: таблицу частот дискретного вариационного ряда, полигон частот, выборочную функцию распределения и ее график.

б) Найти выборочное среднее значение, выборочную дисперсию и выборочный коэффициент вариации.

3. При тестировании общего интеллекта 20 школьников были получены следующие баллы: 11, 17, 17, 12, 13, 12, 15, 15, 14, 16, 13, 14, 13, 15, 16, 16, 15, 15, 14, 14.

а) Построить: таблицу частот дискретного вариационного ряда, полигон частот, выборочную функцию распределения и ее график.

б) Найти выборочное среднее значение, выборочную дисперсию и выборочный коэффициент вариации.

4. Известны следующие результаты (в баллах) сдачи вступительных экзаменов в университет группы абитуриентов: 18, 16, 20, 17, 19, 20, 17, 17, 12, 15, 20, 18, 19, 18, 18, 16, 18, 14, 17, 19, 16, 14, 19, 12, 15, 16, 20, 14.

а) Построить таблицу частот интервального вариационного ряда, разбив данные на 4 равных интервала, и гистограмму частот.

б) Найти выборочное среднее значение, выборочную дисперсию и выборочный коэффициент вариации.

5. Имеются следующие данные выборочного обследования затраты времени (в часах) 100 студентов на дорогу до университета:

Затраты времени	Число студентов
(0; 0,5]	7
(0,5; 1]	18
(1; 1,5]	32
(1,5; 2]	37
(2; 2,5]	6

а) Построить гистограмму частот и гистограмму накопленных частостей.

б) Найти выборочное среднее значение, выборочную дисперсию и выборочный коэффициент вариации.

СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

§ 1. Точечные оценки и их свойства

Основная задача выборочного исследования заключается в том, чтобы дать наибольший объем сведений о неизвестном распределении случайной величины (признака) ξ во всей генеральной совокупности, из которой извлечена случайная выборка, зная лишь наблюдения величины ξ для случайной выборки. Такая задача будет идеально решена, если по результатам наблюдений случайной выборки будет найдена теоретическая функция распределения $F(x)$ величины ξ. В общем случае это трудная задача. Тогда пытаются оценить неизвестные параметры распределения случайной величины ξ с помощью полученных наблюдений случайной выборки. Под параметрами распределения ξ понимают, прежде всего, параметры функции распределения $F(x)$ случайной величины ξ и числовые характеристики ξ. Например, в испытаниях Бернулли неизвестным параметром может быть вероятность p появления интересующего нас события, а при нормальном законе распределения случайной величины ξ неизвестными параметрами могут быть математическое ожидание $\mathsf{M}\xi$ и дисперсия $\mathsf{D}\xi$, определяющие этот закон.

Понятие параметров расширяется, если рассматривать двумерную или многомерную случайную величину ξ. Например, для двумерной ξ параметрами являются: коэффициент корреляции, корреляционные отношения, коэффициенты линейной регрессионной модели и т. д. Это так называемые параметры модели. Таким образом, под параметрами в дальнейшем будут пониматься как параметры распределения, так и параметры модели. Смысл понятия параметра будет постоянно углубляться в дальнейшем.

Под параметром может пониматься как одно число, так и набор из нескольких чисел. В силу случайной изменчивости наблюдаемых данных нельзя, основываясь только на них, указать точное значение параметра и приходится довольствоваться лишь его приближенным значением.

Любая функция от результатов наблюдений $x_1, x_2, \ldots, x_n$ исследуемой случайной величины ξ называется *статистикой*. Обозначим через Θ неизвестный параметр. Пусть это некоторое не известное

нам число. Тогда статистика $\widehat{\Theta}_n$, используемая в качестве приближенного значения неизвестного параметра Θ, называется *статистической точечной оценкой* (или просто *оценкой*) параметра Θ. Так, например, $\overline{x}(n)$, $s^2(n)$, $\widehat{A}(n)$ и $\widehat{E}(n)$ являются статистиками и могут рассматриваться как оценки соответственно параметров $\mathsf{M}\xi$, $\mathsf{D}\xi$, A и E случайной величины ξ.

Все статистики и статистические оценки являются случайными величинами, так как их значения при переходе от одной случайной выборки к другой непредсказуемы. Для оценки параметра Θ можно использовать многие статистики. Однако для практики важно, чтобы эти оценки были в каком-то определенном смысле надежными. Поэтому предъявляются определенные требования к оценкам. Обычно требуется выполнение следующих трех свойств оценок: состоятельности, несмещенности и эффективности.

Оценка $\widehat{\Theta}_n$ неизвестного параметра Θ называется *состоятельной*, если по мере роста числа наблюдений n (т. е. при $n \to \infty$) с вероятностью, близкой к единице, оценка Θ_n стремится к параметру Θ.

Из предыдущей главы следует, что все выборочные характеристики являются состоятельными оценками соответствующих теоретических характеристик случайной величины ξ. В частности, выборочное среднее значение $\overline{x}(n)$ — состоятельная оценка $\mathsf{M}\xi$, выборочная дисперсия $s^2(n)$ — состоятельная оценка $\mathsf{D}\xi$, выборочный коэффициент асимметрии $\widehat{A}(n)$ — состоятельная оценка коэффициента асимметрии A величины ξ, выборочный коэффициент эксцесса $\widehat{E}(n)$ — состоятельная оценка коэффициента эксцесса E величины ξ, выборочный коэффициент корреляции $\widehat{r}(n)$ — состоятельная оценка теоретического коэффициента корреляции $r(X, Y)$. Из теоремы Бернулли также следует, что относительная частота является состоятельной оценкой неизвестной вероятности p интересующего нас события в испытаниях Бернулли.

Требование состоятельности необходимо на практике для того, чтобы увеличение объема n выборки приближало нас к истинному значению параметра. Однако, во-первых, свойство состоятельности может проявляться лишь при больших значениях n, до которых на практике не всегда добираются, а во-вторых, в ряде ситуаций для одного и того же параметра можно предложить несколько состоятельных оценок. Поэтому одного свойства состоятельности недостаточно для полной характеристики надежности оценки. Другим свойством оценки должна быть ее несмещенность.

Оценка $\widehat{\Theta}_n$ неизвестного параметра Θ называется *несмещенной*, если ее математическое ожидание $\mathsf{M}\widehat{\Theta}_n = \Theta$.

Например, выборочное среднее значение $\overline{x}(n)$ — несмещенная оценка математического ожидания $\mathsf{M}\xi$ случайной величины ξ, а относительная частота — несмещенная оценка неизвестной вероятности p в испытаниях Бернулли. Однако выборочная дисперсия $s^2(n)$ уже не является несмещенной оценкой дисперсии $\mathsf{D}\xi$ случайной величины ξ. То же самое можно сказать и об оценках $\widehat{A}(n)$, $\widehat{E}(n)$ и, при малых объемах n, об оценке $\widehat{r}(n)$.

Если же ввести так называемую исправленную выборочную дисперсию

$$\widetilde{s}^2(n) = \frac{n}{n-1} \cdot s^2(n) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2,$$

то $\widetilde{s}^2(n)$ уже будет несмещенной и состоятельной оценкой $\mathsf{D}\xi$. При малых n выборочный коэффициент корреляции $\widehat{r}(n)$ — смещенная оценка теоретического коэффициента $r(X, Y)$.

В отличие от состоятельности, являющимся асимптотическим свойством (при $n \to \infty$), несмещенность оценки устраняет при каждом конечном объеме n выборки систематическую погрешность оценивания. Поэтому требование несмещенности особенно существенно при малом количестве наблюдений. Для небольших выборок используется $\widetilde{s}^2(n)$ вместо $s^2(n)$.

Решающим свойством, определяющим качество оценки, является эффективность оценки. Оценка $\widehat{\Theta}_n$ параметра Θ называется *эффективной*, если она среди всех прочих оценок параметра Θ обладает наименьшей мерой случайного разброса относительно истинного значения Θ.

В том случае, когда Θ — число, в качестве такой меры чаще всего берется математическое ожидание квадрата отклонения $\widehat{\Theta}_n$ от Θ, т. е. $\mathsf{M}(\widehat{\Theta}_n - \Theta)^2$, что для несмещенных оценок $\widehat{\Theta}_n$ совпадает с их дисперсией $\mathsf{D}\widehat{\Theta}_n$, так как тогда $\mathsf{D}\widehat{\Theta}_n = \mathsf{M}(\widehat{\Theta}_n - \Theta)^2$. В более общем случае вместо $\mathsf{M}(\widehat{\Theta}_n - \Theta)^2$ берут $\mathsf{M}W(\Theta_n, \Theta)$, где $W(\Theta_n, \Theta) \geqslant 0$ — некоторая функция «штрафа».

Например, выборочное среднее $\overline{x}(n)$ — эффективная оценка математического ожидания $\mathsf{M}\xi$ случайной величины ξ, а выборочная исправленная дисперсия $\widetilde{s}^2(n)$ не является эффективной оценкой дисперсии $\mathsf{D}\xi$ при нормальном законе распределения величины ξ. Заметим, что относительная частота является эффективной оценкой неизвестной вероятности p в испытаниях Бернулли.

Бывают ситуации, когда требования несмещенности и эффективности оценки оказываются несовместимыми и тогда, как правило, выбирают эффективную оценку.

В качестве меры сравнительной эффективности двух несмещенных оценок $\widehat{\Theta}_n^{(1)}$ и $\widehat{\Theta}_n^{(2)}$ параметра Θ принимают отношение дисперсий

$$e_n = \frac{\mathsf{D}\widehat{\Theta}_n^{(1)}}{\mathsf{D}\widehat{\Theta}_n^{(2)}}.$$

При $e_n > 1$ оценка $\widehat{\Theta}_n^{(1)}$ менее эффективна, чем оценка $\widehat{\Theta}_n^{(2)}$, а при $e_n < 1$ оценка $\widehat{\Theta}_n^{(1)}$ более эффективна, чем оценка $\widehat{\Theta}_n^{(2)}$.

Если оценка смещенная и (или) неэффективная, то желательно, чтобы она была хотя бы асимптотически (при $n \to \infty$) несмещенной и эффективной.

Оценка $\widehat{\Theta}_n$ параметра Θ называется *асимптотически* (при $n \to \infty$) *несмещенной*, если ее смещенность исчезает при $n \to \infty$, т. е. $\lim\limits_{n\to\infty} \mathsf{M}(\widehat{\Theta}_n) = \mathsf{M}\Theta$.

Пусть $\widehat{\Theta}_n^{(1)}$ и $\widehat{\Theta}_n^{(2)}$ — две несмещенные оценки параметра Θ. Оценка $\widehat{\Theta}_n^{(1)}$ параметра Θ называется *асимптотически* (при $n \to \infty$) *более эффективной*, чем оценка $\widehat{\Theta}_n^{(2)}$ параметра Θ, если существует $\lim\limits_{n\to\infty} e_n < 1$, где e_n — определенное выше отношение дисперсий.

Оценка $\widehat{\Theta}_n$ параметра Θ называется *асимптотически* (при $n \to \infty$) *эффективной*, если $\lim\limits_{n\to\infty} \mathsf{M}(\widehat{\Theta}_n - \Theta)^2$ является наименьшим среди всех $\lim\limits_{n\leftarrow\infty} (\widetilde{\Theta}_n - \Theta)^2$ для всех других оценок $\widetilde{\Theta}_n$ параметра Θ (при условии существования указанных пределов).

Например, $s^2(n)$ является асимптотически (при $n \to \infty$) несмещенной и асимптотически эффективной (при $n \to \infty$) оценкой дисперсии $\mathsf{D}\,\xi$ для нормально распределенной случайной величины ξ.

§ 2. Метод максимального правдоподобия. Интервальные оценки. Понятие о робастном оценивании

В предыдущем параграфе рассмотрены различные случаи использования статистик (т. е. функций от результатов наблюдений) в качестве точечных оценок $\widehat{\Theta}_n$ неизвестных параметров Θ. Например, $\overline{x}(n)$ — точечная оценка $\mathsf{M}\xi$, а $s^2(n)$ — точечная оценка $\mathsf{D}\xi$ исследуемой случайной величины (признака) ξ. Однако осталось неясным, каким методом это установлено. Наиболее часто таким методом является метод максимального правдоподобия.

1. Метод максимального правдоподобия

Пусть $x_1, x_2, \ldots, x_n$ — выборка, состоящая из n независимых наблюдений исследуемой случайной величины ξ, закон распределения вероятностей $f(x, \Theta)$ которой зависит от неизвестного параметра Θ. Под функцией $f(x, \Theta)$ будем понимать вероятность $\mathsf{P}(\xi = x, \Theta)$, если ξ — дискретная случайная величина, и плотность вероятности $p(x, \Theta)$, если ξ — непрерывная случайная величина.

Рассмотрим функцию $L(x_1, x_2, \ldots, x_n, \Theta) = f(x_1, \Theta) \cdot f(x_2, \Theta) \cdot \ldots \cdot f(x_n, \Theta)$.

Функция $L(x_1, x_2, \ldots, x_n, \Theta)$ называется *функцией правдоподобия*. Свое название эта функция получила из-за того, что при заданном значении Θ чем больше значение функции $L(x_1, x_2, \ldots, x_n, \Theta)$, тем правдоподобнее (или более вероятна) система наблюдений $x_1, x_2, \ldots, x_n$.

Метод максимального правдоподобия состоит в том, что в качестве оценки $\widehat{\Theta}_n$ неизвестного истинного значения параметра Θ рекомендуется выбирать такое значение, при котором функция правдоподобия $L(x_1, x_2, \ldots, x_n, \Theta)$ (как функция параметра Θ) достигает максимума.

Поскольку $\widehat{\Theta}_n$ зависит от выборки $x_1, x_2, \ldots, x_n$, то $\widehat{\Theta}_n$ — случайная величина. Оценка $\widehat{\Theta}_n$, полученная методом максимального правдоподобия, называется *оценкой максимального правдоподобия* неизвестного параметра Θ.

Например, относительная частота наступления интересующего нас события в испытаниях Бернулли является оценкой максимального правдоподобия неизвестной вероятности этого события.

На практике часто удобнее использовать не функцию правдоподобия $L(x_1, x_2, \ldots, x_n, \Theta)$, а *логарифмическую функцию правдоподобия*

$$l(x_1, x_2, \ldots, x_n, \Theta) = \ln L(x_1, x_2, \ldots, x_n, \Theta) = \sum_{i=1}^{n} \ln f(x_i, \Theta).$$

Так как в силу монотонного возрастания логарифма функции $L(x_1, \ldots, x_n, \Theta)$ и $l(x_1, \ldots, x_n, \Theta)$ достигают максимума при одном и том же значении Θ, то для нахождения оценки максимального правдоподобия можно искать такое значение Θ, при котором функция $\ln(x_1, x_2, \ldots, x_n, \Theta)$ достигает максимума.

Используя логарифмическую функцию правдоподобия, можно, например, установить, что в случае нормально распределенной величины ξ с неизвестными параметрами $a = \mathsf{M}\xi$ и $\sigma^2 = \mathsf{D}\xi$ оценками максимального правдоподобия для a является среднее выборочное

значение $\overline{x}(n)$, а для σ^2 — выборочная дисперсия $s^2(n)$. Если же исследуемая случайная величина ξ подчинена закону распределения Пуассона с неизвестным параметром $\lambda > 0$, то оценкой максимального правдоподобия параметра λ будет среднее выборочное значение $\overline{x}(n)$.

Метод максимального правдоподобия дает хорошие результаты для многих практических задач в том смысле, что оценки максимального правдоподобия неизвестного параметра Θ, полученные для таких задач, являются состоятельными, асимптотически (при $n \to \infty$) несмещенными, асимптотически (при $n \to \infty$) эффективными и асимптотически (при $n \to \infty$) нормальными.

Однако метод максимального правдоподобия не является наилучшим во всех ситуациях, прежде всего, из-за того, что для его применения необходимо знание типа анализируемого закона распределения $f(x, \Theta)$, что во многих случаях оказывается практически нереальным. В таких случаях применяют другие методы, например, так называемый метод моментов.

2. Интервальные оценки

Вычисленная на основании выборочных данных точечная оценка $\widehat{\Theta}_n$ неизвестного параметра Θ является лишь приближенным значением неизвестного параметра Θ. Однако во многих случаях важно не получение точечной оценки $\widehat{\Theta}_n$, а указание интервала вида $(\widehat{\Theta}_n^{(1)}, \widehat{\Theta}_n^{(2)})$, который бы с заранее заданной вероятностью (близкой к единице) накрывал неизвестное истинное значение параметра Θ. В этом случае говорят об *интервальной оценке параметра* Θ. При этом заранее задаваемая вероятность (близкая к единице) называется *доверительной вероятностью*, а сам интервал $(\widehat{\Theta}_n^{(1)}, \widehat{\Theta}_n^{(2)})$ называется *доверительным интервалом*.

Доверительный интервал по своей природе является случайным как по своему расположению, так и по своей длине, поскольку величины $\widehat{\Theta}_n^{(1)}$ и $\widehat{\Theta}_n^{(2)}$ — функции выборочных данных $x_1, x_2, \ldots, x_n$. Длина доверительного интервала уменьшается с ростом объема выборки n и увеличивается с приближением доверительной вероятности к единице.

Особенно важным является построение доверительного интервала с заданной доверительной вероятностью в случае выборок наблюдений небольшого объема. Для большого числа наблюдений точность приближенного значения $\widehat{\Theta}_n$ на практике бывает достаточной в случае несмещенности, состоятельности и эффективности оценки $\widehat{\Theta}_n$.

Доверительную вероятность обозначим через γ. Ее принято выбирать равной 0,95; 0,99; 0,999.

В большинстве случаев процедура построения интервальной оценки неизвестного параметра Θ следующая.

По сделанной выборке наблюдений находится точечная оценка $\widehat{\Theta}_n$ неизвестного параметра Θ. Затем по выбранной доверительной вероятности γ по определенным правилам находят такое число $\varepsilon > 0$, чтобы выполнялось соотношение $\mathsf{P}(\widehat{\Theta}_n - \varepsilon < \Theta < \widehat{\Theta}_n + \varepsilon) = \gamma$ или, что равносильно, $\mathsf{P}(|\Theta - \widehat{\Theta}_n| < \varepsilon) = \gamma$.

Эти соотношения следует читать так: «Доверительная вероятность того, что доверительный интервал $(\widehat{\Theta}_n - \varepsilon, \widehat{\Theta}_n + \varepsilon)$ накроет параметр Θ, равна γ».

Число $\varepsilon > 0$ называется *точностью интервальной оценки* параметра Θ, а интервал $(\widehat{\Theta}_n - \varepsilon, \widehat{\Theta}_n + \varepsilon)$, как уже сказано выше, называют доверительным интервалом.

Заметим, что построенный доверительный интервал является симметричным относительно точечной оценки $\widehat{\Theta}_n$. Однако не во всех случаях удается построить доверительный интервал с таким свойством. Правило нахождения числа $\varepsilon > 0$ зависит от природы параметра Θ и объема выборки n.

В качестве примера рассмотрим задачу интервального оценивания неизвестных параметров a и σ^2 нормально распределенной случайной величины ξ в случае небольшой выборки наблюдений (объем выборки $n < 30$).

В случае *интервальной оценки параметра a* доверительный интервал задается равенством

$$\mathsf{P}\left\{\overline{x} - t_{\text{кр}} \cdot \frac{\widetilde{s}}{\sqrt{n-1}} < a < \overline{x} + t_{\text{кр}} \cdot \frac{\widetilde{s}}{\sqrt{n-1}}\right\} = \gamma = 1 - \alpha,$$

где $\overline{x}$ — среднее выборочное значение, $\widetilde{s}^2$ — исправленная выборочная дисперсия, n — объем выборки, $t_{\text{кр}}$ — значение из таблицы t-распределения Стьюдента, найденное по числам $\frac{\alpha}{2}$ и $(n-1)$ (см. табл. VII).

Пример 1. Пусть распределение роста некоторой совокупности людей подчиняется нормальному распределению с неизвестными параметрами (a, σ^2). Согласно проведенным 25 наблюдениям для случайно отобранной группы людей этой совокупности получены выборочное среднее $\overline{x} = 1{,}72$ м и выборочное среднее квадратическое отклонение $\widetilde{s} = 0{,}18$ м. Требуется с доверительной вероятностью $\gamma = 0{,}95$ найти интервальную оценку для параметра a.

Решение. Так как в нашем случае число степеней свободы $n-1=24$, то из табл. VII t-распределения Стьюдента находим, что $t_{\text{кр}}=2{,}064$. Тогда из формулы доверительного интервала для a имеем, что

$$1{,}72-2{,}064\cdot\frac{0{,}18}{\sqrt{24}}<a<1{,}72+2{,}064\cdot\frac{0{,}18}{\sqrt{24}}.$$

Отсюда получаем, что $1{,}72-0{,}08<a<1{,}72+0{,}08$, т. е. $a\in(1{,}64;\ 1{,}8)$.

Таким образом, с вероятностью 0,95 можно гарантировать, что интервал (1,64 м; 1,8 м) накрывает параметр a. ■

В случае *интервальной оценки параметра* σ^2 доверительный интервал задается равенством

$$\mathrm{P}\left\{\frac{(n-1)\widetilde{s}^2}{\chi_1^2}<\sigma^2<\frac{(n-1)\widetilde{s}^2}{\chi_2^2}\right\}=\gamma=1-\alpha,$$

где $\widetilde{s}^2$ — исправленная выборочная дисперсия, n — объем выборки, $\gamma=1-\alpha$ — доверительная вероятность, χ_1^2 и χ_2^2 — значения из таблицы χ^2-распределения с $(n-1)$ степенями свободы, найденные соответственно при $Q=\frac{\alpha}{2}$ и при $Q=1-\frac{\alpha}{2}$ (см. табл. VI).

Пример 2. Пусть заданы условия примера 1. Требуется с доверительной вероятностью $\gamma=1-\alpha=0{,}95$ определить интервальную оценку для неизвестной дисперсии σ^2.

Решение. Так как $\widetilde{s}=0{,}18$ м, то $\widetilde{s}^2=0{,}03$ м^2. Поскольку $\alpha=0{,}05$, то $\frac{\alpha}{2}=0{,}025$, $1-\frac{\alpha}{2}=0{,}975$ и, значит, по таблице χ^2-распределения с $n-1=24$ степенями свободы находим, что $\chi_1^2=39{,}36$, а $\chi_2^2=12{,}4$. Следовательно, из формулы доверительного интервала для σ^2 имеем, что

$$\frac{24\cdot0{,}03}{39{,}36}<\sigma^2<\frac{24\cdot0{,}03}{12{,}4}.$$

Отсюда получаем, что

$$0{,}018<\sigma^2<0{,}058.$$

Таким образом, с вероятностью 0,95 можно утверждать, что интервал (0,018 м^2, 0,058 м^2) накрывает параметр σ^2. ■

При достаточно больших объемах выборок при построении доверительных интервалов для неизвестных параметров a и σ^2 нормального распределения случайной величины ξ используются более простые формулы. Они основаны на том, что точечные оценки параметров a и σ^2 имеют асимптотически (при $n\to\infty$) нормальные распределения.

3. Робастное (устойчивое) оценивание

Иногда приходится иметь дело с результатами наблюдений, значительно отклоняющимися от основного массива, т. е. с выбросами.

Пусть, например, при тестировании десяти учащихся были получены следующие результаты (в баллах): 15, 13, 18, 37, 12, 16, 45, 17, 11, 12. В приведенных данных имеются два значения 37 и 45, которые значительно больше остальных значений, содержащихся на отрезке [11, 18]. Здесь значения 37 и 45 являются выбросами. Их называют «засорением» или «грубыми ошибками».

Основные причины появления грубых ошибок следующие:

а) неправильная организация наблюдений, ошибки группировки,

б) грубые ошибки при регистрации и обработке данных.

Так как основные выборочные характеристики (выборочное среднее значение, выборочная дисперсия и т. п.) сильно зависят от «выбросов», то при наличии «выбросов» их нельзя брать в качестве точечных оценок неизвестных параметров исследуемой случайной величины ξ, поскольку их свойства (состоятельность, несмещенность, эффективность) становятся неудовлетворительными. Получение устойчивых выборочных характеристик носит название *робастного оценивания*.

Выявление грубых ошибок и оценка степени засорения выборки данных проводится либо визуальным анализом, либо с помощью специальных статистических критериев. После обнаружения выбросов в данных используются два основных подхода. Первый подход основан на отбрасывании грубых ошибок из выборки данных. По усеченной совокупности данных, остающейся после отбрасывания грубых ошибок, по специальным формулам находятся точечные оценки неизвестных параметров.

Другой подход предполагает замену грубых ошибок на модифицированные (винзорированные) значения с устраненными или уменьшенными ошибками. Затем по специальным формулам находят устойчивые оценки неизвестных параметров.

Задачи к главе 4

1. Зарплата работников некоторой фирмы является случайной величиной, распределенной по нормальному закону с неизвестным параметром a и параметром $\sigma = 2$. Найти с доверительной вероятностью $\alpha = 0{,}95$ доверительный интервал для неизвестного парамет-

ра a, если известна выборочная средняя зарплата $\overline{x}$ для случайной выборки работников фирмы в количестве: а) $n=9$, б) $n=16$, в) $n=25$.

2. Среднедушевой доход семей некоторого города является случайной величиной, распределенной по нормальному закону с неизвестными параметрами a и σ. Найти с доверительной вероятностью 0,99 доверительный интервал для неизвестного параметра a, если для среднедушевого дохода известны выборочное среднее значение $\overline{x}$ и исправленная выборочная дисперсия $\widetilde{s}^2=4{,}41$ для случайной выборки семей города в количестве: а) $n=10$, б) $n=17$, в) $n=21$.

3. Найти наименьший объем n случайной выборки из нормальной генеральной совокупности с параметром $\sigma=1$, при котором с доверительной вероятностью $\alpha \geqslant 0{,}9$ погрешность среднего значения $\overline{x}$, найденного для этой выборки, будет меньше: а) 0,2, б) 0,1.

4. При тестировании общего интеллекта группы из 25 человек был получен средний балл $\overline{x}=45$ баллов. Считая, что общий интеллект распределен нормально с параметром $\sigma=16$ баллов, найти доверительную вероятность того, что точность ε среднего балла $\overline{x}$ равна: а) 5 баллов, б) 3 балла, в) 8 баллов.

5. Рост (в см) студентов университета является случайной величиной, распределенной по нормальному закону с неизвестной дисперсией σ^2. Найти интервальную оценку для σ^2 с доверительной вероятностью 0,99, если для случайной выборки студентов в количестве $n=25$ человек была получена исправленная выборочная дисперсия: а) $\widetilde{s}^2=2{,}5$ см^2, б) $\widetilde{s}^2=3$ см^2.

ГЛАВА 5

СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

Из предыдущей главы известно, что по наблюдениям случайной выборки можно находить точечные и интервальные оценки неизвестных параметров интересующего нас случайного признака ξ для всей генеральной совокупности. Однако часто мы располагаем некоторыми предварительными (априорными) догадками или предположениями относительно численной величины этих параметров или о законе распределения вероятностей признака ξ. В таком случае можно проверить гипотезу о том, что наша догадка действительно верна.

Статистической гипотезой (в дальнейшем для краткости будем говорить просто «гипотезой») называется любое предположение о модели закона распределения вероятностей исследуемого случайного признака, о числовых значениях параметров этого закона распределения, о модели зависимости между анализируемыми признаками.

Например, статистическими гипотезами будут:

а) гипотеза о том, что исследуемая случайная величина подчиняется нормальному закону распределения вероятностей,

б) гипотеза о том, что математическое ожидание исследуемой случайной величины равно заданному числу,

в) гипотеза о том, что дисперсии двух исследуемых случайных величин равны между собой,

г) гипотеза о том, что две исследуемые случайные величины являются корреляционно зависимыми.

Гипотеза «на Луне имеется вода» не является статистической гипотезой, поскольку в ней не идет речь о случайном признаке.

Различают простые и сложные гипотезы. Гипотезу, которая однозначно определяет все параметры распределения, называют *простой*, а гипотезу, состоящую из конечного или бесконечного числа простых гипотез, называют *сложной*. Например, гипотеза о том, что исследуемый нормально распределенный признак имеет параметр $a = 2$ и параметр $\sigma^2 = 9$, является простой, а гипотеза о том, что исследуемый нормально распределенный признак имеет параметр $a > 2$ и параметр $\sigma^2 = 9$, является сложной, так как она состоит из бесконечного числа простых гипотез о числовом значении параметра $a > 2$.

Наряду с выдвинутой гипотезой одновременно рассматривают и противоречащую ей гипотезу. Выдвинутую гипотезу называют *нулевой* или *основной* и обозначают H_0. Противоречащую H_0 гипотезу называют *альтернативной* или *конкурирующей* и обозначают H_1. Например, если в качестве гипотезы H_0 рассматривается гипотеза о том, что математическое ожидание $\mathsf{M}\xi$ исследуемого случайного признака ξ равно 5: $\mathsf{M}\xi = 5$, то альтернативной гипотезой H_1 может быть одна из трех гипотез: $\mathsf{M}\xi \neq 5$, $\mathsf{M}\xi < 5$, $\mathsf{M}\xi > 5$.

Для проверки выдвинутых гипотез H_0 и H_1 у нас имеются только выборочные данные наблюдений исследуемых одного или нескольких признаков. Цель проверки состоит в том, чтобы определить, какая из гипотез H_0 или H_1 не противоречит имеющимся выборочным данным. Проверка осуществляется с помощью того или иного статистического критерия. *Статистический критерий* — это правило, в соответствии с которым принимается или отклоняется гипотеза H_0, причем с заданной степенью достоверности (с заданной вероятностью). Процедура проверки выдвинутых гипотез H_0 и H_1 с помощью статистического критерия называется *статистической проверкой гипотез*.

Если в результате статистической проверки принята, например, гипотеза H_0, то это означает лишь, что данные наблюдений выборки не противоречат гипотезе H_0. Принятие гипотезы H_0 не означает, что гипотеза H_0 является наилучшей и единственно подходящей. Просто H_0 не противоречит имеющимся выборочным данным, однако таким свойством могут наряду с H_0 обладать и другие гипотезы. Таким образом, статистически проверенная гипотеза не является раз и навсегда установленным фактом, а является лишь достаточно правдоподобным, не противоречащим опытным данным утверждением.

Поскольку результат статистической проверки выдвинутых гипотез H_0 и H_1 основан на выборочных данных, то возможны ошибочные решения. Различают ошибки 1-го и 2-го родов. *Ошибка 1-го рода* состоит в том, что будет отвергнута гипотеза H_0 (и принята H_1), в то время, когда гипотеза H_0 верна для генеральной совокупности.

Ошибка 2-го рода состоит в том, что будет принята гипотеза H_0, в то время, когда она не верна (т. е. верна H_1) для генеральной совокупности.

Вероятность ошибки 1-го рода обозначается α и называется *уровнем значимости* (критерия). Вероятность ошибки 2-го рода обозначается β, а величину $(1-\beta)$ называют *мощностью критерия*.

При фиксированном объеме выборочных данных уровень значимости α заранее задается. Наиболее распространенными на практи-

ке значениями α являются 0,01; 0,05; 0,1, что соответственно указывает на вероятность получения достоверного вывода $(1-\alpha)$, равную 0,99; 0,95; 0,90. Например, задание $\alpha = 0{,}05$ означает, что при многократном использовании данного статистического критерия в пяти случаях из ста будет ошибочно отвергаться гипотеза H_0.

Ошибки 1-го и 2-го родов зависят друг от друга. Если объем выборки фиксирован, то чем меньше будет α, тем больше будет β, и наоборот. Другими словами, чем меньше α, тем меньше мощность $(1-\beta)$, и наоборот. При слишком малом объеме выборки n при заданном уровне значимости α мощность $(1-\beta)$ может оказаться слишком маленькой. В таком случае либо увеличивают объем выборки, либо увеличивают уровень значимости α, что позволит увеличить мощность.

Рассмотрим теперь процедуру статистической проверки гипотез. Для каждого статистического критерия задается некоторая функция S от результатов наблюдения $S = S(x_1, x_2, \ldots, x_n)$, которая называется *статистикой критерия*. Статистика критерия может обозначаться и по-другому. Эта функция S, как и всякая функция от результатов наблюдения, сама является случайной величиной. В предположении справедливости гипотезы H_0 статистика критерия S подчинена хорошо изученному и затабулированному закону распределения. Множество всех значений статистики S разбивается при заданном уровне значимости α на два непересекающихся подмножества. Одно подмножество содержит все те значения S, при которых гипотеза H_0 принимается, и называется *областью принятия гипотезы* H_0. Другое подмножество содержит все те значения S, при которых гипотеза H_0 отвергается, и называется *критической областью*.

Область принятия гипотезы H_0 и критическая область являются интервалами, а точки, которые разделяют эти интервалы, называют *критическими значениями статистики* S.

Различают три типа критических областей: левостороннюю критическую область, правостороннюю критическую область и двустороннюю критическую область. В зависимости от типа критической области статистические критерии подразделяют на левосторонние, правосторонние и двусторонние. Левосторонняя критическая область представляет собой интервал вида $(-\infty, S^{\text{л}}_{\text{кр},\alpha}]$, где критическое значение $S^{\text{л}}_{\text{кр},\alpha}$ определяется условием: вероятность $\mathsf{P}\{S \leqslant S^{\text{л}}_{\text{кр},\alpha}\} = \alpha$ — уровню значимости.

Правосторонняя критическая область представляет собой интервал вида $[S^{\text{п}}_{\text{кр},\alpha}, +\infty)$, где критическое значение $S^{\text{п}}_{\text{кр},\alpha}$ определяется условием: вероятность $\mathsf{P}\{S \geqslant S^{\text{п}}_{\text{кр},\alpha}\} = \alpha$ — уровню значимости.

Двусторонняя критическая область представляет собой объединение двух интервалов $(-\infty, S^{\text{л}}_{\text{кр},\alpha}]$ и $[S^{\text{п}}_{\text{кр},\alpha}, +\infty)$, где критические значения $S^{\text{л}}_{\text{кр},\alpha}$ и $S^{\text{п}}_{\text{кр},\alpha}$ удовлетворяют условию

$$\mathsf{P}\{S \leqslant S^{\text{л}}_{\text{кр},\alpha}\} + \mathsf{P}\{S \geqslant S^{\text{п}}_{\text{кр},\alpha}\} = \alpha.$$

В частности, если двусторонняя критическая область симметрична относительно нуля, то она задается неравенством $|S| \geqslant S_{\text{кр},\alpha}$, где критическое значение $S_{\text{кр},\alpha}$ определяется условием $\mathsf{P}\{S \leqslant -S_{\text{кр},\alpha}\} = \mathsf{P}\{S \geqslant S_{\text{кр},\alpha}\} = \frac{\alpha}{2}$.

Критические значения $S^{\text{л}}_{\text{кр},\alpha}$, $S^{\text{п}}_{\text{кр},\alpha}$ и $S_{\text{кр},\alpha}$ на практике всегда находятся по таблицам распределения статистики S рассматриваемого статистического критерия.

Приведем общую логическую схему проверки гипотез на практике.

Общая логическая схема проверки гипотез

1. Формулируют нулевую гипотезу H_0 и альтернативную гипотезу H_1.
2. Задают α — уровень значимости критерия.
3. Выбирают подходящий статистический критерий.
4. По заданному α и по таблице распределения статистики S рассматриваемого статистического критерия находят необходимые критические значения $S^{\text{л}}_{\text{кр},\alpha}$, $S^{\text{п}}_{\text{кр},\alpha}$, $S_{\text{кр},\alpha}$.
5. Вычисляют для полученной выборки наблюдений значение $S_{\text{набл}}$ статистики критерия S.
6. Сравнивают выборочное значение $S_{\text{набл}}$ с критическими значениями и делают выбор гипотезы H_0 или H_1.

Если используется *левосторонний критерий*, то на уровне α при $S_{\text{набл}} > S^{\text{л}}_{\text{кр},\alpha}$ гипотеза H_0 принимается, а при $S_{\text{набл}} < S^{\text{л}}_{\text{кр},\alpha}$ гипотеза H_0 отвергается и принимается гипотеза H_1.

Если используется *правосторонний критерий*, то на уровне α при $S_{\text{набл}} < S^{\text{п}}_{\text{кр},\alpha}$ гипотеза H_0 принимается, а при $S_{\text{набл}} \geqslant S^{\text{п}}_{\text{кр},\alpha}$ гипотеза H_0 отвергается и принимается гипотеза H_1.

Если же используется двусторонний критерий, то на уровне α при $S^{\text{л}}_{\text{кр},\alpha} < S_{\text{набл}} < S^{\text{п}}_{\text{кр},\alpha}$ принимается гипотеза H_0, а при $S_{\text{набл}} \leqslant S^{\text{л}}_{\text{кр},\alpha}$ или при $S_{\text{набл}} \geqslant S^{\text{п}}_{\text{кр},\alpha}$ гипотеза H_0 отвергается и принимается гипотеза H_1.

Статистические критерии подразделяются на параметрические и непараметрические.

Критерий называют *параметрическим*, если он использует предположение о принадлежности закона распределения исследуемой случайной величины (признака) в генеральной совокупности к некоторому известному параметрическому семейству, чаще всего к семейству нормальных законов распределения.

Критерий называют *непараметрическим* (или *свободным от распределения*), если он не использует предположения о принадлежности закона распределения исследуемой случайной величины в генеральной совокупности к некоторому известному параметрическому семейству, например, к семейству нормальных законов распределения.

Примеры параметрических и непараметрических критериев будут приведены в следующей главе. Сейчас отметим, что непараметрические критерии имеют определенные преимущества перед параметрическими из-за меньших требований к их применению, так как на практике чаще всего точный закон распределения исследуемой случайной величины (признака) нам не известен. В то же время по сравнению с параметрическими критериями, если они применимы, непараметрические критерии дают менее точные выводы.

Замечание. Процедура доверительного оценивания неизвестного параметра Θ является как бы обращением процедуры проверки статистической гипотезы о значении параметра Θ. При проверке статистической гипотезы о параметре Θ по известному значению Θ строится область $A(\Theta)$ принятия гипотезы H_0, в которую с заданной вероятностью попадает статистика критерия. При доверительном оценивании по таким множествам $A(\Theta)$ находится интервал, который с заданной вероятностью накрывает само значение Θ.

Наконец отметим, что каждый статистический критерий применяется при вполне определенных допущениях или предположениях о законе распределения исследуемой случайной величины. Например, критерии Стьюдента и Фишера (см. главу 6) требуют нормального закона распределения. На практике все условия применения интересующего нас критерия точно проверить не удается. Поэтому важно, чтобы критерий был относительно нечувствителен к небольшим отклонениям от принятых допущений об условиях его применения. Такого рода критерий называется робастным (устойчивым). Именно такие критерии являются наиболее важными для практики.

НЕКОТОРЫЕ СТАТИСТИЧЕСКИЕ КРИТЕРИИ

§ 1. Биномиальный критерий и критерий знаков

1. Биномиальный критерий

Пусть рассматривается некоторая совокупность объектов, обладающих дихотомическим признаком. Это означает, что все объекты совокупности можно разбить на две группы в зависимости от того, обладает или не обладает объект определенным свойством признака.

Например, при изучении успеваемости совокупности школьников можно школьников разделить на отличников и неотличников. При опросе общественного мнения совокупность людей можно разделить на две части: в одну часть отнести тех, кто при опросе ответил «да», а в другую часть — тех, кто ответил «нет». Разделение на две части, два типа производится с помощью независимых повторных испытаний Бернулли.

Исход одного типа считается «успехом», а исход другого типа — «неудачей». Вероятность p «успеха» и вероятность $q = 1 - p$ «неудачи» нам неизвестны для всей совокупности. Известно только, что p и q — постоянны. С целью получения оценки вероятности p проводится n испытаний Бернулли и определяется число успехов m для них. Пусть сначала число n мало ($n < 30$). Выдвигается гипотеза H_0: $p = p_0$, где p_0 — некоторое заданное число, $p_0 \in (0, 1)$, против альтернативы H_0: $p > p_0$.

Требуется проверить нулевую гипотезу H_0 и ее альтернативу при заданном уровне значимости α.

Для такой проверки используется правосторонний биномиальный критерий.

Правосторонний биномиальный критерий

1. В качестве статистики берется случайная величина S — число успехов в испытаниях Бернулли.

2. Для заданной выборки наблюдений объема n имеем значение $S_{\text{набл}} = m$.

3. Далее по заданным p_0, n, α из таблицы биномиального распределения находят критическое значение $S_{\text{кр}}$.

Тогда правосторонний биномиальный критерий утверждает, что на уровне α гипотеза H_0 принимается, если $m < S_{\text{кр}}$, и гипотеза H_0 отвергается на уровне α, если $m \geqslant S_{\text{кр}}$.

Пример 1. Пусть имеется некоторая совокупность учащихся, состоящая из отличников и неотличников, и пусть по случайной выборке учащихся объема $n = 8$ с одним отличником ($m = 1$) необходимо на уровне значимости $\alpha = 0{,}005$ проверить гипотезу H_0 о том, что вероятность отличника во всей совокупности учащихся $p = 0{,}1$ против альтернативы H_1: $p > 0{,}1$.

Решение. По заданным $\alpha = 0{,}005$, $n = 8$, $p = 0{,}1$ из табл. XIV находим критическое значение $S_{\text{кр}} = 4$. Поскольку $S_{\text{набл}} = m = 1 < S_{\text{кр}} = 4$, то на уровне значимости $\alpha = 0{,}005$ гипотеза H_0 принимается. ■

Если объем выборки большой ($n \geqslant 30$), то в качестве статистики в правостороннем биномиальном критерии берут

$$\widetilde{S} = \frac{S - np_0}{\sqrt{np_0(1-p_0)}}, \quad \text{если } np_0(1-p_0) < 9,$$

$$\widetilde{S} = \frac{S - np_0 + \frac{1}{2}}{\sqrt{np_0(1-p_0)}}, \quad \text{если } np_0(1-p_0) \geqslant 9.$$

Статистика $\widetilde{S}$ асимптотически (при $n \to \infty$) имеет нормальное распределение $N(0, 1)$. Тогда критическое значение $\widetilde{S}_{\text{кр}}$ при заданном α находят из табл. III значений функции $\Phi(x)$ по формуле

$$\Phi(\widetilde{S}_{\text{кр}}) = \frac{1}{2} - \alpha.$$

Пример 2. На уровне $\alpha = 0{,}05$ требуется проверить гипотезу H_0 о том, что в некоторой совокупности учащихся вероятность отличника $p = 0{,}3$, против альтернативы H_1 о том, что $p > 0{,}3$. Для проверки гипотез H_0 и H_1 имеется случайная выборка учащихся объема $n = 30$, в которой 10 отличников ($m = 10$).

Решение. По формуле $\Phi(\widetilde{S}_{\text{кр}}) = 0{,}45$ из табл. XIV находим $\widetilde{S}_{\text{кр}} = 1{,}65$. Подсчитаем $np_0(1 - p_0) = 30 \cdot 0{,}3 \cdot 0{,}7 = 2{,}7 < 9$. Вычисляем наблюденное значение $\widetilde{S}_{\text{набл}}$ статистики $\widetilde{S}$ критерия для заданной выборки учащихся:

$$\widetilde{S}_{\text{набл}} = \frac{S - np_0}{\sqrt{np_0(1-p_0)}} = \frac{10 - 30 \cdot 0{,}3}{\sqrt{30 \cdot 0{,}3 \cdot 0{,}7}} = \frac{1}{3 \cdot \sqrt{0{,}7}} \approx 0{,}4.$$

Поскольку $\widetilde{S}_{\text{набл}} \approx 0{,}4 < \widetilde{S}_{\text{кр}} = 1{,}65$, то H_0 принимается. ■

Кроме сформулированного выше правостороннего биномиального критерия существуют также левосторонний и двусторонний биномиальные критерии для малых ($n < 30$) и больших ($n \geqslant 30$) выборок.

Если рассматривается гипотеза H_0: $p = p_0$ против альтернативы H_1: $p < p_0$, то левосторонний биномиальный критерий для малых выборок утверждает, что на заданном уровне α принимается H_0, если наблюденное значение $m > S_{\text{кр}}$ — критического значения, найденного из соответствующей таблицы по заданным p_0, n, α. В случае же $m \leqslant S_{\text{кр}}$ гипотеза H_0 отвергается на уровне α.

Если же рассматривается гипотеза H_0: $p = p_0$ против альтернативы H_1: $p \neq p_0$, то двусторонний биномиальный критерий для малых выборок утверждает, что на заданном уровне α принимается H_0, если $S^{(1)}_{\text{кр}} < m < S^{(2)}_{\text{кр}}$, где критические значения $S^{(1)}_{\text{кр}}$ и $S^{(2)}_{\text{кр}}$ находятся из соответствующих таблиц по заданным p_0, n, α. Гипотеза H_0 на уровне α отвергается, если $m \leqslant S^{(1)}_{\text{кр}}$ или $m \geqslant S^{(2)}_{\text{кр}}$.

В случае большой выборки для левостороннего и двустороннего критериев, как и для правостороннего критерия, вместо статистики S берут статистику $\widetilde{S}$.

Все перечисленные биномиальные критерии являются достаточно простыми для применения и непараметрическими, так как они справедливы без всяких дополнительных допущений относительно рассматриваемых совокупностей.

2. Критерий знаков

Пусть из интересующей нас совокупности испытуемых извлечена случайная выборка объема n, и пусть для каждого испытуемого получены два замера исследуемого признака — до и после некоторого воздействия на них, некоторой их обработки. На основании полученных повторных парных наблюдений требуется проверить гипотезу о наличии сдвига в распределении признака из-за обработки во всей совокупности. Такая гипотеза может быть проверена с помощью критерия знаков. Сформулируем его.

Обозначим наблюдения признака до обработки через $x_1, x_2, \ldots, x_n$, а после обработки — через $y_1, y_2, \ldots, y_n$. Пусть $z_1 = y_1 - x_1$, $z_2 = y_2 - x_2, \ldots, z_n = y_n - x_n$. Предположим, что для разностей $z_1, z_2, \ldots, z_n$ справедлива следующая статистическая модель: $z_i = \Theta + \varepsilon_i$, $i = 1, 2, \ldots, n$, где Θ — интересующая нас неизвестная постоянная (неизвестный эффект обработки), а $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ — ненаблюдаемые взаимно независимые случайные величины, извлеченные из непре-

рывной совокупности, имеющей медиану, равную нулю. Это аддитивная модель отражения результатов обработки.

Сформулируем нулевую гипотезу H_0: $\Theta = 0$, альтернативную гипотезу H_1: $\Theta > 0$ и зададимся уровнем значимости α. Для заданных α, n и вероятности $p = 1/2$ по табл. IX биномиального распределения найдем критическое значение $S_{\text{кр}}$ и сравним его с подсчитанным числом $S_{\text{набл}}$ положительных величин среди $z_1, z_2, \ldots, z_n$. При справедливости вышеуказанной статистической модели и при малых выборках ($n < 30$) имеет место следующий *правосторонний критерий знаков*: на уровне α гипотеза H_0 принимается, если $S_{\text{набл}} < S_{\text{кр}}$, и гипотеза H_0 отклоняется, если $S_{\text{набл}} \geqslant S_{\text{кр}}$.

Аналогичным образом можно сформулировать левосторонний и двусторонний критерии знаков для малых выборок, а также критерии знаков и для больших выборок ($n \geqslant 30$).

Пример 3. Для случайной выборки из десяти учащихся из некоторой совокупности учащихся была проведена контрольная работа по некоторому предмету до и после специального тренинга. До тренинга были получены следующие оценки: 2, 3, 4, 5, 3, 3, 2, 2, 4, 3. После тренинга учащимися в том же порядке были получены такие оценки: 3, 4, 5, 5, 3, 3, 3, 2, 4, 3. На уровне значимости $\alpha = 0{,}01$ с помощью критерия знаков требуется проверить гипотезу H_0 о том, что нет эффекта тренинга, против альтернативы H_1 о том, что есть положительный эффект тренинга.

Решение. Из приведенных оценок находим, что число положительных разностей оценок $S_{\text{набл}} = 4$. По заданным $\alpha = 0{,}01$, $n = 10$, $p = 0{,}5$ из таблицы получаем, что $S_{\text{кр}} = 9$. Так как $S_{\text{набл}} = 4 < S_{\text{кр}} = 9$, то на уровне значимости $\alpha = 0{,}01$ принимается гипотеза H_0 о том, что тренинг не дал эффекта. ■

Критерий знаков является частным случаем биномиального критерия и, следовательно, является непараметрическим критерием.

Если необходимо проверить гипотезу H_0: $\Theta = \Theta_0$ против альтернативы H_1: $\Theta > \Theta_0$, где Θ_0 — некоторое число, отличное от нуля, то вместо $z_1, z_2, \ldots, z_n$ рассматриваются $z_1 - \Theta_0, z_2 - \Theta_0, \ldots, z_n - \Theta_0$. Все остальные рассуждения критерия знаков сохраняются.

Замечание. Критерий знаков применяется и для одной выборки наблюдений $z_1, z_2, \ldots, z_n$, если все наблюдения подчиняются выше сформулированной статистической модели. Тогда для проверки гипотезы H_0: $\Theta = \Theta_0$, где Θ_0 — некоторое заданное число против альтернативы H_1: $\Theta > \Theta_0$ находят разности $z_1 - \Theta_0, z_2 - \Theta_0, \ldots, z_n - \Theta_0$, подсчитывают число положительных таких разностей и далее действуют так же, как и выше.

§ 2. Критерии проверки гипотез о числовых значениях параметров нормального распределения

Как показывает практика, нормальному закону распределения подчиняются самые разнообразные случайные величины (признаки). Примерами могут служить случайные ошибки измерений и рост или вес случайно взятого человека. Считается также нормальным распределение общих способностей (общего интеллекта). Кроме того, в силу центральной предельной теоремы распределение целого ряда распространенных на практике статистик и статистических оценок является асимптотически (при $n \to \infty$) нормальным. Все это указывает на важность для практики нормального закона распределения. Не следует, однако, считать этот закон универсальным. Например, его применение к таким психологическим категориям, как личностная и мотивационная сфера, является спорным. Несмотря на это, при обработке экспериментальных данных на первом этапе часто из априорных соображений приходится считать, что выборка наблюдений получена из нормально распределенной совокупности. И тогда возникает вопрос о числовых значениях параметров a и σ^2 нормального распределения.

Точнее, возникает вопрос о проверке гипотез о числовых значениях неизвестных параметров a и σ^2 нормального распределения случайной величины (признака). Напомним, что параметр a задает математическое ожидание, а параметр σ^2 — дисперсию случайной величины (признака).

1. Критерии проверки гипотез о параметре a

Пусть получена случайная выборка наблюдений $x_1, x_2, \ldots, x_n$ нормально распределенной случайной величины (признака) с неизвестными в общем случае параметрами a и σ^2. Необходимо проверить гипотезу H_0 о том, что $a = a_0$, где a_0 — заданное число, при заданном уровне значимости α. Для альтернативы H_1 возможен один из следующих вариантов: $a > a_0$, $a < a_0$, $a \neq a_0$. В зависимости от выбора варианта альтернативы H_1 получаем правосторонний, левосторонний или двусторонний критерий проверки гипотезы о численном значении параметра a. При этом приходится различать два случая: σ известно и σ неизвестно.

В случае известного параметра σ рассматривается статистика

$$\gamma = (\overline{x} - a_0)\frac{\sqrt{n}}{\sigma},$$

где $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ — среднее значение выборки.

Известно, что при гипотезе H_0 статистика γ имеет стандартное нормальное распределение: $\gamma \sim N(0, 1)$.

Для заданной выборки наблюдений $x_1, x_2, \ldots, x_n$ находим значение $\gamma_{\text{набл}}$ статистики γ.

Если имеется односторонняя гипотеза H_1: $a > a_0$ или $a < a_0$, то из таблицы значений функции $\Phi(x)$ находим критическое значение $\gamma_{\text{кр}}$ статистики γ по формуле $\Phi(\gamma_{\text{кр}}) = (1 - 2\alpha)/2$. Тогда *правосторонний критерий* утверждает, что на уровне значимости α принимается гипотеза H_0, если $\gamma_{\text{набл}} < \gamma_{\text{кр}}$, и гипотеза H_0 отвергается, если $\gamma_{\text{набл}} \geqslant \gamma_{\text{кр}}$.

В случае гипотезы H_1: $a < a_0$ *левосторонний критерий* утверждает, что на уровне значимости α гипотеза H_0 принимается, если $\gamma_{\text{набл}} < -\gamma_{\text{кр}}$ и гипотеза H_0 отвергается, если $\gamma_{\text{набл}} \leqslant -\gamma_{\text{кр}}$.

Наконец, в случае двусторонней гипотезы H_1: $a \neq a_0$ из табл. III значений функции $\Phi(x)$ находим критическое значение $\gamma_{\text{кр}}$ статистики γ по формуле $\Phi(\gamma_{\text{кр}}) = (1 - \alpha)/2$. Тогда *двусторонний критерий* утверждает, что на уровне значимости α гипотеза H_0 принимается, если $|\gamma_{\text{набл}}| < \gamma_{\text{кр}}$, и гипотеза H_0 отвергается, если $|\gamma_{\text{набл}}| \geqslant \gamma_{\text{кр}}$.

Другими словами, двусторонний критерий говорит, что H_0 принимается при уровне значимости α, если значение a_0, попадает в доверительный интервал для a с коэффициентом доверия $(1 - \alpha)$.

Пример 1. С помощью определенной методики проводится исследование общего интеллекта детей некоторого детского учреждения. Предполагается, что общий интеллект распределен нормально с неизвестным параметром a и известным параметром $\sigma = 4$ балла. Для случайной выборки из девяти детей было определено $\overline{x} = 45$ баллов. Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о том, что $a = 50$ баллов, против альтернативы H_1: $a \neq 50$ баллов.

Решение. Используем двусторонний критерий. По формуле $\Phi(\gamma_{\text{кр}}) = (1 - \alpha)/2 = (1 - 0{,}05)/2 = 0{,}475$ из таблицы значений функции $\Phi(x)$ получаем, что $\gamma_{\text{кр}} = 1{,}96$.

Далее

$$\gamma_{\text{набл}} = (\overline{x} - a_0)\frac{\sqrt{n}}{\sigma} = (45 - 50)\frac{\sqrt{9}}{4} = -3{,}75.$$

Так как $|\gamma_{\text{набл}}| = 3{,}75 > \gamma_{\text{кр}} = 1{,}96$, то на уровне $\alpha = 0{,}05$ гипотеза H_0 отвергается. ■

В случае неизвестного параметра σ ограничимся рассмотрением двустороннего критерия.

В случае неизвестного параметра σ рассматривается статистика

$$t = (\overline{x} - a_0)\frac{\sqrt{n-1}}{s},$$

где s^2 — выборочная дисперсия, т. е. $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$, а $\overline{x}$ — выборочное среднее значение.

Известно, что статистика t имеет распределение Стьюдента с $(n-1)$ степенями свободы. Поэтому критерий проверки гипотезы H_0: $a = a_0$ на заданном уровне значимости α в этом случае называют *критерием Стьюдента*.

По заданной выборке наблюдений находим число $t_{\text{набл}}$. По заданным уровню $\alpha/2$ и числу степеней свободы $(n-1)$ из табл. VII распределения Стьюдента находим критическое значение $t_{\text{кр}}$.

Рассмотрим гипотезу H_0: $a = a_0$ и альтернативу H_1: $a \neq a_0$. Тогда двусторонний критерий Стьюдента утверждает, что на уровне значимости α гипотеза H_0 принимается, если $|t_{\text{набл}}| < t_{\text{кр}}$, и гипотеза H_0 отвергается, если $|t_{\text{набл}}| \geqslant t_{\text{кр}}$.

Замечание. Нетрудно заметить, что, отыскивая двустороннюю критическую область параметра a при уровне значимости α, тем самым находят и доверительный интервал для неизвестного параметра a с доверительной вероятностью $(1-\alpha)$. Этот доверительный интервал содержит те значения параметра a, которые совместимы с полученными наблюдениями при проверке соответствующих гипотез на уровне α.

Пример 2. С целью изучения общего интеллекта первоклассников некоторой школы было проведено тестирование по определенной методике случайной выборки из десяти учащихся. При этом были получены выборочные среднее $\overline{x} = 52$ и среднее квадратическое отклонение $s = 4$. В предположении о нормальности распределения общего интеллекта первоклассников требуется при уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 о том, что неизвестный параметр a этого распределения равен 50 против альтернативы H_1: $a \neq 50$.

Решение. Сначала найдем число

$$t_{\text{набл}} = (\overline{x} - a_0)\frac{\sqrt{n-1}}{s} = (52-50)\frac{\sqrt{9}}{4} = 1{,}5.$$

По заданным $\alpha/2 = 0{,}025$ и $(n-1) = 10-1 = 9$ по таблице распределения Стьюдента находим, что $t_{\text{кр}} = 2{,}26$. Поскольку $t_{\text{набл}} = 1{,}5 < t_{\text{кр}} = 2{,}26$, то гипотеза H_0: $a = 50$ принимается на уровне $\alpha = 0{,}05$. ■

К рассмотренным выше задачам сводится и так называемая задача о парных данных, о которой речь была уже ранее, при изложении критерия знаков. Однако здесь допущения о задаче с парными данными будут более сильные, чем при рассмотрении критерия знаков.

Итак, пусть из некоторой совокупности получена случайная выборка испытуемых (или объектов) объема n и для каждого испытуемого (объекта) дважды проведены измерения определенной характеристики: до и после некоторого воздействия. Пусть для i-го испытуемого (объекта) получена пара наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$. Таким образом, имеем n пар наблюдений: (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n). Положим $z_1 = y_1 - x_1$, $z_2 = y_2 - x_2$, $\ldots$, $z_n = y_n - x_n$ и предположим, что для $z_1, z_2, \ldots, z_n$ имеет место следующая статистическая модель:

1) $z_1, z_2, \ldots, z_n$ взаимно независимы,

2) $z_i = \Theta + \varepsilon_i$, $i = 1, 2, \ldots, n$, где Θ — неизвестная постоянная величина (результат воздействия, эффект обработки), а $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ — ненаблюдаемые, независимые случайные величины,

3) случайные величины $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ распределены по нормальному закону $N(0, \sigma^2)$, где дисперсия σ^2 обычно неизвестна.

В отношении неизвестной величины Θ ставится вопрос о проверке гипотезы H_0: $\Theta = 0$ (нет эффекта обработки) или $\Theta = \Theta_0$, где Θ_0 задано.

В случае неизвестной дисперсии σ^2 для проверки гипотезы H_0 используется соответствующий критерий Стьюдента. Например, для гипотезы H_0: $\Theta = 0$ рассматривается статистика

$$t = \overline{z} \cdot \frac{\sqrt{n}}{\widetilde{s}},$$

где $\overline{z}$ — выборочное среднее для $z_1, z_2, \ldots, z_n$, а $\widetilde{s}^2$ — исправленная выборочная дисперсия. Вычисляется выборочное значение $t_{\text{набл}}$ и сравнивается с критическим значением $t_{\text{кр}}$, найденным по таблице распределения Стьюдента по заданному уровню значимости α и по числу степеней свободы $2(n-1)$. Тогда двусторонний критерий Стьюдента гласит, что на уровне α гипотеза H_0 принимается, если $|t_{\text{набл}}| < t_{\text{кр}}$, и отвергается в противном случае в пользу альтернативы H_1: $\Theta \neq 0$.

2. Сравнение параметров a_1 и a_2 двух нормальных распределений

Пусть рассматриваются две нормально распределенные случайные величины (признаки) с параметрами (a_1, σ_1^2) и (a_2, σ_2^2) соответственно. Относительно параметров σ_1^2 и σ_2^2 предполагается выполненным один из следующих вариантов:

1) обе дисперсии известны и равны между собой,

2) обе дисперсии известны, но не равны между собой,

3) обе дисперсии неизвестны, но предполагается, что они равны между собой,

4) обе дисперсии неизвестны и их равенство не предполагается.

Для параметров a_1 и a_2 необходимо проверить гипотезу H_0: $a_1 = a_2$ против альтернативы H_1: $a_1 \neq a_2$.

Проверка гипотез H_0 и H_1 проводится с помощью заданного уровня значимости α случайной выборки независимых наблюдений $x_1, x_2, \ldots, x_n$ объема n первой случайной величины с параметрами (a_1, σ_1^2) и случайной выборки независимых наблюдений $y_1, y_2, \ldots, y_m$ объема m второй случайной величины с параметрами (a_2, σ_2^2).

Например, проверка таких гипотез необходима при исследовании уровня интересующего нас признака, нормально распределенного в контрольной и рабочей совокупностях объектов.

Ограничимся формулировкой соответствующего критерия Стьюдента лишь для случая 3.

В предположении, что параметры σ_1^2 и σ_2^2 неизвестны, но равны между собой рассматривается статистика

$$t = \frac{(\overline{x} - \overline{y})\sqrt{mn}}{s\sqrt{m+n}},$$

где

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i, \quad s^2 = \frac{s_1^2(n-1) + s_2^2(m-1)}{n+m-2},$$

s_1^2 — выборочная дисперсия выборки $x_1, x_2, \ldots, x_n$, а s_2^2 — выборочная дисперсия выборки $y_1, y_2, \ldots, y_m$.

Статистика t имеет распределение Стьюдента с $(n+m-2)$ степенями свободы при гипотезе H_0.

Для заданных выборок находим выборочное значение $t_{\text{набл}}$ для статистики t, а затем на заданном уровне значимости $\alpha/2$ по таблице распределения Стьюдента с $(n+m-2)$ степенями свободы находим критическое значение статистики t.

Двусторонний критерий Стьюдента для двух выборок гласит: если $|t_{\text{набл}}| < t_{\text{кр}}$, то на уровне значимости $\alpha/2$ принимается гипотеза H_0: $a_1 = a_2$. Если же $|t_{\text{набл}}| \geqslant t_{\text{кр}}$, то на уровне значимости α принимается альтернатива H_1: $a_1 \neq a_2$.

Аналогичным образом можно сформулировать левосторонний и правосторонний критерии Стьюдента для двух выборок. Если гипотеза H_0 принимается, то говорят, что различие выборочных средних $\overline{x}$ и $\overline{y}$ статистически незначимо.

Пример 3. Перед спортивным соревнованием производится взвешивание пяти случайно отобранных членов одной команды и шести случайно отобранных членов другой команды. При взвешивании пяти человек из первой команды получены $\overline{x} = 65$ кг, $s_1^2 = 10$ кг, а при взвешивании шести человек из второй команды получены $\overline{y} = 66$ кг, $s_2^2 = 13{,}33$ кг. Считая распределение веса спортсменов обеих команд нормальным, с неизвестным одинаковым параметром σ^2, проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о том, что средние уровни веса двух команд спортсменов не различаются.

Решение. Пусть средний вес спортсменов первой команды равен a_1, а средний вес спортсменов второй команды равен a_2. Необходимо проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0: $a_1 = a_2$ против альтернативы H_1: $a_1 \neq a_2$.

Воспользуемся двусторонним критерием Стьюдента для двух выборок. Находим значение s^2 по формуле

$$s^2 = \frac{s_1^2(n-1) + s_2^2(m-1)}{n+m-2} = \frac{10 \cdot 4 + 13{,}33 \cdot 5}{5+6-2} = \frac{66{,}65+40}{9} \approx 11{,}85.$$

Затем находим выборочное значение $t_{\text{набл}}$ по формуле

$$t_{\text{набл}} = \frac{(\overline{x}-\overline{y})\sqrt{mn}}{s\sqrt{m+n}} = \frac{(65-66)\sqrt{5 \cdot 6}}{\sqrt{11{,}85} \cdot \sqrt{5+6}} \approx -0{,}48.$$

При $\alpha = 0{,}05$ и числе степеней свободы $n+m-2 = 5+6-2 = 9$ по таблице распределения Стьюдента получаем $t_{\text{кр}} = 2{,}26$. Поскольку $t_{\text{набл}} = 0{,}48 < t_{\text{кр}} = 2{,}26$, то на уровне значимости $\alpha/2 = 0{,}025$ гипотеза H_0 принимается. ■

3. Критерии проверки гипотез о параметре σ^2

Пусть по случайной выборке наблюдений $x_1, x_2, \ldots, x_n$ случайной величины (признака), распределенной нормально с неизвестными параметрами (a, σ^2), на уровне значимости α необходимо проверить гипотезу H_0: $\sigma^2 = \sigma_0^2$, где σ_0^2 — некоторая заданная величина.

На практике гипотеза H_0 проверяется, если нужно определить степень рассеяния или проверить точность методики исследования, точность инструментов.

При проверке гипотезы H_0 используют статистику

$$\gamma = \frac{ns^2}{\sigma_0^2},$$

где s^2 — выборочная дисперсия, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$, $\overline{x}$ — выборочное среднее.

Известно, что при гипотезе H_0 статистика γ распределена по закону χ^2 (хи-квадрат) с $(n-1)$ степенями свободы. В зависимости от выбора альтернативы H_1: $\sigma^2 > \sigma_0^2$, H_1: $\sigma^2 < \sigma_0^2$ или H_1: $\sigma^2 \neq \sigma_0^2$ выбирают правосторонний, левосторонний или двусторонний критерий проверки гипотезы H_0 на уровне значимости α.

В случае H_1: $\sigma^2 > \sigma_0^2$ по табл. VI распределения χ^2 по заданным α и числу степеней свободы $(n-1)$ находят критическое значение $\gamma_{\text{кр}}$ статистики γ и сравнивают выборочное значение $\gamma_{\text{набл}}$ с $\gamma_{\text{кр}}$.

Правосторонний критерий. Если $\gamma_{\text{набл}} < \gamma_{\text{кр}}$, то на уровне значимости α гипотеза H_0 принимается. В противном случае принимается гипотеза H_1 на уровне значимости α.

В случае H_1: $\sigma^2 < \sigma_0^2$ по табл. VI χ^2-распределения по заданным $(1-\alpha)$ и $(n-1)$ находят $\gamma_{\text{кр}}$ и сравнивают его с $\gamma_{\text{набл}}$.

Левосторонний критерий. Если $\gamma_{\text{набл}} > \gamma_{\text{кр}}$, то на уровне значимости α принимается гипотеза H_0. В противном случае на уровне α принимается гипотеза H_1.

Наконец, для гипотезы H_1: $\sigma^2 \neq \sigma_0^2$ из таблиц χ^2-распределения находят два критических значения — $\gamma_{\text{кр}}^{(1)}$ и $\gamma_{\text{кр}}^{(2)}$, причем $\gamma_{\text{кр}}^{(1)}$ находится по значениям $(1-\alpha/2)$ и $(n-1)$, $\gamma_{\text{кр}}^{(2)}$ находится по значениям $\alpha/2$ и $(n-1)$.

Двусторонний критерий. Если $\gamma_{\text{кр}}^{(1)} < \gamma_{\text{набл}} < \gamma_{\text{кр}}^{(2)}$, то на уровне значимости α гипотеза H_0 принимается. Если же $\gamma_{\text{набл}} < \gamma_{\text{кр}}^{(1)}$ или $\gamma_{\text{набл}} > \gamma_{\text{кр}}^{(2)}$, то гипотеза H_0 отвергается на уровне значимости α.

Пример 4. При измерении роста случайно отобранных десяти учащихся некоторого класса было получено $s^2 = 11$ см^2. Считая нормальным распределение роста учащихся класса, проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о том, что дисперсия роста учащихся класса $\sigma^2 = 13$ см^2 против альтернативы H_1: $\sigma^2 \neq 13$ см^2.

Решение. Воспользуемся двусторонним критерием проверки гипотезы H_0. Сначала вычисляем по формуле

$$\gamma_{\text{набл}} = \frac{ns^2}{\sigma_0^2} = \frac{10 \cdot 11}{13} \approx 8{,}462.$$

Затем по табл. VI χ^2-распределения по $\alpha/2 = 0{,}025$ и числу степеней свободы $n-1 = 10-1 = 9$ находим $\gamma_{\text{кр}}^{(2)} = 19{,}023$, а по $1-\alpha/2 = 0{,}975$ и числу $n-1 = 9$ находим $\gamma_{\text{кр}}^{(1)} = 2{,}7$. Так как $2{,}7 < \gamma_{\text{набл}} = 8{,}462 < 19{,}023$, то на уровне значимости $\alpha = 0{,}05$ гипотеза H_0 принимается. ■

4. Сравнение параметров σ_1^2 и σ_2^2 двух нормальных распределений

На практике иногда приходится проверять гипотезу о равенстве двух дисперсий, если требуется сравнить рассеяние двух признаков или точность двух методов, двух инструментов. Ясно, что выбирается тот метод, тот прибор, который обеспечивает меньшую дисперсию.

Пусть имеются две случайные величины (признаки), которые нормально распределены с неизвестными дисперсиями σ_1^2 и σ_2^2 соответственно, и получены случайная выборка наблюдений объема n первой случайной величины и независимая случайная выборка наблюдений объема m второй случайной величины. Пусть, кроме того, найдены выборочные дисперсии s_1^2 и s_2^2 для случайных выборок.

Требуется при заданном уровне значимости α проверить гипотезу H_0: $\sigma_1^2 = \sigma_2^2$.

Проверка такой гипотезы проводится с помощью одного из критериев Фишера: правостороннего, левостороннего или двустороннего, в зависимости от типа альтернативы H_1: $\sigma_1^2 > \sigma_2^2$, H_1: $\sigma_1^2 < \sigma_2^2$ или H_1: $\sigma_1^2 \neq \sigma_2^2$.

Будем считать для определенности, что $s_1^2 > s_2^2$. Тогда в качестве статистики критерия Фишера берется величина

$$F = \frac{s_1^2}{s_2^2},$$

которая при гипотезе H_0 имеет распределение Фишера—Снедекора со степенями свободы $\nu_1 = n - 1$ и $\nu_2 = m - 1$. Сначала по заданным s_1^2 и s_2^2 вычисляют наблюденное значение $F_{\text{набл}}$ статистики F по формуле

$$F_{\text{набл}} = \frac{s_1^2}{s_2^2}.$$

Для альтернативы H_1: $\sigma_1^2 > \sigma_2^2$ из табл. VIII распределения Фишера—Снедекора по заданным числам α, ν_1, ν_2 находят критическое значение $F_{\text{кр}}$ статистики F.

Правосторонний критерий Фишера. Если $F_{\text{набл}} < F_{\text{кр}}$, то на уровне значимости α гипотеза H_0: $\sigma_1^2 = \sigma_2^2$ принимается, а если $F_{\text{набл}} \geqslant F_{\text{кр}}$, то на уровне значимости α гипотеза H_0 отвергается и принимается гипотеза H_1: $\sigma_1^2 > \sigma_2^2$.

Для альтернативы $\sigma_1^2 \neq \sigma_2^2$ по таблице распределения Фишера—Снедекора находят критическое значение $F_{\text{кр}}^{(1)}$ по заданным $\alpha/2$, $\nu_1 = n - 1$, $\nu_2 = m - 1$ и критическое значение $F_{\text{кр}}^{(2)}$ по заданным $(1 - \alpha/2)$, $\nu_1 = n - 1$, $\nu_2 = m - 1$.

Двусторонний критерий Фишера. Если $F_{кр}^{(1)} < F_{набл} < F_{кр}^{(2)}$, то на уровне значимости α гипотеза H_0: $\sigma_1^2 = \sigma_2^2$ принимается. В противном случае на уровне значимости α принимается альтернатива H_1: $\sigma_1^2 \neq \sigma_2^2$.

Пример 5. При измерении роста случайно отобранных 12 учащихся одного класса было получено значение $s_1^2 = 14$ см^2, а при измерении роста случайно отобранных пятнадцати учащихся другого класса было получено значение $s_2^2 = 12$ см^2. На уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0: $\sigma_1^2 = \sigma_2^2$ против альтернативы H_1: $\sigma_1^2 > \sigma_2^2$, если считать, что рост учащихся одного класса и рост учащихся другого класса распределены нормально с дисперсиями σ_1^2 и σ_2^2 соответственно.

Решение. Воспользуемся правосторонним критерием Фишера. По заданным s_1^2 и s_2^2 найдем значение

$$F_{набл} = \frac{s_1^2}{s_2^2} = \frac{14}{12} = 1{,}167.$$

По заданным $\alpha = 0{,}05$, $v_1 = n - 1 = 12 - 1 = 11$, $v_2 = m - 1 = 15 - 1 = 14$ из таблицы распределения Фишера—Снедекора находим $F_{кр} = 2{,}53$. Так как $F_{набл} = 1{,}167 < F_{кр} = 2{,}53$, то на уровне значимости $\alpha = 0{,}05$ гипотеза H_0 принимается. ■

Отметим в заключение, что все приведенные в этом параграфе статистические критерии являются параметрическими, так как все они основаны на предположении о нормальном законе распределения исследуемой случайной величины. Кроме того, приведенные в этом параграфе статистические критерии очень чувствительны к выбросам в наблюдениях, т. е. к наблюдениям, резко отличающимся от остальных. Другими словами, они являются необастными (неустойчивыми) по отношению к выбросам, и необходимо исключать из рассмотрения такого рода наблюдения с помощью специальной процедуры.

§ 3. Критерии согласия

Критерии согласия предназначены для проверки гипотез о модельном виде закона распределения исследуемой случайной величины (признака). Они проверяют согласие опытных независимых наблюдений и предполагаемой модели закона распределения. Следует

различать простые и сложные гипотезы об этом законе. Простая гипотеза однозначно задает закон распределения, а сложная гипотеза определяет закон распределения с точностью до его параметров, т. е. задает лишь некоторое параметрическое семейство законов распределения с неизвестными параметрами. Например, сложной является гипотеза о том, что исследуемая случайная величина распределена по нормальному закону с неизвестными параметрами a и σ^2 или распределена по закону Пуассона с неизвестным параметром λ. Приведем критерии согласия χ^2 (хи-квадрат) (К. Пирсона) для простой и сложной гипотез, а также критерий Колмогорова—Смирнова для простой гипотезы.

1. Критерий согласия хи-квадрат Пирсона

Этот критерий применяется как для непрерывной, так и для дискретной случайной величины ξ. Обозначим неизвестную функцию распределения исследуемой случайной величины ξ через $F_\xi(x)$ и выдвинем простую гипотезу H_0: $F_\xi(x) = F(x)$, где $F(x)$ — однозначно заданная функция, против альтернативы H_1: $F_\xi(x) \neq F(x)$.

Для проверки этих гипотез задаются уровень значимости α и случайная выборка независимых наблюдений исследуемой случайной величины $x_1, x_2, \ldots, x_n$, причем все элементы выборки разбиты на l различных значений в случае дискретной ξ или на l интервалов в случае непрерывной ξ с частотами $n_1, n_2, \ldots, n_l$ так, что $n_1 + n_2 + \ldots + n_l = n$, $1 \leqslant l \leqslant n$.

Так как гипотеза H_0 — простая, то функция $F(x)$ полностью нам известна и, следовательно, известны вероятности $p_1, p_2, \ldots, p_l$ всех различных элементов для дискретной ξ или вероятности $p_1, p_2, \ldots, p_l$ попадания в каждый из l интервалов для непрерывной ξ, причем $p_1 + p_2 + \ldots + p_l = 1$.

Рассматривается статистика

$$\gamma = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}.$$

Известно, что асимптотически (при $n \to \infty$) распределение статистики γ приближается к χ^2-распределению с $(l-1)$ степенями свободы, если выполнена гипотеза H_0.

По заданной выборке наблюдений находим выборочное значение $\gamma_{\text{набл}}$ статистики γ. По таблице χ^2-распределения с $(l-1)$ степенями свободы находим два критических значения статистики γ: $\gamma^{(2)}_{\text{кр}}$ для $\alpha/2$ и $\gamma^{(1)}_{\text{кр}}$ для $(1-\alpha/2)$.

Двусторонний критерий хи-квадрат Пирсона. Если $\gamma_{кр}^{(1)} < \gamma_{набл} < \gamma_{кр}^{(2)}$, то на уровне значимости α принимается гипотеза H_0. В противном случае на уровне значимости α принимается гипотеза H_1.

Сформулируем еще и правосторонний критерий проверки гипотезы H_0. По таблице χ^2-распределения для заданного α и $(l-1)$ степеней свободы находим критическое значение $\gamma_{кр}$ статистики γ.

Правосторонний критерий χ^2 (*хи-квадрат*) (*К. Пирсона*). Если $\gamma_{набл} < \gamma_{кр}$, то на уровне значимости α принимается гипотеза H_0, а если $\gamma_{набл} \geqslant \gamma_{кр}$, то на уровне α гипотеза H_0 отклоняется.

Асимптотический характер (при $n \to \infty$) распределения статистики γ требует осторожного применения на практике этих критериев. На них можно полагаться лишь при больших объемах n выборки. Практика применения этих критериев рекомендует выбирать n столь большим, чтобы все $np_i \geqslant 10$, а число l таким, чтобы каждая частота в интервале $n_i \geqslant 5$, $i = 1, 2, \ldots, l$.

Заметим, что формулу статистики γ можно упростить следующим образом:

$$\sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{l} \frac{n_i^2 - 2n \cdot n_i \cdot p_i + n^2 p_i^2}{np_i} = \sum_{i=1}^{l} \left(\frac{n_i^2}{np_i} - 2n_i + np_i \right) =$$

$$= \frac{1}{n} \sum_{i=1}^{l} \frac{n_i^2}{p_i} - 2n + n = \frac{1}{n} \sum_{i=1}^{l} \frac{n_i^2}{p_i} - n.$$

Здесь были использованы следующие равенства:

$$\sum_{i=1}^{l} n_i = n, \quad \sum_{i=1}^{l} p_i = 1.$$

Критерии χ^2 (хи-квадрат) (К. Пирсона) применяются и для проверки сложной гипотезы о виде закона распределения с k неизвестными параметрами исследуемой случайной величины. В этом случае вместо неизвестных параметров закона распределения берут их оценки максимального правдоподобия. Например, в случае нормального закона распределения с неизвестными параметрами a и σ^2 вместо a берут выборочное среднее значение $\overline{x}$, а вместо σ^2 берут выборочную дисперсию s^2. Далее рассматривается та же статистика γ, что и для простой гипотезы, и используется тот факт. что при гипотезе H_0 для больших n ($n \to \infty$) распределение статистики γ приближается к χ^2-распределению с $(l-k-1)$ степенями свободы. Затем находят выборочное значение $\gamma_{набл}$ статистики γ и, если, например, нужно воспользоваться правосторонним критерием хи-

квадрат, по заданному уровню значимости и по числу степеней свободы $(l-k-1)$ из таблицы χ^2-распределения находят критическое значение $\gamma_{\text{кр}}$ статистики γ. Формулировка правостороннего критерия хи-квадрат (К. Пирсона) в этом случае такая же, как и выше.

Пример 1. На первом курсе учится 150 студентов. После сдачи четырех экзаменов в экзаменационную сессию получены следующие результаты: 5 студентов не сдали ни одного экзамена, 6 студентов сдали один экзамен, 9 студентов сдали два экзамена, 40 студентов сдали три экзамена и 90 студентов сдали четыре экзамена. Пусть сдачи четырех экзаменов являются независимыми и вероятность сдачи студентом любого экзамена одна и та же и равна неизвестному значению p. Требуется на уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 о том, что число сданных экзаменов из четырех подчиняется биномиальному закону распределения.

Решение. Для нашего примера $n = 150$, $n_1 = 5$, $n_2 = 6$, $n_3 = 9$, $n_4 = 40$, $n_5 = 90$, $l = 5$, $k = 1$.

Так как значение вероятности p неизвестно, вместо p берем его статистическую оценку — относительную частоту (частость) $\widehat{p}$, вычисляемую по формуле

$$\widehat{p} = \frac{0 \cdot 5 + 1 \cdot 6 + 2 \cdot 9 + 3 \cdot 40 + 4 \cdot 90}{4 \cdot 150} = 0{,}84.$$

Подсчитаем вероятность p каждого из значений $x = 0,\ 1,\ 2,\ 3,\ 4$ по формуле вероятности числа успехов в 150 испытаниях Бернулли:

$$p(x) = C_4^x \widehat{p}(1-\widehat{p})^{4-x} = C_4^x (0{,}84)^x (0{,}16)^{4-x}.$$

Имеем $p_1 = \mathsf{P}(0) = 0{,}00066$, $p_2 = \mathsf{P}(1) = 0{,}01376$, $p_3 = \mathsf{P}(2) = 0{,}10838$, $p_4 = \mathsf{P}(3) = 0{,}37933$, $p_5 = \mathsf{P}(4) = 0{,}49787$.

Находим выборочное значение $\gamma_{\text{набл}}$ статистики γ:

$$\gamma_{\text{набл}} = \frac{1}{n}\sum_{i=1}^{l}\frac{n_i^2}{p_i} - n = \frac{1}{150}\left(\frac{5^2}{p_1} + \frac{6^2}{p_2} + \frac{9^2}{p_3} + \frac{40^2}{p_4} + \frac{90^2}{p_5}\right) - 150 = 261{,}53.$$

Из таблицы χ^2-распределения при $\alpha = 0{,}05$ и числе степеней свободы $l - k - 1 = 5 - 1 - 1 = 3$ находим критическое значение $\gamma_{\text{кр}} = 7{,}8$. Поскольку $\gamma_{\text{набл}} > \gamma_{\text{кр}}$, то на уровне значимости $\alpha = 0{,}05$ гипотеза H_0 отклоняется. ■

Следует отметить, что критерии согласия хи-квадрат Пирсона являются непараметрическими, так как не накладываются никакие ограничения на функцию $F(x)$.

2. Критерий согласия Колмогорова—Смирнова

Критерий согласия Колмогорова—Смирнова применяется лишь для проверки простой гипотезы H_0: $F_\xi(x) = F(x)$, где $F(x)$ — однозначно заданная и непрерывная функция. Таким образом, этот критерий неприменим для дискретных распределений: биномиального, распределения Пуассона и т. п.

Пусть заданы уровень значимости α и случайная выборка независимых наблюдений x_1, x_2, …, x_n интересующей нас случайной величины (признака) ξ.

Проверка гипотезы H_0 с помощью рассматриваемого критерия согласия основана на проверке близости выборочной функции распределения $F_n(x)$ и предполагаемой функции распределения $F(x)$ случайной величины ξ.

Рассматривается статистика Колмогорова

$$d_n = \sup_x |F_n(x) - F(x)|.$$

При малых n для статистики d_n составлены таблицы процентных точек, а при больших n используется таблица предельного (при $n \to \infty$) распределения статистики $\sqrt{n}\, d_n$.

На практике критерием Колмогорова—Смирнова пользуются следующим образом. Из полученной случайной выборки наблюдений строится вариационный ряд $x_{(1)}$, $x_{(2)}$, …, $x_{(n)}$ и вычисляется выборочное значение $d_{n,\text{набл}}$ статистики d_n для этого ряда по формуле

$$d_n = \max_{1 \leqslant k \leqslant n} \left[\frac{k}{n} - F(x_{(k)}),\ F(x_{(k)}) - \frac{k-1}{n} \right].$$

Выборочное значение $d_{n,\text{набл}}$ сравнивают с полученным из таблиц критическим значением $d_{n,\text{кр}}$.

Критерий согласия Колмогорова—Смирнова. Если $d_{n,\text{набл}} < d_{n,\text{кр}}$, то на уровне значимости α гипотеза H_0 принимается. В противном случае гипотеза H_0 отвергается на уровне значимости α.

Задачи к главе 6

1. В случайной выборке из пяти детей, записанных в первый класс школы, оказались четыре ребенка, умеющих читать. На уровне значимости $\alpha = 0{,}02$ проверить гипотезу H_0 о том, что во

всей совокупности детей, записанных в первый класс школы, вероятность ребенка, умеющего читать, $p = 0{,}05$ против альтернативы H_1: $p > 0{,}05$.

2. В случайной выборке из восьми детей подготовительной группы детского сада оказались три ребенка, умеющих считать до десяти. На уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 о том, что во всей подготовительной группе вероятность ребенка, умеющего считать до десяти, $p = 0{,}1$ против альтернативы H_1: $p > 0{,}1$.

3. С помощью специального теста психолог дважды измеряет (в баллах) уровень тревожности у 18 случайно выбранных студентов до и после специального тренинга. Результаты измерения до тренинга были следующие: 31, 35, 38, 34, 39, 40, 27, 27, 32, 24, 20, 34, 35, 34, 33, 33, 31, 32. Результаты измерения после тренинга для студентов в том же порядке были следующие: 33, 35, 35, 30, 36, 36, 38, 27, 25, 30, 25, 34, 36, 27, 35, 33, 33, 34. На уровне значимости $\alpha = 0{,}05$ с помощью критерия знаков проверить гипотезу H_0 о том, что нет эффекта от тренинга, против альтернативы H_1 о том, что есть положительный эффект от тренинга.

4. Для сравнения показателей уровня агрессивности до и после психотренинга некоторой группы подростков с помощью специального теста эти показатели были измерены (в баллах) дважды у 16 случайно выбранных подростков группы. На уровне значимости $\alpha = 0{,}1$ с помощью критерия знаков проверить гипотезу H_0 о том, что в группе подростков нет эффекта от психотренинга, против альтернативы H_1 о том, что имеется положительный эффект от психотренинга, если среди девяти из шестнадцати отобранных подростков наблюдался положительный эффект от психотренинга.

5. В случайной выборке из десяти семей города оказались три семьи с двумя сыновьями. На уровне значимости $\alpha = 0{,}01$ проверить гипотезу H_0 о том, что в городе вероятность семьи с двумя сыновьями $p = 0{,}3$, против альтернативы H_1: $p > 0{,}3$.

6. Известно, что число реализованных за некоторый период времени путевок в туристическом агентстве является случайной величиной, имеющей нормальное распределение с неизвестными параметрами a и σ^2. На уровне значимости $\alpha = 0{,}01$ проверить гипотезу H_0: $a = 10$ путевок против альтернативы H_1: $a \neq 10$ путевок, если при наблюдениях в течение случайно выбранных десяти дней были получены среднее число реализованных путевок $\overline{x} = 11$ и среднее квадратическое отклонение $s = 2$ путевки.

7. Известно, что возраст студентов вуза является нормально распределенной случайной величиной с неизвестными параметрами a и σ^2. На уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0: $a = 20$ лет против альтернативы H_1: $a \neq 20$ лет, если для случайно выбранных 26 студентов средний возраст $\overline{x} = 21$ год при среднем квадратическом отклонении $s = 0{,}5$ года.

8. Известно, что математические способности учащихся выпускных классов средней школы являются нормально распределенной случайной величиной с неизвестными параметрами a и σ^2. Для каждого из случайно выбранных шестнадцати учащихся дважды проведены измерения (в баллах) математических способностей: до и после определенного тренинга. На уровне $\alpha = 0{,}05$ проверить гипотезу H_0 о том, что нет эффекта от тренинга, против альтернативы H_1 о том, что есть эффект от тренинга, если разности полученных измерений после и до тренинга для выборки учащихся имеют среднее значение $\overline{z} = 4$ балла и исправленную дисперсию $\widetilde{s}^2 = 2$ балла.

9. Известно, что математические способности учащихся выпускных классов двух школ распределены нормально с неизвестными параметрами (a_1, σ^2) и (a_2, σ^2) соответственно. На уровне $\alpha = 0{,}01$ проверить гипотезу H_0: $a_1 = a_2$ против альтернативы H_1: $a_1 \neq a_2$, если для случайной выборки из десяти выпускников первой школы $\overline{x} = 65$ баллов, $s_1 = 2$ балла, а для случайной выборки из восьми выпускников второй школы $\overline{y} = 60$ баллов, $s_2 = 3$ балла.

10. Считая распределение веса (в кг) детей подготовительной группы детского сада нормальным с неизвестным параметром σ^2, на уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0: $\sigma^2 = 1$ кг^2 против альтернативы H_1: $\sigma^2 > 1$ кг^2, если при измерении веса случайно выбранных девяти детей подготовительных групп была получена выборочная дисперсия $s^2 = 1{,}5$ кг^2.

11. Известно, что распределения веса (в кг) детей подготовительных групп двух детских садов являются нормальными с неизвестными параметрами σ_1^2 и σ_2^2 соответственно. На уровне $\alpha = 0{,}1$ проверить гипотезу H_0: $\sigma_1^2 = \sigma_2^2$ против альтернативы H_1: $\sigma_1^2 > \sigma_2^2$, если при измерении веса случайно выбранных восьми детей первого детского сада получена выборочная дисперсия $s_1^2 = 1{,}5$ кг^2, а при измерении веса случайно выбранных девяти детей второго детского сада получена выборочная дисперсия $s_2^2 = 1$ кг^2.

12. Результаты измерения роста (в см) случайно выбранных 100 студентов некоторого университета разбиты на восемь интервалов и приведены в следующей таблице. В этой же таблице указаны

середина каждого интервала и частота попадания в каждый интервал.

Номер интервала	Интервалы	Середина интервала	Частота в интервале
1	[155, 159)	157	3
2	[159, 163)	161	4
3	[163, 167)	165	10
4	[167, 171)	169	18
5	[171, 175)	173	30
6	[175, 179)	177	19
7	[179, 183)	181	11
8	[183, 187)	185	5

На уровне значимости $\alpha = 0{,}05$ с помощью критерия хи-квадрат проверить гипотезу H_0 о том, что рост всех студентов университета распределен нормально.

ГЛАВА 7

НЕПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ О СДВИГЕ

В этой главе для количественных признаков изложены критерий ранговых сумм Уилкоксона и критерий Манна—Уитни для анализа однородности двух независимых случайных выборок наблюдений, а также критерий знаковых рангов Уилкоксона для анализа повторных парных наблюдений. Все эти критерии не требуют, чтобы законы распределения генеральной совокупности наблюдений принадлежали известному параметрическому семейству, например, семейству нормальных распределений с неизвестными параметрами. Поэтому эти критерии не зависят от конкретного вида закона распределения. В этом смысле они свободны от распределений и называются непараметрическими. Область применения непараметрических критериев более широкая, чем область применения параметрических критериев, требующих знания параметрического семейства законов распределения совокупности наблюдений.

Однако параметрические критерии более точны и более чувствительны по сравнению с непараметрическими при больших (более 100) объемах выборок.

§ 1. Критерий ранговых сумм Уилкоксона и критерий Манна—Уитни для двухвыборочных задач

На практике часто встречается задача сравнения двух независимых случайных выборок: одна — из контрольной генеральной совокупности, а другая — из рабочей (экспериментальной) совокупности. Необходимо проверить гипотезу о том, имеется ли различие в уровне исследуемого признака в контрольной и рабочей совокупностях. Например, такая задача возникает при сравнении индивидуальных психологических характеристик хронически больных детей и здоровых детей, людей разного возраста или разной культуры. Другой пример дает сравнение двух методов обучения или профессиональной подготовки, двух лекарств, двух рационов питания и т. д.

Если статистически значимого различия в уровне исследуемого признака в двух выборках нет, то говорят, что отсутствует сдвиг уровня признака. Это значит, что две выборки можно объединить и

рассматривать как единую выборку из одной однородной совокупности, т. е. обе выборки являются (статистически) однородными.

Задача о сдвиге уровня признака в случае нормального распределения контрольной и рабочей совокупностей, как уже известно из предыдущей главы, решается с помощью (параметрического) критерия Стьюдента. Непараметрическими критериями эта задача решается без знания законов распределения контрольной и рабочей совокупностей.

1. Критерий ранговых сумм Уилкоксона

Иначе этот критерий называется *критерием Уилкоксона для независимых выборок*.

Пусть заданы две независимые выборки наблюдений: $x_1, x_2, \ldots$ $\ldots, x_m$ и $y_1, y_2, \ldots, y_n$, в общем случае разных объемов m и n. Предположим, что для этих наблюдений справедлива следующая статистическая модель:

$$x_i = \varepsilon_i, \quad i = 1, 2, \ldots, m, \quad y_j = \Theta + \varepsilon_{m+j}, \quad j = 1, 2, \ldots, n,$$

где $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m+n}$ — взаимно независимые случайные величины, причем $\varepsilon_{m+1}, \ldots, \varepsilon_{m+n}$ — неизвестные случайные ошибки наблюдений, а функции распределения совокупностей, из которых взяты выборки, являются непрерывными. Здесь Θ — неизвестный сдвиг уровня в изучаемом признаке. Необходимо проверить гипотезу H_0: $\Theta = 0$, т. е. сдвиг уровня признака отсутствует, на заданном уровне значимости α.

Для проверки на уровне значимости α гипотезы H_0 следует:

1) проранжировать в порядке возрастания все $(m+n)$ наблюдений (при совпадении наблюдений берутся их средние ранги, см. с. 82 о ранжировании),

2) в общей совокупности рангов найти ранг R_i каждого наблюдения y_j, $j = 1, 2, \ldots, n$,

3) определить сумму рангов $W_{\text{набл}}$ всех наблюдений $y_1, y_2, \ldots, y_n$: $W_{\text{набл}} = R_1 + R_2 + \ldots + R_n$,

4) из таблицы по заданным числам α, m, n найти критическое значение $W_{\text{кр}}$ для статистики W — суммы рангов, относящихся к $y_1, y_2, \ldots, y_n$, критерия Уилкоксона.

Правосторонний критерий ранговых сумм Уилкоксона (т. е. рассматривается альтернатива H_1: $\Theta > 0$) на уровне значимости α принимает гипотезу H_0, если $W_{\text{набл}} < W_{\text{кр}}$, и отвергает гипотезу H_0, если $W_{\text{набл}} \geqslant W_{\text{кр}}$.

Левосторонний критерий ранговых сумм Уилкоксона (т. е. рассматривается альтернатива H_1: $\Theta < 0$) на уровне значимости α принимает гипотезу H_0, если $W_{\text{набл}} > n(m+n+1) - W_{\text{кр}}$, и отвергает гипотезу H_0, если $W_{\text{набл}} \leqslant n(m+n+1) - W_{\text{кр}}$.

Существует также двусторонний критерий ранговых сумм Уилкоксона, когда рассматривается альтернатива H_1: $\Theta \neq 0$.

Пример. Была проведена одна и та же контрольная работа для случайной выборки из шести учеников класса А и для случайной выборки пяти учеников параллельного класса Б. Известно, что ученики первой выборки выполнили соответственно 8, 6, 9, 5, 5, 4 заданий, а ученики второй выборки выполнили соответственно 7, 5, 7, 6, 6 заданий.

Необходимо определить на уровне значимости $\alpha = 0{,}05$, являются ли ученики класса Б более подготовленными, чем ученики класса А.

Решение. Пусть гипотеза H_0 означает, что ученики классов А и Б подготовлены одинаково, а гипотеза H_1 означает, что ученики класса Б подготовлены лучше. Для проверки этих гипотез воспользуемся правосторонним критерием ранговых сумм Уилкоксона. Составим объединенную выборку:

$$8,\ 6,\ 9,\ 5,\ 5,\ 4,\qquad 7,\ 5,\ 7,\ 6,\ 6.$$

Присвоим каждому элементу выборки порядковый номер при ее упорядочении по возрастанию (подчеркнуты номера, отвечающие совпадающим наблюдениям):

$$10,\ \underline{5},\ 11,\ \underline{2},\ \underline{3},\ 1,\ \underline{8},\ \underline{4},\ \underline{9},\ \underline{6},\ \underline{7}.$$

Присвоим наблюдениям ранги:

$$10;\ 6;\ 11;\ 3;\ 3;\ 1;\ 8{,}5;\ 3;\ 8{,}5;\ 6;\ 6.$$

Выпишем ранги числа выполненных заданий учеников класса Б:

$$8{,}5;\ 3;\ 8{,}5;\ 6;\ 6.$$

Сумма этих рангов дает $W_{\text{набл}} = 32$.

По заданным $\alpha = 0{,}05$, $m = 6$, $n = 5$ из таблицы X находим критическое значение $W_{\text{кр}} = 40$. Так как $W_{\text{набл}} < W_{\text{кр}}$, то на уровне значимости $\alpha = 0{,}05$ гипотеза H_0 принимается. ■

Если нет средних рангов, то стандартизованная статистика критерия ранговых сумм Уилкоксона

$$W^* = \frac{W - \mathsf{M}W}{\sqrt{\mathsf{D}W}}, \quad \mathsf{M}W = \frac{n(n+m+1)}{2}, \quad \mathsf{D}W = \frac{nm(n+m+1)}{12},$$

является асимптотически (т. е. при $m, n \to \infty$) нормально распределенной, т. е. $W^* \sim N(0, 1)$. Поэтому для больших m и n вычисляется $W^*_{\text{набл}}$ и сравнивается с критическим значением стандартного нормального распределения.

Если необходимо проверить гипотезу H_0: $\Theta = \Theta_0$, где Θ_0 — некоторое заданное ненулевое число, то процедура критерия Уилкоксона применяется к наблюдениям $y_1 - \Theta_0$, $y_2 - \Theta_0$, ..., $y_n - \Theta_0$.

Ранее уже говорилось, что одинаковым наблюдениям приписываются средние ранги. Чем меньше средних рангов, тем точнее выводы, полученные из критерия ранговых сумм Уилкоксона. При большом количестве средних рангов применение этого критерия является сомнительным.

Для неизвестного параметра Θ модели наблюдений можно указать точечную оценку $\widehat{\Theta}$ и построить доверительный интервал. Например, оценкой $\widehat{\Theta}$ параметра Θ служит медиана $m \cdot n$ всевозможных ранжированных в порядке возрастания разностей $(y_i - x_j)$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$.

2. Критерий Манна—Уитни

Для рассматриваемой статистической модели наблюдений можно также применять непараметрический критерий Манна—Уитни, являющийся некоторым обобщением критерия знаков. В качестве статистики U критерия Манна—Уитни выступает сумма всех положительных разностей $(y_i - x_j)$ для всех возможных пар (x_j, y_i), $j = 1, 2, \ldots, m$; $i = 1, 2, \ldots, n$.

Для проверки гипотезы H_0: $\theta = 0$ об отсутствии сдвига признака на заданном уровне значимости α необходимо:

1) проранжировать в порядке возрастания все $(m + n)$ наблюдения (при совпадении наблюдений берутся их средние ранги);

2) из общего проранжированного ряда наблюдений определить сумму рангов R_x для всех наблюдений первой выборки и сумму рангов R_y для всех наблюдений второй выборки;

3) найти два числа U_x и U_y по формулам:

$$U_x = mn + \frac{1}{2}n(n+1) - R_x, \quad U_y = mn + \frac{1}{2}m(m+1) - R_y \quad (U_x + U_y = mn);$$

4) в качестве наблюдаемого значения $U_{\text{набл}}$ статистики U критерия Манна—Уитни берется меньшее из чисел U_x и U_y;

5) по таблице критических значений статистики U критерия Манна—Уитни (см. табл. XV) находится $U_{\text{кр}}$.

Критерий Манна—Уитни. Если $U_{\text{набл}} < U_{\text{кр}}$, то на уровне значимости α принимается гипотеза H_0. Если же $U_{\text{набл}} > U_{\text{кр}}$, то гипотеза H_0 отвергается на уровне значимости α.

В случае совпадения некоторых наблюдений двух выборок выводы критерия Манна—Уитни тем точнее, чем меньше таких совпадений.

Применение критериев ранговых сумм Уилкоксона и критериев Манна—Уитни дает одинаковые результаты в силу того, что для всех m и n статистики W и U этих критериев связаны равенством

$$W = U + \frac{1}{2}n(n+1).$$

На практике применение критериев Уилкоксона и Манна—Уитни используют компьютерные программы в пакетах SPSS, STATISTICA и др.

§ 2. Критерий знаковых рангов Уилкоксона для повторных парных наблюдений

Часто необходимо обнаружить результат какого-либо воздействия (некоторой обработки). Для этого составляется случайная выборка испытуемых и для каждого из них дважды проводятся замеры интересующей нас характеристики до воздействия и после него. Полученные наблюдения называются повторными парными данными, а полученные две выборки наблюдений называются зависимыми. Такие данные появляются и тогда, когда на одной и той же группе испытуемых исследуются два разных воздействия.

Например, речь может идти об измеряемом сдвиге из-за влияния какого-то тренинга или о ситуационном сдвиге, когда сопоставляются уровни изучаемого признака в условиях «покоя» и «стресса». Может также идти речь об измерении одного и того же признака в группе испытуемых в два разных момента времени, в двух разных условиях, двумя разными способами и т. п.

В случае повторных парных наблюдений для проверки гипотезы о сдвиге распределения признака во всей генеральной совокупности из-за обработки, т. е. гипотезы об эффекте обработки, используется *критерий знаковых рангов Уилкоксона* или, по-другому, *критерий Уилкоксона для зависимых выборок*.

Итак, известны $2n$ наблюдений интересующего нас признака, по два наблюдения на каждого из n испытуемых случайной выборки: одно наблюдение «до обработки» выбранных испытуемых и второе наблюдение «после обработки». Обозначим наблюдения «до обработки» через $x_1, x_2, \ldots, x_n$, а наблюдения «после обработки» через $y_1, y_2, \ldots, y_n$. Рассмотрим разности $z_1 = y_1 - x_1$, $z_2 = y_2 - x_2$, $\ldots$ $\ldots$, $z_n = y_n - x_n$ и предположим, что для них справедлива следующая статистическая модель:

$$z_i = \Theta + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

где Θ — неизвестный интересующий нас параметр (эффект обработки), а $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ — неизвестные взаимно независимые случайные ошибки наблюдений, которые имеют непрерывные функции распределения, симметричные относительно нуля. Заметим, что эта модель не требует независимости двух выборок наблюдений.

По этим данным на заданном уровне значимости α необходимо проверить гипотезу H_0: $\Theta = 0$ для всей генеральной совокупности испытуемых.

Для применения одностороннего критерия знаковых рангов Уилкоксона при проверке гипотезы H_0 необходимо:

1) ранжировать от меньшего к большему модули всех наблюдений z_i: $|z_1|, |z_2|, \ldots, |z_n|$ (при наличии одинаковых модулей берутся средние ранги),

2) найти сумму рангов $S_{\text{набл}}$ только положительных наблюдений среди $z_1, z_2, \ldots, z_n$,

3) из табл. XI по заданным n и α найти критическое значение $S_{\text{кр}}$.

Правосторонний критерий знаковых рангов Уилкоксона (альтернатива H_1: $\Theta > 0$) на уровне значимости α принимает гипотезу H_0, если $S_{\text{набл}} < S_{\text{кр}}$, и отвергает H_0, если $S_{\text{набл}} \geqslant S_{\text{кр}}$.

Левосторонний критерий знаковых рангов Уилкоксона (альтернатива H_1: $\Theta < 0$) на уровне значимости α принимает гипотезу H_0, если $S_{\text{набл}} > \frac{1}{2}n(n+1) - S_{\text{кр}}$, и отвергает H_0, если $S_{\text{набл}} \leqslant \frac{1}{2}n(n+1) - S_{\text{кр}}$.

Пример. На уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 о том, что занятие физкультурой не влияет на общее самочувствие группы детей детского сада, против альтернативы H_1 о том, что физкультура положительно влияет на самочувствие детей. С этой целью для случайной выборки из восьми детей группы по специ-

альной шкале были получены следующие показатели самочувствия до и после занятия физкультурой:

3,2; 1,6; 5,7; 2,8; 5,5; 1,2; 6,1; 2,9

3,8; 1,0; 8,4; 3,6; 5,0; 3,5; 7,3; 4,8.

Решение. Найдем разность наблюдений «после» и «до».

0,6; −0,6; 2,7; 0,8; −0,5; 2,3; 1,2; 1,9.

Ранжируя модули этих чисел в порядке возрастания, получаем ряд рангов: 1; 2,5; 2,5; 4; 5; 6; 7; 8. Сумма рангов положительных разностей $S_{\text{набл}} = 2{,}5 + 8 + 4 + 7 + 5 + 6 = 32{,}5$.

По $\alpha = 0{,}05$ и $n = 8$ из табл. XI находим критическое значение $S_{\text{кр}} = 30$. Используем правосторонний критерий Уилкоксона. Так как $S_{\text{набл}} > S_{\text{кр}}$, то на уровне $\alpha = 0{,}05$ отвергается гипотеза H_0 и принимается гипотеза H_1. ■

Кроме односторонних критериев знаковых рангов Уилкоксона существует и двусторонний критерий Уилкоксона, когда рассматривается альтернатива H_1: $\Theta \neq 0$.

Если рассматривается гипотеза H_0: $\Theta = \Theta_0$, где Θ_0 — некоторое заданное ненулевое число, то критерии знаковых рангов Уилкоксона следует применять не для z_i, а для $\widetilde{z}_i = z_i - \Theta_0$, $i = 1, 2, \ldots, n$.

Можно построить для неизвестного параметра Θ рассматриваемой модели точечную оценку $\widehat{\Theta}$ и доверительный интервал.

Например, в качестве оценки $\widehat{\Theta}$ неизвестного «эффекта обработки» Θ берется медиана всех наблюдений $\frac{1}{2}(z_i + z_j)$, где $i \leqslant j$. Эта оценка является нечувствительной к грубым ошибкам (выбросам) в исходных данных.

Для повторных парных наблюдений, как известно, можно применять и критерий знаков. Однако критерий знаков в большинстве случаев менее чувствителен (менее мощный), чем критерий знаковых рангов Уилкоксона.

Для применения знаковых рангов Уилкоксона имеется программа в статистических пакетах SPSS, STATISTICA и др.

Задачи к главе 7

1. Были измерены уровни агрессивности в случайной выборке из пяти детей одной группы детсада и в случайной выборке из четырех детей второй группы того же детсада. Для первой случайной

выборки эти показатели (в баллах) 30, 18, 34, 20, 25, а для второй случайной выборки эти показатели (в баллах) 32, 24, 28, 16. На уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 о том, что не существует различия в уровне агрессивности детей двух групп детсада, против альтернативы H_1 о том, что уровень агрессивности детей второй группы выше, чем у детей первой группы.

2. Были измерены показатели интеллекта (в баллах) в случайной выборке из семи детей выпускного класса сельской школы и в случайной выборке из шести детей выпускного класса городской школы. Для первой случайной выборки эти показатели 100, 105, 120, 95, 102, 110, 115, а для второй случайной выборки эти показатели 97, 100, 125, 120, 104, 118. На уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 о том, что не существует различия в показателях интеллекта детей выпускных классов городской и сельской школ, против альтернативы H_1 о том, что уровень интеллекта детей городской школы выше, чем у детей сельской школы.

3. Были измерены уровни тревожности в случайной выборке из шести детей группы детсада до и после просмотра мультфильма о бандитах. Показатели тревожности до просмотра мультфильма (в баллах) были 21, 15, 18, 16, 19, 17, а после просмотра мультфильма стали 21, 16, 18, 17, 20, 19. На уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 о том, что не существует различия в уровне тревожности детей группы, против альтернативы H_1 о том, что уровень тревожности выше после просмотра мультфильма.

4. Были измерены показатели интеллекта (в баллах) в случайной выборке из семи учеников первого класса на момент поступления в школу и после двух месяцев учебы. Сначала показатели интеллекта были 30, 18, 23, 25, 27, 20, 21, а после двух месяцев учебы стали 32, 19, 24, 25, 28, 20, 22. На уровне значимости $\alpha = 0{,}1$ проверить гипотезу H_0 о том, что нет различия в показателях интеллекта детей первого класса, против альтернативы H_1 о том, что уровень интеллекта детей первого класса выше после двух месяцев учебы, чем на момент поступления в школу.

ГЛАВА 8

ОДНОФАКТОРНЫЙ АНАЛИЗ

Рассмотрим обобщение двухвыборочной задачи, т. е. задачи для двух независимых выборок, из предыдущей главы на случай $k \geqslant 3$ независимых случайных выборок, по одной из k генеральных совокупностей. Будем считать, что имеется некоторый фактор A, постоянно действующий на эти $k \geqslant 3$ совокупностей. Значения фактора A для каждой из $k \geqslant 3$ совокупностей называют уровнями фактора A или способами обработки. *Задача однофакторного анализа* состоит в том, чтобы исследовать и сравнить действие фактора A на каждую из к совокупностей. Это действие определяется математическим ожиданием (эффектом обработки) изучаемого количественного признака.

Приведем примеры применения однофакторного анализа. Пусть, например, проводится одна и та же контрольная работа для $k \geqslant 3$ случайных выборок учащихся. В качестве фактора A, влияющего на результат выполнения контрольной работы, может выступать: факт принадлежности учащихся $k \geqslant 3$ разным классам или школам, способ проведения контрольной работы (например, три его уровня — устно, письменно, с дополнительной информацией), возраст учащихся (например, три его уровня — старший, средний, младший), мотивация испытуемых (например, три ее уровня — низкая, средняя и высокая мотивации), условия работы (например, три уровня условий работы — в условиях, когда каждый учащийся сидит за отдельным столом, в условиях, когда за каждым столом сидят два учащихся, в условиях, когда столы, за которыми сидят учащиеся, отделены друг от друга большими промежутками) и т. д.

Однофакторный анализ для $k \geqslant 3$ независимых выборок проводится либо с помощью непараметрических критериев (критерий Краскела—Уоллиса, критерий Джонкхиера), либо с помощью параметрического F-критерия Фишера—Снедекора (однофакторный дисперсионный анализ).

§ 1. Непараметрические критерии Краскела—Уоллиса и Джонкхиера

Пусть заданы: случайная выборка наблюдений объема n_1 изучаемого признака для первой генеральной совокупности, случайная

выборка наблюдений объема n_2 для второй совокупности и т. д., наконец, задана случайная выборка наблюдений объема n_k для k-й совокупности и $k \geqslant 3$.

Наблюдения x_{ij} изучаемого признака принято располагать в так называемой однофакторной таблице следующего вида:

Номер наблюдения	Уровни фактора A			
	A_1	A_2	...	A_k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
...	...	...	...	...
...	$x_{n_1,1}$	$x_{n_2,2}$	...	$x_{n_k,k}$

В этой таблице всего $N = n_1 + n_2 + \ldots + n_k$ наблюдений x_{ij}, причем столбец таблицы, расположенный под A_1, задает выборку наблюдений изучаемого признака для первой совокупности, столбец, расположенной под A_2, задает выборку наблюдений изучаемого признака для второй совокупности, и т. д., наконец, столбец таблицы, расположенный под A_k, задает выборку наблюдений изучаемого признака для k-й совокупности.

Предположим, что наблюдения рассматриваемого признака подчиняются следующей статистической линейной (аддитивной) модели:

$$x_{ij} = a_j + \varepsilon_{ij}, \quad i = 1, 2, \ldots, n_k, \quad j = 1, 2, \ldots, k, \quad k \geqslant 3,$$

где a_j — неизвестный результат способа обработки A_j (эффект j-й обработки) фактора A, ε_{ij} — неизвестные случайные независимые ошибки наблюдений, имеющие одну и ту же непрерывную функцию распределения. Случайные ошибки ε_{ij} неизвестны, так как они появляются по различным неконтролируемым причинам. Например, при написании контрольной работы случайные ошибки могут появляться из-за различия испытуемых, их знаний, их настроения, характера, места испытаний, разного настроения преподавателей и т. д. Если линейная модель наблюдений неверна, т. е. она неточно описывает наблюдения, то появляются дополнительные ошибки.

Чтобы проверить, влияет ли способ обработки на результирующий признак, формулируется основная гипотеза H_0: $a_1 = a_2 = \ldots = a_k$ и альтернативная гипотеза H_1 о том, что не все числа $a_1, a_2, \ldots, a_k$ равны между собой. Гипотеза H_0 означает, что величина результата не зависит от способа обработки, т. е. все k выборок однородны и образуют объединенную выборку из общей совокупности, а гипотеза H_1 означает, что величина результата зависит от способа обработки.

Для проверки сформулированных гипотез используется непараметрический правосторонний критерий Краскела—Уоллиса.

Для его применения необходимо:

1) проранжировать по возрастанию все N наблюдений,

2) если r_{ij} — ранг наблюдения x_{ij} в совместной ранжировке и нет совпадающих наблюдений x_{ij}, то найти сумму рангов для каждой выборки в отдельности:

$$R_j = \sum_{i=1}^{n_j} r_{ij}, \quad j = 1, 2, \ldots, k,$$

3) для заданных наблюдений вычислить наблюдаемое значение $H_{\text{набл}}$ статистики $H = \dfrac{12}{N(N+1)} \cdot \sum\limits_{j=1}^{k} \dfrac{R_j^2}{n_j} - 3(N+1)$,

4) задать уровень значимости α,

5) по заданным числам α, k, n_1, n_2, …, n_k из табл. XII найти критическое значение $H_{\text{кр}}$ статистики H.

Тогда *правосторонний критерий Краскела—Уоллиса* утверждает, что на уровне значимости α гипотеза H_0 принимается при $H_{\text{набл}} < H_{\text{кр}}$ и гипотеза H_0 отклоняется при $H_{\text{набл}} \geqslant H_{\text{кр}}$.

Пример. Необходимо узнать на уровне значимости $\alpha = 0{,}05$, имеется ли различие в подготовленности учеников четвертых классов по математике из трех различных школ города, если для случайных выборок из пяти учеников четвертых классов каждой школы по числу выполненных из пятнадцати контрольных заданий были получены результаты, записанные в следующей однофакторной таблице:

№ ученика	№ школы		
	1	2	3
1	4	8	6
2	5	9	7
3	15	12	13
4	10	11	14
5	1	2	3

Решение. Поскольку результаты уже проранжированы, то имеем:

$$R_1 = 4 + 5 + 15 + 10 + 1 = 35,$$
$$R_2 = 8 + 9 + 12 + 11 + 2 = 42,$$
$$R_3 = 6 + 7 + 13 + 14 + 3 = 43.$$

Так как $k=3,\ n_1=n_2=n_3=5,\ N=15$, то

$$H_{\text{набл}}=\frac{12}{15\cdot 16}\left(\frac{35^2}{5}+\frac{42^2}{5}+\frac{43^2}{5}\right)-3\cdot 16=0{,}38.$$

Из табл. XII при $\alpha=0{,}05$, $n_1=n_2=n_3=5$ находим $H_{\text{кр}}=5{,}78$. Поскольку $H_{\text{набл}}<H_{\text{кр}}$, то на уровне $\alpha=0{,}05$ принимается гипотеза H_0 о том, что нет различия в подготовленности учеников четвертых классов трех школ. ■

Сделаем ряд замечаний по поводу критерия Краскела—Уоллиса. Если имеются одинаковые наблюдения x_{ij}, то для них используются средние ранги. В этом случае критерий Краскела—Уоллиса носит приближенный характер.

Заметим, что табл. XII критических значений критерия Краскела—Уоллиса дает критические значения для случая $n_1\geqslant n_2\geqslant n_3$. Если же наборы n_1, n_2, n_3 расположены в другом порядке, то нужно переставить выборки в порядке убывания их объемов и только тогда пользоваться таблицей критических значений. Это возможно в силу того, что перенумерация выборок не меняет гипотезу H_0.

Известно, что при $(n_1, n_2, \ldots, n_k)\to\infty$ распределение статистики H критерия Краскела—Уоллиса асимптотически приближается к χ^2-распределению с $(k-1)$ степенями свободы, если верна гипотеза H_0. В этом случае критическое значение $H_{\text{кр}}$ находят по заданным числам α и $(k-1)$ по таблице χ^2-распределения. Этим пользуются при больших выборках, для которых нет критических значений в таблице критических значений.

Заметим, что при $k=2$ критерий Краскела—Уоллиса сводится к критерию ранговых сумм Уилкоксона.

В том случае, когда на заданном уровне значимости α отклоняется гипотеза H_0 и принимается гипотеза H_1, непараметрический критерий Джонкхиера позволяет проверить гипотезу H_0 против альтернативы с упорядочением H_2 вида $a_1\leqslant a_2\leqslant\ldots\leqslant a_k$, где хотя бы одно из неравенств строгое. Гипотеза H_2 означает, что случайные выборки наблюдений упорядочены по возрастанию влияния фактора A (эффекта обработки).

Статистика J критерия Джонкхиера строится следующим образом.

1) Для каждой пары из k выборок, в которой номер первой выборки пары меньше номера второй выборки этой пары, составляется статистика Манна—Уитни.

2) Статистика J представляет собой сумму всевозможных таких статистик Манна—Уитни. Таким образом при $k=2$ статистика J совпадает со статистикой Манна—Уитни.

3) Наблюденное значение $J_{\text{набл}}$ статистики J сравнивается с критическим значением $J_{\text{кр}}$ статистики J, которое находится по таблице по заданным уровню значимости α и числам k, n_1, n_2, …, n_k.

Правосторонний критерий Джонкхиера утверждает, что на уровне значимости α принимается гипотеза H_0 об однородности выборок в случае $J_{\text{набл}} < J_{\text{кр}}$ и принимается гипотеза H_2 о возрастании эффектов обработки в случае $J_{\text{набл}} \geqslant J_{\text{кр}}$.

Для оценивания величин эффектов обработки необходимо записать рассматриваемую модель наблюдений x_{ij} по-другому, вводя вместо a_j — влияния j-й обработки на результат — влияние j-й обработки на отклонение x_{ij} от среднего уровня.

Если ввести средний уровень

$$\mu = \frac{1}{k}\sum_{i=1}^{k} a_i,$$

то $\tau_j = a_j - \mu$ является отклонением от среднего уровня при j-й обработке. Ясно, что $\tau_1 + \tau_2 + \ldots + \tau_k = 0$. Тогда предполагаемую статистическую модель наблюдений x_{ij} можно записать в виде

$$x_{ij} = \mu + \tau_j + \varepsilon_{ij}, \quad i = 1, 2, \ldots, n_j, \quad j = 1, 2, \ldots, k,$$

а вопрос о различии обработок сводится к вопросу о различии между числами τ_1, τ_2, …, τ_k. Гипотеза H_0 об однородности выборок означает равенство $\tau_1 = \tau_2 = \ldots = \tau_k = 0$, а альтернатива H_2 об упорядоченности по возрастанию эффектов обработки имеет вид $\tau_1 \leqslant \tau_2 \leqslant \ldots \leqslant \tau_k$. Различие между эффектами i-й и j-й обработок характеризуется сдвигом $a_i - a_j = \tau_i - \tau_j$. Можно дать оценку такого сдвига.

Пусть z_{ij} — медиана всевозможных разностей наблюдений i-й и j-й выборок. Заметим, что $z_{ij} = -z_{ji}$ и $z_{ii} = 0$. Тогда оценкой сдвига $(\tau_i - \tau_j)$ служит величина $\Delta_{ij} = \frac{1}{N}\sum_{p=1}^{k} n_p (z_{ip} - z_{jp})$, где $N = n_1 + n_2 + \ldots + n_k$. Эту величину называют *оценкой Спетволля*.

Иногда в задачах однофакторного анализа необходимо оценить не сами величины τ_i, а некоторые их линейные комбинации. С этой целью вводится понятие контрастов. *Контрастом* Θ пара-

метров $\tau_1, \tau_2, \ldots, \tau_k$ называется всякая их линейная комбинация с заданными коэффициентами:

$$\Theta = c_1\tau_1 + c_2\tau_2 + \ldots + c_k\tau_k,$$

где $c_1, c_2, \ldots, c_k$ — заданные числа, причем $c_1 + c_2 + \ldots + c_k = 0$.

Простейшими примерами контрастов являются сдвиги $(\tau_i - \tau_j)$. Поэтому умение оценивать контраст по заданной таблице однофакторного анализа позволяет оценивать, в частности, и сдвиги любых двух неизвестных параметров $\tau_1, \tau_2, \ldots, \tau_k$.

Оценкой контраста Θ служит величина

$$\widehat{\Theta} = \sum_{i=1}^{k} \sum_{j=1}^{k} d_{ij} \cdot \Delta_{ij},$$

где $d_{ij} = \frac{1}{k} \cdot c_i$, а Δ_{ij} — оценка Спетволля сдвига $(\tau_i - \tau_j)$.

Отметим, что в случае равных объемов выборок $n_1 = n_2 = \ldots = n_k$ существуют так называемые методы множественных сравнений, позволяющие отобрать те из обработок, которые отличаются между собой.

Компьютерные программы для критериев Краскела—Уоллиса и Джонкхиера имеются в статистических пакетах SPSS, STATISTICA и др.

В заключение заметим, что однофакторный анализ можно проводить и для $k \geqslant 3$ зависимых выборок. Это делается с помощью критерия Фридмана.

§ 2. Однофакторный дисперсионный анализ

Пусть, как и в предыдущем параграфе, независимые наблюдения x_{ij} рассматриваемого количественного признака расположены в однофакторной таблице и представляются статистической линейной (аддитивной) моделью:

$$x_{ij} = \mu + \tau_j + \varepsilon_{ij}, \quad i = 1, 2, \ldots, n_j, \quad j = 1, 2, \ldots, k,$$

где неизвестные параметры μ, τ_j модели имеют тот же смысл, что и в § 1. Однако неизвестные случайные ошибки ε_{ij} теперь должны удовлетворять более сильным требованиям. А именно, случайные ошибки ε_{ij} должны быть независимыми и нормально распределенными с общим нулевым средним и общей неизвестной дисперсией σ^2.

Дополнительная информация о распределении случайных ошибок ε_{ij} позволяет использовать более точные методы для проверки гипотез и для получения оценок неизвестных параметров однофакторной модели.

Рассматривается гипотеза H_0: $\tau_1 = \tau_2 = \tau_k = 0$ и ее альтернатива H_1 о том, что не все параметры $\tau_1, \tau_2, \ldots, \tau_k$ равны нулю.

Проверка гипотезы H_0 и альтернативы H_1 при заданном уровне значимости α и получение оценок неизвестных параметров μ, $\tau_1, \tau_2, \ldots, \tau_k, \sigma^2$ при сформулированных в этом параграфе условиях на модель наблюдений называется *однофакторным дисперсионным анализом*. Термин «дисперсионный анализ» происходит от того, что проверка гипотез H_0 и H_1 основана на сравнении выборочных дисперсий. Часто вместо термина «дисперсионный анализ» используется обозначение ANOVA (от английских слов «Analysis of variance»). Однофакторный ANOVA является обобщением t-критерия Стьюдента на случай $k \geqslant 3$ независимых выборок.

Для проведения однофакторного дисперсионного анализа (однофакторного ANOVA) сначала находят выборочное среднее для каждой выборки (каждого столбца таблицы однофакторного анализа):

$$\overline{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}, \quad j = 1, 2, \ldots, k,$$

и общее среднее всех наблюдений

$$\overline{x} = \frac{1}{N} \sum_{j=1}^{k} n_j \overline{x}_j, \quad N = n_1 + n_2 + \ldots + n_k.$$

Затем рассматривается полная сумма S_{Π}^* квадратов отклонений x_{ij} от общего среднего $\overline{x}$:

$$S_{\Pi}^* = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \overline{x})^2.$$

После несложных преобразований S_{Π}^* получаем, что

$$S_{\Pi}^* = \sum_{j=1}^{k} n_j (\overline{x}_j - \overline{x})^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2.$$

Обозначив здесь первое слагаемое через S_A^*, а второе слагаемое — через $S_{\text{ост}}^*$, имеем равенство

$$S_{\Pi}^* = S_A^* + S_{\text{ост}}^*.$$

Сумма S_A^* характеризует степень расхождения между выборками, и поэтому ее называют *рассеянием наблюдений за счет фактора A*.

Сумма $S_{\text{ост}}^*$ характеризует воздействие на наблюдения случайных ошибок наблюдений, и поэтому ее называют *остаточным рассеянием*.

Следовательно, полное рассеяние $S_{\text{П}}^*$ наблюдений представляет собой сумму рассеяния S_A^* по фактору A и остаточного рассеяния $S_{\text{ост}}^*$.

Назовем $(k-1)$ числом степеней свободы для S_A^*, а $(N-k)$ — числом степеней свободы для $S_{\text{ост}}^*$. Число степеней свободы $S_{\text{П}}^*$ равно $(k-1)+(N-k)=N-1$. Тогда выборочные полная, факторная и остаточная дисперсии соответственно равны:

$$S_{\text{П}}^2=\frac{S_{\text{П}}^*}{(N-1)},\quad S_A^2=\frac{S_A^*}{(k-1)},\quad S_{\text{ост}}^2=\frac{S_{\text{ост}}^*}{(N-k)}.$$

При выполнении гипотезы H_0 каждая из выборочных дисперсий S_A^2 и $S_{\text{ост}}^2$ является несмещенной точечной оценкой неизвестной дисперсии σ^2.

Известно, что статистика

$$F=\frac{S_A^2}{S_{\text{ост}}^2}$$

при выполнении гипотезы H_0 имеет F-распределение Фишера—Снедекора. На основании этого имеем следующую процедуру проверки гипотез H_0 и H_1.

---

1) Для заданных выборок находим наблюдаемое значение $F_{\text{набл}}$ статистики F.

2) Задаваясь уровнем значимости α, по числам α, $(k-1)$ и $(N-k)$ из табл. VIII находим критическое значение $F_{\text{кр}}$.

Тогда на заданном уровне значимости α гипотеза H_0 принимается в случае $F_{\text{набл}}<F_{\text{кр}}$ и гипотеза H_0 отклоняется в случае $F_{\text{набл}}\geqslant F_{\text{кр}}$.

Пример. Пусть измерения веса каждых четырех студентов, взятых из трех разных групп, дали следующие результаты (в кг):

1-я группа — 51, 52, 56, 57,
2-я группа — 52, 54, 56, 58,
3-я группа — 42, 44, 50, 52.

Предполагая, что вес студентов распределен нормально во всех трех группах с одинаковыми дисперсиями σ^2, на уровне значимости $\alpha=0{,}05$ проверить гипотезу H_0 о том, что средние значения веса

студентов во всех трех группах одинаковы против альтернативы H_1, что это не так.

Решение. По данным примерам находим:

$$\overline{x}_1 = 54, \quad \overline{x}_2 = 55, \quad \overline{x}_3 = 47,$$

$$\overline{x} = \frac{1}{12}(4 \cdot 54 + 4 \cdot 55 + 4 \cdot 47) = \frac{1}{3}(54 + 55 + 47) = \frac{1}{3} \cdot 156 = 52,$$

$$n_1 = n_2 = n_3 = 4, \quad k = 3, \quad N = 12, \quad S_A^2 = \frac{1}{2} \cdot 152 = 76,$$

$$S_{\text{ост}}^2 = \frac{1}{9} \cdot 114 = 12{,}67.$$

Тогда

$$F_{\text{набл}} = \frac{76}{12{,}67} \approx 6.$$

Из табл. VIII критических значений F-критерия для $\alpha = 0{,}05$, $v_1 = 2$, $v_2 = 9$ находим $F_{\text{кр}} = 4{,}26$.

Так как $F_{\text{набл}} > F_{\text{кр}}$, то гипотезу H_0 о равенстве средних значений веса студентов во всех трех группах отвергаем и принимаем альтернативу H_1, что это не так. ■

Для рассматриваемой модели наблюдений можно указать доверительный интервал для величины каждого эффекта обработки $a_j = \mu + \tau_j$, $j = 1, 2, \ldots, k$, с коэффициентом доверия $1 - 2\alpha$:

$$\overline{x}_j - \frac{\sqrt{S_{\text{ост}}^*}}{\sqrt{n_j}} t_{1-\alpha} < a_j < \overline{x}_j + \frac{\sqrt{S_{\text{ост}}^*}}{\sqrt{n_j}} \cdot t_{1-\alpha}, \quad j = 1, 2, \ldots, k,$$

где $t_{1-\alpha}$ — квантиль уровня $(1 - \alpha)$ распределения Стьюдента с $(N - k)$ степенями свободы (см. табл. VII). Можно указать также доверительный интервал для каждого из контрастов в отдельности и совместный доверительный интервал для всех интересующих нас контрастов одновременно. В частности, на заданном уровне значимости α можно указать значимо различные эффекты обработки (выборки) с помощью метода Шеффе множественных сравнений.

При применении однофакторного ANOVA и метода Шеффе множественных сравнений на практике используются программы статистических пакетов SPSS, STATISTICA и др.

§ 3. Понятие о двухфакторном дисперсионном анализе

Рассмотрим задачу о действии двух независимых факторов A и B на измеряемую величину интересующего нас признака, причем фактор A имеет m уровней $A_1, A_2, \ldots, A_m$, а фактор B имеет n уровней $B_1, B_2, \ldots, B_n$.

Например, пусть в одной и той же группе из m испытуемых производится замер одного и того же количественного признака, скажем эффективности работы, несколько раз — в разное время, в разных условиях, разными способами. Тогда фактором A выступает индивидуальность испытуемого и мы имеем m уровней фактора A. Если замеры результативности работы испытуемых производятся в различные моменты времени, то фактором B служит фактор времени с n уровнями, отвечающими n моментам времени. Если же результативность работы изучается в зависимости от стимулирования работы, то фактором B служит фактор стимулирования, например, с тремя уровнями (нет стимула, небольшой стимул, солидный стимул).

Нас интересует влияние на результат каждого из факторов A и B в отдельности и одновременного воздействия на результат двух факторов A и B.

В двухфакторных экспериментах наблюдения рассматриваемого признака располагаются обычно в виде так называемой двухфакторной таблицы, у которой m строк соответствуют m уровням фактора A, а n столбцов — n уровням фактора B. В (i, j)-ячейку, расположенную на пересечении i-й строки и j-го столбца, записываются наблюдения, полученные при одновременном исследовании i-го уровня A_i и j-го уровня B_j. Число наблюдений в разных ячейках может быть как одинаковым, так и разным. Однако предполагается всегда, что в каждой ячейке имеется, по крайней мере, одно наблюдение.

Самая простая двухфакторная таблица получается в случае, когда в каждой ячейке имеется ровно одно наблюдение. В этом случае имеются $m \cdot n$ наблюдений x_{ij}, которые составляют двухфакторную таблицу следующего вида:

Фактор A	Фактор B			
	B_1	B_2	$\ldots$	B_n
A_1	x_{11}	x_{12}	$\ldots$	x_{1n}
A_2	x_{21}	x_{22}	$\ldots$	x_{2n}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
A_m	x_{m1}	x_{m2}	$\ldots$	x_{mn}

Исследования двухфакторных таблиц проводятся при условии, что имеет место следующая статистическая линейная (аддитивная) модель наблюдений:

$$x_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}, \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n,$$

где μ — неизвестное общее среднее значение, τ_i и β_j — неизвестные отклонения от μ в результате действия факторов A и B, ε_{ij} — случайные (ненаблюдаемые) ошибки, являющиеся независимыми и нормально распределенными с общим нулевым средним и с общей неизвестной дисперсией σ^2. При этом

$$\sum_{i=1}^{m} \tau_i = \sum_{j=1}^{n} \beta_j = 0.$$

Для такого вида линейной модели действия факторов A и B возможны следующие гипотезы:

$H_0(A)$: все $\tau_i = 0$ и ее альтернатива $H_1(A)$: не все $\tau_i = 0$,
$H_0(B)$: все $\beta_j = 0$ и ее альтернатива $H_1(B)$: не все $\beta_j = 0$.

Заметим, что гипотеза $H_0(A)$ означает отсутствие постоянного действия фактора A, а гипотеза $H_0(B)$ означает отсутствие постоянного действия фактора B.

Двухфакторный дисперсионный анализ, т. е. двухфакторный ANOVA, на заданном уровне значимости α позволяет не только проверить каждую из сформулированных гипотез, но и построить доверительные интервалы для неизвестных параметров линейной модели.

Основной недостаток рассмотренной линейной модели действия факторов A и B в том, что она не учитывает взаимодействия постоянных факторов A и B. Факторы A и B в этой модели являются независимыми.

Если же предположить, что в каждой (i, j)-ячейке число наблюдений $K > 1$, то можно исследовать более полезную и широко распространенную статистическую модель наблюдений следующего вида:

$$x_{ij} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

где μ, τ_i, β_j имеют тот же смысл, что и выше, $\gamma_{ij} \neq \tau_i + \beta_j$ — смещение результата из-за взаимодействия i-го уровня A_i и j-го уровня B_j, ε_{ijk} — случайные независимые ошибки, распределенные по нормальному закону с общим нулевым средним значением и неизвестной общей дисперсией σ^2, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$; $k = 1, 2, \ldots, K$.

При таких условиях двухфакторный дисперсионный анализ (в отличие от непараметрического анализа) может проверить не только гипотезы $H_0(A)$, $H_0(B)$ и их альтернативы, но и гипотезу об отсутствии взаимодействия факторов A и B — гипотезу $H_0(AB)$: все $\gamma_{ij} = 0$ и ее альтернативу $H_1(AB)$: не все $\gamma_{ij} = 0$. Заметим, что выполнение гипотезы $H_0(AB)$ на уровне значимости α означает справедливость линейной гипотезы действия факторов A и B.

Отметим, что в общем случае двухфакторного анализа с одним наблюдением в ячейке ($K = 1$) невозможно рассмотреть модель с взаимодействием γ_{ij} факторов A и B общего вида. Необходимо накладывать дополнительные условия на структуру γ_{ij}. В этом случае чаще всего предполагают, что $\gamma_{ij} = c \cdot \tau_i \cdot \beta_j$, где c — некоторое число.

Дисперсионный анализ можно применять и при малых отклонениях от требования нормальности закона распределения случайных ошибок. Однако требования об одинаковости дисперсии σ^2 для всех случайных ошибок и об отсутствии выбросов в наблюдениях являются существенными.

Многофакторный ANOVA с числом факторов больше 2 принципиально не отличается от двухфакторного варианта. Специфика его только в наличии проблемы взаимодействия более чем двух факторов. Эта проблема заключается в том, что количество взаимодействий с увеличением числа факторов растет в геометрической прогрессии. Так, например, если двухфакторный ANOVA должен проверить три гипотезы, то трехфакторный ANOVA — уже семь гипотез. Кроме того, содержательная интерпретация многофакторного взаимодействия при числе факторов больше двух или сильно затруднена, или вообще невозможна.

Кроме рассмотренных вариантов ANOVA также часто используется ANOVA с повторными измерениями (непараметрический его аналог — критерий Фридмана для зависимых выборок) и многомерный ANOVA. В первом случае разным градациям фактора соответствует одна и та же группа объектов (зависимые выборки). Многомерный ANOVA применяется для изучения действия факторов не на один признак, а на несколько признаков сразу. Его сокращенно называют MANOVA (Multivariate ANOVA). MANOVA проверяет не только гипотезы о влиянии факторов на каждый признак в отдельности, но и гипотезу о влиянии факторов на всю совокупность признаков.

При проведении ANOVA и MANOVA на практике современные статистические пакеты программ (SPSS, STATISTICA и др.) избавляют пользователей от громоздких расчетов.

Задачи к главе 8

1. Необходимо проверить на уровне значимости $\alpha = 0{,}1$, имеется ли различие в степени уверенности в себе для детей трех различных групп детсада, если для случайных выборок из четырех детей каждой группы результаты (в баллах) проверки степени уверенности в себе записаны в следующей однофакторной таблице:

№ ребенка	№ группы		
	1	2	3
1	23	25	21
2	35	28	25
3	40	38	30
4	43	30	29

2. Необходимо проверить на уровне значимости $\alpha = 0{,}05$, имеется ли различие в степени общительности для детей трех различных групп детсада, если для случайных выборок из четырех детей каждой группы результаты (в баллах) проверки степени общительности записаны в следующей однофакторной таблице:

№ ребенка	№ группы		
	1	2	3
1	45	25	22
2	40	28	27
3	43	30	28
4	35	32	21

3. Пусть измерения общего интеллекта в трех случайных выборках, в каждой из которых по четыре ученика, дали следующие результаты (в баллах):

1-я выборка — 52, 60, 58, 48,
2-я выборка — 45, 50, 46, 49,
3-я выборка — 34, 32, 36, 38.

Предполагая, что три выборки взяты из трех параллельных классов школы и что общий интеллект учеников всех трех классов распределен нормально с одинаковой дисперсией σ^2, на уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 о том, что средний общий интеллект учеников всех трех классов одинаков, против альтернативы H_1, что это не так.

ГЛАВА 9

СТАТИСТИЧЕСКИЙ АНАЛИЗ КОРРЕЛЯЦИОННОЙ ЗАВИСИМОСТИ

Корреляционный анализ является одним из наиболее распространенных методов статистического исследования взаимозависимости двух или более признаков. Его назначение в том, чтобы на основании полученной случайной выборки наблюдений анализируемой многомерной случайной величины (многомерного признака) ответить на следующие основные вопросы:

Как с учетом природы анализируемого многомерного признака выбрать подходящую меру связи признаков?

Как по выборочным данным получить точечную и интервальную оценки величины связи признаков?

Как проверить гипотезу о том, что полученные оценки силы связи действительно подтверждают статистическую связь признаков?

Как проверить наличие связи для каждой пары компонент многомерного признака?

Ответы на эти вопросы будут даны в этой главе в зависимости от того, в какой шкале измерены значения исследуемых признаков. Если установлен факт зависимости анализируемых признаков, то затем выявляется вид и математическая форма этой зависимости. Это делается с помощью методов и моделей регрессионного анализа, о чем будет идти речь в следующей главе.

§ 1. Мера силы корреляционной связи двух количественных признаков

Пусть X и Y — количественные признаки, т. е. их значения измеряются в интервальной шкале или в шкале отношений, и пусть X — объясняющий признак (он может быть случайным или неслучайным), а Y — результирующий случайный признак. Нас интересует степень тесноты статистической связи между результатом Y и объясняющим признаком X. Будем предполагать, что имеет место следующая, наиболее часто встречающаяся модель зависимости Y от X:

$$Y = f(X) + \varepsilon(X),$$

где неизвестная функция $f(x)$ равна условному математическому

ожиданию $\mathsf{M}[Y \mid X = x]$ для всех значений x случайной величины X и называется *функцией регрессии* Y по X, а $\varepsilon(X)$ — неизвестный случайный остаток с математическим ожиданием $\mathsf{M}[\varepsilon \mid X = x] = 0$, с неизвестной дисперсией $\mathsf{D}[\varepsilon \mid X = x] = \sigma^2(x)$ для всех значений $X = x$, причем $\varepsilon(X)$ не коррелирует с $f(X)$, если X — случайный признак.

Будем считать также, что зависимость Y от X является корреляционной, т. е. $f(x) \neq \text{const}$ для всех значений $X = x$.

Для получения оценки степени корреляционной зависимости Y от X имеется n пар наблюдений (x_1, y_1), (x_2, y_2), …, (x_n, y_n) признаков Y и X. Рассмотрим три случая: 1) пара (X, Y) образует двумерную нормально распределенную случайную величину; 2) признаки X и Y связаны линейной регрессионной зависимостью, т. е.

$$Y(X) = \Theta_0 + \Theta_1 X + \varepsilon(X),$$

где Θ_0 и Θ_1 — некоторые неизвестные параметры модели, X и $\varepsilon(X)$ некоррелированы, и законы распределения X, Y, $\varepsilon(X)$ не обязаны быть нормальными; 3) регрессия Y по X является нелинейной, т. е. функция $f(x)$ — нелинейная.

1. Случай двумерного нормального распределения

Пусть (X, Y) — двумерная случайная величина, распределенная по нормальному закону с плотностью

$$\varphi(x_1, x_2) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \cdot e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{a})^T \cdot \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{a})},$$

где $\boldsymbol{x}$ — вектор с компонентами x_1, x_2, $\boldsymbol{a}$ — вектор математических ожиданий $a_1 = \mathsf{M}X$, $a_2 = \mathsf{M}Y$, $\boldsymbol{\Sigma}$ — ковариационная матрица:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix},$$

ковариация $\sigma_{ij} = \mathsf{M}[(X - a_i)(Y - a_j)]$, $i,\ j = 1,\ 2$, $|\boldsymbol{\Sigma}|$ — определитель матрицы $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{-1}$ — обратная матрица к матрице $\boldsymbol{\Sigma}$.

Можно показать, что в этом случае для всех значений $X = x$

$$f(x) = \mathsf{M}[Y \mid X = x] = \Theta_0 + \Theta_1 x,$$

где $\Theta_1 = r \cdot \sqrt{\sigma_{22}/\sigma_{11}}$, $\Theta_0 = a_2 - \Theta_1 a_1$, r — коэффициент корреляции между признаками X и Y, и что условная дисперсия Y по X для всех значений $X = x$ имеет вид $\mathsf{D}[Y \mid X = x] = \sigma_{22}^2 \cdot (1 - r^2)$.

Таким образом, если анализируемые признаки X и Y подчинены двумерному нормальному закону распределения, то функция

регрессии одной из них по другой всегда является линейной функцией. Кроме того, в этом случае независимость признаков X и Y равносильна их некоррелируемости ($r = 0$).

Для случайной выборки n пар наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, можно найти выборочный коэффициент корреляции Пирсона

$$\widehat{r} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2 \cdot \sum\limits_{i=1}^{n} (y_i - \overline{y})^2}},$$

где $\overline{x}$ и $\overline{y}$ — выборочные средние для наблюдений признаков X и Y соответственно.

Зная выборочный коэффициент корреляции $\widehat{r}$, можно проверить гипотезу о независимости признаков X и Y , т. е. гипотезу H_0: $r = 0$.

В этом случае ($r = 0$) говорят о статистической незначимости коэффициента r. Здесь гипотеза H_1: $r \neq 0$, а r— коэффициент корреляции Пирсона для X и Y.

Для проверки гипотезы H_0 используется тот факт, что статистика

$$t = \frac{\widehat{r}\sqrt{n-2}}{\sqrt{1-\widehat{r}^2}}$$

распределена при H_0 по закону Стьюдента с $(n-2)$ степенями свободы. Для проверки гипотезы H_0 на заданном уровне значимости α необходимо вычислить наблюденное значение $t_{\text{набл}}$ статистики t и сравнить его с найденным по числам $\alpha/2$ и $(n-2)$ из табл. VII критическим значением $t_{\text{кр}}$. Гипотеза H_0 принимается, если $|t_{\text{набл}}| < t_{\text{кр}}$, и H_0 отвергается, если $|t_{\text{набл}}| \geqslant t_{\text{кр}}$, на уровне значимости α.

Если же $r \neq 0$, то при малых n ($n < 100$) для него можно построить доверительный интервал, используя нормальность следующего преобразования Р. Фишера:

$$z = \frac{1}{2} \ln \frac{1+\widehat{r}}{1-\widehat{r}}.$$

Нахождение z по данному значению $\widehat{r}$ и, наоборот, нахождение $\widehat{r}$ по заданной величине z производятся с помощью специальной табл. XIII.

Доверительный интервал, покрывающий неизвестный коэффициент корреляции r с доверительной вероятностью $(1-\alpha)$, имеет вид:

$$r_1 = \operatorname{th} z_1 < r < \operatorname{th} z_2 = r_2,$$

$$z_{1,2} = \frac{1}{2} \ln \frac{1+\widehat{r}}{1-\widehat{r}} \mp \frac{1+U_{\alpha/2}}{\sqrt{n-3}} - \frac{\widehat{r}}{2(n-1)},$$

где квантиль $U_{\alpha/2}$ находится по табл. IV, а th z (тангенс гиперболический z) является обратной функцией к функции

$$z = \frac{1}{2} \ln \frac{1+\widehat{r}}{1-\widehat{r}}.$$

Пример 1. Для случайной выборки из 28 учащихся выпускных классов школы получен выборочный коэффициент корреляции $\widehat{r} = 0{,}7$, характеризующий тесноту связи между общим интеллектом и математическими способностями учащихся выборки.

а) Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о независимости общего интеллекта и математических способностей для всех учащихся выпускных классов школы.

б) С доверительной вероятностью 0,95 построить доверительный интервал для коэффициента корреляции r.

Предполагается, что общий интеллект и математические способности выпускников школы образуют двумерную нормальную генеральную совокупность.

Решение. а) Находим

$$t_{\text{набл}} = \frac{\widehat{r}\sqrt{n-2}}{\sqrt{1-\widehat{r}^2}} = \frac{0{,}7 \cdot \sqrt{26}}{\sqrt{1-0{,}49}} = \frac{0{,}7 \cdot \sqrt{26}}{\sqrt{0{,}51}} \approx \frac{0{,}7 \cdot 5{,}1}{0{,}7} = 5{,}1.$$

Из табл. VII по заданным $\alpha/2$ и $n-2=26$ находим $t_{\text{кр}} = 2{,}056$. Поскольку $t_{\text{набл}} > t_{\text{кр}}$, то на уровне значимости $\alpha = 0{,}05$ гипотеза H_0 отклоняется.

б) По таблице z-преобразований Р. Фишера для $\widehat{r} = 0{,}7$ находим $z = 0{,}8673$. Тогда

$$z_1 = 0{,}8673 - \frac{1{,}96}{\sqrt{25}} = 0{,}8673 - 0{,}392 = 0{,}4753,$$

$$z_2 = 0{,}8673 + \frac{1{,}96}{\sqrt{25}} = 0{,}8673 + 0{,}392 = 1{,}2593.$$

По найденным z_1 и z_2 из той же таблицы получаем $r_1 = 0{,}442$, $r_2 = 0{,}851$. Итак, с доверительной вероятностью 0,95 коэффициент корреляции

$$0{,}442 < r < 0{,}851. \blacksquare$$

2. Случай линейной регрессионной зависимости

Пусть теперь признаки X и Y связаны линейной регрессионной зависимостью

$$Y = \Theta_0 + \Theta_1 X + \varepsilon(X),$$

причем, в отличие от предыдущего случая, здесь не требуется дву-

мерной нормальности анализируемой пары признаков X и Y и не требуется, чтобы признак X был случайным. В рассматриваемом случае из некоррелируемости X и Y (т. е. при $r=0$) не следует в общем случае независимость признаков X и Y. Поэтому необходима осторожность при использовании r в качестве характеристики степени тесноты связи X и Y. Коэффициенты корреляции r и $\widehat{r}$ сохраняют прежний смысл, но проверить гипотезу H_0: $r=0$ и построить доверительный интервал для r уже нельзя.

3. Случай нелинейной регрессионной зависимости

Если регрессионная зависимость Y от X нелинейная, т. е. нелинейной является функция регрессии $f(x)$ для всех значений x признака X, то, как известно, коэффициент корреляции r уже не дает характеристику степени тесноты связи X и Y. Такой характеристикой в этом случае является корреляционное отношение $\rho(Y/X)$. Статистической оценкой неизвестного корреляционного отношения $\rho(Y/X)$ является выборочное корреляционное отношение $\widehat{\rho}(Y/X)$, которое по заданным выборочным двумерным наблюдениям (x_1, y_1), (x_2, y_2), …, (x_n, y_n) находится по формуле

$$\widehat{\rho}(Y/X)=S_{\overline{y}(x)}/S_y, \quad S^2_{\overline{y}(x)}=\frac{1}{n}\sum_{j=1}^{s}(y_j-\overline{y})^2, \quad S^2_y=\frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{n_j}(y_{ji}-\overline{y}_j)^2,$$

где s — число интервалов группирования наблюдений величины X, n_j — число значений X, попавших в j-й интервал группирования, $\overline{y}_j$ — среднее значение ординат точек, попавших в j-й интервал группирования, $\overline{y}$ — общее среднее число наблюдений величины Y, т. е.

$$\overline{y}=\frac{1}{n}\sum_{j=1}^{s}n_j\overline{y}_j.$$

Выборочное корреляционное отношение $\widehat{\rho}(Y/X)$ обладает теми же свойствами, что и теоретическое корреляционное отношение $\rho(Y/X)$. В частности, $0\leqslant\widehat{\rho}(Y/X)\leqslant 1$ и $\widehat{\rho}(Y/X)\geqslant|\widehat{r}(Y, X)|$, где $\widehat{r}(Y, X)$ — выборочный коэффициент корреляции Пирсона. Число $\widehat{\rho}^2(Y/X)$ называется выборочным коэффициентом детерминации Y по X. Этот коэффициент показывает для заданной выборки ту долю общей вариации результирующего признака Y, которая объясняется изменением функции регрессии $f(x)$. Пусть условные распределения результирующего признака (Y/X) при каждом значении x

признака X описываются нормальным законом о постоянной (неизвестной) дисперсией σ^2.

Тогда можно на заданном уровне значимости α проверить гипотезу H_0: $\rho(Y/X)=0$ об отсутствии корреляционной связи между признаками Y и X.

Справедливость гипотезы H_0 означает, что значение найденного выборочного корреляционного отношения $\widehat{\rho}(Y/X)$ является статистически незначимым, т. е. гипотеза H_0 не противоречит выборке наблюдений. Здесь альтернатива H_1 такая: $\rho(Y/X)\neq 0$.

Для проверки гипотезы H_0 следует найти наблюденное значение $\gamma_{\text{набл}}$ статистики

$$\gamma=\frac{\widehat{\rho}^2(Y/X)}{1-\widehat{\rho}^2(Y/X)}\cdot\frac{n-s}{s-1},$$

где s — число интервалов группирования (или число различных) значений признака X. При гипотезе H_0 статистика γ имеет F-распределение с числом степеней свободы числителя $(s-1)$ и знаменателя $(n-s)$. При заданных α, n, s из таблицы F-распределения (см. табл. VIII) находится $\gamma_{\text{кр}}$. Если $\gamma_{\text{набл}}<\gamma_{\text{кр}}$, то на уровне α принимается гипотеза H_0.

Пример 2. Проверить на уровне значимости $\alpha=0{,}05$ гипотезу H_0 об отсутствии корреляционной связи между признаком Y и признаком X, если для $n=15$ двумерных наблюдений Y и X, расположенных в $s=2$ равных интервалах группирования, было найдено $\widehat{\rho}^2(Y/X)=0{,}15$.

Решение. Для данных задачи $\gamma_{\text{набл}}\approx 2{,}29$. Из таблицы F-распределения (см. табл. VIII) находим $\gamma_{\text{кр}}=4{,}67$. Так как $\gamma_{\text{набл}}<\gamma_{\text{кр}}$, то гипотеза H_0 на уровне $\alpha=0{,}05$ принимается. ■

При сформулированных выше условиях по найденному выборочному коэффициенту $\widehat{\rho}(Y/X)$ и по заданному уровню значимости α можно построить доверительный интервал для неизвестного значения корреляционного отношения $\rho(Y/X)$. Он получается существенно несимметричным относительно точечной оценки $\widehat{\rho}(Y/X)$.

Выборочные коэффициенты $\widehat{r}$ и $\widehat{\rho}(Y/X)$ часто выступают в качестве коэффициентов надежности психодиагностических методик. Методика признается надежной, если коэффициент надежности больше 0,8. Следует отметить, что надежность методики понимается в трех различных смыслах: 1) как ее однородность, т. е. насколько взаимосогласованы между собой задания методики; 2) как стабильность (устойчивость) признака, измеряемого этой методикой; 3) как

константность методики, т. е. независимость результатов опытов по данной методике от личности экспериментаторов.

Для проверки однородности методики используется метод расщепления. Согласно этому методу задания делятся на четные и нечетные, отдельно обрабатываются, затем для двух выборок результатов выполнения четных и нечетных заданий находят выборочный коэффициент корреляции, который и называется *коэффициентом надежности*.

Для проверки стабильности измеряемого признака используется метод «тест-ретест». Он заключается в том, что с помощью той же методики через некоторое время проводят обследование той же выборки испытуемых и находят выборочный коэффициент корреляции между результатами первого и повторного обследований, который и называется *коэффициентом стабильности методики*. Наконец, выборочный коэффициент корреляции результатов двух опытов с данной методикой в относительно одинаковых условиях для одной и той же выборки испытуемых, но с разными экспериментаторами называют *коэффициентом константности*.

Психодиагностическая методика считается надежной, если все три коэффициента: коэффициент надежности, коэффициент стабильности и коэффициент константности — не ниже 0,8.

§ 2. Мера силы множественных корреляционных связей

Часто необходимо по заданной случайной выборке проанализировать корреляционную зависимость результирующего количественного признака от двух или более других объясняющих количественных признаков. Например, необходимо определить корреляционную связь общего интеллекта учащихся с математическими и гуманитарными способностями учащихся или корреляционную зависимость интеллекта учащихся от личностных психологических характеристик. Такая корреляционная связь называется *множественной*. Ясно, что анализ корреляционной связи каждой пары из трех признаков не позволяет установить множественную корреляционную связь.

В дальнейшем для простоты изложения ограничимся случаем корреляционной зависимости результирующего признака X_1 от двух объясняющих признаков X_2 и X_3. При этом будем считать, что трехмерная случайная величина (X_1, X_2, X_3) описывается трехмерным нормальным законом распределения. Тогда можно показать,

что функция регрессии X_1 по X_2 и X_3 является линейной, т. е.

$$F(x_2, x_3) = \mathsf{M}[X_1 \mid X_2 = x_2, X_3 = x_3] = \Theta_0 + \Theta_1 x_2 + \Theta_2 x_3,$$

где коэффициенты Θ_0, Θ_1 и Θ_2 выражаются через параметры нормального распределения.

В рассматриваемом случае теперь коэффициенты корреляции Пирсона двух признаков r_{ij}, i, $j = 1, 2, 3$, *называют коэффициентами парной корреляции*. Они учитывают тот факт, что связь каждой пары признаков находится под воздействием связей всех других признаков между собой и с признаками из данной пары. Для трех признаков X_1, X_2, X_3 можно определить корреляционную матрицу R, которая является матрицей третьего порядка и состоит из всех коэффициентов парной корреляции для этих признаков, причем на ее главной диагонали стоят единицы.

Кроме коэффициентов парной корреляции вводятся еще так называемые *частные коэффициенты корреляции*. Они устанавливают силу чистой линейной связи каждой пары признаков при условии, что связи всех других признаков с признаками данной пары не действуют, нивелированы, зафиксированы, т. е. очищены от влияния третьего признака на данную пару признаков. Обозначим частные коэффициенты корреляции через: $r_{12.3}$ для признаков X_1 и X_2, $r_{13.2}$ для признаков X_1 и X_3 и $r_{23.1}$ для признаков X_2 и X_3. Они выражаются через элементы корреляционной матрицы R по формулам:

$$r_{12.3} = \frac{-R_{12}}{\sqrt{R_{11} \cdot R_{22}}}, \quad r_{13.2} = \frac{-R_{13}}{\sqrt{R_{11} \cdot R_{33}}}, \quad r_{23.1} = \frac{-R_{23}}{\sqrt{R_{22} \cdot R_{33}}},$$

где R_{ij} — алгебраическое дополнение к соответствующему элементу $r_{i,j}$, i, $j = 1, 2, 3$, матрицы R. Отметим, что при условии выполнения нормального закона для X_1, X_2, X_3 частные коэффициенты корреляции не зависят от зафиксированных уровней мешающей переменной.

Кроме введенных парных и частных коэффициентов корреляции, вводится также *множественный коэффициент корреляции* $R_{1.23}$ по формуле:

$$R_{1.23} = \sqrt{1 - \frac{|R|}{R_{11}}},$$

где $|R|$ — определитель матрицы R, R_{11} — алгебраическое дополнение элемента r_{11} в матрице R.

Множественный коэффициент корреляции $R_{1.23}$ характеризует силу связи результата X_1 с набором признаков X_2 и X_3. Число $R^2_{1.23}$ называют множественным коэффициентом детерминации. Он характеризует ту часть вариации признака X_1, которая объясняется

вариацией признаков X_2 и X_3. Для множественного коэффициента корреляции $R_{1.23}$ справедливы следующие свойства:

1) $0 \leqslant R_{1.23} \leqslant 1$,

2) $R_{1.23} = 0$ соответствует полному отсутствию корреляционной зависимости признака X_1 от признаков X_2 и X_3,

3) $R_{1.23} = 1$ соответствует точной линейной связи признака X_1 с признаками X_2 и X_3: $X_1 = \Theta_0 + \Theta_1 X_2 + \Theta_2 X_3$,

4) $R_{1.23}$ не меньше любого парного или любого частного коэффициента корреляции.

Приведенные формулы для частных коэффициентов корреляции и для множественных коэффициентов корреляции переносятся на их выборочные аналоги, если матрицу R в этих формулах заменить на матрицу выборочных коэффициентов корреляции $\widehat{r}_{ij}$, $i, j = 1, 2, 3$.

То же относится и к выборочному коэффициенту $\widehat{R}_{1.23}$. При проверке статистически значимого отличия от нуля частного коэффициента корреляции и при построении для него доверительного интервала следует действовать так же, как и в случае парного коэффициента корреляции, но с заменой n на $(n-1)$.

Для выяснения вопроса о том, можно ли считать выборочный множественный коэффициент детерминации $\widehat{R}^2_{1.23}$ статистически значимо отличающимся от нуля, т. е. для проверки гипотезы H_0: $\widehat{R}^2_{1.23} = 0$ на заданном уровне значимости α, используется статистика

$$\gamma = \frac{n-3}{2} \cdot \frac{\widehat{R}^2_{1.23}}{1 - \widehat{R}^2_{1.23}}.$$

Наблюденное значение $\gamma_{\text{набл}}$ этой статистики сравнивается с критическим значением $\gamma_{\text{кр}}$, найденным по таблице F-распределения с числом степеней свободы числителя 2 и знаменателя $(n-3)$ (см. табл. VIII). Здесь альтернатива H_1 такая: $\widehat{R}^2_{1.23} \neq 0$.

Если $\gamma_{\text{набл}} \geqslant \gamma_{\text{кр}}$, то гипотеза H_0 об отсутствии множественной корреляционной связи X_1 с признаками X_2 и X_3 отвергается на уровне значимости α.

Пример. Известно, что выборочный множественный коэффициент детерминации $R^2_{1.23} = 0{,}373$ характеризует корреляционную связь X_1 — числа детей в семье с уровнем дохода семьи X_2 и возрастом вступления в брак отца X_3 для выборки $n = 27$ семей. Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 об отсутствии корреляционной связи X_1 с набором X_2 и X_3: $R_{1.23} = 0$ для всего множества семей, из которого взята выборка.

Решение. Имеем $\gamma_{\text{набл}} = 7{,}14$. Из таблицы F-распределения с числом степеней числителя 2 и знаменателя 24 находим $\gamma_{\text{кр}} = 3{,}4$. Так как $\gamma_{\text{набл}} > \gamma_{\text{кр}}$, то на уровне значимости $\alpha = 0{,}05$ гипотеза H_0 отклоняется. ■

Теперь рассмотрим общую схему, когда случайная величина (X_1, X_2, X_3) не подчинена нормальному закону, однако признаки X_1, X_2, X_3 связаны статистической линейной моделью вида

$$X_1 = \Theta_0 + \Theta_1 X_2 + \Theta_2 X_3 + \varepsilon(X_2, X_3).$$

Здесь Θ_0, Θ_1, Θ_2 — неизвестные параметры модели, а $\varepsilon(X_2, X_3)$ — неизвестный случайный остаток, для которого математическое ожидание $\mathsf{M}\varepsilon(X_2, X_3) = 0$, а ковариационная матрица $V = \sigma^2 E$ (здесь E— единичная матрица второго порядка). Признаки X_2, X_3 могут быть случайными или неслучайными. Если X_2, X_3 — случайные признаки, то предполагается, кроме того, что случайный остаток $\varepsilon(X_2, X_3)$ не коррелирует с X_2 и X_3.

В этом случае определения и свойства частных и множественных коэффициентов корреляции сохраняются на практике, однако эти коэффициенты дают лишь приближенный результат. При даже незначительном отклонении статистической модели от выше указанной линейной модели не рекомендуется использовать введенные коэффициенты.

В заключение отметим, что для вычисления выборочной корреляционной матрицы $\widehat{R}$ и выборочных частных и множественных коэффициентов корреляции можно использовать программы пакетов SPSS, STATISTICA и др.

§ 3. Коэффициенты ранговой корреляции

Для изучения силы связи признаков, значения которых измерены в порядковой шкале, применяются коэффициенты ранговой корреляции. Они являются непараметрическими, так как для их применения не требуется никакой информации о распределении признака в генеральной совокупности.

Пусть рассматривается случайная выборка n испытуемых и испытуемые упорядочены (ранжированы) по степени проявления в них того или другого признака (свойства или качества). Обозначим через $x_1, x_2, \ldots, x_n$ результат ранжировки n испытуемых по признаку X, а через $y_1, y_2, \ldots, y_n$ — результат ранжировки n испытуемых по признаку Y. Ранжировка по Y может также рассматриваться и как ранжировка опять по признаку X, но, например, другим экспертом.

Напомним, что процесс ранжирования испытуемых состоит в присвоении каждому из них ранга — порядкового номера в зависимости от того, в какой степени он обладает признаком X. Если для двух или более испытуемых степень проявления признака X одинаковая, то для каждого из таких испытуемых ранг равен среднему арифметическому тех номеров мест, которые они занимают. Такие ранги называют связными.

Например, для каждого из чисел 13, 18, 20, 16, 21, 18, 13, 13 ранг определяется следующим образом. Строим из чисел вариационный ряд: 13, 13, 13, 16, 18, 18, 20, 21. Тогда: ранг числа 13 равен $\frac{1}{3}(1+2+3)=2$, ранг числа 16 равен 4, ранг числа 18 равен $\frac{1}{2}(5+6)=5{,}5$, ранг числа 20 равен 7 и ранг числа 21 равен 8.

Следовательно, для заданного ряда чисел получаем следующий ряд рангов: 2; 2; 2; 4; 5,5; 5,5; 7; 8. Это и есть результат ранжирования заданного ряда чисел.

Для изучения силы связи двух признаков X и Y, измеренных в порядковой шкале, используются коэффициент $\tau^{(s)}$ ранговой корреляции Спирмена и коэффициент $\tau^{(k)}$ ранговой корреляции Кендалла.

Пусть заданы две ранжировки $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$ случайной выборки n испытуемых по степени обладания признаками X и Y соответственно.

В том случае, когда нет связных рангов, *выборочный коэффициент ранговой корреляции Спирмена* $\widehat{r}_s$ задается формулой

$$\widehat{r}_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (x_i - y_i)^2.$$

При наличии связных рангов есть другая, более громоздкая для вычислений формула $\widehat{r}_s$.

Коэффициент $\widehat{r}_s$ может принимать значения лишь из $[-1, 1]$.

При $\widehat{r}_s = 1$ существует полное согласование рангов, т. е. каждый испытуемый имеет один и тот же ранг в обоих ранжировках.

При $\widehat{r}_s = -1$ имеется полная рассогласованность рангов в обоих ранжировках, т. е. в одной ранжировке ранги расположены в обратном порядке по сравнению с другой ранжировкой. При $\widehat{r}_s = 0$ связь между ранжировками отсутствует.

Если случайная выборка n испытуемых взята из некоторой генеральной совокупности $N > n$ испытуемых, то для всей совокупности определяется теоретический коэффициент r_s ранговой корреляции

Спирмена по той же формуле, что и для $\widehat{r}_s$ при отсутствии связных рангов, но с заменой n на N. Свойства коэффициентов r_s и $\widehat{r}_s$ одинаковы.

При $n > 10$ можно проверить гипотезу H_0: $r_s = 0$ об отсутствии ранговой корреляции в совокупности N испытуемых при заданном уровне значимости α, сравнив $|\widehat{r}_s|$ с критическим значением

$$T_{\text{кр}} = t_{\alpha/2} \cdot (n-2)\sqrt{\frac{1-(\widehat{r}_s)^2}{(n-2)}},$$

где значение $\tau_{\alpha/2}(n-2)$ берется из таблицы распределения Стьюдента по числу $\alpha/2$ и числу степеней свободы $(n-2)$. Если $|\widehat{r}_s| < T_{\text{кр}}$, то на уровне значимости α гипотеза H_0 принимается. Это значит, что ранговая корреляционная связь между признаками X и Y во всей генеральной совокупности незначима на уровне значимости α.

Пример 1. Предположим, что группа из 10 учеников некоторого класса проранжирована в соответствии с их способностями к математике и музыке следующим образом:

Ученики	А	Б	В	Г	Д	Е	Ж	З	И	К
Математика	1	2	3	4	5	6	7	8	9	10
Музыка	8	9	3	7	4	1	5	2	6	10

Необходимо: а) найти выборочный коэффициент Спирмена $\widehat{r}_s$, б) на уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 об отсутствии ранговой корреляционной связи математических и музыкальных способностей учеников класса.

Решение. а) Имеем

$$\widehat{r}_s = 1 - \frac{6}{10(100-1)} \cdot 182 \approx -0{,}103.$$

Для рассматриваемой выборки связь математических и музыкальных способностей слабая.

б) По $\alpha = 0{,}05$ и $n-2 = 8$ из таблицы распределения Стьюдента находим $t_{\text{кр}} = 2{,}306$ и, значит,

$$T_{\text{кр}} = 2{,}306 \cdot (n-2)\sqrt{\frac{1-(0{,}103)^2}{8}} \approx 0{,}812.$$

Так как $|\widehat{r}_s| < T_{\text{кр}}$, то гипотеза H_0 принимается на уровне значимости $\alpha = 0{,}05$. ■

Выборочный коэффициент ранговой корреляции Спирмена $\widehat{r}_s$ часто применяется как коэффициент надежности, характеризующий однородность психодиагностической методики. С этой целью расщепляют вопросы на четные и нечетные, ранжируют ответы на четные вопросы и отдельно на нечетные вопросы, а затем вычисляется $\widehat{r}_s$, который и будет коэффициентом надежности. Принято считать, что он должен быть не менее 0,8.

Другой используемой мерой связи двух ранжировок для выборки n испытуемых является *выборочный коэффициент ранговой корреляции Кендалла* $\widehat{\tau}$. Он задается следующей формулой в случае отсутствия связных рангов:

$$\widehat{\tau} = 1 - \frac{4K}{n(n-1)},$$

где K означает число инверсий в ряде рангов $(y_1, y_2, \ldots, y_n)$, если в качестве первого ряда рангов выступает ряд $(1, 2, \ldots, n)$.

Пусть, например, $(y_1, y_2, y_3, y_4) = (4, 3, 1, 2)$. Тогда инверсии (нарушения порядка) такие: (4 прежде 3) — одна инверсия, (4 прежде 1) — одна инверсия, (4 прежде 2) — одна инверсия, (3 прежде 1) — одна инверсия, (3 прежде 2) — одна инверсия. Единица, как и требуется, стоит прежде 2, и потому пара чисел (1, 2) инверсии не образует. В итоге в данном примере число инверсий $K = 5$.

В случае связных рангов формула $\widehat{\tau}$ несколько более сложная. Свойства коэффициента $\widehat{\tau}$ такие же, как и свойства $\widehat{r}_s$. В частности всегда $\widehat{\tau} \in [-1; 1]$.

Если выборка n испытуемых берется из генеральной совокупности $N > n$ испытуемых, то теоретический коэффициент ранговой корреляции Кендалла τ для всей совокупности в случае отсутствия связных рангов определяется по той же формуле, что и $\widehat{\tau}$, но с заменой n на N. Свойства τ и $\widehat{\tau}$ одинаковы.

При $n > 10$ можно проверить гипотезу H_0: $\tau = 0$ об отсутствии ранговой корреляции в совокупности N испытуемых для заданного уровня значимости α, если сравнить $|\widehat{\tau}|$ с критическим значением

$$T_{\text{кр}} = U_{1-\alpha/2} \cdot \sqrt{\frac{2 \cdot (2n+5)}{9n(n-1)}},$$

где $U_{1-\alpha/2}$ находится из табл. IV значений функции квантилей нормального распределения.

Пример 2. По данным примера 1 необходимо: а) найти коэффициент $\widehat{\tau}$, б) на уровне значимости $\alpha = 0{,}05$ проверить гипотезу об отсутствии ранговой корреляционной связи математических и музыкальных способностей учеников класса.

Решение. а) Имеем

$$\widehat{\tau} = 1 - \frac{4 \cdot 24}{90} \approx -0{,}067.$$

б) По $\alpha = 0{,}05$ из табл. IV находим $U_{1-\alpha/2} = 1{,}96$ и, значит,

$$T_{\text{кр}} = 1{,}96 \cdot \sqrt{\frac{2 \cdot 25}{9 \cdot 90}} \approx 0{,}49.$$

Так как $|\widehat{r}_s| < T_{\text{кр}}$, то гипотеза H_0 принимается на уровне значимости α. ■

Как видно из примера, численные значения $\widehat{r}_s$ и $\widehat{\tau}$ отличаются друг от друга, поскольку они имеют разные масштабы. Если $n > 10$ и значения $|\widehat{r}_s|$, $|\widehat{\tau}|$ не слишком близки к единице, то $\widehat{r}_s = 1{,}5\widehat{\tau}$.

Хотя на практике легче подсчитать $\widehat{r}_s$, чем $\widehat{\tau}$, однако сравнение коэффициентов $\widehat{r}_s$ и $\widehat{\tau}$ явно в пользу $\widehat{\tau}$.

Прежде всего заметим, что $\widehat{\tau}$ является несмещенной оценкой τ, а $\widehat{r}_s$ — смещенной оценкой r_s. Далее, если, например, необходимо заново ранжировать ту же выборку испытуемых и вычислять $\widehat{r}_s$ и $\widehat{\tau}$ в силу появления новых испытуемых или обнаружения ошибок, то вычисление $\widehat{r}_s$ требует полного пересчета данных, в то время как вычисление $\widehat{\tau}$ этого не требует. Наконец, распределение $\widehat{r}_s$ при $n \leqslant 10$ неизвестно и таблицы критических значений для $\widehat{r}_s$ в этом случае ненадежны. Для $n \leqslant 10$ изучено распределение $\widehat{\tau}$ и получены надежные таблицы критических значений для $\widehat{\tau}$.

С помощью выборочных коэффициентов ранговой корреляции Спирмена и Кендалла далеко не всегда можно отличить зависимость от независимости признаков. Однако считается, что большие (по модулю) эти коэффициенты свидетельствуют в пользу зависимости признаков (положительной или отрицательной, смотря по знаку коэффициента).

До сих пор рассматривалась ранговая корреляция для двух признаков. Однако иногда необходимо измерить связь между несколькими (более чем двумя) ранжировками. Для этой цели служит *выборочный коэффициент конкордации* (*согласованности*) $\widehat{W}(m)$, определяемый в случае отсутствия связных рангов формулой

$$\widehat{W}(m) = \frac{12}{m^2(n^3 - n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{m} x_i^{(j)} - \frac{m(n+1)}{2} \right)^2,$$

где m — число анализируемых признаков, n — число испытуемых в выборке, $x_i^{(j)}$ — ранг i-го испытуемого по j-му признаку.

Свойства коэффициента $\widehat{W}(m)$ следующие:

1) $0 \leqslant \widehat{W}(m) \leqslant 1$,

2) $\widehat{W}(m) = 1$ тогда и только тогда, когда все m ранжировок совпадают,

3) при $m > 3$ и при значительных различиях между всеми m ранжировками $\widehat{W}(m)$ близок к нулю,

4) $\widehat{W}(2) = \frac{1}{2}[\widehat{r}_s + 1]$, т. е. коэффициент конкордации для двух признаков является некоторой линейной функцией коэффициента Спирмена $\widehat{r}_s$.

Заметим, что при $m \geqslant 3$ коэффициент $\widehat{W}(m)$ не может принимать отрицательных значений, так как в этом случае ранжировки не могут полностью не совпадать в том смысле, как это бывает при $m = 2$.

Если случайная выборка n испытуемых берется из генеральной совокупности $N > n$ испытуемых, для которых анализируются m ранжировок без связных рангов, то можно определить теоретический коэффициент конкордации $W(m)$ по формуле, аналогичной формуле $\widehat{W}(m)$, но только с заменой n на N. При $n > 7$ и при заданном уровне значимости α можно проверить гипотезу H_0: $W(m) = 0$ об отсутствии ранговой связи между всеми m признаками. Для этого следует найти по таблице χ^2-распределения с $(n-1)$ степенью свободы критическое значение $\chi^2_\alpha(n-1)$.

При $m(n-1)\widehat{W}(m) < \chi^2_\alpha(n-1)$ гипотеза H_0 принимается на уровне значимости α.

Пример 3. Три эксперта упорядочили восемь испытуемых следующим образом:

1-й эксперт — 1, 2, 3, 4, 5, 6, 7, 8,
2-й эксперт — 2, 3, 1, 5, 4, 7, 6, 8,
3-й эксперт — 1, 3, 2, 4, 5, 8, 7, 6.

Необходимо: а) найти коэффициент $\widehat{W}(3)$, б) проверить гипотезу H_0: $W(3) = 0$ об отсутствии множественной ранговой связи на уровне значимости $\alpha = 0{,}05$ для всей совокупности, из которой взята выборка восьми испытуемых.

Решение. а) Имеем: $m = 3$, $n = 8$,

$$\widehat{W}(3) = \frac{12}{9 \cdot 8(64-1)}\left[\left(4 - \frac{27}{2}\right)^2 + \left(8 - \frac{27}{2}\right)^2 + \left(6 - \frac{27}{2}\right)^2 + \left(13 - \frac{27}{2}\right)^2 + \right.$$

$$\left. + \left(14 - \frac{27}{2}\right)^2 + \left(21 - \frac{27}{2}\right)^2 + \left(20 - \frac{27}{2}\right)^2 + \left(22 - \frac{27}{2}\right)^2\right] \approx 0{,}921.$$

б) $m(n-1)\widehat{W}(m) = 3 \cdot 7 \cdot 0{,}921 = 19{,}341 > \chi^2_\alpha(n-1) = 14{,}067$.

Таким образом, гипотеза H_0 отвергается на уровне значимости $\alpha = 0{,}05$. ■

В заключение отметим, что в случае задания для выборки из n испытуемых более чем двух ранжировок, т. е. при ранжировании испытуемых по значениям более чем двух признаков, кроме вычисления выборочных парных коэффициентов ранговой корреляции Кендалла $\widehat{\tau}$ можно также по определенным формулам находить и выборочные частный коэффициент ранговой корреляции Кендалла и множественный коэффициент ранговой корреляции Кендалла.

§ 4. Анализ связи номинальных признаков

Связь между двумя номинальными признаками, т. е. признаками, значения которых измеряются в номинальной шкале, обычно называют *сопряженностью*.

Пусть номинальный признак A имеет m градаций (уровней или групп) A_1, A_2, …, A_m, а номинальный признак B имеет k градаций (уровней или групп) B_1, B_2, …, B_k, причем или $m > 1$, или $k > 1$, или одновременно $m > 1$, $k > 1$. Рассмотрим случайную выборку n испытуемых и построим для нее таблицу с m строками и k столбцами следующего вида:

A \ B	B_1	B_2	…	B_k	Σ
A_1	n_{11}	n_{12}	…	n_{1k}	n_{1*}
A_2	n_{21}	n_{22}	…	n_{2k}	n_{2*}
…	…	…	…	…	…
A_m	n_{m1}	n_{m2}	…	n_{mk}	n_{m*}
Σ	n_{*1}	n_{*2}	…	n_{*k}	n

Такая таблица называется *таблицей сопряженности признаков A и B*. В этой таблице: n_{ij} — количество испытуемых выборки, обладающих одновременно градациями A_i и B_j, т. е. частота появления комбинации A_i и B_j, $n_{i*} = n_{i1} + n_{i2} + \ldots + n_{ik}$ обозначает частоту признака A_i, $n_{j*} = n_{1j} + n_{2j} + \ldots + n_{kj}$ обозначает частоту признака B_j, наконец,

$$n = n_{1*} + n_{2*} + \ldots + n_{m*} = n_{*1} + n_{*2} + \ldots + n_{*k}$$

обозначает общее количество n испытуемых.

Например, в качестве признака A может выступать цвет волос с четырьмя градациями A_1, A_2, A_3, A_4 (светлый, русый, черный, рыжий), а в качестве признака B может выступать цвет глаз с четырьмя градациями B_1, B_2, B_3, B_4 (голубой, серый, зеленый, карий). Другим примером признака A могут быть умственные способности с двумя градациями (нормальные, слабые), а признаком B может быть темперамент с тремя градациями (живой, вялый, ровный). По заданной таблице сопряженности двух номинальных признаков A и B для некоторой случайной выборки n испытуемых из генеральной совокупности $N > n$ испытуемых можно определить несколько выборочных коэффициентов сопряженности, характеризующих связь признаков A и B. Наиболее употребимыми на практике являются коэффициент сопряженности Крамера, коэффициент сопряженности К. Пирсона и коэффициент сопряженности А. А. Чупрова. Не существует среди них никакого универсального коэффициента сопряженности и нет рекомендаций по их выбору. Положим

$$\chi^2 = n\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_{i*}\cdot n_{*j}} - 1\right).$$

Тогда *выборочный коэффициент сопряженности Крамера* равен

$$\widehat{c}^{(k)} = \left(\frac{\chi^2}{n\cdot\min(m-1,\,k-1)}\right)^{\frac{1}{2}},$$

выборочный коэффициент сопряженности К. Пирсона имеет вид

$$\widehat{c}^{(\Pi)} = \left(\frac{\chi^2}{n+\chi^2}\right)^{\frac{1}{2}},$$

выборочный коэффициент сопряженности А. А. Чупрова равен

$$\widehat{T} = \left(\frac{\chi^2}{\sqrt{n(m-1)(k-1)}}\right)^{\frac{1}{2}}.$$

Коэффициенты $\widehat{c}^{(k)} = \widehat{c}^{(\Pi)} = \widehat{T} = 0$ свидетельствуют о независимости признаков A и B. Далее, $0 \leqslant \widehat{c}^{(k)} \leqslant 1$ и $0 \leqslant \widehat{T} \leqslant 1$, причем $\widehat{c}^{(k)} = \widehat{T} = 1$ дает возможность однозначно восстановить значение другого признака по заданному значению одного признака. Наибольшее значение коэффициента $\widehat{c}^{(\Pi)}$ зависит каждый раз от размеров таблицы $m \times k$. Это свойство делает $\widehat{c}^{(\Pi)}$ неудобным для применения в тех

случаях, когда необходимо сравнивать коэффициенты $\widehat{c}^{(\Pi)}$ для таблиц сопряженности разных размеров. Кроме того, $\widehat{c}^{(\Pi)}$ не достигает единицы.

Зная выборочный коэффициент сопряженности Крамера $\widehat{c}^{(k)}$, можно на заданном уровне значимости α проверить *гипотезу* H_0 *о независимости признаков* A *и* B *во всей генеральной совокупности испытуемых*. Для этого необходимо сравнить χ^2 с найденным по заданным уровню значимости α и числу степеней свободы $(m-1)(k-1)$ из таблицы χ^2-распределения критическим значением $\chi^2_{\text{кр}}$ (см. табл. VI). Если $\chi^2 < \chi^2_{\text{кр}}$, то гипотеза H_0 на уровне α принимается.

В том случае, когда гипотеза H_0 отвергается на уровне α, и, значит, связь между признаками A и B в генеральной совокупности действительно существует, то с доверительной вероятностью $P = 1 - 2\alpha$ для неизвестного коэффициента $c^{(k)}$ для всей генеральной совокупности испытуемых можно построить доверительный интервал вида

$$(\widehat{c}^{(k)} - U_{1-\alpha} \cdot \sigma,\ \widehat{c}^{(k)} + U_{1-\alpha} \cdot \sigma),$$

где U_q — q-квантиль стандартного нормального распределения (находится из табл. IV), а

$$\sigma = \left(\frac{1}{n \cdot \min(m-1,\ k-1)} \right)^{\frac{1}{2}}.$$

Пример 1. Задана следующая таблица сопряженности умственных способностей (признак A) и темперамента (признак B) для случайной выборки 42 детей детского сада. Здесь градации A: A_1 — нормальные способности, A_2 — слабые способности, а градации B: B_1 — живой темперамент, B_2 — ровный темперамент, B_3 — вялый темперамент.

A \ B	B_1	B_2	B_3	Σ
A_1	12	18	6	36
A_2	3	2	1	6
Σ	15	20	7	42

Необходимо: а) найти коэффициент $\widehat{c}^{(k)}$, б) проверить на уровне значимости $a = 0{,}05$ гипотезу H_0: $\widehat{c}^{(k)} = 0$ о независимости признаков A и B для детей детсада.

Решение. а) У нас $m=2$, $k=3$, $n=42$, $\sigma=\left(\frac{1}{42}\right)^{\frac{1}{2}}\approx 0{,}154$.

$$\chi^2=42\left(\frac{12^2}{15\cdot 36}+\frac{18^2}{20\cdot 36}+\frac{6^2}{7\cdot 36}+\frac{3^2}{15\cdot 6}+\frac{2^2}{20\cdot 6}+\frac{1^2}{7\cdot 6}-1\right)\approx 0{,}84.$$

Значит, $\widehat{c}^{(k)}=\chi\cdot\sigma=0{,}92\cdot 0{,}154=0{,}142$.

б) По $a=0{,}05$ и по числу степеней свободы $(m-1)(k-1)=2$ из табл. VI находим $\chi_{\text{кр}}=5{,}991$. Так как $\chi^2<\chi^2_{\text{кр}}$, то на уровне значимости $\alpha=0{,}05$ гипотеза H_0 принимается. ■

Пример 2. Задана следующая таблица сопряженности, в которой градации признака A — принадлежность школе № 1 или школе № 2, а градации признака B — сдали или не сдали выпускники школы вступительные экзамены в вузы, для выборки 187 выпускников школ:

A \ B	B_1	B_2	Σ
A_1	82	18	100
A_2	44	43	87
Σ	126	61	187

Проверить гипотезу H_0 о независимости признаков A и B для всех выпускников двух школ на уровне значимости $\alpha=0{,}05$ и построить доверительный интервал для неизвестного коэффициента сопряженности $c^{(k)}$.

Решение. Имеем $n=187$, $m=k=2$, $\sigma=\left(\frac{1}{187}\right)^{\frac{1}{2}}\approx 0{,}073$,

$$\chi^2=187\left[\frac{82^2}{126\cdot 100}+\frac{18^2}{100\cdot 61}+\frac{44^2}{126\cdot 87}+\frac{43^2}{61\cdot 87}-1\right]\approx 20{,}944,$$

$$\widehat{c}^{(k)}=\chi\cdot\sigma\approx 4{,}576\cdot 0{,}073\approx 0{,}334.$$

Из табл. VI по $\alpha=0{,}05$ и числу степеней свободы $(m-1)(k-1)=1$ находим $\chi^2_{\text{кр}}=3{,}841$. Так как $\chi^2>\chi^2_{\text{кр}}$, гипотеза H_0 отклоняется на уровне значимости $\alpha=0{,}05$.

Используя формулу доверительного интервала для неизвестного (теоретического) коэффициента сопряженности $c^{(k)}$, получаем, что с доверительной вероятностью $1-2\alpha=0{,}9$ значение $c^{(k)}$ находится в пределах от $0{,}334-1{,}64\cdot 0{,}073=0{,}214$ до $0{,}334+1{,}64\cdot 0{,}073=0{,}454$. ■

Заметим в заключение, что для анализа таблиц сопряженности 2×2 используется ряд других коэффициентов, например коэффициенты φ и Юла. Если же признаки измерены в разных шкалах, то методы анализа связи таких признаков становятся гораздо сложнее.

В статистических пакетах программ (SPSS, STATISTICA и др.) имеются компьютерные программы вычисления всех введенных в этой главе коэффициентов корреляции.

Задачи к главе 9

1. Для случайной выборки из 30 детей детсада была применена некоторая психодиагностическая методика с целью определения связи общего интеллекта с художественными способностями. Для определения однородности методики методом расщепления был получен выборочный коэффициент корреляции (надежности методики) 0,8. Предполагая, что общий интеллект и художественные способности детей детсада образуют двумерную нормальную генеральную совокупность, проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о нулевом коэффициенте надежности для всей совокупности и с доверительной вероятностью 0,95 построить доверительный интервал для коэффициента надежности.

2. Для случайной выборки из 27 студентов вуза была применена некоторая психодиагностическая методика с целью определения связи общего интеллекта с техническими способностями и методом «тест-ретест» был получен выборочный коэффициент корреляции (коэффициент стабильности методики) 0,6. Предполагая, что общий интеллект и технические способности студентов вуза образуют двумерную нормальную совокупность, проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о нулевом коэффициенте стабильности для всей совокупности студентов вуза и с доверительной вероятностью 0,95 построить доверительный интервал для коэффициента стабильности.

3. Известно, что для учеников одного класса общий интеллект, математические и гуманитарные способности образуют трехмерную нормальную совокупность. По случайной выборке из 8 учеников класса была получена матрица выборочных парных коэффициентов корреляции

$$\widehat{R} = \begin{pmatrix} 1 & 0{,}94 & 0{,}02 \\ 0{,}94 & 1 & 0{,}8 \\ 0{,}92 & 0{,}8 & 1 \end{pmatrix}.$$

а) Найти матрицу выборочных частных коэффициентов корреляции.

б) Найти выборочный множественный коэффициент детерминации $\widehat{R}^2_{1.23}$.

в) Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0: $R^2_{1.23} = 0$ для совокупности учеников класса.

4. Два эксперта проранжировали десять детсадов с точки зрения эффективности подготовки выпускников детсада к школе. Получены два ряда рангов:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10
2, 3, 1, 4, 6, 5, 9, 7, 8, 10.

Оценить степень согласованности мнений двух экспертов, использовав в качестве измерителя коэффициент Спирмена $\widehat{r}_s$ и коэффициент Кендалла $\widehat{\tau}$. Сравнить эти коэффициенты.

5. Семь средних школ района были проранжированы по трем признакам: процент выпускников, поступивших в вузы; процент выпускников, закончивших школу без троек; процент учителей школы, закончивших госуниверситеты. Получены три ряда рангов:

1, 2, 3, 4, 5, 6, 7
3, 1, 2, 5, 4, 7, 6
2, 1, 3, 4, 6, 5, 7.

С помощью коэффициента конкордации оценить степень согласованности в ранжировании школ и проверить гипотезу H_0 об отсутствии множественной ранговой связи трех признаков во всей совокупности школ района на уровне значимости $\alpha = 0{,}05$.

6. Задана 2×2-таблица сопряженности признака A — уровня жилищных условий с градациями A_1 (плохие жилищные условия), A_2 (хорошие жилищные условия) и признака B — общими умственными способностями с градациями B_1 (нормальные способности), B_2 (слабые способности) для выборки 168 детей детсада:

A \ B	B_1	B_2	Σ
A_1	33	38	71
A_2	39	58	97
Σ	72	96	168

Необходимо:

а) найти коэффициент $\widetilde{n}^{(k)}$,

б) проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0: $c^{(k)} = 0$ о независимости признаков A и B для детей детсада,

в) построить доверительный интервал для неизвестного коэффициента сопряженности $c^{(k)}$.

7. Задана 2×2-таблица сопряженности признака A — успеваемость с градациями A_1 (успешная), A_2 (неуспешная) и признака B — семейное положение с градациями B_1 (холост), B_2 (женат) для выборки 61 студентов-мужчин курса:

A \ B	B_1	B_2	Σ
A_1	33	17	50
A_2	7	4	11
Σ	40	21	61

Необходимо:

а) найти коэффициент $\widetilde{n}^{(k)}$,

б) на уровне значимости $\alpha = 0{,}1$ проверить гипотезу H_0: $c^{(k)} = 0$ о независимости признаков A и B для совокупности студентов-мужчин курса.

ГЛАВА 10

РЕГРЕССИОННЫЙ АНАЛИЗ

Регрессионный анализ применяется для исследования формы зависимости между одной результирующей количественной переменной и одной или несколькими объясняющими количественными переменными. В дальнейшем будем обозначать результирующую переменную через Y, а объясняющую переменную (объясняющие переменные) — через X. Основная цель регрессионного анализа — это прогноз некоторого результата (обучения, работы) Y по ряду предварительно измеренных характеристик $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$.

§ 1. Простая линейная регрессия

1. Определения

Начнем с наиболее простого случая, когда имеется результирующая случайная переменная Y и одна объясняющая переменная X. Для построения формы зависимости переменной Y от X задана некоторая исходная случайная выборка данных (x_i, y_i), где y_i — значение зависимой переменной Y при заданном значении x_i, $i = 1, 2, \ldots, n$. Предположим, что каждое наблюдаемое в опыте значение y_i можно мысленно представить в виде

$$y_i = a + bx_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Здесь $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ — ненаблюдаемые некоррелируемые случайные величины, a и b — неизвестные параметры.

Это значит, что предполагается справедливой следующая модель зависимости случайной переменной Y от переменной X:

$$Y = a + bX + \varepsilon,$$

где a и b — неизвестные параметры, а ε — некоторая случайная величина. Такая модель зависимости Y от X называется *простой линейной регрессией* Y по X. Случайная компонента ε называется *остатком* или *ошибкой*. Она отражает тот факт, что Y зависит не только от X, а имеются другие факторы, хотя и незначительно, но влияющие на значения случайной переменной Y. Например, присутствие компоненты ε можно объяснить наличием погрешностей при измерении значений переменной Y.

В классическом варианте простой линейной регрессии накладываются следующие требования:

1) независимая переменная X является неслучайной или случайной,

2) ненаблюдаемые случайные ошибки (отклонения, возмущения) $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ являются некоррелируемыми и одинаково распределенными случайными величинами, для которых одинаковы средние значения $\mathsf{M}\varepsilon_i = 0$ и одинаковы дисперсии $\mathsf{D}\varepsilon_i = \sigma^2$ при всех $i = 1, 2, \ldots, n$ (здесь σ^2 — постоянная, но неизвестная дисперсия),

3) если X — случайная переменная, то ошибки ε_i не коррелируют со значениями x_i, $i = 1, 2, \ldots, n$.

Если X не случайная переменная, то часто дополнительно к этим требованиям выдвигается требование распределения всех случайных величин ε_i, $i = 1, 2, \ldots, n$, по нормальному закону $N(0, \sigma^2)$. Тогда говорят, что имеет место *гауссовская модель простой линейной регрессии*. Для нее случайные ошибки $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ являются независимыми.

На практике для выдвижения гипотезы о существовании простой линейной регрессии Y по X используют так называемую диаграмму рассеяния (или корреляционное поле). Так называют множество n точек O_i с координатами (x_i, y_i), $i = 1, 2, \ldots, n$, изображенными на плоскости с прямоугольными осями координат X и Y. Обычно визуальное изучение диаграммы рассеяния позволяет выдвинуть или отбросить гипотезу о простой линейной регрессии Y по X. Если диаграмма рассеяния группируется вдоль некоторой прямой, то можно предполагать существование линейной регрессии Y по X (см. рис. 12).

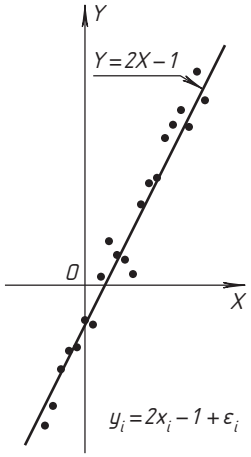


Рис. 12

2. Оценки параметров

По исходной случайной выборке наблюдений (x_i, y_i), $i = 1, 2, \ldots$ $\ldots, n$, можно получить статистические оценки неизвестных параметров a и b простой линейной регрессии. Параметры a и b являются случайными величинами, поскольку соответствуют случайной выборке. Пусть $\widehat{a}$ — оценка параметра a, а $\widehat{b}$ — оценка параметра b.

Для получения оценок параметров a и b пользуются так называемым *методом наименьших квадратов* (МНК), который минимизирует сумму квадратов отклонений наблюдаемых значений y_i от расчетных (сумму квадратов ошибок)

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2.$$

Минимум этой суммы квадратов ищется по переменным a и b.

Для того чтобы эта сумма квадратов достигала минимума, необходимо равенство нулю ее частных производных по a и b:

$$\begin{cases} -2 \sum\limits_{i=1}^{n} [y_i - a - bx_i] = 0, \\ -2 \sum\limits_{i=1}^{n} [y_i - a - bx_i] x_i = 0. \end{cases}$$

Из этой системы так называемых нормальных уравнений следует, что

$$\begin{cases} \sum\limits_{i=1}^{n} y_i - na - b \sum\limits_{i=1}^{n} x_i = 0, \\ \sum\limits_{i=1}^{n} y_i x_i - a \sum\limits_{i=1}^{n} x_i - b \sum\limits_{i=1}^{n} x_i^2 = 0. \end{cases}$$

Для получения оценок $\widehat{a}$ и $\widehat{b}$ необходимо решить эту систему уравнений. Если ввести в рассмотрение выборочные средние значения $\overline{x}$ и $\overline{y}$:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

то из первого уравнения системы получаем после деления на n, что $\overline{y} = \widehat{a} + \widehat{b}\overline{x}$.

Подставив $\widehat{a} = \overline{y} - \widehat{b}\overline{x}$ во второе уравнение системы, получаем

$$\sum_{i=1}^{n} y_i x_i = (\overline{y} - \widehat{b}\overline{x}) \sum_{i=1}^{n} x_i + \widehat{b} \sum_{i=1}^{n} x_i^2 = (\overline{y} - \widehat{b}\overline{x}) n\overline{x} + \widehat{b} \sum_{i=1}^{n} x_i^2.$$

Отсюда находим, что

$$\widehat{b} = \frac{\sum\limits_{i=1}^{n} y_i x_i - n\overline{x}\cdot\overline{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}.$$

Используя частные производные второго порядка по a и b от суммы квадратов ошибок, можно доказать, что на самом деле найденные $\widehat{a}$ и $\widehat{b}$ дают минимум суммы квадратов ошибок.

Величины $\widehat{a}$ и $\widehat{b}$ называются полученными по МНК оценками неизвестных параметров a и b модели простой линейной регрессии. Пусть s_x^2 — выборочная дисперсия для выборки $x_1, x_2, \ldots, x_n$, s_y^2 — выборочная дисперсия для выборки $y_1, y_2, \ldots, y_n$, а $\widehat{r}$ — выборочный коэффициент корреляции для этих выборок. Тогда нетрудно увидеть, что

$$\widehat{b} = \widehat{r}\cdot\frac{s_y}{s_x}.$$

Если наблюдения X и Y стандартизированы, т. е. вместо x_i берутся $\frac{x_i-\overline{x}}{s_x}$ и вместо y_i берутся $\frac{y_i-\overline{y}}{s_y}$, то отсюда $\widehat{b}=\widehat{r}$, т. е. $\widehat{b}$ дает оценку степени корреляционной зависимости Y от X. Эти формулы дают другой способ вычисления $\widehat{b}$, не решая системы нормальных уравнений.

Пример 1. При тестировании по двум тестам X и Y случайной выборки первоклассников в количестве 15 человек были получены (в баллах) среднее значение $\overline{x}=5{,}13$, среднее значение $\overline{y}=7{,}27$, выборочная дисперсия $s_x^2=6{,}15$, выборочная дисперсия $s_y^2=4{,}42$, выборочный коэффициент корреляции $\widehat{r}=0{,}96$.

Найти оценки $\widehat{a}$ и $\widehat{b}$ неизвестных параметров a и b простой линейной регрессии Y по X для первоклассников.

Решение. Сначала находим

$$\widehat{b}=\widehat{r}\cdot\frac{s_y}{s_x}=0{,}96\cdot\frac{\sqrt{4{,}42}}{\sqrt{6{,}15}}\approx 0{,}96\cdot\frac{2{,}1}{2{,}48}\approx 0{,}81.$$

Оценку $\widehat{a}$ получаем из формулы $\overline{y}=\widehat{a}+\widehat{b}\overline{x}$:

$$\widehat{a}=\overline{y}-\widehat{b}\overline{x}\approx 7{,}27-0{,}81\cdot 5{,}13\approx 3{,}11. \blacksquare$$

Оценки $\widehat{a}$ и $\widehat{b}$ являются случайными величинами, так как они зависят от случайных выборок, поэтому необходимо знать их свойства. Известно, что при выполнении сформулированных выше тре-

бований на модель зависимости Y от X *оценки* $\widehat{a}$ *и* $\widehat{b}$, *полученные с помощью МНК, обладают следующими свойствами*.

1) Оценки $\widehat{a}$ и $\widehat{b}$ являются несмещенными, т. е. $\mathsf{M}(\widehat{a}) = a$, $\mathsf{M}(\widehat{b}) = b$.

2) Оценки $\widehat{a}$ и $\widehat{b}$ являются состоятельными, так как (по вероятности)

$$\widehat{a}_n \to a, \quad \widehat{b}_n \to b, \quad n \to \infty,$$

если только выборочная дисперсия $s_x^2(n) \to c$ при $n \to \infty$, где число $c \neq 0$.

3) Оценки $\widehat{a}$ и $\widehat{b}$ являются эффективными. Это значит, что они имеют наименьшую дисперсию по сравнению с любыми другими оценками параметров a и b, линейными относительно величин y_i, и ошибок модели ε_i, $i = 1, 2, \ldots, n$.

4) Оценки $\widehat{a}$ и $\widehat{b}$ являются некоррелируемыми случайными величинами, т. е. их ковариация $\operatorname{cov}(\widehat{a}, \widehat{b}) = 0$.

Если простая линейная регрессия является гауссовской, то кроме перечисленных свойств для оценок $\widehat{a}$ и $\widehat{b}$ справедливы и другие свойства. Для гауссовской модели простой линейной регрессии оценки $\widehat{a}$ и $\widehat{b}$, полученные МНК, имеют наименьшую дисперсию не только среди линейных, но среди всех несмещенных оценок. Кроме того, обе случайные величины $\widehat{a}$ и $\widehat{b}$ распределены по нормальному закону и независимы как случайные величины. Более точно можно показать, что

$$\widehat{a} \sim N\left(a, \frac{\sigma^2}{n}\right), \quad \widehat{b} \sim N\left(b, \frac{\sigma^2}{(x_i - \overline{x})^2}\right).$$

В модели простой линейной регрессии кроме a и b участвует еще один неизвестный параметр σ^2 — дисперсия ошибок наблюдения, который необходимо оценить. Можно показать, что для σ^2 имеется несмещенная оценка

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} [y_i - \widehat{a} - \widehat{b}x_i]^2.$$

Оценки $\widehat{a}$ и $\widehat{b}$ тем надежнее (качественнее), чем меньше их дисперсии, т. е. разброс вокруг a, b соответственно и σ^2. Таким образом, необходимо на практике не только получать по заданной выборке (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) оценки $\widehat{a}$ и $\widehat{b}$ коэффициентов a, b простой линейной регрессии, но и дать анализ надежности полученных оценок, указав их дисперсии.

3. Проверка гипотезы об отсутствии линейной регрессии

В случае гауссовской простой линейной регрессии можно построить доверительные интервалы для оценок $\widehat{a}$ и $\widehat{b}$ и проверить гипотезу H_0: $b=0$ об отсутствии всякой линейной регрессии между Y и объясняющей переменной X.

Рассмотрим гипотезу H_0: $b=0$ и конкурирующую гипотезу H_1: $b\neq 0$. Известно, что при гипотезе H_0 статистика

$$t=\frac{\widehat{b}\cdot\sqrt{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}{\widehat{\sigma}}=\frac{\widehat{b}\cdot s_x\cdot\sqrt{n}}{\widehat{\sigma}},$$

где $\overline{x}$ — выборочное среднее значение для наблюдений $x_1, x_2, \ldots, x_n$, распределена по закону Стьюдента с $(n-2)$ степенями свободы.

Для этой статистики находят выборочное значение $t_{\text{набл}}$.

Пусть задан уровень значимости α. По числу α и числу $(n-2)$ из таблицы критических значений статистики Стьюдента (см. табл. VII) находят критическое значение $t_{\text{кр}}$.

Если $|t_{\text{набл}}|<t_{\text{кр}}$, то гипотеза H_0 принимается, а при $|t_{\text{набл}}|\geqslant t_{\text{кр}}$ гипотеза H_0 отклоняется и принимается гипотеза H_1.

Пример 2. Рассмотрим данные примера 1 и будем считать, что имеет место гауссовская простая линейная регрессия Y по X с оценкой дисперсии ошибки регрессии $\widehat{\sigma}^2=0{,}56$. Требуется на уровне значимости $\alpha=0{,}05$ проверить гипотезу H_0 об отсутствии простой линейной регрессии Y по X, т. е. $b=0$.

Решение. Имеем $t_{\text{набл}}=\dfrac{0{,}81\cdot\sqrt{6{,}15}\cdot\sqrt{15}}{0{,}56}=13{,}82$.

На уровне $\alpha=0{,}05$ и при $n-2=13$ из табл. VII находим $t_{\text{кр}}=2{,}16$. Поскольку $t_{\text{набл}}=13{,}82>t_{\text{кр}}=2{,}16$, то принимается альтернатива H_1: $b\neq 0$. ■

Для проверки общего качества выборочной функции $\widehat{f}(x)=\widehat{a}+\widehat{b}x$ простой линейной регрессии используют обычно так называемый выборочный коэффициент детерминации $\widehat{R}=\widehat{r}^2$. Он характеризует ту долю общей вариации (разброса) зависимой переменной Y, которая обусловлена изменением только выборочной линейной функции регрессии $\widehat{f}(x)=\widehat{a}+\widehat{b}x$. Разумеется, что общая вариация Y зависит от вариации $\widehat{f}(x)$ и от вариации случайного остатка ε.

Для примера 1 выборочный коэффициент корреляции $\widehat{r} = 0{,}96$ и, значит, выборочный коэффициент детерминации $\widehat{R} = \widehat{r}^2 \approx 0{,}92$. Это значит, что доля дисперсии зависимой переменной Y, объясненная с помощью выборочной функции простой линейной регрессии, является достаточно большой.

4. Прогноз значений зависимой переменной

Используя модель простой линейной регрессии, можно дать (точечный или интервальный) прогноз значения зависимой переменной Y по заданной величине независимой переменной X.

Пусть заданы исходные наблюдения (x_i, y_i), $i = 1, 2, \ldots, n$, и пусть известно, что зависимость между переменными Y и X описывается моделью простой линейной регрессии. Пусть, кроме того, задано еще одно значение x_{n+1} независимой переменной X, однако соответствующее значение y_{n+1} зависимой переменной неизвестно, но известно, что $y_{n+1} = a + bx_{n+1} + \varepsilon_{n+1}$, причем $\mathsf{M}\varepsilon_{n+1} = 0$, $\mathsf{D}\varepsilon_{n+1} = \sigma^2$ и случайные остатки $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{n+1}$ являются некоррелированными и одинаково распределенными.

Оказывается, что наилучший (в смысле среднего квадрата ошибки) линейный относительно $y_1, y_2, \ldots, y_n$ и несмещенный точечный прогноз для неизвестного значения y_{n+1} дает оценка

$$\widehat{y}_{n+1} = \widehat{a} + \widehat{b}x_{n+1}.$$

В случае гауссовской простой линейной регрессии можно указать и интервальный прогноз для неизвестного значения y_{n+1}. С доверительной вероятностью $\gamma = 1 - \alpha$ доверительный интервал для неизвестного прогнозного значения y_{n+1} при значении x_{n+1} определяется границами

$$\widehat{a} + \widehat{b}x_{n+1} \pm U_q \cdot \widehat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}},$$

где U_q — q-квантиль стандартного нормального распределения, которая находится по табл. IV при $q = \dfrac{1+\gamma}{2}$.

Пример 3. При тестировании по двум тестам X и Y случайной выборки пяти студентов из группы студентов были получены (в баллах) следующие пары значений: (1, 2), (2, 5), (3, 3), (4, 8), (5, 7).

Считая, что для двух тестов в группе студентов имеет место гауссовская простая линейная регрессия Y по X, построить доверительный интервал с доверительной вероятностью $\gamma = 0{,}95$ для неизвестного значения $y(6)$ при $x = 6$.

Решение. Имеем $\overline{x} = 3$, $\overline{y} = 5$, $s_x \approx 1{,}41$, $s_y \approx 2{,}28$. Тогда

$$\widehat{r} \approx 0{,}81, \quad \widehat{b} \approx 1{,}31, \quad \widehat{a} \approx 1{,}07.$$

Следовательно, выборочная линейная функция регрессии

$$\widehat{f}(x) = 1{,}07 + 1{,}31x,$$

и, значит, $\widehat{f}(6) = 8{,}93$, $\widehat{\sigma} = 1{,}745$.

Из табл. IV находим, что $U_q = 1{,}96$. Кроме того,

$$\sqrt{1 + \frac{1}{5} + \frac{(6-3)^2}{\sum\limits_{i=1}^{5}(x_i - 3)^2}} = 1{,}451.$$

Тогда доверительный интервал имеет вид:

$$(\widehat{f}(6) - 1{,}96 \cdot 1{,}745 \cdot 1{,}451;\ \widehat{f}(6) + 1{,}96 \cdot 1{,}745 \cdot 1{,}451).$$

Расчет дает такой интервал: (3,97; 13,89). ■

На практике используются и некоторые обобщения классической модели линейной регрессии, когда отказываются от некоррелируемости случайных остатков ε_i, $i = 1, 2, \ldots, n$, и от некоррелируемости случайных остатков со случайной величиной X.

Замечание. Если дисперсии всех остатков ε_i различны и (или) остатки ε_i коррелируют между собой, то МНК дает несмещенные и состоятельные оценки $\widehat{a}$ и $\widehat{b}$ неизвестных параметров a и b модели, но эффективности оценок не будет. Вместо МНК можно применять обобщенный МНК, который дает другие оценки параметров a и b, являющиеся в классе линейных несмещенных оценок уже эффективными.

Если случайные остатки ε_i коррелируют со случайной величиной X, то МНК дает смещенные и несостоятельные оценки неизвестных параметров a и b. Поэтому для анализа линейных регрессионных моделей такого типа используется другой метод — так называемый метод инструментальных переменных, который дает уже несмещенные и состоятельные оценки параметров a и b, однако в общем случае эти оценки не являются эффективными.

§ 2. Непараметрическая линейная регрессия и множественная линейная регрессия. Понятие о нелинейной регрессии

Классическая простая линейная регрессия может быть использована лишь при предположениях, что результаты отдельных измерений представляют собой некоррелируемые случайные величины и что случайные ошибки опыта распределены одинаково с общим нулевым математическим ожиданием и с общей дисперсией. Эти достаточно жесткие предположения на практике проверить нелегко. Если же эти предположения нарушены, то оценки коэффициентов регрессии не будут обладать такими желательными свойствами, как несмещенность, состоятельность и эффективность.

Поэтому при невозможности проверить предпосылки применения классической простой линейной регрессии следует использовать непараметрическую линейную регрессию, основанную на рангах наблюдений.

1. Модель непараметрической линейной регрессии

Пусть для n различных значений $x_1, x_2, \ldots, x_n$ величины X получены наблюдения $y_1, y_2, \ldots, y_n$ случайной величины Y, где y_i — значение Y в точке x_i, $i = l, 2, \ldots, n$. Будем считать в дальнейшем, что $x_1 < x_2 < \ldots < x_n$. Предполагается, что $y_i = a + bx_i + \varepsilon_i$, $i = 1, 2, \ldots, n$, где $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ — независимые случайные величины, извлеченные из одной и той же непрерывной совокупности.

Исследование зависимости Y от X основано на рангах наблюдений Y. В этом случае ничего определенного о величине коэффициента регрессии a сказать нельзя, поскольку изменение всех наблюдений y_i на одно и то же число не изменяет ранги $y_1, y_2, \ldots, y_n$. Поэтому речь идет только о получении оценки коэффициента регрессии b.

Используя коэффициент τ ранговой корреляции Кендалла, можно показать, что оценка $\widehat{b}$ неизвестного параметра b модели задается формулой

$$\widehat{b} = \text{медиана } \{S_{ij}\}$$

где $S_{ij} = \dfrac{y_j - y_i}{x_j - x_i}$.

Все значения S_{ij}, $1 \leqslant i < j \leqslant n$, задают всевозможные значения углового коэффициента прямой. Число таких значений S_{ij} равно $N = C_n^2$, где C_n^2 — число сочетаний из n по два.

Пусть $S^{(1)} \leqslant S^{(2)} \leqslant \ldots \leqslant S^{(N)}$ обозначают упорядоченные значения всех S_{ij}. Тогда если N — нечетное число: $N = 2k + 1$, то $\widehat{b} = S^{(k+1)}$, а если N — четное число: $N = 2k$, то

$$\widehat{b} = \frac{1}{2}[S^{(k)} + S^{(k+1)}].$$

Замечание. Оценка $\widehat{b}$ менее чувствительна к грубым ошибкам, чем классическая оценка b, полученная МНК. Можно указать и доверительный интервал для оценки $\widehat{b}$. Он не зависит от закона распределения случайных ошибок $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Имеется также непараметрический критерий Тейла для проверки гипотезы об отсутствии какой-либо линейной связи между случайной переменной Y и неслучайной переменной X.

Пример. Пусть на контрольной работе по математике 5 учеников одного класса получили следующие баллы (по десятибалльной системе):

Ученики	1	2	3	4	5
Баллы	8	9	3	7	4

В предположении, что справедлива модель непараметрической линейной регрессии зависимости баллов по контрольной работе от учеников класса, дать оценку $\widehat{b}$ параметра b модели.

Решение. Имеем:

$$S_{12} = 1, \quad S_{13} = -\frac{5}{2}, \quad S_{14} = -\frac{1}{3}, \quad S_{15} = -1,$$

$$S_{23} = -6, \quad S_{24} = -1, \quad S_{25} = -\frac{5}{3},$$

$$S_{34} = 4, \quad S_{35} = \frac{1}{2}, \quad S_{45} = -3.$$

Расположим эти числа по возрастанию:

$$-6, \ -3, \ -\frac{5}{2}, \ -\frac{5}{3}, \ -1, \ -1, \ -\frac{1}{3}, \ \frac{1}{2}, \ 1, \ 4.$$

Искомую оценку $\widehat{b}$ задает медиана этого ряда

$$\widehat{b} = \frac{1}{2}(-1 - 1) = -1. \blacksquare$$

2. Множественная линейная регрессия

На практике поведение количественной случайной результирующей переменной Y объясняется обычно не одной количественной

переменной X, а несколькими количественными переменными $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$, где $m > 1$. Регрессионная модель, включающая несколько объясняющих переменных, называется *множественной регрессией*.

Например, такая модель необходима, когда требуется определить зависимость общего интеллекта от математических способностей и от гуманитарных способностей.

Чаще всего предполагается, что справедлива модель *множественной линейной регрессии*, т. е. что

$$Y = a + b_1 X^{(1)} + b_2 X^{(2)} + \ldots + b_m X^{(m)} + \varepsilon, \quad m > 1,$$

где a, b_1, b_2, …, b_m — неизвестные параметры регрессии, а ε — случайная ошибка, $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ — неслучайные переменные.

В классической модели линейной множественной регрессии на случайную ошибку ε накладываются ограничения, подобные ограничениям в простой линейной регрессии.

Если заданы n наблюдений вида y_i, $x_i^{(1)}$, $x_i^{(2)}$, …, $x_i^{(m)}$, $i = 1, 2, \ldots$ …, n, то предполагается, что

$$y_i = a + b_1 x_i^{(1)} + b_2 x_i^{(2)} + \ldots + b_m x_i^{(m)} + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

причем ε_1, ε_2, …, ε_n — некоррелируемые случайные величины с одинаковым средним значением $\mathsf{M}\varepsilon_i = 0$ и одинаковой (неизвестной) дисперсией $\mathsf{D}\varepsilon_i = \sigma^2$ для всех $i = 1, 2, \ldots, n$. Эти данные пишут в виде матрицы порядка $n \times (m + 1)$.

Иногда дополнительно предполагается, что ошибки ε_i нормально распределены по закону $N(0, \sigma^2)$, и тогда линейная множественная регрессия называется *гауссовской*. При гауссовской линейной множественной регрессии случайные ошибки ε_1, ε_2, …, ε_n являются независимыми.

Основной задачей линейной множественной регрессии, как и в случае простой линейной регрессии, является задача построения точечного и интервального прогноза для неизвестного значения результирующей переменной Y.

Как и в случае простой линейной регрессии, коэффициенты a, b_1, b_2, …, b_m линейной множественной регрессии неизвестны и, по заданным n наблюдениям приходится получать их оценки $\widehat{a}$, $\widehat{b}_1$, $\widehat{b}_2$, …, $\widehat{b}_m$. Для нахождения этих оценок необходимо потребовать, чтобы $n \geqslant m + 1$ и чтобы ранг матрицы X наблюдений был равен m. Если эти условия не выполняются, то либо принципиально невозможно получить оценки неизвестных параметров модели, либо эти оценки нельзя получить классическим методом наименьших квадратов (МНК).

При выполнении всех вышесформулированных условий на модель линейной множественной регрессии для получения оценок неизвестных параметров модели используется МНК.

При выполнении предпосылок линейной множественной регрессии МНК дает несмещенные, состоятельные (при некоторых условиях) и эффективные оценки $\widehat{a}$, $\widehat{b}_1$, $\widehat{b}_2$, …, $\widehat{b}_m$ параметров a, b_1, b_2, …, b_m линейной множественной регрессии только в том случае, когда матрица исходных наблюдений переменных $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ имеет ранг, равный m.

Кроме получения оценок $\widehat{a}$, $\widehat{b}_1$, $\widehat{b}_2$, …, $\widehat{b}_m$, как и в случае простой линейной регрессии, необходимо оценить степень разброса (т. е. дисперсию) значений полученных коэффициентов. В случае гауссовской линейной множественной регрессии можно построить доверительные интервалы для оценок $\widehat{a}$, $\widehat{b}_1$, $\widehat{b}_2$, …, $\widehat{b}_m$ и проверить нулевую гипотезу для каждого из параметров a, b_1, b_2, …, b_m.

Если исходные данные для Y, $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ стандартизированы, то уравнение линейной множественной регрессии имеет вид

$$Y = \beta_1 X^{(1)} + \beta_2 X^{(2)} + \ldots + \beta_m X^{(m)} + \varepsilon, \quad m > 1,$$

где параметры β_1, β_2, …, β_m называются стандартными коэффициентами регрессии или β-коэффициентами.

При наличии двух или более объясняющих переменных $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ каждый β-коэффициент зависит от корреляции данной объясняющей переменной с результирующей переменной Y и от ее корреляции с другими объясняющими переменными. Знак β-коэффициента соответствует знаку центр корреляции данной объясняющей переменной и переменной Y, а величина β-коэффициента тем меньше, чем сильнее связана данная объясняющая переменная с другими переменными. Для любой переменной X модуль ее β-коэффициента не превосходит модуля коэффициента корреляции X с Y, т. е. $|\beta_x| \leqslant |r(X, Y)|$. Произведение $\beta_x \cdot r(X, Y)$ — это вклад данной переменной X в дисперсию результирующей переменной Y. Этот вклад тем выше, чем выше ее корреляция с Y и чем ниже ее корреляция с другими объясняющими переменными. При незначительном вкладе переменной X в оценку Y эта переменная X является несущественной для предсказания неизвестных значений Y и ее отбрасывают.

Для объяснения доли вариации зависимой переменной Y, обусловленной вариацией выборочной функции регрессии $\widehat{f}(x^{(1)}, x^{(2)}, \ldots, x^{(m)}) = \widehat{a} + \widehat{b}_1 x^{(1)} + \widehat{b}_2 x^{(2)} + \ldots + \widehat{b}_m x^{(m)}$, т. е. вариацией объясняющих переменных $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$, используется выборочный коэффи-

циент множественной детерминации $\widehat{R}^2$, задаваемый формулой

$$\widehat{R}^2 = 1 - \frac{\sum\limits_{i=1}^{n}(y_i - a - b_1 x_i^{(1)} - \ldots - b_m x_i^{(m)})^2}{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}.$$

$\widehat{R}$ — это выборочный коэффициент множественной корреляции переменной Y с переменными $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$.

Говорят, что объясняющие переменные $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ линейной множественной регрессии *обладают свойством мультиколлинеарности*, если между переменными $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ существует линейная функциональная зависимость, т. е. значение хотя бы одной из переменных является линейной комбинацией наблюденных значений остальных переменных. В этом случае ранг матрицы наблюдений $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ меньше m и из системы нормальных уравнений МНК не позволяет получить оценки $\widehat{a}$, $\widehat{b}_1$, $\widehat{b}_2$, …, $\widehat{b}_m$. Самыми простыми мерами в отношении устранения мультиколлинеарности могут быть: 1) увеличение выборки, 2) исключение тех из переменных $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$, которые сильно коррелируют с остальными. На практике анализ явления мультиколлинеарности начинают с анализа матрицы $\widehat{R}$ выборочных коэффициентов парных корреляций. Считается, что наличие значений этих коэффициентов по модулю больше 0,75 свидетельствует о присутствии мультиколлинеарности.

Множественный регрессионный анализ можно проводить с помощью компьютера, используя для этого соответствующую программу из статистических пакетов SPSS, STATISTICA и др.

3. Нелинейная регрессия

До сих пор обсуждалась лишь линейная регрессия. Однако может случиться, что взаимосвязь зависимой переменой Y и одной или более независимыми переменными X будет нелинейной. Например, может оказаться, что на диаграмме рассеяния наблюдения (x_i, y_i), $i = 1, 2, \ldots, n$, переменной Y, зависящей от переменной X, группируются не в окрестности какой-то прямой, а в окрестности параболы $y = ax^2 + bx + c$, $a \neq 0$.

В таком случае либо преобразуют данные и применяют линейную регрессию, либо используют методы нелинейной регрессии.

Например, если $Y = ae^{bX+\varepsilon}$, где $a > 0$ и b — неизвестные параметры, а ε — случайный остаток, то $\ln Y = \ln a + bX + \varepsilon$. Для полученной

линейной регрессии с помощью МНК можно найти оценки $\ln \hat{a}$ и $\hat{b}$, значит, и оценку $\hat{a}$. Для зависимости вида

$$Y = \frac{1}{a + bX + \varepsilon}$$

замена $Y_1 = 1/Y$ приводит к линейной регрессии $Y_1 = a + bX + \varepsilon$, а для зависимости вида $Y = aX^b \cdot e^{\varepsilon}$, где $Y > 0$, $X > 0$, $a > 0$, логарифмирование приводит к линейной регрессии $\ln Y = \ln a + b \ln X + \varepsilon$.

Хотя список примеров, сводящихся к линейной регрессии, можно продолжить, однако во многих случаях актуальна непосредственная оценка нелинейной функции регрессии $f(X, \Theta)$, если $Y = f(X, \Theta) + \varepsilon$, где Θ — набор неизвестных параметров, а ε — случайный остаток. Здесь остро стоит проблема выбора правильного вида функции $f(X, \Theta)$, поскольку неточности при выборе $f(X, \Theta)$ существенно сказываются на адекватности всей модели нелинейной регрессии. Оценки неизвестных параметров Θ находятся с помощью нелинейного МНК. Если заданы наблюдения (x_i, y_i), $i = 1, 2, \ldots, n$, то параметры Θ оценивают таким образом, чтобы сумма квадратов ошибок

$$\sum_{i=1}^{n} [y_i - f(x_i, \Theta)]^2$$

была минимальной. Для решения такой задачи существуют специальные методы.

В заключение отметим, что регрессионный анализ работает не только самостоятельно, но нередко является частью методов многомерной статистики (факторного, дискриминантного, кластерного анализов, многомерного шкалирования).

Задачи к главе 10

1. При тестировании по двум тестам X и Y случайной выборки пяти учащихся из группы учащихся были получены (в баллах) следующие пары значений: (1, 3), (2, 4), (3, 6), (4, 5), (5, 8). Считая, что для двух тестов в группе учащихся имеет место гауссовская простая линейная регрессия Y по X, построить доверительный интервал с доверительной вероятностью $\gamma = 0{,}95$ для неизвестного значения $y(6)$ при $x = 6$.

2. Известно, что вес Y и рост X для группы мужчин образуют гауссовскую простую линейную регрессию Y по X. Для случайной

выборки из десяти мужчин были получены $\overline{x} = 170$ см, $\overline{y} = 62$ кг, $s_x = 2$, $s_y = 1{,}5$, $\widehat{r} = 0{,}8$. Найти оценки $\widehat{a}$ и $\widehat{b}$ неизвестных параметров a и b регрессии и проверить гипотезу H_0: $b = 0$ об отсутствии простой линейной регрессии Y по X в группе мужчин на уровне значимости $\alpha = 0{,}05$, если оценка дисперсии ошибки регрессии $\widehat{\sigma}^2 = 0{,}45$.

3. Известно, что математические способности Y и художественные способности X для группы детей образуют гауссовскую простую линейную регрессию Y по X. Для случайной выборки из семнадцати детей по некоторой методике были получены $\overline{x} = 30$ баллов, $\overline{y} = 80$ баллов, $s_x = 1{,}9$ балла, $s_y = 1{,}17$ балла, $\widehat{r} = -0{,}54$. Найти оценки $\widehat{a}$ и $\widehat{b}$ неизвестных параметров a и b регрессии и проверить гипотезу H_0: $b = 0$ об отсутствии простой линейной регрессии Y по X в группе детей на уровне значимости $\alpha = 0{,}05$, если оценка дисперсии ошибки регрессии $\widehat{\sigma}^2 = 0{,}51$.

ГЛАВА 11

АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

§ 1. Определение и структура временны́х рядов

Временны́м рядом называют ряд наблюдений $x(t_1)$, $x(t_2)$, …, $x(t_n)$ некоторого случайного процесса, т. е. некоторой случайной величины $x(t)$, зависящей от времени t, полученных в последовательные моменты времени t_1, t_2, …, t_n.

Примеры временных рядов дают:

а) наблюдения за весом или ростом новорожденного ребенка в течение нескольких недель или месяцев,

б) число школьников в некоторой школе за несколько лет,

в) результаты тестирования некоторого испытуемого по одному и тому же тесту общего интеллекта, полученные в последовательные моменты времени,

г) изменение доходов или расходов определенной семьи в каждом месяце некоторого года.

Имеются два принципиальных отличия временного ряда от ряда наблюдений, образующего случайную выборку. Первое отличие временного ряда от случайной выборки в том, что различные наблюдения временного ряда не являются статистически независимыми друг от друга. Например, временные ряды резко отличаются от последовательных подбрасываний монеты. Другое отличие временного ряда от случайной выборки в том, что наблюдения временного ряда не являются одинаково распределенными с точки зрения теории вероятностей, т. е. функция распределения случайной величины $x(t)$ зависит от времени t.

Будем считать в дальнейшем, как это обычно предполагается, что наблюдения во временных рядах сделаны через равные промежутки времени. Приняв этот промежуток за единицу, значения временного ряда записываются в виде $x(1)$, $x(2)$, …, $x(n)$, где $x(t)$ — наблюдение в момент времени t. Таким образом, порядок следования наблюдений временного ряда является весьма существенным. Числа $x(t)$ называются *элементами временного ряда*.

Основной задачей анализа временного ряда является изучение его свойств и предсказание поведения временного ряда в недалеком будущем, т. е. краткосрочное прогнозирование значений $x(t)$ для $t = n + 1$, $n + 2$, $n + 3$.

Факторы, воздействующие на значения элементов временного ряда, подразделяют на закономерные (детерминированные) и случайные (нерегулярные). В свою очередь эти факторы выделяют *детерминированную составляющую* $d(t)$ и *случайную составляющую* $\varepsilon(t)$ каждого элемента $x(t)$ временного ряда при $t = 1, 2, \ldots, n$.

Составляющая $d(t)$ отражает действие каких-то определенных факторов или причин, совокупное влияние которых является устойчивым в течение длительного промежутка времени.

Составляющая $\varepsilon(t)$ выглядит хаотично и непредсказуемо. Она не поддается учету и контролю.

Наиболее часто используются аддитивная и мультипликативная модели временного ряда.

Аддитивной моделью временного ряда называется представление ряда в виде суммы детерминированной и случайной компонент, т. е.

$$x(t) = d(t) + \varepsilon(t), \quad t = 1, 2, \ldots, n.$$

Мультипликативной моделью временного ряда называется представление ряда в виде произведения детерминированной и случайной составляющих, т. е.

$$x(t) = d(t)\,\varepsilon(t), \quad t = 1, 2, \ldots, n.$$

В дальнейшем ограничимся рассмотрением лишь аддитивной модели временного ряда. В этой модели предполагается обязательное участие случайной компоненты $\varepsilon(t)$. Участие детерминированной компоненты $d(t)$ является необязательным.

В детерминированной компоненте $d(t)$ временного ряда выделяют три составляющих ее части: тренд $\mathrm{tr}(t)$, сезонную компоненту $s(t)$ и циклическую компоненту $c(t)$. Тогда можно записать, что $d(t) = \mathrm{tr}(t) + s(t) + c(t)$, $t = 1, 2, \ldots, n$.

Трендом $\mathrm{tr}(t)$, $t = 1, 2, \ldots, n$, временного ряда называется плавная, не циклического типа компонента, описывающая основную тенденцию в изменении временного ряда под влиянием долговременных факторов.

К таким факторам, например, можно отнести рост населения, изменение структуры возрастного состава, изменение структуры населения с точки зрения их образования, занятости, доходов, географического расселения и т. д.

Действие таких факторов происходит постепенно и поэтому их вклад описывается с помощью простых функций, например, линейной или квадратичной функций.

Сезонная компонента $s(t)$, $t = 1, 2, \ldots, n$, временного ряда описывает поведение ряда, изменяющееся регулярно в течение заданного периода (года, месяца, недели, дня и т. п.).

Сезонная компонента служит главным источником краткосрочных колебаний временного ряда. Например, некоторые виды заболеваний носят ярко выраженный сезонный характер. Сезонные эффекты присущи производству и потреблению некоторых видов товаров.

Циклическая компонента $c(t)$, $t = 1, 2, \ldots, n$, временного ряда описывает влияние на значения элементов временного ряда долговременных циклов экономической, демографической или астрофизической природы (подъем или спад экономической активности, демографические «ямы», циклы солнечной активности и т. п.).

Анализ временного ряда обычно начинается с построения и изучения его графика. По графику временного ряда во многих случаях можно определить наличие и характер тренда, наличие сезонных и циклических компонент.

Для объяснения поведения временного ряда и осуществления прогноза его дальнейшего поведения необходимо подобрать модель поведения случайной составляющей $\varepsilon(t)$ в аддитивной модели временного ряда. При этом чаще всего используются так называемые стационарные временные ряды.

§ 2. Стационарные временные ряды

Различают стационарные в узком смысле временные ряды и стационарные в широком смысле временные ряды.

Временной ряд $x(t)$ называется *стационарным* (или *стационарным в узком смысле*), если совместное распределение вероятностей наблюдений $x(t_1)$, $x(t_2)$, …, $x(t_m)$ такое же, как и для наблюдений $x(t_1 + \tau)$, $x(t_2 + \tau)$, …, $x(t_m + \tau)$, при любых m, t_1, t_2, …, t_m, τ.

Это определение говорит о том, что свойства стационарного в узком смысле временного ряда не зависят от изменения начала отсчета времени. Кроме того, из приведенного определения следует, что для случайной величины $x(t)$ не зависят от t ее закон распределения и все ее основные числовые характеристики, в том числе математическое ожидание (среднее значение) $\mathrm{M}x(t) = a$ и дисперсия $\mathrm{D}x(t) = \sigma^2$.

Для стационарных в узком смысле временных рядов $x(t)$ среднее значение $\mathrm{M}x(t)$ и дисперсия $\mathrm{D}x(t)$могут быть обычным образом оценены по наблюдениям $x(1)$, $x(2)$, …, $x(n)$. В качестве оценки $\mathrm{M}x(t)$ берется

$$\widehat{a} = \frac{1}{n} \sum_{t=1}^{n} x(t),$$

а в качестве оценки $\mathrm{D}x(t)$ —

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}[x(t) - \widehat{a}]^2.$$

Оценка $\widehat{a}$ среднего значения $\mathrm{M}x(t)$ является несмещенной, а при некоторых дополнительных условиях на ряд $x(t)$ является и состоятельной.

Из стационарности в узком смысле ряда $x(t)$ вытекает также, что ковариация между значениями $x(t)$ и $x(t \pm \tau)$ будет зависеть только от величины сдвига по времени τ и не зависит от времени t. Эта ковариация называется *автоковариацией* и определяется так:

$$\mathrm{cov}[x(t),\, x(t+\tau)] = \mathrm{M}\{[x(t) - a][x(t+\tau) - a]\}.$$

Степень тесноты статистической связи между наблюдениями временного ряда, отделенными по времени на τ единиц, определяется коэффициентом корреляции

$$r(\tau) = \frac{\mathrm{cov}[x(t),\, x(t+\tau)]}{\sqrt{\mathrm{D}x(t) \cdot \mathrm{D}x(t+\tau)}}.$$

Этот коэффициент измеряет корреляционную зависимость между членами одного и того же временного ряда, и поэтому его называют *коэффициентом автокорреляции*.

При исследовании зависимости $r(\tau)$ от времени τ функцию $r(\tau)$ называют *автокорреляционной функцией*, а ее график — *коррелограммой*.

Отметим основные свойства функции $r(\tau)$.

1) Автокорреляционная функция $r(\tau)$ является безразмерной, т. е. не зависит от масштаба измерения временного ряда. Ее значения могут находиться между (-1) и $(+1)$.

2) Из стационарности следует, что $r(\tau) = r(-\tau)$, т. е. функция $r(\tau)$ — четная и, значит, при ее анализе можно ограничиться рассмотрением только $\tau > 0$.

3) Функция $r(\tau)$ является монотонно убывающей (по абсолютной величине), поскольку чем больше величина сдвига τ, тем слабее взаимосвязь членов временного ряда $x(t)$ и $x(t+\tau)$ и, значит, тем меньше должна быть абсолютная величина $r(\tau)$.

По наблюдениям $x(1)$, $x(2)$, …, $x(n)$ статистическая оценка $\widehat{r}(\tau)$ автокорреляционной функции $r(\tau)$ определяется формулой

$$\widehat{r}(\tau) = \left[\frac{1}{n-\tau}\sum_{i=1}^{n-\tau}\{x(t) - \widehat{a}\}\{x(t+\tau) - \widehat{a}\}\right] \Big/ \left[\frac{1}{n}\sum_{i=1}^{n}\{x(t) - \widehat{a}\}^2\right],$$

где $\tau = 1,\, 2,\, \ldots,\, n-1$.

Функцию $\widehat{r}(\tau)$ называют *выборочной автокорреляционной функцией* или *сериальной корреляцией*.

Расширением класса стационарных в узком смысле временных рядов является класс стационарных в широком смысле временных рядов.

Временной ряд $x(t)$ называется *стационарным в широком смысле*, если его среднее значение и дисперсия не зависят от времени t, а автокорреляционная функция зависит лишь от сдвига τ.

Из этого определения следует, что и автоковариация не зависит от времени t, а зависит лишь от сдвига τ.

Очевидно, что всякий стационарный в узком смысле временной ряд является одновременно стационарным в широком смысле, но не наоборот.

Среди стационарных в широком смысле временных рядов наиболее распространенными являются ряды, называемые «белым шумом».

Временной ряд $x(t)$ называется *белым шумом*, если его среднее значение $\mathsf{M}x(t) \equiv 0$ для всех t и если автоковариация

$$\operatorname{cov}[x(t), x(t+\tau)] = \begin{cases} \sigma^2, & \text{при } \tau = 0, \\ 0, & \text{при } \tau \neq 0. \end{cases}$$

Здесь σ^2 не зависит от t.

Выше уже отмечалось, что стационарный в широком смысле временной ряд в общем случае не обязан быть стационарным в узком смысле. Однако имеются исключения. К ним относятся *нормальные* (или *гауссовские*) временные ряды, т. е. такие стационарные в широком смысле ряды, для которых случайная величина $x(t)$ при всех t распределена нормально. Гауссовские временные ряды, стационарные в широком смысле, одновременно являются и стационарными в узком смысле. В частности, белый шум является гауссовским, если для него $x(t)$ распределена нормально с параметрами 0 и σ^2.

В дальнейшем под *стационарным временным рядом* будет всегда пониматься стационарный в широком смысле ряд. Наиболее распространенным на практике случаем нарушения стационарности временного ряда является наличие зависимости среднего значения $\mathsf{M}x(t)$ от времени t. Если каким-то образом удается оценить $\mathsf{M}x(t)$, то преобразование $Y(t) = X(t) - \mathsf{M}x(t)$ превращает временной ряд в стационарный.

После того, как получен стационарный в широком смысле временной ряд, выбирают ту или иную модель этого стационарного временного ряда. На практике чаще всего используются два типа

линейной параметрической модели: модель авторегрессии некоторого порядка и модель скользящего среднего некоторого порядка.

Поскольку эти модели предназначены для описания поведения стационарных случайных составляющих $\varepsilon(t)$ временного ряда $x(t)$, то далее будем анализируемый стационарный временной ряд обозначать через $\varepsilon(t)$ и полагать, что его среднее значение $\mathsf{M}\varepsilon(t) = 0$ при всех значениях t. Кроме того, условимся в дальнейшем белый шум обозначать через $\delta(t)$, т. е. $\mathsf{M}\delta(t) \equiv 0$ и

$$\operatorname{cov}[\delta(t), \delta(t+\tau)] = \begin{cases} \sigma^2, & \text{при } \tau = 0, \\ 0, & \text{при } \tau \neq 0. \end{cases}$$

Говорят, что стационарный временной ряд $\varepsilon(t)$ описывается *моделью авторегрессии 1-го порядка*, если для всех t

$$\varepsilon(t) = \alpha\varepsilon(t-1) + \delta(t),$$

где α — число, для которого $|\alpha| < 1$, и $\delta(t)$ — белый шум.

Модель авторегрессии первого порядка обладает так называемым *марковским свойством*. Это значит, что поведение временного ряда в будущем определяется лишь его состоянием в настоящем и воздействиями на него в будущем, но не зависит от его предыдущего развития.

Как уже отмечено выше, среднее значение $\mathsf{M}\varepsilon(t) = 0$ для всех t. Кроме того, в случае модели авторегрессии 1-го порядка можно показать, что дисперсия

$$\mathsf{D}\varepsilon(t) = \frac{\sigma^2}{1-\alpha^2}.$$

Отсюда видно, что если $|\alpha|$ близок к единице, то $\mathsf{D}\varepsilon(t)$ будет намного больше $\mathsf{D}\delta(t) = \sigma^2$. Это значит, что слабые возмущения $\delta(t)$ могут порождать большие колебания $\varepsilon(t)$.

Автокорреляционная функция $r(\tau)$ для модели авторегрессии 1-го порядка определяется формулой $r(\tau) = \alpha^\tau$.

Отсюда следует простой вероятностный смысл параметра α: $\alpha = r(1)$, т. е. α задает коэффициент парной корреляции между двумя соседними членами ряда $\varepsilon(t)$.

Из формулы $r(\tau) = \alpha^\tau$ следует также, что степень тесноты корреляционной связи между членами временного ряда $\varepsilon(t)$ убывает с ростом сдвига τ.

Оценка параметров α и σ^2 модели авторегрессии 1-го порядка дается с помощью заданного временного ряда $x(1)$, $x(2)$, …, $x(n)$, а не с помощью его остатка $\varepsilon(t)$, который является ненаблюдаемым.

Для этого сначала находят оценку $\widehat{d}(t)$ детерминированной составляющей $d(t)$ временного ряда $x(t)$ (о методе получения $\widehat{d}(t)$ см. в следующем параграфе) и потом получают оценки $\widehat{\varepsilon}(t)$ случайных остатков $\varepsilon(t)$: $\widehat{\varepsilon}(t) = x(t) - \widehat{d}(t)$. При этом будем считать для простоты, что среднее значение

$$\overline{\varepsilon}(t) = \frac{1}{n} \sum_{t=1}^{n} \widehat{\varepsilon}(t) = 0.$$

Тогда оценка $\widehat{\alpha}$ параметра α задается формулой

$$\widehat{\alpha} = \frac{\frac{1}{n-1} \sum\limits_{t=1}^{n-1} \{\widehat{\varepsilon}(t) - \alpha\} \cdot \widehat{\varepsilon}(t+1)}{\frac{1}{n} \sum\limits_{t=1}^{n} \widehat{\varepsilon}^2(t)},$$

а оценка $\widehat{\sigma}^2$ параметра σ^2 задается формулой

$$\widehat{\sigma}^2 = (1 - \widehat{\alpha}^2) \frac{1}{n} \sum_{t=1}^{n} \widehat{\varepsilon}^2(t).$$

Для автокорреляционной функции $r(\tau)$ оценкой служит $\widehat{r}(\tau) = \widehat{\alpha}^{\tau}$.

Обобщением модели авторегрессии 1-го порядка является модель авторегрессии порядка p. Ограничимся описанием модели авторегрессии 2-го порядка.

Говорят, что стационарный временной ряд $\varepsilon(t)$ является *рядом авторегрессии 2-го порядка*, если для всех t

$$\varepsilon(t) = \alpha_1 \varepsilon(t-1) + \alpha_2 \varepsilon(t-2) + \delta(t),$$

где α_1 и α_2 — некоторые числа, для которых $|\alpha_1| < 2$, $|\alpha_1| + |\alpha_2| < 1$, а $\delta(t)$ — белый шум.

Для модели авторегрессии 2-го порядка $\mathsf{M}\varepsilon(t) = 0$ для всех t, а автокорреляционная функция $r(\tau)$ задается следующей рекуррентной формулой:

$$r(\tau) = \alpha_1 r(\tau - 1) + \alpha_2 r(\tau - 2).$$

С помощью заданного временного ряда $x(t)$ можно также дать оценку параметров α_1, α_2, σ^2 модели авторегрессии 2-го порядка.

Остановимся теперь коротко на временных рядах скользящего среднего 1-го порядка.

Пусть, как и ранее, $\delta(t)$ обозначает белый шум: $\mathsf{M}\delta(t) = 0$, $\mathsf{D}\delta(t) = \sigma^2$ для всех t.

Говорят, что стационарный временной ряд $\varepsilon(t)$ является *рядом скользящего среднего 1-го порядка*, если для всех t

$$\varepsilon(t) = \delta(t) + \theta\delta(t-1).$$

Можно показать, что для такого ряда среднее значение $\mathsf{M}\varepsilon(t) = 0$, дисперсия $\mathsf{D}\varepsilon(t) = \sigma^2(1+\theta^2)$ и автокорреляционная функция

$$r(\tau) = \begin{cases} \dfrac{\Theta}{1+\Theta^2}, & \text{при } \tau = 1, \\ 0, & \text{при } \tau \geqslant 2. \end{cases}$$

Указанное свойство $r(\tau)$ позволяет различать модель скользящего среднего 1-го порядка для временного ряда, используя график выборочной автокорреляционной функции $\hat{r}(\tau)$.

Можно ввести также понятие модели скользящего среднего порядка p для временных рядов. На практике встречаются и комбинации моделей авторегрессии и скользящего среднего. Например, стационарный ряд $\varepsilon(t)$ называют *рядом авторегрессии-скользящего среднего порядка* $(1,1)$, если для всех t

$$\varepsilon(t) = \alpha\varepsilon(t-1) + \delta(t) + \theta\delta(t-1),$$

где $\delta(t)$ — белый шум, α и θ — заданные числа, причем $|\alpha| < 1$.

Мы не будем останавливаться на свойствах таких рядов.

Если временной ряд не является стационарным, то с помощью вычисления разностей первого, второго и т. д. порядков необходимо пытаться получить ряды разностей соответственно первого, второго и так далее порядков. Например, разности первого порядка — это величины $\Delta x(t) = x(t) - x(t-1)$. Полученные ряды разностей уже могут являться стационарными рядами, которые можно моделировать.

Моделирование при помощи нестационарных рядов может привести, например, к так называемой ложной корреляции двух временных рядов, когда высокая корреляция двух нестационарных временных рядов может быть принята за причинную связь, в то время как этого нет на самом деле.

§ 3. Анализ детерминированной составляющей временного ряда

На начальном этапе статистического анализа временного ряда выявляются и оцениваются детерминированная составляющая $d(t)$ ряда и ее части: тренд $\mathrm{tr}(t)$, сезонная компонента $s(t)$ и циклическая компонента $c(t)$.

Предположим, что рассматриваемый временной ряд $x(1), x(2), \ldots$ $\ldots, x(n)$ может быть описан аддитивной моделью

$$x(t) = \mathrm{tr}(t) + s(t) + c(t) + \varepsilon(t), \quad t = 1, 2, \ldots, n,$$

где $\varepsilon(t)$ — случайная составляющая ряда.

После изучения графика временного ряда необходимо выделить (оценить) во временном ряде тренд, сезонную и циклическую компоненты. После их исключения временной ряд должен стать стационарным в широком смысле, для которого можно уже подбирать подходящую модель (модель авторегрессии некоторого порядка или модель скользящего среднего некоторого порядка или их комбинации). Модель считается подобранной, если остаточная компонента является белым шумом. Далее проводится анализ остатков и прогнозирование будущих значений временного ряда и указание точности этого прогноза.

1. Метод скользящего среднего

Задача выделения (оценки) детерминированной составляющей $d(t)$ временного ряда называется иногда *задачей элиминирования случайного остатка* $\varepsilon(t)$ или *задачей сглаживания временного ряда* $x(1), x(2), \ldots, x(n)$. Такая задача решается двумя способами.

Первый способ основан на предположении, что известна функция $f(t, \theta_1, \theta_2, \ldots, \theta_k)$, где $\theta_1, \theta_2, \ldots, \theta_k$ — некоторые неизвестные параметры, описывающая составляющую $d(t)$, и требуется лишь получить статистические оценки $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_k$ неизвестных параметров. Например, это может быть линейная функция $f(t, \theta_1, \theta_2) = \theta_1 + \theta_2 t$ с неизвестными параметрами θ_1, θ_2. Для получения оценок неизвестных параметров модели $f(t, \theta_1, \theta_2, \ldots, \theta_k)$ неслучайной составляющей используют метод наименьших квадратов (МНК), как и в регрессионном анализе, или некоторые обобщения МНК, если остаток $\varepsilon(t)$ удовлетворяет требованиям регрессионной модели.

Второй способ уже не связан гипотезой об общем виде составляющей $d(t)$ и дает лишь алгоритм (процедуру) расчета оценки $\widehat{d}(t)$ в любой наперед заданной точке t. Этот способ называют *методом скользящего среднего*. Опишем простейшую форму (линейную форму) метода скользящего среднего. Пусть сначала число наблюдений n — нечетное: $n = 2m + 1$, и пусть имеется временной ряд $x(1), x(2), \ldots, x(n)$. Берем три первых значения ряда и находим их среднее значение:

$$\widehat{x}(2) = \frac{1}{3}[x(1) + x(2) + x(3)].$$

Затем берем значения $x(2)$, $x(3)$, $x(4)$ и находим их среднее значение

$$\hat{x}(3) = \frac{1}{3}[x(2) + x(3) + x(4)].$$

Далее находят среднее значение $\hat{x}(4)$ для $x(3)$, $x(4)$, $x(5)$ и т. д. Найденный временной ряд

$$\hat{x}(2), \hat{x}(3), \hat{x}(4), \ldots, \hat{x}(n-1)$$

и дает «сглаженные» значения $\widehat{d}(t)$ (т. е. оценки $d(t)$) в соответствующие моменты времени t, за исключением $t=1$ и $t=n$.

Название метода скользящего среднего происходит от того, что при вычислении средних значений как бы «скользят» по временному ряду от его начала к концу, каждый раз отбрасывая один элемент ряда в начале и добавляя один следующий. Иногда, чтобы получить сглаженные значения $\widehat{d}(t)$ в моменты времени $t=1$ и $t=n$, полагают $\hat{x}(1) = \hat{x}(2)$, а $\hat{x}(n) = \hat{x}(n-1)$.

Если число наблюдений n — четное: $n=2m$, то при вычислении средних значений мы будем получать сглаженные значения $\widehat{d}(t)$ не в моменты времени $t=1, 2, \ldots, n$, а для моментов времени t, лежащих посередине между точками наблюдения. Чтобы получить сглаженное значение $\widehat{d}(t)$ в точках наблюдений $t=1, 2, \ldots, n$, необходимо взять среднее арифметическое двух сглаженных значений для двух окаймляющих эту точку промежуточных моментов времени.

Заметим, что в общем случае метод скользящего среднего для оценки составляющей $d(t)$ изменяет исходную остаточную случайную компоненту $\varepsilon(t)$.

Пример 1. Новорожденный весом 3 кг ежемесячно в течение последующих пяти месяцев прибавлял в весе согласно следующей таблице:

t	0	1	2	3	4	5
$x(t)$	3	4	5	5,7	6,7	7,4

а) Методом скользящего среднего с длиной усреднения, равной трем, получить сглаженные значения составляющей $\widehat{d}(t)$ заданного временного ряда для $t=1, 2, 3, 4$.

б) Считая случайный остаток $\varepsilon(t)$ белым шумом, а детерминированную составляющую $d(t) = at + b$ с неизвестными параметрами a и b для заданного временного ряда, методом наименьших квадратов (МНК) получить оценки $\widehat{a}$ и $\widehat{b}$ неизвестных параметров α и b.

Решение. а) Имеем: $\widehat{d}(1) = \frac{1}{3}(3+4+5) = 4$, $\widehat{d}(2) = \frac{1}{3}(4+5+5{,}7) = 4{,}9$, $\widehat{d}(3) = \frac{1}{3}(5+5{,}7+6{,}7) = 5{,}8$, $\widehat{d}(4) = \frac{1}{3}(5{,}7+6{,}7+7{,}4) = 6{,}6$.

б) Из условия минимальности функции

$$\varphi(a, b) = \sum_{i=0}^{5} [x(t_i) - at_i - b]^2$$

получаем систему уравнений

$$\begin{cases} \frac{\partial \varphi}{\partial a} = -2 \sum_{i=0}^{5} [x(t_i) - at_i - b] \cdot t_i = 0, \\ \frac{\partial \varphi}{\partial b} = -2 \sum_{i=0}^{5} [x(t_i) - at_i - b] = 0, \end{cases}$$

которая записывается в виде

$$\begin{cases} a \cdot \sum_{i=0}^{5} t_i^2 + b \cdot \sum_{i=0}^{5} t_i = \sum_{i=0}^{5} t_i \cdot x(t_i), \\ a\bar{t} + b = \bar{x}. \end{cases}$$

Для данных примера 1 имеем:

$$\sum_{i=0}^{5} t_i^2 = 55, \quad \sum_{i=0}^{5} t_i = 15, \quad \sum_{i=0}^{5} t_i x(t_i) = 94{,}9, \quad \bar{t} = \frac{1}{6} \sum_{i=0}^{5} t_i = 2{,}5,$$

$$\bar{x} = \frac{1}{6} \sum_{i=0}^{5} x(t_i) = 5{,}3.$$

Следовательно, система для нахождения оценок a и b принимает вид:

$$\begin{cases} 55a + 15b = 94{,}9, \\ 2{,}5a + b = 5{,}3. \end{cases}$$

Решая эту систему по правилу Крамера, получаем оценки параметров a и b: $\hat{a} = 0{,}88$; $\hat{b} = 3{,}1$. Значит $\hat{d}(t) = 0{,}88t + 3{,}1$. ■

Если на начальном этапе анализа временного ряда не удается оценить детерминированную составляющую $d(t)$ в целом, то обычно пробуют оценить и удалить ее компоненты: тренд $\mathrm{tr}(t)$, сезонную компоненту $s(t)$ и циклическую компоненту $c(t)$.

Наиболее часто рассматриваются следующие модели тренда: линейная модель: $\mathrm{tr}(t) = a_0 + a_1 t$, полиномиальная модель: $\mathrm{tr}(t) = a_0 + a_1 t + a_2 t^2 + \ldots + a_m t^m$, экспоненциальная модель: $\ln[\mathrm{tr}(t)] = a_0 + a_1 t$, логистическая модель: $\mathrm{tr}(t) = \frac{a}{1 + be^{-ct}}$, модель Гомперца: $\ln[\mathrm{tr}(t)] = a - br^t$, $0 < r < 1$.

Выбор модели тренда в первую очередь определяется графиком временного ряда.

Для оценки и удаления тренда из временного ряда используется метод наименьших квадратов (МНК), как и в регрессионном анализе, хотя для временных рядов нарушаются условия применения классического МНК, поскольку для временных рядов характерна взаимозависимость его членов. Однако, если правильно выбрана модель тренда и если среди наблюдений нет больших выбросов, то классический МНК дает для тренда разумные оценки. В других случаях для оценки тренда используются модифицированные МНК. После получения оценки тренда $\widehat{\mathrm{tr}}(t)$ в случае аддитивной модели временного ряда $x(t)$ тренд $\mathrm{tr}(t)$ можно удалить из ряда $x(t)$ с помощью замены $x(t) - \widehat{\mathrm{tr}}(t) = y(t)$. Новый временной ряд $y(t)$ уже не будет содержать тренда.

2. Оценка сезонной компоненты

Пусть теперь в аддитивной модели временного ряда $x(t)$ присутствует сезонная компонента $s(t)$, и пусть известен период p компоненты $s(t)$. Необходимо оценить значения $s(t)$ по наблюдениям $x(t)$.

Рассмотрим вначале случай отсутствия циклической компоненты $c(t)$ в аддитивной модели $x(t)$. Если $\widehat{\mathrm{tr}}(t)$ — оценка тренда и если считать для простоты, что $n = (m+1)p$ (длина ряда содержит целое число периодов p), то для каждого i-го сезона, $1 \leqslant i < p$, берем относящиеся к нему разности:

$$x(i) - \widehat{\mathrm{tr}}(i), \quad x(i+p) - \widehat{\mathrm{tr}}(i+p), \quad \ldots, \quad x(i+mp) - \widehat{\mathrm{tr}}(i+mp).$$

Тогда в качестве оценки значения $\widehat{s}(i)$ сезонной компоненты $s(i)$ берем среднее значение

$$\widehat{s}(i) = \frac{1}{m+1} \sum_{l=0}^{m} [x(i + l \cdot p) - \widehat{\mathrm{tr}}(i + lp)], \quad i = 1, 2, \ldots, p.$$

Оценки сезонной компоненты считаются недостаточно точными, если число периодов в исследуемом ряде меньше пяти-шести. Это означает, в частности, что для получения более-менее точной оценки сезонной компоненты необходимы наблюдения за пять-шесть лет.

После получения оценок сезонной компоненты в аддитивной модели временного ряда удаляют сезонные эффекты из рассматриваемого ряда, вычитая их из начальных значений ряда. Подобная процедура называется обычно *сезонным выравниванием временного ряда*.

При наличии во временном ряде циклической компоненты $c(t)$ для получения оценки сезонного эффекта $s(t)$ необходимо сначала оценить не только тренд, но и циклическую компоненту. На практике лучше одновременно оценить тренд и циклическую компоненту. Это делается с помощью метода скользящего среднего, самый простой вариант которого уже был описан выше для получения оценки составляющей $d(t)$.

Если известен период p сезонной компоненты $s(t)$, то порядок оценки значений $s(t)$ по наблюдениям $x(t)$ в целом такой же, как и в случае отсутствия $c(t)$. Только теперь вместо оценки тренда МНК используется скользящее среднее в качестве совместной оценки $\mathrm{tr}(t)$ и $c(t)$.

Обозначим через $\widehat{x}(t)$ скользящее среднее с периодом p, построенное для ряда $x(t)$. Для простоты занумеруем величины $\widehat{x}(t)$ так: $\widehat{x}(1)$, $\widehat{x}(2)$, ..., $\widehat{x}(k)$ и изменим нумерацию ряда $x(t)$ так, чтобы $\widehat{x}(t)$ соответствовал член $x(t)$. Кроме того, для простоты предположим, что $k = (m+1)p$. Для каждого сезона i, $1 \leqslant i \leqslant p$, рассмотрим все относящиеся к нему разности

$$x(i) - \widehat{x}(i), \quad x(i+p) - \widehat{x}(i+p), \quad \ldots, \quad x(i+mp) - \widehat{x}(i+mp).$$

Тогда в качестве оценки сезонной компоненты $s(i)$ можно взять среднее значение

$$\widehat{s}(i) = \frac{1}{m+1} \sum_{j=1}^{m+1} [x(i + j \cdot p) - \widehat{x}(i + j \cdot p)], \quad i = 1, 2, \ldots, p.$$

Для аддитивной модели временного ряда удаление сезонной компоненты проводится путем вычитания оцененной сезонной компоненты из исходного временного ряда.

Кроме перечисленных, имеются и другие методы получения оценок детерминированной составляющей $d(t)$ временного ряда и ее компонент: тренда, сезонной и циклической компонент.

Иногда необходимо преобразовать значения временного ряда $x(t)$ с помощью функции $y = f(x)$ и получить новый временной ряд $y(t) = f[x(t+1)]$. Чаще всего это делается либо для того, чтобы приблизить распределение $x(t)$ к нормальному, либо для того, чтобы избавиться от нестационарности наблюдаемого ряда, либо для того, чтобы сделать дисперсию временного ряда близкой к постоянной. Чаще других используется логарифмическое преобразование $y = \log_a x$, $a > 0$, $a \neq 1$, значений ряда $x(t)$, если все эти значения положительные.

3. Прогноз временного ряда

Коротко остановимся на методах прогнозирования значений временного ряда, когда имеющийся в наличии ряд экстраполируется вперед на один или несколько (l) временных тактов вперед.

Универсального метода прогнозирования не существует. Выбор метода прогнозирования в первую очередь зависит от длины такта (l), от длины анализируемого временного ряда, от наличия сезонной составляющей в ряде, от используемой модели стационарных и нестационарных рядов. При $l \leqslant 3$ обычно прогноз называют *краткосрочным*, при $l \leqslant 6$ прогноз называют *среднесрочным* и при $l > 6$ прогноз называют *долгосрочным*. При $n \leqslant 50$ ряд считается коротким, а при $n > 50$ — длинным.

Эффективно можно решать лишь задачи кратко- и среднесрочного прогноза. Долгосрочный прогноз требует использования и анализа специальных экспертных оценок.

Отметим, что в прогнозе временных рядов существенно используется не только прогноз детерминированной составляющей $d(t)$, но и взаимозависимость и прогноз случайных остатков. В этом принципиальное отличие прогноза временных рядов от прогноза, основанного на регрессионной модели, в котором используется лишь функция регрессии и не учитываются значения случайных остатков.

Из теории известно, что наилучшим (т. е. с наименьшей стандартной ошибкой) линейным прогнозом временного ряда в момент времени $(t+l)$, где $l = 1, 2, \ldots$, является условное математическое ожидание случайной величины $x(t+l)$ при условии, что все значения $x(t)$ известны до момента t включительно. Отсюда видно, что формулы для прогнозных значений временного ряда $x(t)$ получаются разные в зависимости от модели случайного остатка $\varepsilon(t)$.

Пусть, например, случайный остаток $\varepsilon(t)$ описывается моделью авторегрессии первого порядка, т. е.

$$\varepsilon(t) = \alpha\varepsilon(t-1) + \delta(t),$$

где α — заданное число, $|\alpha| < 1$, и $\delta(t)$ — белый шум. Пусть, кроме того, задан временной ряд $x(1), x(2), \ldots, x(t)$. Тогда прогнозные значения ряда в последующие три момента времени упрощенно задаются формулами вида

$$\begin{aligned} x(t+1) &= 1{,}8x(t) - 0{,}8x(t-1), \\ x(t+2) &= 1{,}8x(t+1) - 0{,}8x(t), \\ x(t+3) &= 1{,}8x(t+2) - 0{,}8x(t+1). \end{aligned}$$

Если же случайный остаток $\varepsilon(t)$ описывается моделью скользящего среднего первого порядка, т. е.

$$\varepsilon(t) = \delta(t) + \theta\delta(t-1),$$

где θ — заданное число и $\delta(t)$ — белый шум, и задан временной ряд $x(1), x(2), \ldots, x(t)$, то прогнозные значения ряда в последующие три момента времени приближенно задаются формулами вида

$$x(t+1) = 0{,}9x(t) + 0{,}1x(t-1),$$
$$x(t+2) = 0{,}9x(t) + 0{,}1x(t+1),$$
$$x(t+3) = 0{,}9x(t) + 0{,}1x(t+2).$$

Прогнозы, использующие модели случайного остатка $\varepsilon(t)$, являются наиболее эффективными. Кроме указанного метода прогноза, существуют и так называемые *адаптивные методы прогноза*, позволяющие сравнительно быстро обновлять ранее сделанные прогнозы по мере поступления новых данных.

ГЛАВА 12

МЕТОДЫ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ

§ 1. Дискриминантный анализ с обучением

Пусть задано n объектов (или испытуемых) $O_1, O_2, \ldots, O_n$, каждый из которых характеризуется определенным набором значений $m \geqslant 2$ количественных (т. е. измеренных в интервальной шкале или в шкале отношений) признаков $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$. Это значит, что имеются исходные данные вида $x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}$ для каждого объекта с номером $i = 1, 2, \ldots, n$. Другими словами, задана матрица исходных данных $\|x_i^{(j)}\|$ порядка $m \times n$. Признаки $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$ называются *дискриминантными переменными*.

Под *классификацией заданного множества n объектов* понимается разделение рассматриваемого неоднородного множества n объектов на несколько однородных классов (групп), причем число $2 \leqslant k < n$ этих классов и сами классы могут быть заранее известными или неизвестными.

Различают классификацию с обучением (или с учителем) и классификацию без обучения (или без учителя).

При классификации с обучением имеется априори заданное число k классов и каждый класс представлен своей случайной выборкой, т. е. заданы случайные выборки $x_{i1}, x_{i2}, \ldots, x_{ini}$, $j = 1, 2, \ldots \ldots, k$, причем i-я выборка определяет значения анализируемых признаков на n_i объектах ($n_i < n$) и берется из i-го класса. Эти заданные k выборок называются *обучающими* и классификация при таких условиях называется *классификацией с обучением*. В случае неизвестного числа классов или отсутствия обучающих выборок говорят о *классификации без обучения*.

Классы, на которые разбивают множество объектов, можно представить как значение некоторой зависимой (классифицирующей) переменной Y, измеренной в номинальной шкале.

Дискриминантный анализ дает методы различения (дискриминации) объектов наблюдения по определенным признакам. Например, речь может идти о разбиении некоторой неоднородной совокупности людей на несколько однородных классов (групп) по результатам тестирования. Другим примером может служить разделение опрашиваемых претендентов при выборе кандидатов на определенную

должность на две группы: «подходит» и «не подходит». Еще одним примером применения дискриминантного анализа может быть его применение при разделении школ, детсадов или предприятий на два класса: «с высокой эффективностью работы» и «с низкой эффективностью работы». Успешность или неуспешность работы или учебы, их прогноз тоже можно определять с помощью дискриминантного анализа.

В этом параграфе речь будет идти только о так называемом *параметрическом дискриминантном анализе с обучающими выборками*. Это означает следующее. Заранее заданы $2 \leqslant k < n$ классов и из каждого класса имеется случайная выборка наблюдений исходных признаков $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$. При этом под каждым классом понимается некоторая генеральная совокупность, причем все k классов описываются одной и той же известной одномодальной плотностью распределения вероятностей $p(x, \theta_j)$, зависящей однако от неизвестных разных параметров θ_j, где $j = 1, 2, \ldots, k$. Таким образом, классы — это генеральные совокупности, отличающиеся друг от друга только значениями неизвестных параметров θ_j в плотностях распределения вероятности. Исследуемая генеральная совокупность объектов представляет собой смесь генеральных совокупностей, причем ее плотность распределения вероятностей

$$p(x) = \sum_{j=1}^{k} \pi_j \cdot p(x, \Theta_j),$$

где π_j — удельный вес элементов j-го класса в исследуемой генеральной совокупности. Числа π_j либо определяются заранее самим содержанием задачи классификации, либо оцениваются по формуле

$$\pi_j = \frac{n_j}{n},$$

где n_j — объем j-й обучающей выборки, $n = n_1 + n_2 + \ldots + n_k$.

Дискриминантный анализ с обучением дает правило классификации, по которому новые объекты рассматриваемой генеральной совокупности относятся к одному из существующих классов.

Это правило классификации задается таким образом, чтобы оно минимизировало потери (или вероятность) неправильной классификации объектов.

Особенно простым является правило классификации в случае равных потерь при выполнении следующих условий:

1) число классов $k = 2$,

2) удельные веса обоих классов одинаковы, т. е. $\pi_1 = \pi_2 = \frac{1}{2}$,

3) исследуемая генеральная совокупность представляет собой смесь двух m-мерных нормально распределенных генеральных совокупностей, имеющих разные векторы математических ожиданий a_1 и a_2 и одинаковую ковариационную матрицу Σ.

При таких условиях объект с наблюдениями $X = (X^{(1)}, X^{(2)}, \ldots, X^{(m)})$ следует отнести к первому классу только в том случае, когда значение функции

$$D(X) = \left[X - \frac{1}{2}(\widehat{a}_1 + \widehat{a}_2)\right]^{\mathrm{T}} \cdot \widehat{\Sigma}^{-1} \cdot (\widehat{a}_1 - \widehat{a}_2) \geqslant 0,$$

и ко второму классу в других случаях. Здесь: $\widehat{a}_1$ и $\widehat{a}_2$ — векторы выборочных средних значений соответственно для первой и второй обучающих выборок, буква T означает операцию транспонирования, $\widehat{\Sigma}$ — оценка общей ковариационной матрицы Σ, вычисляемая по формуле

$$\widehat{\Sigma} = \frac{1}{n_1 + n_2 - 2}\left[n_1\widehat{\Sigma}_1 + n_2\widehat{\Sigma}_2\right],$$

где n_1 и n_2 — соответственно объем первой и второй обучающих выборок, а $\widehat{\Sigma}_1$ и $\widehat{\Sigma}_2$ — соответственно матрицы выборочных ковариаций для первой и второй обучающих выборок, $\widehat{\Sigma}^{-1}$ — обратная матрица к $\widehat{\Sigma}$.

Функция

$$D(X) = \left[X - \frac{1}{2}(\widehat{a}_1 + \widehat{a}_2)\right]^{\mathrm{T}} \cdot \widehat{\Sigma}^{-1} \cdot (\widehat{a}_1 - \widehat{a}_2),$$

по значениям которой принимается решение об отнесении наблюдения X к первому или второму классу, называется *линейной дискриминанатной функцией Фишера*, а сам дискриминантный анализ с обучением в таком случае называется *линейным*.

В том частном случае, когда проводится классификация на два класса одномерных ($m = 1$) стандартизованных нормальных наблюдений X, к первому классу относятся только те наблюдения X, для которых

$$\left[X - \frac{1}{2}(\widehat{a}_1 + \widehat{a}_2)\right]^{\mathrm{T}} (\widehat{a}_1 - \widehat{a}_2) \geqslant 0.$$

В случае $m = 2$ с помощью геометрической интерпретации можно пояснить смысл линейной дискриминантной функции $D(X)$. В этом случае каждый объект характеризуется двумя признаками $X^{(1)}$ и $X^{(2)}$ и может быть изображен точкой на плоскости с прямоугольными координатами $X^{(1)}$ и $X^{(2)}$. Если объекты принадлежат двум

разным множествам M_1 и M_2, то, чтобы наилучшим образом разделить M_1 и M_2, необходимо построить новую систему координат таким образом, чтобы проекции объектов на новые оси координат были максимально разделены. Этого результата можно достичь за счет подбора коэффициентов линейной дискриминантной функции

$$D(X) = a + b_1 X^{(1)} + b_2 X^{(2)}.$$

Правило классификации тогда означает, что объекты, расположенные над прямой $a + b_1 X^{(1)} + b_2 X^{(2)} = 0$, находятся ближе к центру множества M_1 и, следовательно, могут быть отнесены к первому классу, а объекты, расположенные ниже этой прямой, находятся ближе к центру множества M_2, т. е. относятся ко второму классу.

Пример. Эффективность работы школ области оценивалась по двум показателям: отношению числа отличников в конце учебного года к числу всех учащихся (%) и отношению числа выпускников школы, поступивших в вузы, к числу всех выпускников школы (%). В результате были выделены группа 1 из пяти школ с высокой эффективностью работы и группа 2 из четырех школ с низкой эффективностью работы. Для этих групп получены векторы выборочных средних значений

$$\widehat{a}_1 = (44\ 28)^{\mathrm{T}}, \quad \widehat{a}_2 = (14\ 38)^{\mathrm{T}}$$

и матрицы выборочных ковариаций

$$\widehat{\Sigma}_1 = \begin{pmatrix} 84 & 30 \\ 30 & 58 \end{pmatrix}, \quad \widehat{\Sigma}_2 = \begin{pmatrix} 80 & 26 \\ 26 & 54 \end{pmatrix}.$$

Необходимо провести классификацию школы, показатели которой задаются вектором: а) $X = (34\ 24)^{\mathrm{T}}$, б) $X = (17\ 60)^{\mathrm{T}}$, в) $X = (38\ 10)^{\mathrm{T}}$, если считать, что относительное количество школ с высокой эффективностью работы в области достигает 50% и что показатели группы школ 1 и 2 были обучающими выборками, извлеченными из нормально распределенных генеральных совокупностей с одинаковой ковариационной матрицей Σ.

Решение. Из условий задачи следует, что нужно воспользоваться линейной дискриминантной функцией Фишера $D(X)$. Проведем необходимые вычисления. Имеем:

$$\widehat{a}_1 - \widehat{a}_2 = (30\ -10)^{\mathrm{T}}, \quad \frac{1}{2}(\widehat{a}_1 + \widehat{a}_2) = (29\ 33), \quad \widehat{\Sigma} = \begin{pmatrix} 105{,}72 & 36{,}29 \\ 36{,}29 & 72{,}29 \end{pmatrix},$$

$$\widehat{\Sigma}^{-1} = \begin{pmatrix} 0{,}011 & -0{,}006 \\ -0{,}006 & 0{,}017 \end{pmatrix}, \quad \widehat{\Sigma}^{-1} \cdot (\widehat{a}_1 - \widehat{a}_2) = (0{,}39\ -0{,}35)^{\mathrm{T}}.$$

Тогда:

а) $X - \frac{1}{2}(\hat{a}_1 + \hat{a}_2) = (5\ -9)^{\mathrm{T}},\ D(X) = 0{,}51 > 0.$

Школа относится к группе 1.

б) $X - \frac{1}{2}(\hat{a}_1 + \hat{a}_2) = (-12\ 27)^{\mathrm{T}},\ D(X) = -0{,}49 < 0.$

Школа относится к группе 2.

в) $X - \frac{1}{2}(\hat{a}_1 + \hat{a}_2) = (9\ -23)^{\mathrm{T}},\ D(X) = 11{,}56 > 0.$

Школа относится к группе 1. ■

В общем случае линейный дискриминантный анализ является множественным, т. е. функция $D(X)$ содержит $m \geqslant 2$ дискриминантных переменных:

$$D(X) = a + b_1 X^{(1)} + b_2 X^{(2)} + \ldots + b_m X^{(m)}.$$

Дискриминантная функция может быть как линейной, так и нелинейной. Выбор ее вида определяется геометрическим расположением разделяемых классов в пространстве дискриминантных переменных. Дискриминантные переменные — это признаки, использующиеся для отличения одного класса от другого. Дискриминантные переменные должны быть линейно независимыми. На практике их выбор осуществляется на основе логического анализа исходной информации и специальных критериев. В рассмотренном выше примере в качестве дискриминантных переменных выступали процентное отношение числа отличников школы и процентное отношение числа выпускников, поступивших в вузы, к числу всех выпускников.

Изменение числа дискриминантных переменных сильно влияет на результат дискриминантного анализа. Чтобы судить о целесообразности включения или удаления дискриминантной переменной, используются специальные критерии, позволяющие оценить статистическую значимость ухудшения или улучшения разбиения после включения или удаления каждой из отобранных переменных.

Для оценки относительного вклада каждой переменной $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ в значение дискриминантной функции $D(X)$ необходимо использовать стандартизированные значения исходных переменных, для которых средние значения равняются нулю, а дисперсии — единице. Помимо вклада каждой исходной переменной в дискриминантную функцию $D(X)$, можно анализировать и степень корреляционной зависимости между $D(X)$ и каждой переменной $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$. Для этой цели служат коэффициенты корреляции, которые называются *структурными коэффициентами*.

Положительные или отрицательные структурные коэффициенты ориентируют объекты в различных направлениях. Структурные коэффициенты показывают вклад каждой дискриминантной переменной в различительную способность $D(X)$. Чем больше структурный коэффициент, тем больше зависимость $D(X)$ от данной переменной.

Основными проблемами дискриминантного анализа с обучением являются выбор дискриминантных переменных и выбор вида дискриминантной функции, если не подходит линейная функция. Кроме того, в тех случаях, когда множества, используемые в качестве обучающих выборок, расположены близко друг к другу и когда классифицируемый объект сильно удален от центров обоих множеств, возрастает вероятность ошибочной классификации новых объектов. Одним из возможных выходов в таком случае является пересмотр набора дискриминантных переменных.

Дискриминантный анализ с обучением иногда называют *распознаванием образов*. Он может быть использован для прогнозирования поведения объектов исследуемой совокупности на основе имеющихся типов поведения аналогичных объектов, входящих в состав объективно существующих или специально сформированных множеств (обучающих выборок).

В заключение отметим, что дискриминантный анализ без обучения даже в том простом его варианте, когда априорно известны число k классов и удельные веса $\pi_1, \pi_2, \ldots, \pi_k$ классов в исследуемой совокупности и когда все классы имеют одну и ту же одномодальную плотность распределения вероятностей $p(x, \Theta_j)$, зависящую от неизвестных параметров Θ_j, $j = 1, 2, \ldots, k$, является значительно более сложным, чем дискриминантый анализ с обучением. Сложность в том, что в этом случае неизвестные параметры оцениваются по классифицируемым наблюдениям с помощью, например, метода максимального правдоподобия. После получения оценок параметров используется то же правило классификации, что и ранее.

Основным показателем качества дискриминации на практике является процент совпадения действительной классификации известных объектов и их классификации с помощью дискриминантной функции.

Компьютерные программы статистических пакетов SPSS, STATISTICA позволяют провести дискриминантный анализ. Они позволяют, в частности, автоматически отсеять малозначимые переменные для дискриминантного анализа и получить графическое изображение дискриминантых функций.

§ 2. Кластерный анализ

1. Постановка задачи

Пусть исходные статистические данные заданы либо в виде матрицы $\|x_i^{(j)}\|$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, наблюдений для n объектов (или испытуемых) $O_1, O_2, \ldots, O_n$, каждый из которых характеризуется значениями m количественных признаков $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$, т. е. задана матрица X порядка $m \times n$, либо в виде матрицы $\|\gamma_{ij}\|$, $i, j = 1, 2, \ldots, n$, парных сравнений O_i и O_j, если заданы характеристики $\|\gamma_{ij}\|$ попарных сравнений каждых двух объектов. В качестве таких характеристик могут выступать меры сходства или различия объектов. В зависимости от природы исходных признаков в качестве мер сходства и различия объектов могут быть использованы коэффициенты парной корреляции Пирсона, коэффициенты ранговой корреляции Спирмена и Кендалла, а также коэффициенты сопряженности.

Задание матрицы попарных сравнений в общем случае не позволяет воспользоваться для классификации методами дискриминантного анализа. Если же исходные данные задают матрицу $\|x_i^{(j)}\|$, то в отличие от предыдущего параграфа, посвященного дискриминантному анализу с обучением, в общем случае теперь отсутствуют обучающие выборки и априорная информация о законах распределения совокупности всех наблюдений и отдельных ее классов. Общая *постановка задачи классификации объектов* в настоящем параграфе такая же, как и в § 1: необходимо все анализируемое множество n объектов $O_1, O_2, \ldots, O_n$ разбить на некоторое число (заранее известное или нет) однородных (в определенном смысле) классов или групп. Эту задачу удобно интерпретировать геометрически. Если исходные данные представлены матрицей $\|x_i^{(j)}\|$, то каждый из объектов $O_1, O_2, \ldots, O_n$ представляет собой точку в m-мерном признаковом пространстве переменных $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$. Если же исходные данные представлены в форме матрицы парных сравнений, то нам неизвестны координаты таких точек, но зато даны попарные расстояния (близости) между объектами. Естественно считать, что геометрическая близость двух или несколько точек в этом пространстве означает «физическую» близость соответствующих объектов, их однородность. Тогда задача классификации геометрически означает разбиение анализируемой совокупности точек-наблюдений на некоторое число (заранее известное или нет) классов таким образом, чтобы точки одного класса находились бы на небольших расстояниях друг от друга. Это легко изобразить геометрически при $m = 2$.

Полученные в результате разбиения классы называют *кластерами* (*таксонами, образами*), а методы их нахождения называют *кластерным анализом* (*численной таксономией*, *распознаванием образов*).

Заметим, что множество исходных наблюдений может не расслаиваться на кластеры, например, это множество образует один общий кластер. Кластерный анализ можно применить, например, в том случае, когда необходимо разбить группу сотрудников учреждения на два класса, исходя из их показателей по трем признакам: профессионализм, трудолюбие, общительность.

Наиболее трудным в кластерном анализе является определение понятия однородности объектов. В общем случае понятие однородности объектов O_i и O_j определяется величиной d_{ij}, являющейся либо расстоянием $d(O_i, O_j)$ между этими объектами, либо степенью близости (сходства) этих объектов. Выбор метрики или меры близости решающим образом влияет на разбиение объектов на классы. В каждом конкретном случае этот выбор должен производиться по-своему, в зависимости от цели исследования, природы наблюдений и априорных сведений о распределении наблюдений. Рассмотрим наиболее часто используемые расстояния и меры близости между объектами и между кластерами.

2. Расстояния между объектами

В случаях, когда все наблюдаемые признаки $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ однородны по своему физическому смыслу и одинаково важны в вопросах классификации или когда они независимы и нормально распределены с одной и той же дисперсией, используется обычное евклидово расстояние между объектами $O_i(x_i^{(1)}, \ldots, x_i^{(m)})$ и $O_j(x_j^{(1)}, \ldots, x_j^{(m)})$, определяемое формулой

$$d_{\mathrm{E}}(O_i, O_j) = \sqrt{\sum_{k=1}^{m} [x_i^{(k)} - x_j^{(k)}]^2}.$$

В тех случаях, когда каждому признаку $X^{(k)}$ приписан некоторый вес ω_k: $0 \leqslant \omega_k \leqslant 1$, $k = 1, 2, \ldots, m$, пропорциональный степени важности компоненты с точки зрения классификации, используется взвешенное евклидово расстояние

$$d_{\mathrm{BE}}(O_i, O_j) = \sqrt{\sum_{k=1}^{m} \omega_k [x_i^{(k)} - x_j^{(k)}]^2}.$$

Если же объекты задаются дихотомическими признаками, то в качестве меры различия объектов используется хеммингово расстояние, или метрика города между объектами O_i и O_j:

$$d_{\text{Н}}(O_i, O_j) = \sum_{k=1}^{m} |x_i^{(k)} - x_j^{(k)}|.$$

Хеммингово расстояние равно числу несовпадений значений признаков в рассматриваемых i-м и j-м объектах.

3. Меры близости кластеров

В различных кластер-процедуах используются различные расстояния между кластерами и меры близости двух кластеров.

Пусть S_i — i-й кластер (класс, группа), содержащий n_i объектов, $\overline{x_i}$ — среднее арифметическое векторных наблюдений, входящих в S_i, т. е. «центр тяжести» S_i, а $d(S_k, S_e)$ — расстояние между кластерами S_k и S_e. Наиболее употребительными расстояниями и мерами близости между кластерами являются следующие.

1) Расстояние, измеряемое по принципу «ближайшего соседа»:

$$d(S_k, S_l) = \min_{O_i \in S_k,\, O_j \in S_l} d(O_i, O_j).$$

2) Расстояние, измеряемое по принципу «дальнего соседа»:

$$d(S_k, S_l) = \max_{O_i \in S_k,\, O_j \in S_l} d(O_i, O_j).$$

3) Расстояние, измеряемое по «центрам тяжести» кластеров:

$$d(S_k, S_l) = d(\overline{x}_k, \overline{x}_l).$$

4) Расстояние, измеряемое по принципу «средней связи». Оно определяется как среднее арифметическое всех попарных расстояний между представителями рассматриваемых кластеров:

$$d_{cp}(S_k, S_l) = \frac{1}{n_k \cdot n_l} \sum_{O_i \in S_k} \sum_{O_j \in S_l} d(O_i, O_j).$$

4. Оценка качества разбиения

Чтобы сравнить качества различных способов разбиения заданного множества объектов на классы, в задачах кластерного анализа вводится так называемый функционал качества разбиения $\theta(S)$,

определенный на множестве всех возможных разбиений. Под наилучшим разбиением S^* понимается такое разбиение, при котором достигается экстремум выбранного функционала качества $\theta(S)$.

Выбор функционала качества опирается чаще всего на профессиональные и интуитивные соображения исследователя. Приведем наиболее распространенные функционалы качества в том случае, когда уже выбрана метрика d в m-мерном пространстве признаков и задано $S = (S_1, S_2, \ldots, S_k)$ — некоторое разбиение наблюдений $O_1, O_2, \ldots, O_n$ на известное число k классов $S_1, S_2, \ldots, S_k$.

В этом случае за функционалы качества берутся:

1) сумма внутриклассовых дисперсий

$$Q_1(S) = \sum_{l=1}^{k} \sum_{O_i \in S_l} d^2(O_i, \overline{O}_l),$$

2) сумма попарных внутриклассовых расстояний между элементами

$$Q_2(S) = \sum_{l=1}^{k} \sum_{O_i, O_j \in S_l} d^2(O_i, O_j).$$

3) обобщенная внутриклассовая дисперсия

$$Q_3(S) = \det\left(\sum_{l=1}^{k} n_l \cdot \widehat{\Sigma}_l \right),$$

где $\widehat{\Sigma}_3$ — выборочная ковариационная матрица класса S_l, а n_l — объем класса S_l.

В том случае, когда число классов заранее не известно, функционал качества разбиения $\theta(S)$ выбирается по-другому. Перечисленные выше способы оценки качества разбиения являются формальными и вспомогательными. Основная роль в оценке качества разбиения принадлежит содержательному анализу результатов классификации, т. е. анализу с точки зрения поставленной задачи.

5. Методы иерархического кластерного анализа

Методы кластер-анализа принципиально различны для малых объемов n наблюдений (n имеет порядок несколько десятков) и для больших объемов n наблюдений (n имеет порядок нескольких сотен и тысяч). Кроме того, выбор метода кластер-анализа существенно зависит от того, задано ли число классов априорно или неизвестно и подлежит определению. В дальнейшем ограничимся рассмотрением

случая малых объемов n наблюдений и опишем самые распространенные методы так называемого *иерархического кластерного анализа*, когда число классов неизвестно, но его определение и не требуется, поскольку требуется лишь построить так называемый иерархический граф-дерево исследуемой совокупности, или дендрограмму.

Методы иерархического кластерного анализа подразделяются на агломеративные (объединяющие) и дивизимные (разделяющие). Агломеративные методы последовательно объединяют отдельные объекты в группы (кластеры), а дивизимные методы расчленяют группы на отдельные объекты. При этом агломеративные методы последовательно объединяют сначала самые близкие объекты, а затем все более отдаленные друг от друга, а дивизимные методы сначала разделяют самые далекие объекты, а затем все более приближенные друг к другу объекты.

Методы иерархического кластер-анализа дают более полный и тонкий анализ структуры множества наблюдений по сравнению с другими методами кластер-анализа. Наглядная интерпретация проведенного анализа в виде дендрограммы является отличительной чертой этих методов. Недостатком этих методов является тот факт, что при заданном априорном числе классов эти методы приводят к разбиению, которое, как правило, не дает экстремум функционалов качества.

Агломеративный метод иерархического кластерного анализа на первом шаге рассматривает каждый из n объектов $O_1, O_2, \ldots, O_n$ как отдельный кластер. Далее на каждом шаге происходит объединение двух самых близких объектов, а затем — двух самых близких кластеров. Работа заканчивается, когда все заданные объекты объединены в один кластер. Агломеративные методы при выбранной формуле расстояния между объектами отличаются друг от друга способом вычисления расстояния между кластерами. Таким образом, различают: агломеративный метод «ближайшего соседа», агломеративный метод «дальнего соседа», агломеративный метод «средней связи», агломеративный метод «центров тяжести».

Кроме перечисленных агломеративных методов встречаются агломеративные методы, использующие понятие порога. Задается монотонно возрастающий набор положительных чисел (порогов) $c_1, c_2, \ldots, c_p$. На первом шаге объединяются объекты, расстояние между которыми не превосходит c_1, на втором шаге объединяются объекты или группы объектов, расстояние между которыми не превосходит c_2 и т. д. Очевидно, что при задании достаточно большого c_p все объекты исходного множества объектов будут объединен-

ными в один класс. Недостатком этого метода является тот факт, что могут образовываться пересекающиеся промежуточные классы, которые могут не расцепиться вплоть до последнего шага.

Дивизимные методы иерархического кластерного анализа по логическому построению прямо противоположны агломеративным методам. В дивизимных методах первоначально все объекты принадлежат одному кластеру (классу). В процессе классификации по определенным правилам постепенно от этого кластера отделяются группы схожих между собой объектов. Таким образом, на каждом шаге количество кластеров возрастает, а мера расстояния между кластерами уменьшается. Дивизимные методы являются менее трудоемкими по сравнению с агломеративными методами, поскольку они не требуют пересчета матрицы расстояний на каждом шаге классификации.

Рассмотрим примеры.

Пример 1. Уровень медицинского обслуживания пяти регионов характеризуется двумя показателями: $X^{(1)}$ — число врачей на 10 тысяч жителей и $X^{(2)}$ — число больничных коек на 10 тысяч жителей. Значения показателей представлены в таблице:

№ региона	1	2	3	4	5
$X^{(1)}$	35	36	32	31	30
$X^{(2)}$	126	128	123	112	115

Необходимо с помощью агломеративного иерархического кластерного анализа при использовании обычной евклидовой метрики провести классификацию этих регионов и построить дендрограмму, если расстояние между кластерами определять по принципу: а) «ближайшего соседа», б) «дальнего соседа».

Решение. Построим сначала матрицу D_1 расстояний между всеми пятью регионами. В обычной евклидовой метрике расстояние, например, между первым и вторым регионами равно

$$d_{12} = \sqrt{(35-36)^2 + (126-128)^2} = \sqrt{5} \approx 2{,}24.$$

Найдя аналогичным образом расстояние для каждой пары регионов, получаем матрицу расстояний

$$D_1 = \begin{pmatrix} 0 & 2{,}24 & 4{,}24 & 14{,}56 & 12{,}08 \\ 2{,}24 & 0 & 6{,}4 & 16{,}76 & 14{,}32 \\ 4{,}24 & 6{,}4 & 0 & 11{,}05 & 8{,}25 \\ 14{,}56 & 16{,}76 & 11{,}05 & 0 & 3{,}16 \\ 12{,}08 & 14{,}32 & 8{,}25 & 3{,}16 & 0 \end{pmatrix}.$$

Из матрицы D_1 видно, что регионы 1 и 2 наиболее близки ($d_{12} = 2{,}24$). Поэтому объединим их в один кластер $S_{(1,2)}$. Получаем четыре кластера $S_{(1,2)}$, S_3, S_4, S_5.

а) Расстояния между кластерами находим по принципу «ближайшего соседа».

Например, расстояние между кластерами $S_{(1,2)}$ и S_3 равно $d((1, 2); 3) = \min(d_{13}, d_{23}) = \min(4{,}24; 6{,}4) = 4{,}24$.

Проводя аналогичные вычисления расстояний $d((1, 2); 4)$ и $d((1, 2); 5)$, получаем матрицу расстояний

$$D_2 = \begin{pmatrix} 0 & 4{,}24 & 14{,}56 & 12,\,08 \\ 4{,}24 & 0 & 11{,}05 & 8{,}25 \\ 14{,}56 & 11{,}05 & 0 & 3{,}16 \\ 12{,}08 & 8{,}25 & 3{,}16 & 0 \end{pmatrix}$$

Объединим регионы 4 и 5, имеющие наименьшее расстояние $d_{45} = 3{,}16$. После объединения имеем 3 кластера $S_{(1,2)}$, S_3 и $S_{(4,5)}$.

Снова строим матрицу расстояний. Для этого требуется вычислить расстояния $d((1, 2); 3)$, $d((1, 2); (4, 5))$ и $d(3; (4, 5))$. Имеем: $d((1, 2); 3) = \min(d_{13}, d_{23}) = 4{,}24$; $d(3; (4, 5)) = \min(d_{34}, d_{35}) = 8{,}25$; $d((1, 2); (4, 5)) = \min(d_{14}, d_{15}, d_{24}, d_{25}) = 12{,}08$.

Получаем матрицу расстояний

$$D_3 = \begin{pmatrix} 0 & 4{,}24 & 12{,}08 \\ 4{,}24 & 0 & 8{,}25 \\ 12{,}08 & 8{,}25 & 0 \end{pmatrix}.$$

Далее объединяем кластеры $S_{(1,2)}$ и S_3, расстояние между которыми, как видно из D_3, минимально и равно 4,24. В результате

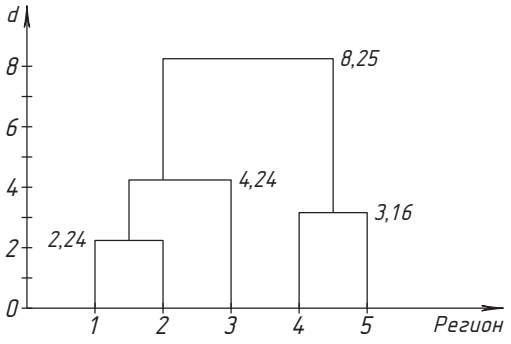


Рис. 13

получаем два кластера $S_{(1,2,3)}$ и $S_{(4,5)}$. Матрица расстояний имеет вид

$$D_4 = \begin{pmatrix} 0 & 8{,}25 \\ 8{,}25 & 0 \end{pmatrix}.$$

Из матрицы D_4 следует, что на расстоянии $d = 8{,}25$ все пять регионов объединяются в один кластер.

Результаты анализа графически представлены в виде дендрограммы, на которой указаны расстояния между объединяемыми на каждом этапе кластерами (см. рис. 13).

Можно показать, что наилучшим является разбиение пяти регионов на два кластера: $S_{(1,2,3)}$ и $S_{(4,5)}$.

б) Теперь проведем классификацию, вычисляя расстояния по принципу «дальнего соседа».

Матрица D_1 остается без изменения, так как опять используется обычная евклидова метрика для вычисления расстояний между регионами.

Сначала объединяем в один кластер регионы 1 и 2, как наиболее близкие ($d_{12} = 2{,}24$). После объединения имеем четыре кластера: $S_{(1,2)}$, S_3, S_4, S_5. Вычислим расстояния между кластерами по принципу «дальнего соседа». Имеем: $d((1, 2); 3) = \max(d_{13}, d_{23}) = 6{,}4$, $d((1, 2); 4) = \max(d_{14}, d_{24}) = 16{,}76$, $d((1, 2); 5) = \max(d_{15}, d_{25}) = 14{,}32$.

Получили новую матрицу попарных расстояний

$$D_2 = \begin{pmatrix} 0 & 6{,}4 & 16{,}76 & 14{,}32 \\ 6{,}4 & 0 & 11{,}05 & 8{,}25 \\ 16{,}76 & 11{,}05 & 0 & 3{,}16 \\ 14{,}32 & 8{,}25 & 3{,}16 & 0 \end{pmatrix}.$$

Объединим регионы 4 и 5 в один кластер как наиболее близкие ($d_{45} = 3{,}16$). После объединения имеем три кластера: $S_{(1,2)}$, S_3 и $S_{(4,5)}$. По принципу «дальнего соседа» строим матрицу расстояний

$$D_3 = \begin{pmatrix} 0 & 6{,}4 & 16{,}76 \\ 6{,}4 & 0 & 11{,}05 \\ 16{,}76 & 11{,}05 & 0 \end{pmatrix}.$$

Объединим кластеры $S_{(1,2)}$ и S_3, расстояние между которыми минимально и равно 6,4. Получаем два кластера: $S_{(1,2,3)}$ и $S_{(4,5)}$. Матрица расстояний имеет вид

$$D_4 = \begin{pmatrix} 0 & 11{,}05 \\ 11{,}05 & 0 \end{pmatrix}.$$

Из матрицы D_4 видно, что на расстоянии $d = 11{,}05$ объединяются все пять регионов в один кластер.

Как и в случае а), наилучшим в случае б) является разбиение пяти регионов на два кластера: $S_{(1,2,3)}$ и $S_{(4,5)}$.

Результаты анализа пункта б) можно представить в виде дендрограммы с указанием расстояний между кластерами по принципу «дальнего соседа». Отличие дендрограммы пункта б) от дендрограммы пункта а) только в том, что расстояние между кластерами $S_{(1,2)}$ и S_3 равно 6,4, и что расстояние между кластерами $S_{(1,2,3)}$ и $S_{(4,5)}$ равно 11,05. ■

В примере 1 разбиение регионов на два кластера получилось одинаковым для случаев а) и б). Однако разбиение объектов на два кластера по принципу «ближайшего соседа» и по принципу «дальнего соседа» может быть и различным. Это же замечание касается и тех случаев, когда расстояние между кластерами измеряется по другим принципам. Все это говорит о том, что для выбора наилучшего варианта иерархического кластерного анализа необходима содержательная априорная информация об исследуемом явлении.

Пример 2. Задана следующая матрица расстояний между пятью объектами:

$$D = \begin{pmatrix} 0 & 4{,}5 & 2{,}2 & 3{,}5 & 3{,}2 \\ 4{,}5 & 0 & 3{,}3 & 1{,}9 & 2 \\ 2{,}2 & 3{,}3 & 0 & 2{,}7 & 2{,}8 \\ 3{,}5 & 1{,}9 & 2{,}7 & 0 & 0{,}7 \\ 3{,}2 & 2 & 2{,}8 & 0{,}7 & 0 \end{pmatrix}.$$

Необходимо с помощью дивизимного иерархического кластерного анализа провести классификацию этих объектов, если расстояние между кластерами определять по принципу «дальнего соседа».

Решение. Наиболее удаленными являются объекты 1 и 2. Оценим расстояния до них от оставшихся объектов:

$$d_{31} < d_{32} \text{ — объект 3 ближе к объекту 1,}$$
$$d_{41} > d_{42} \text{ — объект 4 ближе к объекту 2,}$$
$$d_{51} > d_{52} \text{ — объект 5 ближе к объекту 2.}$$

Получили два кластера: $S_{(1,3)}$ и $S_{(2,4,5)}$. В каждом из них анализируем расстояние между объектами. Тогда происходит разделение того кластера, где достигается максимум расстояния между объектами. Имеем:

$$d_{13} = 2{,}2, \quad d_{24} = 1{,}9, \quad d_{25} = 2, \quad d_{45} = 0{,}7.$$

Наибольшее расстояние $d_{13} = 2{,}2$, значит, объекты 1 и 3 выделяем в отдельные кластеры. В кластере $S_{(2,4,5)}$ максимальное расстояние

$d_{25} = 2$, следовательно, на следующем шаге из этого кластера выделяем объект 2. Наконец, на последнем шаге разделяем кластер $S_{(4,5)}$ на два кластера на расстоянии 0,7. На рис. 14 представлена дендрограмма проведенной классификации. ■

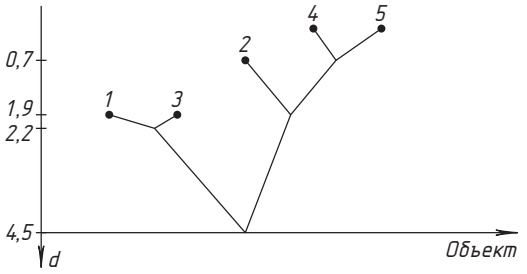


Рис. 14

Заметим еще раз, что в отличие от агломеративных методов дивизимный метод не требует пересчета матрицы расстояний на каждом шаге классификации и тем самым снижает трудоемкость расчетов.

Кластерный анализ можно применять не только для классификации объектов или испытуемых, но и для классификации рассматриваемых психологических признаков.

На практике для устранения влияния масштаба классификационных признаков на результат кластерного анализа рекомендуется данные предварительно нормировать (стандартизировать), чтобы уравнять влияние признаков на расстоянии между объектами. Если признаки измерены в разных шкалах, имеют разный масштаб (т. е. разные средние значения и дисперсии), то на расстояние больше будут влиять признаки, имеющие больший разброс. Если обозначить значение признака $X^{(j)}$, $j = 1, 2, \ldots, n$, для объекта O_i, $i = 1, 2, \ldots, n$, через $X_i^{(j)}$, выборочное среднее значение наблюдений признака $X^{(j)}$ через $\overline{X}^{(j)}$, а выборочное стандартное отклонение наблюдение $X^{(j)}$ для всех объектов — через $\hat{\sigma}^{(j)}$, то исходные данные можно стандартизировать заменой каждого $x_i^{(j)}$ на $z_i^{(j)} = \dfrac{x_i^{(j)} - \overline{x}^{(j)}}{\hat{\sigma}^{(j)}}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

Результатом проведения кластерного анализа является получение групп сходных объектов. Различные методы кластерного анализа позволяют получать кластеры, различающиеся по размеру и по форме. Только серьезная подготовительная работа по подбору признаков, их количества и их взаимосвязи, позволяет подобрать наи-

более подходящий метод классификации. Основная роль по оценке результатов разбиения принадлежит содержательному анализу результатов.

Для иерархического кластерного анализа в статистических пакетах SPSS, STATISTICA и других имеются программы, позволяющие по исходным данным построить таблицу последовательности применения агломеративного метода и дендрограмму.

Задачи к главе 12

1. При анализе уровня медицинского обслуживания населения региона были выделены две группы районов. При этом группа I включает 4 района, а группа II — 5 районов. Для этих групп районов по двум показателям — число врачей на 10 тысяч жителей и число коек на 10 тысяч человек — были получены векторы выборочных средних значений

$$\widehat{a}_1 = \begin{pmatrix} 34 \\ 134 \end{pmatrix}, \quad \widehat{a}_2 = \begin{pmatrix} 23 \\ 35 \end{pmatrix}$$

и матрицы выборочных ковариаций

$$\widehat{\Sigma}_1 = \begin{pmatrix} 1 & 3 \\ 3 & 8 \end{pmatrix}, \quad \widehat{\Sigma}_2 = \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix}.$$

Необходимо провести классификацию района, показатели которого представлены вектором: а) $X = \begin{pmatrix} 31 \\ 123 \end{pmatrix}$, б) $X = \begin{pmatrix} 31 \\ 110 \end{pmatrix}$, в) $X = \begin{pmatrix} 35 \\ 117 \end{pmatrix}$, если считать, что удельный вес районов группы I в регионе достигает 50% и что показатели групп районов I и II были обучающими выборками, извлеченными из нормально распределенных генеральных совокупностей с одинаковой ковариационной матрицей Σ.

2. При анализе уровня подготовки детей подготовительных групп детских садов региона были выделены две группы детсадов. При этом группа I включает 5 детсадов, а группа II — 4 детсада. Для этих групп детсадов по двум показателям — удельный вес (%) детей, умеющих читать, и удельный вес (%) детей, умеющих считать до десяти, были получены векторы выборочных средних значений

$$\widehat{a}_1 = \begin{pmatrix} 68 \\ 34 \end{pmatrix}, \quad \widehat{a}_2 = \begin{pmatrix} 40 \\ 26 \end{pmatrix}$$

и матрицы выборочных ковариаций

$$\widehat{\Sigma}_1 = \begin{pmatrix} 10 & 2 \\ 2 & 1 \end{pmatrix}, \quad \widehat{\Sigma}_2 = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}.$$

Провести классификацию детсада региона, показатели которого представлены вектором а) $X = \begin{pmatrix} 50 \\ 32 \end{pmatrix}$, б) $X = \begin{pmatrix} 52 \\ 68 \end{pmatrix}$, в) $X = \begin{pmatrix} 56 \\ 46 \end{pmatrix}$, если считать, что удельный вес детсадов группы I в регионе достигает 50% и что показатели групп детсадов I и II были обучающими выборками, извлеченными из нормально распределенных генеральных совокупностей с одинаковой ковариационной матрицей Σ.

3. Потребительское поведение пяти семей характеризуется удельными (на душу) расходами (тыс. руб) в летние месяцы на: $X^{(1)}$ — культуру, спорт, отдых и $X^{(2)}$ — питание. Значения этих показателей представлены в следующей таблице:

№ семьи	1	2	3	4	5
$X^{(1)}$	2	4	8	12	13
$X^{(2)}$	10	7	6	11	9

Необходимо с помощью агломеративного кластерного анализа при использовании обычной евклидовой метрики провести классификацию семей и построить дендрограмму, если расстояние между кластерами определять по принципу: а) «ближайшего соседа», б) «дальнего соседа».

4. Работа за четверть для пяти учеников оценивалась по двум показателям: $X^{(1)}$ — число хороших и отличных оценок, $X^{(2)}$ — число пропущенных дней занятий. Значения этих показателей представлены в следующей таблице:

№ ученика	1	2	3	4	5
$X^{(1)}$	4	5	12	14	16
$X^{(2)}$	10	7	6	4	5

Необходимо с помощью агломеративного кластерного анализа при использовании обычной евклидовой метрики провести классификацию учеников и построить дендрограмму, если расстояние между кластерами определять по принципу: а) «ближайшего соседа», б) «дальнего соседа».

МЕТОДЫ СНИЖЕНИЯ РАЗМЕРНОСТИ И ВЫДЕЛЕНИЯ ГЛАВНЫХ ХАРАКТЕРИСТИК

При обработке многомерных наблюдений очень важно уметь выделять признаки объектов, имеющих большое влияние на общее изменение структуры. Это позволяет уменьшить число рассматриваемых признаков объектов, т. е. уменьшить размерность статистической модели, а также избавиться от неинформативных признаков, мало меняющихся при переходе от одного объекта к другому, или от дублирования информации, доставляемой сильно взаимосвязанными признаками. Например, если два признака идеально коррелируют, то можно обойтись одним из них, так как второй не несет в себе никакой дополнительной информации.

Снижение размерности статистической модели часто позволяет наглядно представить исходные данные, упростить вычисления и интерпретацию полученных статистических выводов, а также существенно сжать объем хранимой статистической информации.

К методам снижения размерности относятся метод главных компонент, факторный анализ, многомерное шкалирование.

§ 1. Метод главных компонент

1. Основные понятия

Пусть заданы n объектов O_1, O_2, …, O_n, каждый из которых описывается значениями m количественных признаков $X^{(1)}$, $X^{(2)}$, … …, $X^{(m)}$: j-й объект O_j описывается значениями $x_j^{(1)}$, $x_j^{(2)}$, …, $x_j^{(m)}$, $j = 1, 2, \ldots, n$. Другими словами, задана матрица наблюдений $\|x_j^{(i)}\|$ порядка $n \times m$.

Предполагается, что вектор признаков $X = (X^{(1)}, X^{(2)}, \ldots, X^{(m)})^{\mathrm{T}}$, где T означает операцию транспонирования, является m-мерной случайной величиной, распределенной по нормальному закону, с неизвестными вектором математических ожиданий $a = (a^{(1)}, a^{(2)}, \ldots …, a^{(m)})^{\mathrm{T}}$ и ковариационной матрицей $\Sigma = \|\sigma_{ij}\|$, $i, j = 1, 2, \ldots, m$. На практике в качестве a берется вектор $\widehat{a}$ выборочных средних значений, а в качестве Σ берется матрица $\widehat{\Sigma} = \|\widehat{\sigma}_{ij}\|$, $i, j = 1, 2, \ldots, m$ выборочных ковариаций.

Метод главных компонент определяет и ранжирует m некоторых линейных комбинаций исходных m признаков $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ по их вкладу в суммарную дисперсию этих признаков. Каждая из этих m линейных комбинаций — это «главная компонента».

При этом предполагается, что все главные компоненты $F^{(1)}$, $F^{(2)}$, …, $F^{(m)}$ полностью объясняют суммарную вариацию признаков $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$. В качестве первой главной компоненты $F^{(1)}$ берется такая линейная комбинация исходных признаков, которая обладает наибольшей дисперсией. В качестве второй главной компоненты $F^{(2)}$ берется такая линейная комбинация исходных признаков, которая не коррелирует с $F^{(1)}$ и вносит второй по величине после $F^{(1)}$ вклад в суммарную дисперсию и т. д. Наконец, в качестве m-й главной компоненты $F^{(m)}$ берется такая линейная комбинация исходных признаков, которая не коррелирует со всеми предыдущими главными компонентами $F^{(1)}$, $F^{(2)}$, …, $F^{(m-1)}$ и по сравнению с ними имеет наименьшую дисперсию.

Если все признаки $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ измеряются в одних и тех же единицах (т. е. имеют одинаковую физическую природу), то в дальнейшем будем считать, что все признаки являются центрированными, т. е. $\mathsf{M}X^{(i)} = 0$ для всех $i = 1, 2, \ldots, m$. Этого всегда можно добиться, переходя от $x_j^{(i)}$ к наблюдениям $y_j^{(i)} = x_j^{(i)} - \overline{x}^{(i)}$, где $\overline{x}^{(i)}$ — выборочное среднее значение признака $X^{(i)}$ для n объектов, $i = 1, 2, \ldots, m$.

Если же признаки $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ измеряются в различных единицах (т. е. они различной физической природы), то, чтобы главные компоненты не зависели от выбора масштаба и природы единиц измерения, будем считать в дальнейшем наблюдения $x_j^{(i)}$ безразмерными. Этого всегда можно добиться переходом от $y_j^{(i)}$ к наблюдениям

$$z_j^{(i)} = \frac{y_j^{(i)}}{\sqrt{\widehat{\sigma}_{ii}}}, \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n,$$

где $\widehat{\sigma}_{ii}$ — выборочная дисперсия признака $X^{(i)}$ для n объектов, $i = 1, 2, \ldots, m$. Как известно, для стандартизированных признаков $Z^{(1)}$, $Z^{(2)}$, …, $Z^{(m)}$ всегда $\mathsf{M}Z^{(i)} = 0$ и $\mathsf{D}Z^{(i)} = 1$ для всех $i = 1, \ldots, m$. В этом случае матрица $\widehat{\Sigma}$ выборочных ковариаций является матрицей $\widehat{R}$ выборочных корреляций:

$$\widehat{R} = \begin{pmatrix} 1 & \widehat{r}_{12} & \ldots & \widehat{r}_{1m} \\ \widehat{r}_{21} & 1 & \ldots & \widehat{r}_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ \widehat{r}_{m1} & \widehat{r}_{m2} & \ldots & 1 \end{pmatrix}.$$

Здесь $\hat{r}_{kl}$ — выборочный коэффициент парной корреляции для стандартизованных признаков $X^{(k)}$ и $X^{(l)}$.

В случае двух признаков $X^{(1)}$ и $X^{(2)}$ процесс стандартизации геометрически означает переход от системы координат $X^{(1)}$ и $X^{(2)}$ к системе координат $Z^{(1)}$ и $Z^{(2)}$, начало координат которой находится в центре распределения данных $(\overline{X}^{(1)}, \overline{X}^{(2)})$.

2. Вычисление главных компонент

Вычисление главных компонент проводится следующим образом

По матрице $\|x_j^{(i)}\|$ исходных данных строят матрицу $\hat{\Sigma}$ выборочных ковариаций в случае нестандартизованных данных и матрицу $\hat{R}$ выборочных коэффициентов парной корреляции в случае стандартизованных данных. Затем находят собственные значения и ортонормированную систему собственных векторов матрицы $\hat{\Sigma}$ (соответственно матрицы $\hat{R}$).

Собственные значения λ находятся решением уравнения

$$|\hat{\Sigma} - \lambda E| = 0$$

или

$$|\hat{R} - \lambda E| = 0,$$

где E обозначает единичную матрицу порядка m. Матрицы $\hat{\Sigma}$ и $\hat{R}$ являются симметричными и положительно определенными и, следовательно, эти уравнения имеют m вещественных положительных корней $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_m > 0$. Такая нумерация собственных значений по убыванию диктуется правилом определения главных компонент по убыванию вклада их дисперсий в суммарную дисперсию.

Зная собственные значения λ и решая линейные алгебраические системы уравнений

$$(\hat{\Sigma} - \lambda E)h = 0$$

или

$$(\hat{R} - \lambda E)h = 0$$

находят ортонормированную систему собственных векторов h_1, h_2, …, h_m, соответствующих собственным значениям λ_1, λ_2, …, λ_m. Если X — вектор признаков $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$, то первая главная компонента $F^{(1)} = (h_1, X)$, вторая главная компонента $F^{(2)} = (h_2, X)$ и т. д., m-я главная компонента $F^{(m)} = (h_m, X)$ (здесь круглые скобки обозначают скалярное произведение двух векторов).

Заметим, что матрица H, столбцами которой являются собственные векторы h_1, h_2, …, h_m, является ортогональной, т. е.

$H \cdot H^{\mathrm{T}} = H^{\mathrm{T}} \cdot H = E$ (здесь буква T означает операцию транспонирования). Вектор F главных компонент и вектор X исходных признаков связаны равенством $F = H^{\mathrm{T}} \cdot X$.

3. Основные характеристики главных компонент

Отметим основные числовые характеристики вектора F главных компонент $F^{(1)}$, $F^{(2)}$, ..., $F^{(m)}$. Для вектора F математическое ожидание $\mathsf{M}F = 0$, а ковариационная матрица представляет собой диагональную матрицу, где на главной диагонали стоят собственные значения λ_1, λ_2, ..., λ_m. Кроме того, сумма дисперсий исходных признаков равна сумме дисперсий всех главных компонент, в свою очередь, равной сумме всех собственных значений $\lambda_1 + \lambda_2 + \ldots + \lambda_m$. Анализируя выражение, называемые критерием информативности метода главных компонент,

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_p}{\lambda_1 + \lambda_2 + \ldots + \lambda_m}, \quad p = 1, 2, \ldots, m,$$

показывающее относительную долю дисперсии, вносимой первыми p главными компонентами в сумму всех дисперсий, в зависимости от числа этих компонент, можно разумно определить число p главных компонент, которые целесообразно оставить в рассмотрении и тем самым сократить размерность m исследуемого признакового пространства без особого ущерба. Заметим, что $\lambda_1 + \lambda_2 + \ldots + \lambda_m = m$ в случае выборочной корреляционной матрицы $\widehat{R}$.

После того, как в анализе остается $p \leqslant m$ наиболее значащих главных компонент, субъективно дают для них названия, исходя из так называемой *матрицы «нагрузок»* $A = \|a_i^{(j)}\|$, $i, j = 1, 2, \ldots, m$, главных компонент на исходные признаки. Каждый элемент $a_i^{(j)}$ матрицы A определяет удельный вес влияния пронормированной главной компоненты $F_H^{(j)}$ на исходный признак $X^{(i)}$. Если исходные признаки стандартизованы ($\mathsf{M}X^{(i)} = 0$, $\mathsf{D}X^{(i)} = 1$, $i = 1, 2, \ldots, m$), то $a_i^{(j)}$ одновременно определяет парный коэффициент корреляции между $F_H^{(j)}$ и $X^{(i)}$. Матрица нагрузок A определяется формулой

$$A = H \cdot \Lambda^{\frac{1}{2}},$$

где $\Lambda^{\frac{1}{2}}$ — диагональная матрица, на главной диагонали которой стоят числа $\sqrt{\lambda_1}$, $\sqrt{\lambda_2}$, ..., $\sqrt{\lambda_m}$. Пронормированный вектор F_H главных компонент определяется формулой $F_H = \Lambda^{-\frac{1}{2}} \cdot F$, т. е. $F_H^{(j)} = F^{(j)} \cdot (\lambda_j)^{-\frac{1}{2}}$, $j = 1, 2, \ldots, m$.

Таким образом, $X = AF_H$.

Отметим еще два свойства элементов $a_i^{(j)}$ матрицы нагрузок A.

1) Сумма квадратов элементов любой строки матрицы A равна единице, если данные стандартизованы.

2) Сумма квадратов элементов любого j-го столбца матрицы A равна дисперсии j-й главной компоненты, т. е. вкладу компоненты $F^{(j)}$ в суммарную дисперсию. Квадрат нагрузки фактора $F^{(j)}$ на признак $X^{(i)}$, т. е. число $|a_i^{(j)}|^2$ дает ту часть дисперсии $X^{(i)}$, которая объясняется компонентой $F^{(j)}$.

Эти свойства используются при содержательной интерпретации главных компонент, т. е. для принятия решения о числе значащих главных компонент и определения для них названий.

Пример. Для группы учащихся проводились измерения (в баллах) отклонений от соответствующих средних значений невербального интеллекта $X^{(1)}$ и вербального интеллекта $X^{(2)}$. В результате расчета были получены выборочные дисперсии $\widehat{\sigma}_{11} = 15$, $\widehat{\sigma}_{22} = 25$ и выборочная ковариация $\widehat{\sigma}_{12} = 5$ для $X^{(1)}$ и $X^{(2)}$.

Требуется определить: а) главные компоненты, б) вклад каждой главной компоненты в совокупную дисперсию, в) матрицу нагрузок A.

Решение. Составляем ковариационную матрицу

$$\widehat{\Sigma} = \begin{pmatrix} 15 & 5 \\ 5 & 25 \end{pmatrix}$$

и находим для нее собственные значения λ_1, λ_2, а также ортонормированную систему собственных векторов h_1, h_2.

Собственные значения находим из уравнения

$$\begin{vmatrix} 15-\lambda & 5 \\ 5 & 25-\lambda \end{vmatrix} = \lambda^2 - 40\lambda + 350 = 0.$$

Корнями этого уравнения являются $\lambda = 20 \pm \sqrt{50}$. Положим

$$\lambda_1 = 20 + \sqrt{50} \approx 27{,}1, \quad \lambda_2 = 20 - \sqrt{50} \approx 12{,}9.$$

Для нахождения собственных векторов необходимо решить линейную систему уравнений (x и y — координаты собственного вектора):

$$\begin{cases} (15-\lambda)x + 5y = 0, \\ 5x + (25-\lambda)y = 0, \end{cases}$$

соответственно для $\lambda = \lambda_1$ и $\lambda = \lambda_2$. Нас интересуют какие-либо ненулевые решения этих систем. Например, в качестве собственных векторов можно взять векторы

$$\begin{pmatrix} 1 \\ 1+\sqrt{2} \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1-\sqrt{2} \end{pmatrix}.$$

Чтобы из этих двух собственных векторов получить ортонормированную систему собственных векторов h_1 и h_2, необходимо каждый из векторов разделить на его длину. Получаем, что

$$h_1 = \frac{1}{\sqrt{4+2\sqrt{2}}}\begin{pmatrix} 1 \\ 1+\sqrt{2} \end{pmatrix} \approx \begin{pmatrix} 0{,}4 \\ 0{,}9 \end{pmatrix}, \quad h_2 = \frac{1}{\sqrt{4-2\sqrt{2}}}\begin{pmatrix} 1 \\ 1-\sqrt{2} \end{pmatrix} \approx \begin{pmatrix} 0{,}9 \\ -0{,}4 \end{pmatrix}.$$

Тогда главные компоненты имеют вид:

$$F^{(1)} \approx 0{,}4X^{(1)} + 0{,}9X^{(2)},$$
$$F^{(2)} \approx 0{,}9X^{(1)} - 0{,}4X^{(2)}.$$

Главная компонента $F^{(1)}$ дает

$$\frac{\lambda_1}{\lambda_1+\lambda_2} \approx \frac{27{,}1}{40} \approx 0,\ 68$$

общей дисперсии, а главная компонента $F^{(2)}$ дает 0,32 общей дисперсии. Таким образом, главная компонента $F^{(1)}$ более важная, чем компонента $F^{(2)}$, так как $F^{(1)}$ содержит большую часть информации о специфике интеллекта группы учащихся.

Матрица нагрузок A определяется формулой

$$A = H \cdot \Lambda^{\frac{1}{2}} \approx \begin{pmatrix} 0{,}4 & 0{,}9 \\ 0{,}9 & -0{,}4 \end{pmatrix}\begin{pmatrix} \sqrt{27{,}1} & 0 \\ 0 & \sqrt{12{,}9} \end{pmatrix} \approx$$

$$\approx \begin{pmatrix} 0{,}4 & 0{,}9 \\ 0{,}9 & -0{,}4 \end{pmatrix}\begin{pmatrix} 5{,}2 & 0 \\ 0 & 3{,}6 \end{pmatrix} = \begin{pmatrix} 2{,}08 & 3{,}24 \\ 4{,}68 & -1,\ 44 \end{pmatrix}. \blacksquare$$

4. Геометрическая интерпретация метода главных компонент

Множество n объектов, каждый из которых описывается значениями m признаков, образует множество n точек («облако») в m-мерном пространстве. Метод главных компонент означает, что находятся такие направления (главные компоненты) в этом пространстве, что «облако» данных более вытянуто вдоль первой главной компоненты. Это означает также, что сумма расстояний всех точек до прямой, проходящей через центр «облака» в направлении первой

главной компоненты, является минимальной. Вторая главная компонента обладает тем же свойством, но для проекции «облака» на подпространство, перпендикулярное первой главной компоненте, и т. д. Преобразование исходных m признаков в m главных компонент является линейным ортогональным преобразованием. Главные компоненты — это ортогональные статистически независимые факторы, являющиеся линейными функциями исходных признаков.

Отметим в заключение, что полученные главные компоненты являются выборочными, поскольку были найдены для заданной случайной выборки наблюдений. При дополнительных условиях они могут выступать оценками главных компонент для всей исследуемой генеральной совокупности признаков. Кроме того, существуют методы построения интервальных оценок для неизвестных характеристик главных компонент и статистические критерии проверки гипотез относительно таких характеристик.

Метод главных компонент часто используется не только самостоятельно, но и в комбинации с методами регрессионного анализа, классификации и многомерного шкалирования. При построении регрессионной модели сначала учитываются все главные компоненты, а затем их число уменьшается за счет исключения незначимых главных компонент (для них собственные значения малы). В методах классификации (дискриминантный анализ, кластерный анализ) группировка проводится по одной или нескольким главным компонентам.

§ 2. Факторный анализ

Методы факторного анализа возникли в начале двадцатого века при решении некоторых задач психологии. Определяющими для формирования факторного анализа были работы британского психолога Ч. Спирмена. В настоящее время набор методов факторного анализа достаточно велик и насчитывает десятки различных приемов обработки данных. Теория факторного анализа продолжает развиваться.

Факторный анализ широко используется в психодиагностике, дифференциальной психологии, психогенетике. С помощью факторного анализа измеряют индивидуальные особенности, разрабатывают тесты для изучения связи различных психологических характеристик, проводят стандартизацию тестовых методик и т. д.

Следует отметить, что факторный анализ используется для обработки больших массивов экспериментальных данных (их больше сотни) и поэтому процедура подобной обработки очень трудоемка. По этой причине факторный анализ, как правило, проводится на компьютерах с помощью программ по факторному анализу, имеющихся в стандартных статистических пакетах.

Как правило, описание объектов начинается с заведомо избыточного числа признаков (переменных) $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$, так как некоторые признаки могут дублировать друг друга, т. е. они коррелируют с большим коэффициентом корреляции. В таких случаях сразу возникают вопросы о полноте выбора признаков, их независимости и их минимальности для описания объекта. Если некоторые признаки коррелируют, то это значит, что на них влияет некоторый латентный (скрытый) фактор, неизвестный нам, с помощью которого можно объяснить наблюдаемое сходство данных.

Будем считать в дальнейшем, что наблюдаемые признаки $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ являются количественными (т. е. измеренными в интервальной шкале или в шкале отношений) и центрированными (т. е. $\overline{x}^{(i)} = 0$, $i = 1, 2, \ldots, m$). Этого всегда можно добиться переходом к наблюдениям $y_j^{(i)} = x_j^{(i)} - \overline{x}^{(i)}$, где $\overline{x}^{(i)}$ — среднее значение всех наблюдений $x_j^{(i)}$ i-го признака. Если же признаки $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ имеют различную физическую природу, то будем считать их стандартизованными. Этого всегда можно достичь переходом к наблюдениям $z_j^{(i)} = \dfrac{y_j^{(i)}}{\sigma_i}$, где σ_i — стандартное отклонение признака $x^{(i)}$. Тогда $z_j^{(i)}$ будут иметь нулевые средние значения и единичную дисперсию.

Дадим описание классической линейной модели факторного анализа.

Линейная модель факторного анализа базируется на следующих трех гипотезах.

1) Каждый наблюдаемый признак $X^{(i)}$, $i = 1, 2, \ldots, m$, представляет собой сумму некоторой линейной комбинации общих для всех признаков латентных, некоррелированных факторов $F^{(1)}$, $F^{(2)}$, …, $F^{(r)}$, где $1 \leqslant r \leqslant m$, и некоторой остаточной случайной компоненты $U^{(i)}$, характерной лишь для признака $x^{(i)}$. Остаточная компонента $U^{(i)}$ определяет ту часть признака $x^{(i)}$, которая не может быть объяснена общими факторами $F^{(1)}$, $F^{(2)}$, …, $F^{(r)}$, и включает в себя ошибки измерения $x^{(i)}$.

Высказанное предположение записывается формулой

$$X^{(i)} = \sum_{j=1}^{r} a_j^{(i)} F^{(j)} + U^{(i)}, \quad i = 1, 2, \ldots, m.$$

Здесь $A = \|a_j^{(i)}\|$ — числовая матрица размера $m \times r$, называемая *матрицей нагрузок общих факторов на исследуемые признаки*. Число $a_j^{(i)}$ — это нагрузка j-го общего фактора $F^{(j)}$ на признак $x^{(i)}$, характеризующая вклад $F^{(j)}$ в признак $x^{(i)}$.

2) Вектор F общих факторов $F^{(1)}$, $F^{(2)}$, …, $F^{(r)}$ является r-мерной случайной величиной, нормально распределенной со средним значением $\mathsf{M}F = 0$ и с ковариационной матрицей специального вида.

3) Вектор U остаточных случайных компонент U_1, U_2, …, U_m состоит из взаимонезависимых компонент, не зависит от вектора F общих факторов и подчиняется m-мерному нормальному распределению $N(0, V)$, т. е. его ковариационная матрица V — диагональная, где по диагонали стоят числа $v_{ii}^2 = \mathsf{D}U^{(i)}$ — дисперсии $U^{(i)}$, а $\mathsf{M}U = 0$.

Из сделанных предположений 1)—3) следует, что вектор x наблюдений x_1, x_2, …, x_n исходных n объектов должен быть нормально распределенной m-мерной случайной величиной $N(0, \Sigma)$, т. е. $\mathsf{M}X = 0$, а ковариационная матрица для X имеет вид $\Sigma = A \cdot A^{\mathrm{T}} + V$.

Целью факторного анализа является нахождение и интерпретация латентных общих факторов, минимизация их числа и степени зависимости признаков $X^{(i)}$ от характерных для них остаточных случайных компонент $U^{(i)}$. Разумеется, эта цель достигается лишь приближенно.

Факторный анализ часто используется в психологии. Например, при тестировании общего интеллекта испытуемых естественно предположить, что в качестве латентных общих факторов, от которых зависят оценки испытуемых по всем тестам, выступают такие факторы, как характеристика общей одаренности $F^{(1)}$, характеристика математических способностей $F^{(2)}$, характеристика технических способностей $F^{(3)}$ и характеристика гуманитарных способностей $F^{(4)}$. Факторный анализ используется при конструировании тестов. Он показывает: сколько отдельных шкал входит в состав теста, какие задания входят в тот или иной тест, какие задания должны быть выброшены из теста, если, например, они измеряют что-то отличное от других заданий или сильно подвержены ошибкам измерений. Факторный анализ чистит тесты от незначительных различий в шкалах и выделяет основные факторы. Возможности факторного анализа не ограничиваются анализом

заданий или оценок теста. Например, группу учащихся без специальной подготовки можно оценивать с точки зрения их успешности по различным учебным предметам, по различным видам спорта и т. п. Факторный анализ выявит основные учебные или спортивные способности. Можно, например, с помощью факторного анализа определять независимые показатели времени реакции, взятые из когнитивных тестов.

Отметим, что первая гипотеза линейной факторной модели внешне похожа на описание модели линейной множественной регрессии, в которой под вектором F понимается вектор объясняющих переменных (аргументов). Однако принципиальное отличие этих двух моделей состоит в том, что значения вектора F, выступающего в роли аргумента в модели линейной регрессии могут быть измерены при выборочном обследовании объектов, а в модели линейного факторного анализа вектор F не является непосредственно наблюдаемым.

Имеется также отличие методов линейного факторного анализа от метода главных компонент. В методах линейного факторного анализа заранее принимается, что одними общими факторами нельзя объяснить всю дисперсию рассматриваемых признаков $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ и что некоторая часть их дисперсий объясняется неизвестными случайными факторами $U^{(1)}$, $U^{(2)}$, …, $U^{(m)}$. В методе главных компонент заранее принимается, что вся дисперсия признаков объясняется только главными компонентами.

При всем своем многообразии методы факторного анализа имеют общий алгоритм решения.

По исходной матрице X порядка $m \times n$ наблюдений $x_j^{(i)}$ m признаков $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ для n объектов на первом шаге строится ковариационная матрица S из выборочных коэффициентов ковариации между признаками, если признаки физически однородные, или корреляционную матрицу R из выборочных коэффициентов корреляции между признаками, если признаки физически не являются однородными. Матрицы S и R являются симметричными матрицами порядка $m \times m$. В матрице R на главной диагонали стоят единицы.

На втором шаге корреляционная матрица R преобразуется в так называемую «редуцированную» корреляционную матрицу R_h порядка $m \times m$. Матрица R_h получается из матрицы R лишь заменой единиц на главной диагонали числами (их называют общностями) h_1^2, h_2^2, …, h_m^2, где каждое число $h_j^2 < 1$. Эти числа описывают ту часть общей вариации признаков, которая объясняется только об-

щими факторами. Существует несколько способов получения общностей h_j^2. Самый простой из них заключается в том, что на главной диагонали записывается с положительным знаком наибольший по величине коэффициент корреляции. Во всяком случае на втором шаге возникает проблема нахождения общностей.

На третьем шаге строится матрица нагрузок A. Здесь возникают проблемы определения $r \leqslant m$ общих факторов $F^{(1)}$, $F^{(2)}$, ..., $F^{(r)}$. В общем случае выделенные факторы не обязательно ортогональны и тогда столбцы матрицы A будут линейно зависимыми.

Когда пространственное расположение общих факторов нелогично или трудно поддается интерпретации, то необходим четвертый шаг — поворот факторов. Здесь возникает проблема оптимального расположения факторных осей.

На пятом шаге строятся новая матрица нагрузок после поворота осей и матрица значений факторов.

Среди современных методов факторного анализа наиболее популярны метод главных факторов, метод наибольшего правдоподобия и метод наименьших квадратов. Современные методы факторного анализа являются итеративными, часто предполагающими, что на первом шаге приближенное решение уже найдено каким-либо из способов, а последующими шагами это решение лишь улучшается. Различные алгоритмы различаются тем, как происходит итерация вычислений. При использовании различных алгоритмов полученные результаты могут различаться и только содержательный анализ результатов позволяет отобрать в каком-то смысле наилучший из них.

Остановимся на некоторых теоретических вопросах, возникающих в факторном анализе.

Первый вопрос — это вопрос существования линейной модели факторного анализа, т. е. вопрос проверки выполнения гипотез 1)—3). Теория этого вопроса разработана слабо. Можно указать лишь статистические критерии проверки гипотезы H_0 о том, что исследуемый вектор наблюдений X допускает представление с помощью линейной модели факторного анализа с заданным заранее числом r общих факторов.

Вопрос второй — это вопрос единственности линейной модели факторного анализа. Оказывается, что если исходные данные m, r и Σ допускают построение линейной модели факторного анализа, то определение неизвестных параметров этой модели, т. е. общих факторов $F^{(1)}$, $F^{(2)}$, ..., $F^{(r)}$ и матрицы факторных нагрузок A не единственно. Для однозначного определения этих параметров необходимы дополнительные условия.

При выполнении неравенства $r < \frac{(m-1)}{2}$ матрица факторных нагрузок A получается единственной с точностью до ее ортогонального преобразования или с точностью до вращения факторов $F^{(1)}, F^{(2)}, \ldots, F^{(r)}$. Вращение факторов используется для того, чтобы получить хорошую интерпретацию факторов с точки зрения содержательного анализа исходных данных. Расположение факторных осей после вращения считается оптимальным, если факторные оси проходят через наиболее плотные скопления точек, так что большинство точек находятся либо вблизи осей, либо вблизи начала координат.

Вращение общих факторов может быть ортогональным или косоугольным. При ортогональном вращении углы между общими факторами остаются прямыми и, следовательно, факторы остаются некоррелированными.

Косоугольное вращение может проводиться поочередным вращением каждого фактора на определенный угол или одновременным вращением всех факторов. Такое вращение предусматривает корреляционную зависимость латентных факторов.

Среди методов ортогонального вращения наиболее часто используется метод «варимакс» и метод «квартимакс», а среди методов косоугольного вращения наиболее популярны метод «облимин» и метод «квартимин».

Вопрос о выборе угла вращения общих факторов и вопрос достаточности числа поворотов пространства факторов решаются субъективно с помощью построения графиков расположения наблюдаемых объектов в пространстве повернутых факторов или с использованием специальных критериев для оценки структуры общих факторов.

Еще раз отметим, что решение вопроса о минимальном количестве общих факторов и сравнение различных типов вращения должны базироваться на содержательном анализе полученных результатов. При таком анализе необходимо проверить, согласуются ли полученные данные с вашими ожиданиями, с результатами ранее выполненных исследований, с теоретическими положениями в данной области психологии.

При проведении косоугольного вращения факторного пространства, как было отмечено ранее, получаемые общие факторы обычно коррелируют между собой. Тогда матрица взаимных корреляций факторов может быть подвергнута линейному факторному анализу и можно выделить сгустки факторов, т. е. провести факторный

анализ второго порядка, если считать факторный анализ для исходных признаков анализом первого порядка. Далее можно провести факторный анализ для факторов второго порядка и выполнить факторный анализ третьего порядка. Этот процесс можно продолжать либо до получения нулевых корреляций, либо до получения одного фактора. Такого рода линейный факторный анализ называется иерархическим.

Факторный анализ может выступать как разведочный (эксплораторный) и как проверочный (конфирматорный).

В первом случае основная задача факторного анализа в том, чтобы структурировать связи между переменными и сформулировать рабочие гипотезы о причинах этих связей. В отличие от разведочного факторного анализа, который используется на начальном этапе исследования данных, конфирматорный факторный анализ используется на более поздних этапах исследования, когда необходимо сделать выбор между несколькими конкурирующими гипотезами, описывающими структуру данных. Этот выбор делается с помощью проверки соответствия каждой из гипотез исходным эмпирическим данным.

§ 3. Многомерное шкалирование

В многомерном шкалировании (МШ) рассматривается следующая задача. Заданы характеристики D_{ij} попарных сравнений объектов (или испытуемых) O_i и O_j, $i, j = 1, 2, \ldots, n$, для n объектов $O_1, O_2, \ldots, O_n$. Другими словами, задана симметричная матрица $D = \|D_{ij}\|$ порядка $n \times n$. Числа D_{ij} могут выражать меру сходства или различия объектов O_i и O_j, меру их связи, расстояние между объектами и т. д. Если в качестве объектов выступают признаки, то D_{ij} могут обозначать коэффициенты корреляции или ковариации i-го и j-го признаков. Кроме того, значения D_{ij} могут быть измерены как в интервальной шкале и в шкале отношений, так и в шкале порядка.

Таким образом, здесь исходные данные аналогичны исходным данным кластерного анализа. Однако ни координаты $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)})$ объектов O_i, $i = 1, 2, \ldots, n$, ни даже размерность m признакового пространства $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$ нам не известны.

Задача многомерного шкалирования состоит в том, чтобы построить некоторое геометрическое пространство, чаще всего евклидово пространство, установить его минимальную размерность m и приписать каждому объекту O_i координаты $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)})$ таким образом, чтобы расстояния d_{ij} между i-й и j-й точками этого пространства были в некотором смысле близки к мере сходства или различия D_{ij} для объектов O_i и O_j.

Таким образом, *цель методов МШ* состоит в том, чтобы восстановить признаковое пространство $X^{(1)}$, $X^{(2)}$, …, $X^{(m)}$ и чтобы из информации, заданной матрицей D попарных сравнений n объектов, получить информацию о геометрической конфигурации n точек в соответствующем геометрическом многомерном пространстве, сопоставимых n объектам.

В этом отличие методов МШ от методов главных компонент и факторного анализа, для которых такое исходное геометрическое многомерное пространство задано.

Традиционно в МШ применяется своя терминология. Исходная заданная матрица D называется *матрицей различий или сходств стимулов*. Под *стимулом* понимается некоторый объект или набор свойств объекта, который непосредственно нельзя измерить. *Стимульное пространство* (или *психологическое пространство*) — это геометрическое пространство, точки которого представляют исходные стимулы. Каждая ось координат называется *шкалой*, а *шкальные значения стимула* — это координаты соответствующей точки стимульного пространства.

Как уже говорилось выше, предполагается, что субъективные оценки различий или сходств объектов определенным образом соотносимы с расстояниями между соответствующими точками стимульного пространства. Расстояние d_{ij} между точками может определяться по-разному, однако ясно, что два стимула, сильно различающиеся между собой, будут расположены на далеком расстоянии друг от друга в пространстве, а сходные стимулы расположатся рядом. Для отображения сложного реального мира явлений необходимо использовать многомерное пространство шкал — многомерное шкалирование.

В зависимости от характера исходных данных все методы МШ подразделяются на метрические и неметрические. Возможна также и другая классификация методов МШ в зависимости от того, что выступает объектом исследования, например, анализ стимулов изучает и строит образы объекта, а анализ индивидуальных различий изучает особенности субъективного восприятия стимулов и т. д.

1. Метрическое МШ

Классическим методом метрического МШ является *метод У. Торгерсона*. Метод У. Торгенсона базируется на следующих гипотезах.

1) Мера различия D_{ij} между объектами O_i и O_j равна евклидовому расстоянию d_{ij} между точками в некотором определенном психологическом пространстве со шкалами $X^{(1)}$, $X^{(2)}$, ..., $X^{(m)}$. Таким образом, $D_{ij} = d_{ij}$ для всех значений i, $j = 1, 2, \ldots, n$.

Для выполнения этой гипотезы необходимо, чтобы D_{ij} были измерены в шкале отношений.

2) В искомом психологическом пространстве нуль — начало координат.

Напомним, что евклидово расстояние d_{ij} между точками $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)})$ и $x_i = (x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(m)})$ задается формулой

$$d_{ij} = \sqrt{\sum_{k=1}^{m} [x_i^{(k)} - x_j^{(k)}]^2}.$$

Опишем процедуру метода У. Торгерсона.

Сначала по исходной матрице $D = \|D_{ij}\|$ строится матрица $B = \|b_{ij}\|$, i, $j = 1, 2, \ldots, n$, элементы которой вычисляются по следующей формуле:

$$b_{ij} = \frac{1}{2}\left(-D_{ij}^2 + \frac{1}{n}\sum_{i=1}^{n} D_{ij}^2 + \frac{1}{n}\sum_{j=1}^{n} D_{ij}^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}^2\right),$$

$$i, j = 1, 2, \ldots, n.$$

Затем находятся собственные значения и собственные векторы матрицы B. При этом могут использоваться метод главных компонент или методы факторного анализа.

Отметим следующие свойства матрицы B.

1) Матрица B неотрицательно определена, т. е. все ее собственные значения являются неотрицательными.

2) Ранг матрицы B равен размерности m искомого пространства стимулов $X^{(1)}$, $X^{(2)}$, ..., $X^{(m)}$.

3) Если положительные собственные значения $\lambda_1, \lambda_2, \ldots, \lambda_m$ матрицы B упорядочены по убыванию и если h_r — собственный вектор для собственного значения λ_r матрицы B, то вектор $F^{(r)} = \sqrt{\lambda_r} \cdot h_r$ определяет r-ю ось координат, т. е. r-ю шкалу пространства стимулов $X^{(1)}$, $X^{(2)}$, ..., $X^{(m)}$.

Из приведенных свойств получаем координатное представление искомых стимулов в пространстве со шкалами $F^{(1)}$, $F^{(2)}$, ..., $F^{(m)}$. Это и будет решением задачи метрического МШ. Заметим, что решение этой задачи возможно только с точностью до ортогонального преобразования, так как при ортогональном преобразовании системы координат (т. е. при переносе и повороте осей координат) расстояния между точками евклидова пространства сохраняются. Матрица X координат стимулов для исходных объектов определяется равенством $B = XX^{\mathrm{T}}$, где T означает операцию транспонирования.

На заключительном этапе решаются вопросы снижения размерности пространства стимулов и интерпретируемости результатов. Чаще всего число шкал $r < m$ выбирается таким, чтобы сохранялась не менее 80—90% информативности стимулов. Интерпретируемость результатов МШ определяется компонентным составом векторов, представляющих шкалы. Если значения на шкалах не поддаются логическому объяснению, то делается поворот шкального пространства. Заметим, что по сравнению с другими ортогональными преобразованиями преобразование к главным компонентам наименьшим образом искажает структуру исходных данных.

Метод У. Торгерсона является линейным методом метрического МШ, поскольку в этом методе ищется ортогональное преобразование, приводящее матрицу B к диагональному виду. Метод Торгерсона базируется на жесткой гипотезе, что меры различия D_{ij} совпадают с евклидовыми расстояниями d_{ij} между соответствующими точками стимульного пространства. Добиться выполнения гипотез Торгерсона достаточно сложно. Поэтому имеются модели МШ с менее жесткими требованиями, чтобы $D_{ij} = d_{ij} + \varepsilon$, где ε — некоторая пороговая константа.

Дальнейшее развитие методы метрического МШ получили после разработанных Р. Шепардом и Д. Краскалом нелинейных вариантов методов метрического МШ. Во-первых, вместо евклидовых пространств и расстояний вида d_{ij} в них используются другие метрические пространства и расстояния, а, во-вторых, вместо жесткого требования $D_{ij} = d_{ij}$ используется требование минимальности различных функционалов качества соответствия между мерами различий D_{ij} и расстояниями d_{ij}. Для использования в качестве стимульного пространства некоторого метрического пространства необходимо требовать, чтобы меры различия D_{ij}, $i, j = 1, 2, \ldots, n$, удовлетворяли всем аксиомам этого метрического пространства. Выбор метода метрического МШ диктуется содержательным анализом поставленной задачи.

2. Неметрическое МШ

Методы неметрического МШ применяются для обработки данных, измеренных в порядковой (ранговой) шкале.

Классическим методом неметрического МШ является *метод Шепарда—Краскала*. Он базируется на следующей гипотезе: меры различия D_{ij} между объектами O_i и O_j, измеренные в порядковой шкале, должны иметь такую же монотонную связь, как и евклидовы расстояния d_{ij} между соответствующими точками в некотором определяемом шкальном пространстве. Это значит, что если $D_{ij} < D_{kl}$, то и соответственно $d_{ij} < d_{kl}$.

В общем случае должно быть условие: $D_{ij} = f(d_{ij})$, где f — заранее не известная монотонная функция, которая подбирается методом проб. Например, если f — линейная функция, то $D_{ij} = a_0 + a_1 d_{ij}$. В качестве функции f может выступать и другая элементарная функция, а в качестве d_{ij} может выступать не только евклидова метрика, но и другие метрики, например, метрика города.

Цель неметрического МШ такая же как и цель метрического МШ: построить психологическое пространство минимальной размерности и вычислить координаты каждого стимула в этом пространстве.

Опишем процедуру неметрического метода Шепарда—Краскала.

На первом шаге получают матрицу D различий D_{ij}, содержащие ранговые данные — характеристики непохожести объектов $O^{(1)}$, $O^{(2)}$, …, $O^{(m)}$. На втором шаге берется полученная каким-нибудь методом, например, методом главных компонент, какая-либо стартовая конфигурация n точек в евклидовом пространстве некоторой размерности r. В дальнейшем для получения минимальной размерности пространства проводятся расчеты для нескольких значений r. Если последовательность мер различий D_{ij} монотонно возрастает и соответствующая ей последовательность евклидовых расстояний d_{ij} для выбранной конфигурации точек в r-мерном пространстве тоже монотонно возрастает, то задача решена. Если же условие монотонности нарушается, то строится последовательность чисел $\widehat{d}_{ij}$, возможно более близких к числам d_{ij} и удовлетворяющим условию монотонности. Чтобы построить соответствующую $\widehat{d}_{ij}$ конфигурацию точек в r-мерном пространстве, меняют координаты точек так, чтобы приблизить расстояния d_{ij} к числам $\widehat{d}_{ij}$. Это делается с помощью минимизации различных функционалов качества соответствия d_{ij} и $\widehat{d}_{ij}$. Эти функционалы еще называются *мерами Краскала* или «*стрессом*».

Таким образом, получается новая конфигурация точек в r-мерном пространстве. Описанная итерационная (т. е. последовательная) процедура продолжается до тех пор, пока значение «стресса» не будет достаточно малым, т. е. решение уже не может быть существенно улучшено. Тогда приступают к интерпретации итогов анализа.

Приведем оценку метода Шепарда—Краскала. Если исходная матрица D является матрицей точных расстояний D_{ij} между точками-стимулами в евклидовом пространстве, то метод Шепарда—Краскала с высокой степенью точности восстанавливает не только порядковые, но и количественные отношения между D_{ij}. Если же исходные различия D_{ij} не удовлетворяют каким-либо аксиомам евклидова пространства, то метод Шепарда—Краскала в принципе не может восстановить количественную структуру данных. Такая же картина может наблюдаться и в том случае, когда D_{ij} удовлетворяют всем аксиомам евклидова пространства, но допущены грубые ошибки (выбросы) в исходных данных D_{ij}.

Замечание. Расширениями методов метрического и неметрического МШ является модель индивидуальных различий и модель предпочтений.

Модель индивидуальных различий используется в тех случаях, когда необходимо найти шкалы и представление в координатном пространстве не только стимулов, но и субъектов (экспертов) их оценивающих. При этом координатами субъектов служат значения весовых коэффициентов ω_{ks}, характеризующие уровень значимости шкалы k для субъекта s. Наиболее важной моделью индивидуальных различий является взвешенная евклидова модель, использующая взвешенную евклидову метрику.

В этой модели наряду с матрицей координат стимулов интерпретируются и матрицы индивидуальных различий.

Задачи к главе 13

1. Для группы школьников проводились измерения (в баллах) отклонений от соответствующих средних значений показателя математических способностей $X^{(1)}$ и показателя технических способностей $X^{(2)}$. В результате расчета были получены выборочные дисперсии $\hat{\sigma}_{11} = 5$, $\hat{\sigma}_{22} = 5$ и выборочная ковариация $\hat{\sigma}_{12} = 3$ для $X^{(1)}$ и $X^{(2)}$. Необходимо определить: а) главные компоненты, б) вклад каждой главной компоненты в общую дисперсию, в) матрицу нагрузок A.

2. Группа нескольких детсадов была обследована по двум показателям: $X^{(1)}$ — число детей детсада, умеющих читать, и $X^{(2)}$ — число детей детсада, умеющих считать до десяти. В результате обследова-

ния была получена выборочная ковариационная матрица $\widehat{\Sigma}$:

$$\widehat{\Sigma} = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}.$$

Требуется найти: а) главные компоненты, б) вклад каждой главной компоненты в общую дисперсию, в) матрицу нагрузок A.

3. Группа подростков нескольких школ была обследована по трем показателям (в баллах): уровню тревожности $X^{(1)}$, уровню агрессивности $X^{(2)}$, уровню терпеливости $X^{(3)}$. В результате обследования была получена выборочная ковариационная матрица $\widehat{\Sigma}$:

$$\widehat{\Sigma} = \begin{pmatrix} 2 & 2 & -2 \\ 2 & 5 & -4 \\ -2 & -4 & 5 \end{pmatrix}.$$

Необходимо найти: а) главные компоненты, б) вклад каждой главной компоненты в общую дисперсию, в) матрицу нагрузок A.

4. Группа студентов обследовалась с точки зрения отклонений (в см) от соответствующих средних значений роста студента $X^{(1)}$, роста его отца $X^{(2)}$, роста его матери $X^{(3)}$. В результате расчета была получена оценка ковариационной матрицы $\widehat{\Sigma}$:

$$\widehat{\Sigma} = \begin{pmatrix} 2 & -1 & 2 \\ -1 & 3 & 1{,}5 \\ 2 & 1{,}5 & 4 \end{pmatrix}.$$

Найти: а) главные компоненты, б) вклад каждой главной компоненты в общую дисперсию, в) матрицу нагрузок A.

ЗАКЛЮЧЕНИЕ

Вы познакомились с основными математическими методами, которые используются в современных психолого-педагогических исследованиях. Их применение на практике позволяет профессиональному психологу обосновать свою точку зрения и проверить эффективность того или иного практического метода. Однако изложенные в настоящем пособии математические методы не исчерпывают всего многообразия математических методов, используемых в современной психологии.

Процесс интенсивного применения математических методов для изучения психических процессов и функций привел в 60-х годах XX века к выделению и оформлению новой ветви психологии — математической психологии. Предметом математической психологии является разработка математического аппарата, пригодного для адекватного описания и моделирования психических процессов и явлений. Основная цель математической психологии — логическое развитие психологических теорий от описательных к гипотетико-дедуктивным и далее аксиоматизированным содержательным. Кроме разработки математических моделей явлений в психической области, математическая психология развивает и уточняет традиционные статистические методы и процедуры обработки экспериментальных данных. Разумеется, далеко не все направления и проблемы в психологии поддаются процессу математизации. Например, не поддается математизации психология семьи.

На начальном этапе своего развития математическая психология использовала готовый математический аппарат для моделирования психических процессов. Так, например, для моделирования процессов обучаемости был применен аппарат марковских случайных процессов, для моделирования деятельности человека, управляющего динамической системой, применялась теория автоматического регулирования, для описания коллективного поведения людей применялась теория игр. Настоящий этап развития математической психологии характерен не только использованием сравнительно новых математических теорий (теории дифференциальных уравнений, теории катастроф, теории детерминированного хаоса, многомерной геометрии, теории нечетких множеств и т. д.), но и разработкой

специализированного математического аппарата для исследования и моделирования психических процессов и функций. Кроме того, в настоящее время для построения имитационных моделей психических процессов и явлений широко используются персональные компьютеры и создаются специализированные компьютерные программы.

Из всего сказанного следует, что роль математических методов в психологии будет постоянно возрастать. Поэтому каждому психологу необходимо если не использовать эти методы, то, по крайней мере, ориентироваться в них.

В Московском психолого-педагогическом университете (МГППУ) постоянно действует общемосковский научный семинар по математической психологии, на котором рассматриваются доклады о новых применениях математических методов в психологии. Кроме того, в МГППУ функционирует учебная лаборатория математических моделей в психологии и педагогике, в которой любой сотрудник или студент может получить консультацию по применению математических методов.

ТАБЛИЦЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Таблица I

Значения функции $P_k(\lambda) = \dfrac{\lambda^k}{k!} e^{-\lambda}$

k \ λ	0,1	0,2	0,3	0,4	0,5
0	0,90484	0,81873	0,74082	0,67032	0,60653
1	0,09048	0,16375	0,22225	0,26813	0,30327
2	0,00452	0,01638	0,03334	0,05363	0,07582
3	0,00015	0,00109	0,00333	0,00715	0,01264
4		0,00006	0,00025	0,00072	0,00158
5			0,00002	0,00006	0,00016
6					0,00001

k \ λ	0,6	0,7	0,8	0,9
0	0,54881	0,49659	0,44933	0,40657
1	0,32929	0,34761	0,35946	0,36591
2	0,09879	0,12166	0,14379	0,16466
3	0,01976	0,02839	0,03834	0,04940
4	0,00296	0,00497	0,00767	0,01112
5	0,00036	0,00070	0,00123	0,00200
6	0,00004	0,00008	0,00016	0,00030
7		0,00001	0,00002	0,00004

k \ λ	1,0	2,0	3,0	4,0	5,0
0	0,36788	0,13534	0,04979	0,01832	0,00674
1	0,36788	0,27067	0,14936	0,07326	0,03369
2	0,18394	0,27067	0,22404	0,14653	0,08422
3	0,06131	0,18045	0,22404	0,19537	0,14037
4	0,01533	0,09022	0,16803	0,19537	0,17547
5	0,00307	0,03609	0,10082	0,15629	0,17547
6	0,00051	0,01203	0,05041	0,10419	0,14622
7	0,00007	0,00344	0,02160	0,05954	0,10445
8	0,00001	0,00086	0,00810	0,02977	0,06528
9		0,00019	0,00270	0,01323	0,03627
10		0,00004	0,00081	0,00529	0,01813
11		0,00001	0,00022	0,00193	0,00824
12			0,00006	0,00064	0,00343
13			0,00001	0,00020	0,00132
14				0,00006	0,00047
15				0,00002	0,00016
16					0,00005
17					0,00001

Таблица II

Значения функции $\varphi(x) = \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2}$

x	Сотые доли									
	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
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	0,2420	0,2396	0,2371	0,2347	0,2323	0,2299	0,2275	0,2251	0,2227	0,2203
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	0,0540	0,0529	0,0519	0,0508	0,0498	0,0488	0,0478	0,0468	0,0459	0,0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	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	0,0044	0,0043	0,0042	0,0040	0,0039	0,0038	0,0037	0,0036	0,0035	0,0034
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

Таблица III

Значения функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-u^2/2}\, du$

x	Сотые доли									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0200	0,0239	0,0279	0,0319	0,0359
0,1	398	438	478	517	557	596	636	675	714	753
0,2	793	832	871	910	948	987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	406	443	480	517
0,4	554	591	628	664	700	736	772	808	844	879
0,5	915	950	985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	357	389	422	454	486	517	549
0,7	580	611	642	673	703	734	764	794	823	852
0,8	881	910	939	967	995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	289	315	340	365	389
1,0	0,3413	0,3437	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	643	665	686	708	729	749	770	790	810	830
1,2	849	869	888	907	925	944	962	980	997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	177
1,4	192	207	222	236	251	265	279	292	306	319
1,5	332	345	357	370	382	394	406	418	429	441
1,6	457	463	474	484	495	505	515	525	535	545
1,7	554	564	573	582	591	599	608	616	625	633
1,8	641	649	656	664	671	678	686	693	699	706
1,9	713	719	726	732	738	744	750	756	761	767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	821	826	830	834	838	842	846	850	854	857
2,2	861	864	868	871	875	878	881	884	887	890
2,3	893	896	898	901	904	906	909	911	913	916
2,4	918	920	922	925	927	929	931	932	934	936
2,5	938	940	941	943	945	946	948	949	951	952
2,6	953	955	956	957	959	960	961	962	963	964
2,7	965	966	967	968	969	970	971	972	973	974
2,8	974	975	976	977	977	978	979	979	980	981
2,9	981	982	982	983	984	984	985	985	985	986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Таблица IV

Значения q-квантилей U_q стандартного нормального распределения

q	u_q	q	u_q	q	u_q	q	u_q
0,50	0,000000	0,55	0,125661	0,60	0,253347	0,65	0,385320
51	025069	56	150969	61	279319	66	412463
52	050154	57	176374	62	305481	67	439913
53	075270	58	201893	63	331853	68	467699
54	100434	59	227545	64	358459	69	495850

Таблица IV (*окончание*)

q	u_q	q	u_q	q	u_q	q	u_q
0,70	0,524401	0,85	1,036433	0,97	1,880794	0,985	2,170090
71	553385	86	080319	971	895698	986	197286
72	582842	87	126391	972	911036	987	226212
73	612813	88	174987	973	926837	988	257129
74	643345	89	226528	974	943134	989	290368
0,75	0,674490	0,90	1,281552	0,975	1,959964	0,990	2,326348
76	706303	91	340755	976	977368	991	365618
77	738847	92	405072	977	995393	992	408916
78	772193	93	475791	978	2,014091	993	457263
79	806421	94	554774	979	033520	994	512144
0,80	0,841621	0,95	1,044854	0,980	2,053749	0,995	2,575829
81	877896	96	750686	981	074855	996	652070
82	915365			982	096927	997	747781
83	954165			983	120072	998	878162
84	994458			984	144411	999	3,090232

Таблица V

Равномерно распределенные случайные числа

10	09	73	25	33	76	52	01	35	86	34	67	35	48	76
37	54	20	48	05	64	89	47	42	96	24	80	52	40	37
08	42	26	89	53	19	64	50	93	03	23	20	90	25	60
99	01	90	25	29	09	37	67	07	15	38	31	13	11	65
12	80	79	99	70	80	15	73	61	47	64	03	23	66	53
80	95	90	91	17	39	29	27	49	45	66	06	57	47	17
20	63	61	04	02	00	82	29	16	65	31	06	01	08	05
15	95	33	47	64	35	08	03	36	06	85	26	97	76	02
88	67	67	43	97	04	43	62	76	59	63	57	33	21	35
98	95	11	68	77	12	17	17	68	33	73	79	64	57	53
34	07	27	68	50	36	69	73	61	70	65	81	33	98	85
45	57	18	24	06	35	30	34	26	14	86	79	90	74	39
02	05	16	56	92	68	66	57	48	18	73	05	38	52	47
05	32	54	70	48	90	55	35	75	48	28	46	82	87	09
03	52	96	47	78	35	80	83	42	82	60	93	52	03	44
11	19	92	91	70	98	52	01	77	67	14	90	56	86	07
23	40	30	97	32	11	80	50	54	31	39	80	82	77	32
18	62	38	85	79	83	45	29	96	34	06	28	89	80	83
83	49	12	56	24	88	68	54	02	00	86	50	75	84	01
35	27	38	84	35	99	59	46	73	48	87	51	76	49	69
22	10	94	05	58	60	97	09	34	33	50	50	07	39	98
50	72	56	82	48	29	40	52	42	01	52	77	56	78	51
13	74	67	00	78	18	47	54	06	10	68	71	17	78	17
36	76	66	79	51	90	36	47	64	93	29	60	91	10	62
91	82	60	89	28	93	78	56	13	68	23	47	83	41	13
65	48	11	76	74	17	46	85	09	50	58	04	77	69	74
80	12	43	56	35	17	72	70	80	15	45	31	02	23	74
74	35	09	98	17	77	40	27	72	14	43	23	60	02	10
69	91	62	68	03	66	25	22	91	48	36	93	68	72	03
09	89	32	05	05	14	22	56	85	14	46	42	75	67	88

Таблица V (*окончание*)

73	03	95	71	86	40	21	81	65	44	91	49	91	45	23
21	11	57	82	53	14	38	55	37	63	80	33	69	45	98
45	52	16	42	37	96	28	60	26	55	44	10	48	19	49
76	62	11	39	90	94	40	05	64	18	12	55	07	37	42
96	29	77	88	22	54	38	21	45	98	63	60	64	93	29
68	47	92	76	86	46	16	28	35	54	94	75	08	99	23
26	94	03	68	58	70	29	73	41	35	53	14	03	33	40
85	15	74	79	54	32	97	92	65	75	57	60	04	08	81
11	10	00	20	40	12	86	07	46	97	96	64	48	94	39
16	50	53	44	84	40	21	95	25	63	43	65	17	70	82
37	08	92	00	48	61	19	69	04	46	26	45	74	77	74
42	05	08	23	41	15	47	44	52	66	95	27	07	99	53
22	22	20	64	13	94	55	72	85	73	67	89	75	43	87
28	70	72	58	15	42	48	11	62	13	97	34	40	87	21
07	20	73	17	90	23	52	37	83	17	73	20	88	98	37
51	92	43	37	29	65	39	45	95	93	42	58	26	05	27
59	36	78	38	48	82	39	61	01	18	33	21	15	94	66
54	62	24	44	31	91	92	04	25	92	92	92	74	59	73
16	86	84	87	67	03	07	11	20	59	25	70	14	66	70
68	93	59	14	16	26	25	22	96	63	05	52	28	25	62
04	49	35	24	94	75	24	63	38	24	45	86	25	10	25
00	54	99	76	54	64	05	18	81	59	96	11	96	38	96
35	96	31	53	07	26	89	80	93	54	33	35	13	54	62
59	80	80	83	91	45	42	72	68	42	83	60	94	97	00
46	05	88	52	36	01	39	09	22	86	77	28	14	40	77
61	96	27	93	35	65	33	71	24	72	32	17	90	05	97
54	69	28	23	91	23	28	72	95	29	69	23	46	14	06
77	97	45	00	24	90	10	33	93	33	19	56	54	14	30
13	02	12	48	92	78	56	52	01	06	45	15	51	49	38
93	91	08	36	47	70	61	74	29	41	94	86	43	19	94
87	37	92	52	41	05	56	70	70	07	86	74	31	71	57
20	11	74	52	04	15	95	66	00	00	18	74	39	24	23
01	75	87	53	79	40	41	92	15	85	66	67	43	68	06
19	47	60	72	46	43	66	79	45	43	59	04	79	00	33
36	16	81	08	51	34	88	88	15	53	01	54	03	54	56
85	39	41	18	38	98	08	62	48	26	45	24	02	84	04
97	11	89	63	38	33	18	51	62	32	41	94	15	09	49
84	96	28	52	07	80	95	10	04	06	96	38	27	07	74
20	82	66	95	41	79	75	24	91	40	71	96	12	82	96
05	01	45	11	76	18	63	33	25	37	98	14	50	65	71
44	99	90	88	96	39	09	47	34	07	39	44	13	18	80
80	43	54	85	81	89	69	54	19	94	37	54	87	30	43
20	15	12	33	87	25	01	62	52	98	94	62	46	11	71
69	86	10	25	91	74	85	22	05	39	00	38	75	95	79
31	01	02	46	74	05	45	56	14	27	77	93	89	19	36
74	02	94	39	02	77	55	73	22	70	97	79	01	71	19
54	17	84	56	11	80	99	33	71	43	05	33	51	29	69
11	66	44	98	83	52	07	98	48	27	59	38	17	15	39
48	32	47	75	28	31	24	96	47	10	02	29	53	68	70
69	07	49	41	38	87	63	79	19	76	35	58	40	44	01
52	52	75	80	21	80	81	45	17	48	09	18	82	00	97
56	12	71	92	55	36	04	09	03	24	90	04	58	54	97
09	97	33	34	40	88	46	12	33	56	73	18	95	02	07
32	30	75	75	46	15	02	00	99	94	75	76	87	64	90
10	51	82	16	15	01	84	87	69	38	54	01	64	40	56

χ^2-распределение с ν степенями свободы

ν \ Q	0,995	0,990	0,975	0,950	0,900	0,100	0,05	0,025	0,010	0,005
1	$392704 \cdot 10^{-10}$	$1{,}57088 \cdot 10^{-10}$	$982069 \cdot 10^{-9}$	$393214 \cdot 10^{-8}$	$0{,}0157908 \cdot 10^{-8}$	2,70554	3,84146	5,02389	6,63490	7,87944
2	0,0100251	0,0201007	0,0506356	0,102587	0,210720	4,60517	5,99147	7,37776	9,21034	10,5966
3	0,0717212	0,114832	0,215795	0,351846	0,584375	6,25139	7,81473	9,34840	11,3449	12,8381
4	0,205990	0,297110	0,484419	0,710721	1,053623	7,77944	9,48773	11,1433	13,2767	14,8602
5	0,411740	0,554300	0,831211	1,145476	1,61031	9,23635	11,0705	12,8325	15,0863	16,7496
6	0,675727	0,872085	1,237347	1,63539	2,20413	10,6446	12,5916	14,4494	16,8119	18,5476
7	0,989265	1,239043	1,68987	2,16735	2,83311	12,0170	14,0671	16,0128	18,4753	20,2777
8	1,344419	1,646482	2,17973	2,73264	3,48954	13,3616	15,5073	17,5346	20,0902	21,9550
9	1,734926	2,087912	2,70039	3,32511	4,16816	14,6837	16,9190	19,0228	21,6660	23,5893
10	2,15585	2,55821	3,24697	3,94030	4,86518	15,9871	18,3070	20,4831	23,2093	25,1882
11	2,60321	3,05347	3,81575	4,57481	5,57779	17,2750	19,6751	21,9200	24,7250	26,7569
12	3,07382	3,57056	4,40379	5,22603	6,30380	18,5494	21,0261	23,3367	26,2170	28,2995
13	3,56503	4,10591	5,00874	5,89186	7,04150	19,8119	22,3621	24,7356	27,6883	29,8194
14	4,07466	4,66043	5,62872	6,57063	7,78953	21,0642	23,6848	26,1190	29,1413	31,3193
15	4,60094	5,22935	6,26214	7,26094	8,54675	22,3072	24,9958	27,4884	30,5779	32,8013
16	5,14224	5,81221	6,90766	7,96164	9,31223	23,5418	26,2962	28,8454	31,9999	34,2672
17	5,69724	6,40776	7,56418	8,67176	10,0852	24,7690	27,5871	30,1910	33,4087	35,7185
18	6,26481	7,01491	8,23075	9,39046	10,8649	25,9894	28,8693	31,5264	34,8053	37,1564

Таблица VI (*окончание*)

ν \ Q	0,995	0,990	0,975	0,950	0,900	0,100	0,05	0,025	0,010	0,005
19	6,84398	7,63273	8,90655	10,1170	11,6509	27,2036	30,1435	32,8523	36,1908	38,5822
20	7,43386	8,26040	9,59083	10,8508	12,4426	28,4120	31,4104	34,1696	37,5662	39,9968
21	8,03366	8,89720	10,28293	11,5913	13,2396	29,6151	32,6705	35,4789	38,9321	41,4010
22	8,64272	9,54249	10,9823	12,3380	14,0415	30,8133	33,9244	36,7807	40,2894	42,7956
23	9,26042	10,19567	11,6885	13,0905	14,8479	32,0069	35,1725	38,0757	41,6384	44,1813
24	9,88623	10,8564	12,4011	13,8484	15,6587	33,1963	36,4151	39,3641	42,9798	45,5585
25	10,5197	11,5240	13,1197	14,6114	16,4734	34,3816	37,6525	40,6465	44,3141	46,9278
26	11,1603	12,1981	13,8439	15,3791	17,2919	35,5631	38,8852	41,9232	45,6417	48,2899
27	11,8076	12,8786	14,5733	16,1513	18,1138	36,7412	40,1133	43,1944	46,9630	49,6449
28	12,4613	13,5648	15,3079	16,9279	18,9392	37,9159	41,3372	44,4607	48,2782	50,9933
29	13,1211	14,2565	16,0471	17,7083	19,7677	39,0875	42,5569	45,7222	49,5879	52,3356
30	13,7867	14,9535	16,7908	18,4926	20,5992	40,2560	43,7729	46,9792	50,8922	53,6720
40	20,7065	22,1643	24,4331	26,5093	29,0505	51,8050	55,7585	59,3417	63,6907	66,7659
50	27,9907	29,7067	32,3574	34,7642	37,6886	63,1671	67,5048	71,4202	76,1539	79,4900
60	35,5346	37,4848	40,4817	43,1879	46,4589	74,3970	79,0819	83,2976	88,3794	91,9517
70	43,2752	45,4418	48,7576	51,7393	55,3290	85,5271	90,5312	95,0231	100,425	104,215
80	51,1720	53,5400	57,1532	60,3915	64,2778	96,5782	101,879	106,629	112,329	116,321
90	59,1963	61,7541	65,6466	69,1260	73,2912	107,565	113,145	118,136	124,116	128,299
100	67,3276	70,0648	74,2219	77,9295	82,3581	118,498	124,342	129,561	135,807	140,169

Таблица VII

Распределение Стьюдента с ν степенями свободы

ν \ Q / $2Q$	0,4 0,8	0,25 0,5	0,1 0,2	0,05 0,1	0,025 0,05	0,01 0,02	0,005 0,01	0,0025 0,005
1	0,325	1,000	3,078	6,314	12,700	31,821	63,657	127,32
2	289	0,816	1,886	2,920	4,303	0,905	9,925	14,089
3	277	765	638	353	3,182	4,541	5,841	7,453
4	271	741	533	132	2,775	3,747	4,604	5,598
5	0,207	0,727	1,476	2,015	2,571	365	4,032	4,773
6	205	718	440	1,943	447	143	3,707	317
7	203	711	415	895	365	2,998	499	4,029
8	202	706	397	860	306	896	355	3,833
9	201	703	383	833	262	821	250	690
10	0,260	0,700	1,372	1,812	2,228	2,764	3,169	3,581
11	260	697	363	796	201	718	106	497
12	259	695	356	782	179	681	055	428
13	259	694	350	771	160	650	012	372
14	258	692	345	761	145	024	2,977	326
15	0,258	0,691	1,341	1,763	2,131	2,602	2,947	3,286
16	258	690	337	746	120	583	921	252
17	257	689	333	740	110	567	898	222
18	257	688	330	734	101	552	878	197
19	257	688	323	729	093	539	861	174
20	0,257	0,687	1,325	1,725	2,086	2,528	2,845	3,153
21	257	686	323	721	080	518	831	135
22	256	686	321	717	074	508	819	119
23	256	685	319	714	069	500	807	104
24	256	685	313	711	064	492	797	091
25	0,256	0,684	1,316	1,708	2,060	2,485	2,787	3,078
26	256	684	315	706	056	479	779	067
27	256	684	314	703	052	473	771	057
28	256	683	313	701	048	467	763	047
29	256	683	311	699	045	462	756	038
30	0,256	0,683	1,310	1,697	2,042	2,457	2,750	3,030
40	255	681	303	684	021	423	704	2,971
60	254	679	296	671	000	390	660	915
120	254	677	289	653	1,980	358	617	860
∞	253	674	282	645	960	326	570	807

Таблица VIII

F-распределение с числом степеней свободы числителя v_1 и знаменателя v_2

$Q = 0{,}1$

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	39,86	49,50	53,59	55,83	57,24	58,2	58,9	59,44	59,86	60,1	60,7	61,2	61,7	62,00	62,2	62,5	62,79	63,06	63,33
2	8,53	9,00	9,16	9,24	9,29	9,33	9,3	9,37	9,38	9,3	9,4	9,4	9,44	9,4	9,46	9,4	9,47	9,48	9,49
3	5,54	5,46	5,39	5,34	5,3	5,28	5,2	5,25	5,24	5,2	5,2	5,20	5,18	5,1	5,17	5,1	5,15	5,14	5,13
4	4,54	4,32	4,19	4,11	4,05	4,0	3,9	3,95	3,94	3,9	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
5	4,06	3,78	3,62	3,52	3,45	3,40	3,3	3,34	3,32	3,30	3,2	3,24	3,2	3,19	3,17	3,16	3,14	3,12	3,10
6	3,78	3,46	3,29	3,18	3,11	3,05	3,0	2,98	2,96	2,9	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,6	2,63	2,59	2,58	2,56	2,54	2,5!	2,49	2,47
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
9	3,36	3,01	2,81	2,69	2,61	2,55	2,5	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
10	3,29	2,92	2,73	2,61	2,52	2,46	2,4	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
20	2,97	2,59	2,38	2,25	2,16	2,09	2,01	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,87	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
24	2,93	2,51	2,33	2,19	2,10	2,04	1,98	1,01	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60	1,55	1,49	1,42	1,38	1,34	11,30	1,24	1,17	1,00

Таблица VIII (*продолжение*)

$Q = 0{,}05$

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,14	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	4,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,09	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Таблица VIII (*окончание*)

$Q = 0{,}01$

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	4052	4999,5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6215	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	6,26	13,27	2,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	3,75	0,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	2,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	1,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	0,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	0,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,15	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Примечание. При вычислении стопроцентных точек для $Q > 0{,}9$ следует воспользоваться тождеством $v^2_Q(v_1, v_2) = (v^2_{1-Q}(v_1, v_2))^{-1}$.

Таблица IX

Биномиальное распределение при вероятности «успеха» $p = 1/2$

n	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$		n	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$	
10	7	0,1719	8	0,0547	9	0,0107	18	12	0,1189	13	0,0481	14	0,0154
11	8	0,113	9	0,0327	10	0,0059	19	13	0,0835	14	0,0318	15	0,0096
12	9	0,0730	10	0,0193	11	0,0032	20	13	0,1316	14	0,0577	15	0,0207
13	9	0,1334	10	0,0461	11	0,0112	21	14	0,0946	15	0,0392	16	0,0133
14	10	0,0898	11	0,0287	12	0,0065	22	14	0,1431	15	0,0669	16	0,0233
15	10	0,1509	11	0,0592	12	0,0176	23	15	0,1050	16	0,0466	17	0,0173
16	11	0,1051	12	0,0384	13	0,0106	24	15	0,1537	16	0,0758	18	0,0320
17	12	0,0717	13	0,0245	14	0,0064	25	16	0,1148	17	0,0539	19	0,0216

Таблица X

Критические значения статистики ранговых сумм Уилкоксона W

n	m	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$		n	m	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$	
2	2							4	4	23	0,100	24	0,057	26	0,014
2	3	9	0,100					4	5	26	0,095	27	0,056	30	0,008
2	4	10	0,133	11	0,067			4	6	29	0,086	30	0,057	33	0,010
2	5	12	0,095	13	0,048			4	7	31	0,115	33	0,055	36	0,012
2	6	14	0,071	15	0,036			4	8	34	0,107	36	0,055	40	0,008
2	7	15	0,111	16	0,056	17	0,028	4	9	37	0,099	39	0,053	43	0,010
2	8	17	0,089	18	0,044	19	0,022	4	10	40	0,094	42	0,053	46	0,012
2	9	18	0,109	20	0,036	21	0,018	4	11	43	0,089	45	0,052	50	0,009
2	10	20	0,091	21	0,061	23	0,015	4	12	45	0,106	48	0,052	53	0,010
2	11	21	0,115	23	0,051	25	0,013	4	13	48	0,101	51	0,051	56	0,011
2	12	23	0,099	25	0,044	27	0,011	4	14	51	0,096	54	0,051	60	0,009
2	13	24	0,114	26	0,057	29	0,010	4	15	54	0,092	57	0,050	63	0,010
2	14	26	0,100	28	0,050	31	0,008	4	16	56	0,106	60	0,050	66	0,011
2	15	28	0,088	30	0,044	33	0,007	4	17	59	0,101	63	0,049	70	0,009
2	16	29	0,105	31	0,059	35	0,007	4	18	62	0,098	66	0,049	73	0,010
2	17	31	0,094	33	0,053	36	0,012	4	19	65	0,094	69	0,049	76	0,011
2	18	32	0,105	35	0,047	38	0,011	4	20	67	0,105	72	0,048	80	0,009
2	19	34	0,095	37	0,043	40	0,010	5	5	34	0,111	36	0,048	39	0,008
2	20	35	0,108	38	0,052	42	0,009	5	6	38	0,089	40	0,041	43	0,009
								5	7	41	0,101	43	0,053	47	0,009
3	3	14	0,100	15	0,050			5	8	44	0,111	47	0,047	51	0,009
3	4	16	0,114	17	0,057	18	0,029	5	9	48	0,095	50	0,056	55	0,009
3	5	18	0,125	20	0,036	21	0,018	5	10	51	0,103	54	0,050	59	0,010
3	6	21	0,083	22	0,048	24	0,012	6	6	48	0,090	50	0,047	54	0,008
3	7	23	0,092	24	0,058	27	0,008	6	7	52	0,090	54	0,051	58	0,041
3	8	25	0,097	27	0,042	29	0,012	6	8	56	0,091	58	0,054	63	0,010
3	9	27	0,105	29	0,050	32	0,009	6	9	60	0,091	53	0,044	68	0,009
3	10	29	0,108	31	0,056	35	0,007	6	10	64	0,090	67	0,047	72	0,011
3	11	31	0,113	34	0,044	37	0,011	7	7	63	0,104	66	0,049	71	0,009
3	12	34	0,090	36	0,051	40	0,009	7	8	68	0,095	71	0,047	76	0,010
3	13	36	0,095	38	0,055	42	0,012	7	9	72	0,105	76	0,045	81	0,011
3	14	38	0,099	41	0,046	45	0,010	7	10	77	0,097	80	0,054	87	0,009
3	15	40	0,102	43	0,050	48	0,009	8	8	81	0,097	84	0,052	90	0,010
3	16	42	0,105	45	0,055	50	0,011	8	9	86	0,100	90	0,046	96	0,010
3	17	44	0,108	48	0,046	53	0,010	8	10	91	0,102	95	0,051	102	0,010
3	18	47	0,092	50	0,050	56	0,008	9	9	101	0,095	105	0,047	112	0,009
3	19	49	0,095	52	0,054	58	0,010	9	10	106	0,106	111	0,047	119	0,009
3	20	51	0,098	55	0,047	61	0,009	10	10	123	0,095	127	0,053	136	0,009

Таблица XI

Критические значения статистики знаковых рангов Уилкоксона *S*

x \ n	3	4	5	6	7	8	9	10	11	12
3	0,625									
4	0,375									
5	0,250	0,562								
6	0,125	0,438								
7		0,312								
8		0,188	0,500							
9		0,125	0,406							
10		0,062	0,312							
11			0,219	0,500						
12			0,156	0,422						
13			0,094	0,344						
14			0,062	0,281	0,531					
15			0,031	0,219	0,469					
16				0,156	0,406					
17				0,109	0,344					
18				0,078	0,289	0,527				
19				0,047	0,234	0,473				
20				0,031	0,188	0,422				
21				0,016	0,148	0,371				
22					0,109	0,320				
23					0,078	0,273	0,500			
24					0,055	0,230	0,455			
25					0,039	0,191	0,410			
26					0,023	0,156	0,367			
27					0,016	0,125	0,326			
28					0,008	0,098	0,285	0,500		
29						0,074	0,248	0,461		
30						0,055	0,213	0,423		
31						0,039	0,180	0,385		
32						0,027	0,150	0,348		
33						0,020	0,125	0,312	0,517	
34						0,012	0,102	0,278	0,483	
35						0,008	0,082	0,246	0,449	
36						0,004	0,064	0,216	0,416	
37							0,049	0,188	0,382	
38							0,037	0,161	0,350	
39							0,027	0,138	0,319	0,515
40							0,020	0,116	0,289	0,485

Таблица XI (*окончание*)

x \ n	9	10	11	12	13	14	15	x \ n	13	14	15
41	0,014	0,097	0,260	0,455				81	0,005	0,039	0,126
42	0,010	0,080	0,232	0,425				82	0,004	0,034	0,115
43	0,006	0,065	0,207	0,396				83	0,003	0,029	0,104
44	0,004	0,053	0,183	0,367				84	0,002	0,025	0,094
45	0,002	0,042	0,160	0,339				85	0,002	0,021	0,084
46		0,032	0,139	0,311	0,500			86	0,001	0,018	0,076
47		0,024	0,120	0,285	0,473			87	0,001	0,015	0,068
48		0,019	0,103	0,259	0,446			88	0,001	0,012	0,060
49		0,014	0,087	0,235	0,420			89	0,000	0,010	0,053
50		0,010	0,074	0,212	0,393			90	0,000	0,008	0,047
51		0,007	0,062	0,190	0,368			91	0,000	0,007	0,042
52		0,005	0,051	0,170	0,342			92		0,005	0,036
53		0,003	0,042	0,151	0,318	0,500		93		0,004	0,032
54		0,002	0,034	0,133	0,294	0,476		94		0,003	0,028
55		0,001	0,027	0,117	0,271	0,452		95		0,003	0,024
56			0,021	0,102	0,249	0,428		96		0,002	0,021
57			0,016	0,088	0,227	0,404		97		0,002	0,018
58			0,012	0,076	0,207	0,380		98		0,001	0,015
59			0,009	0,065	0,188	0,357		99		0,001	0,013
60			0,007	0,055	0,170	0,335	0,511	100		0,001	0,011
61			0,005	0,046	0,153	0,313	0,489	101		0,000	0,009
62			0,003	0,039	0,137	0,292	0,467	102		0,000	0,008
63			0,002	0,032	0,122	0,271	0,445	103		0,000	0,006
64			0,001	0,026	0,108	0,251	0,423	104		0,000	0,005
65			0,001	0,021	0,095	0,232	0,402	105		0,000	0,004
66			0,000	0,017	0,084	0,213	0,381	106			0,003
67				0,013	0,073	0,196	0,360	107			0,003
68				0,010	0,064	0,179	0,339	108			0,002
69				0,008	0,055	0,163	0,319	109			0,002
70				0,006	0,047	0,148	0,300	110			0,001
71				0,005	0,040	0,134	0,281	111			0,001
72				0,003	0,034	0,121	0,262	112			0,001
73				0,002	0,029	0,108	0,244	113			0,001
74				0,002	0,024	0,097	0,227	114			0,000
75				0,001	0,020	0,086	0,211	115			0,000
76				0,001	0,016	0,077	0,195	116			0,000
77				0,000	0,013	0,068	0,180	117			0,000
78				0,000	0,011	0,059	0,165	118			0,000
79					0,009	0,052	0,151	119			0,000
80					0,007	0,045	0,138	120			0,000

Таблица XII

Критические значения статистики Краскела—Уоллиса H

n_1	n_2	n_3	n_4	n_5	n_6	α				n_1	n_2	n_3	n_4	n_5	n_6	α			
						0,10	0,05	0,025	0,01							0,10	0,05	0,025	0,01
2	2	2	0	0	0	4,571				6	6	3	0	0	0	4,558	5,625	6,725	7,725
3	2	1	0	0	0	4,286				6	6	4	0	0	0	4,548	5,724	6,812	8,000
3	2	2	0	0	0	4,500	4,714			6	6	5	0	0	0	4,542	5,765	6,848	8,124
3	3	1	0	0	0	4,571	5,143			6	6	6	0	0	0	4,643	5,801	6,889	8,222
3	3	2	0	0	0	4,556	5,361	5,556		7	1	1	0	0	0	4,267			
3	3	3	0	0	0	4,622	5,600	5,956	7,200	7	2	1	0	0	0	4,200	4,706	5,727	
4	2	1	0	0	0	4,500				7	2	2	0	0	0	4,526	5,143	5,818	7,000
4	2	2	0	0	0	4,458	5,333	5,500		7	3	1	0	0	0	4,173	4,952	5,758	7,030
4	3	1	0	0	0	4,056	5,208	5,833		7	3	2	0	0	0	4,502	5,357	6,201	6,839
4	3	2	0	0	0	4,511	5,444	6,000	6,444	7	3	3	0	0	0	4,603	5,620	6,449	7,228
4	3	3	0	0	0	4,709	5,791	6,155	6,745	7	4	1	0	0	0	4,121	4,986	5,791	6,986
4	4	2	0	0	0	4,555	5,455	6,327	7,036	7	4	2	0	0	0	4,549	5,376	6,184	7,321
4	4	3	0	0	0	4,545	5,598	6,394	7,144	7	4	3	0	0	0	4,527	5,623	6,578	7,550
4	4	4	0	0	0	4,654	5,692	6,615	7,654	7	4	4	0	0	0	4,562	5,650	6,707	7,814
5	2	1	0	0	0	4,200	5,000			7	5	1	0	0	0	4,035	5,064	5,953	7,061
5	2	2	0	0	0	4,373	5,160	6,000	6,533	7	5	2	0	0	0	4,485	5,393	6,221	7,450
5	3	1	0	0	0	4,018	4,960	6,044		7	5	3	0	0	0	4,535	5,607	6,627	7,697
5	3	2	0	0	0	4,651	5,251	6,004	6,909	7	5	4	0	0	0	4,542	5,733	6,738	7,931
5	3	3	0	0	0	4,533	5,648	6,315	7,079	7	5	5	0	0	0	4,571	5,708	6,835	8,108
5	4	1	0	0	0	3,987	4,985	5,858	6,955	7	6	1	0	0	0	4,033	5,067	6,067	7,254
5	4	2	0	0	0	4,541	5,273	6,068	7,205	7	6	2	0	0	0	4,500	5,357	6,223	7,490
5	4	3	0	0	0	4,549	5,656	6,410	7,445	7	6	3	0	0	0	4,550	5,689	6,694	7,756
5	4	4	0	0	0	4,668	5,657	6,673	7,760	7	6	4	0	0	0	4,562	5,706	6,787	8,039
5	5	1	0	0	0	4,109	5,127	6,000	7,309	7	6	5	0	0	0	4,560	5,770	6,857	8,157
5	5	2	0	0	0	4,623	5,338	6,346	7,338	7	6	6	0	0	0	4,530	5,730	6,897	8,257
5	5	3	0	0	0	4,545	5,705	6,549	7,578	7	7	1	0	0	0	3,986	4,986	6,057	7,157
5	5	4	0	0	0	4,523	5,666	6,760	7,823	7	7	2	0	0	0	4,491	5,398	6,328	7,491
5	5	5	0	0	0	4,560	5,780	6,740	8,000	7	7	3	0	0	0	4,613	5,688	6,708	7,810
6	2	1	0	0	0	4,200	4,822	5,600		7	7	4	0	0	0	4,563	5,766	6,788	8,142
6	2	2	0	0	0	4,545	5,345	5,745	6,655	7	7	5	0	0	0	4,546	5,746	6,886	8,257
6	3	2	0	0	0	3,909	4,855	5,945	6,873	7	7	6	0	0	0	4,568	5,793	6,927	8,345
6	3	2	0	0	0	4,682	5,348	6,136	6,970	7	7	7	0	0	0	4,594	5,818	6,954	8,378
6	3	3	0	0	0	4,590	5,615	6,436	7,410	8	1	1	0	0	0	4,418			
6	4	1	0	0	0	4,038	4,947	5,856	7,106	8	2	1	0	0	0	4,011	4,909	5,420	
6	4	2	0	0	0	4,494	5,340	6,186	7,340	8	2	2	0	0	0	4,587	5,356	5,817	6,663
6	4	3	0	0	0	4,604	5,610	6,538	7,500	8	3	1	0	0	0	4,010	4,881	6,064	6,804
6	4	4	0	0	0	4,595	5,681	6,667	7,795	8	3	2	0	0	0	4,451	5,316	6,295	7,022
6	5	1	0	0	0	4,128	4,990	5,951	7,182	8	3	3	0	0	0	4,543	5,617	6,588	7,350
6	5	2	0	0	0	4,596	5,338	6,196	7,376	8	4	1	0	0	0	4,038	5,044	5,885	6,973
6	5	3	0	0	0	4,535	5,602	6,667	7,590	8	4	2	0	0	0	4,500	5,393	6,193	7,350
6	5	4	0	0	0	4,522	5,661	6,750	7,936	8	4	3	0	0	0	4,529	5,623	6,562	7,585
6	5	5	0	0	0	4,547	5,729	6,788	8,028	8	4	4	0	0	0	4,561	5,779	6,750	7,853
6	6	1	0	0	0	4,000	4,945	5,923	7,121	8	5	1	0	0	0	3,967	4,869	5,864	7,110
6	6	2	0	0	0	4,438	5,410	6,210	7,467	8	5	2	0	0	0	4,466	5,415	6,260	7,440

Таблица XII (*окончание*)

n_1	n_2	n_3	n_4	n_5	n_6	α				n_1	n_2	n_3	n_4	n_5	n_6	α			
						0,10	0,05	0,025	0,01							0,10	0,05	0,025	0,01
8	5	3	0	0	0	4,514	5,614	6,614	7,706	4	4	1	1	0	0	5,182	5,945	6,955	7,909
8	5	4	0	0	0	4,549	5,718	6,782	7,992	4	4	2	1	0	0	5,568	6,386	7,159	7,909
8	5	5	0	0	0	4,555	5,769	6,843	8,116	4	4	2	2	0	0	5,808	6,731	7,538	8,346
8	6	1	0	0	0	4,015	5,015	5,933	7,256	4	4	3	1	0	0	5,692	6,635	7,500	8,231
8	6	2	0	0	0	4,463	5,404	6,294	7,522	4	4	3	2	0	0	5,901	6,874	7,747	8,621
8	6	3	0	0	0	4,575	5,678	6,658	7,796	4	4	3	3	0	0	6,019	7,038	7,929	8,876
8	6	4	0	0	0	4,563	5,743	6,795	8,045	4	4	4	1	0	0	5,654	6,725	7,648	8,588
8	6	5	0	0	0	4,550	5,750	6,867	8,226	4	4	4	2	0	0	5,914	6,957	7,914	8,871
8	6	6	0	0	0	4,599	5,770	6,932	8,313	4	4	4	3	0	0	6,042	7,142	8,079	9,075
8	7	1	0	0	0	4,045	5,041	6,047	7,308	4	4	4	4	0	0	6,088	7,235	8,228	9,287
8	7	2	0	0	0	4,451	5,403	6,339	7,571	9	9	9	9	0	0	6,251	7,815	9,348	11,34
8	7	3	0	0	0	4,556	5,698	6,671	7,872	2	2	1	1	1	0	5,786			
8	7	4	0	0	0	4,548	5,759	6,837	8,118	2	2	2	1	1	0	6,250	6,750	6,750	
8	7	5	0	0	0	4,551	5,782	6,884	8,242	2	2	2	2	1	0	6,600	7,133	7,333	7,533
8	7	6	0	0	0	4,553	5,781	6,917	8,333	2	2	2	2	2	0	6,982	7,418	7,964	8,291
8	7	7	0	0	0	4,585	5,802	6,980	8,363	3	2	1	1	1	0	6,139	6,583		
8	8	1	0	0	0	4,044	5,039	6,005	7,314	3	2	2	1	1	0	6,511	6,800	7,200	7,601
8	8	2	0	0	0	4,509	5,408	6,351	7,654	3	2	2	2	1	0	6,709	7,309	7,745	8,127
8	8	3	0	0	0	4,555	5,734	6,682	7,889	3	2	2	2	2	0	6,955	7,682	8,182	8,682
8	8	4	0	0	0	4,579	5,743	6,886	8,168	3	3	1	1	1	0	6,311	7,111	7,467	
8	8	5	0	o,	0	4,573	5,761	6,920	8,297	3	3	2	1	1	0	6,600	7,200	7,618	
8	8	6	0	0	0	4,572	5,779	6,953	8,367	3	3	2	2	1	0	6,788	7,591	8,121	8,576
8	8	7	0	0	0	4,571	5,791	6,980	8,419	3	3	2	2	2	0	7,026	7,910	8,538	9,115
8	8	8	0	0	0	4,595	5,805	6,995	8,465	3	3	3	1	1	0	6,788	7,576	8,061	8,424
9	9	9	0	0	0	4,605	5,991	7,378	9,210	3	3	3	2	1	0	6,910	7,769	8,449	9,051
2	2	2	1	0	0	5,357	5,679			3	3	3	2	2	0	7,121	8,044	8,813	9,505
2	2	2	2	0	0	5,667	6,167	6,667	6,667	3	3	3	3	1	0	7,077	8,000	8,703	9,451
3	2	1	1	0	0	5,143				3	3	3	3	2	0	7,210	8,200	9,038	9,876
3	2	2	1	0	0	5,556	5,833	6,250		3	3	3	3	3	0	7,333	8,333	9,233	10,20
3	2	2	2	0	0	5,644	6,333	6,978	7,133	9	9	9	9	9	0	7,779	9,488	11,14	13,28
3	3	1	1	0	0	5,333	6,333	6,333		2	2	1	1	1	1	6,833			
3	3	2	1	0	0	5,689	6,244	6,689	7,200	2	2	2	1	1	1	7,267	7,800		
3	3	2	2	0	0	5,745	6,527	7,055	7,636	2	2	2	2	1	1	7,527	8,018	8,345	8,618
3	3	3	1	0	0	5,655	6,600	7,036	7,400	2	2	2	2	2	1	7,909	8,455	8,864	9,227
3	3	3	2	0	0	5,879	6,727	7,515	8,015	2	2	2	2	2	2	8,154	8,846	9,385	9,846
3	3	3	3	0	0	6,026	7,000	7,667	8,538	3	2	1	1	1	1	7,133	7,467	7,667	
4	2	1	1	0	0	5,250	5,833			3	2	2	1	1	1	7,418	7,945	8,236	8,509
4	2	2	1	0	0	5,533	6,133	6,533	7,000	3	2	2	2	1	1	7,727	8,348	8,727	9,136
4	2	2	2	0	0	5,755	6,545	7,064	7,391	3	2	2	2	2	1	7,987	8,731	9,218	9,692
4	3	1	1	0	0	5,067	6,178	6,711	7,067	3	2	2	2	2	2	8,198	9,033	9,648	10,22
4	3	2	1	0	0	5,591	6,309	6,955	7,455	3	3	1	1	1	1	7,400	7,909	8,564	8,564
4	3	2	2	0	0	5,750	6,621	7,326	7,871	3	3	2	1	1	1	7,697	8,303	8,667	9,045
4	3	3	1	0	0	5,689	6,545	7,326	7,758	3	3	2	2	1	1	7,872	8,615	9,128	9,628
4	3	3	2	0	0	5,872	6,795	7,564	8,333	3	3	2	2	2	1	8,077	8,923	9,549	10,15
4	3	3	3	0	0	6,016	6,984	7,775	8,659	3	3	2	2	2	2	8,305	9,190	9,914	10,61

Таблица XIII

Преобразование Фишера выборочного коэффициента корреляции $\hat{r}$

$\hat{r}$	0,000	0,002	0,004	0,006	0,008	$\hat{r}$	0,000	0,002	0,004	0,006	0,008
0,00	0,0000	0,0020	0,0040	0,0060	0,0080	0,50	0,5493	0,5520	0,5547	0,5573	0,5600
1	0,0100	0,0120	0,0140	0,0160	0,0180	1	0,5627	0,5654	0,5682	0,5709	0,5736
2	0,0200	0,0220	0,0240	0,0260	0,0280	2	0,5763	0,5791	0,5818	0,5846	0,5874
3	0,0300	0,0320	0,0340	0,0360	0,0380	3	0,5901	0,5929	0,5957	0,5985	0,6013
4	0,0400	0,0420	0,0440	0,0460	0,0480	4	0,6042	0,6070	0,6098	0,6127	0,6155
0,05	0,0500	0,0520	0,0541	0,0561	0,0581	0,55	0,6184	0,6213	0,6241	0,6270	0,6299
6	0,0601	0,0621	0,0641	0,0661	0,0681	6	0,6328	0,6358	0,6387	0,6416	0,6446
7	0,0701	0,0721	0,0741	0,0761	0,0782	7	0,6475	0,6505	0,6535	0,6565	0,6596
8	0,0802	0,0822	0,0842	0,0862	0,0882	8	0,6625	0,6656	0,6686	0,6716	0,6746
9	0,0902	0,0923	0,0943	0,0963	0,0983	9	0,6777	0,6807	0,6838	0,6369	0,6900
0,10	0,1003	0,1024	0,1044	0,1064	0,1084	0,60	0,6931	0,6963	0,6994	0,7026	0,7057
1	0,1104	0,1125	0,1145	0,1165	0,1186	1	0,7089	0,7121	0,7153	0,7185	0,7213
2	0,1206	0,1226	0,1246	0,1267	0,1287	2	0,7250	0,7283	0,7315	0,7348	0,7348
3	0,1307	0,1328	0,1348	0,1368	0,1389	3	0,7414	0,7447	0,7481	0,7514	0,7548
4	0,1409	0,1430	0,1450	0,1471	0,1491	4	0,7582	0,7616	0,7650	0,7684	0,7718
0,15	0,1511	0,1532	0,1552	0,1573	0,1593	0,65	0,7753	0,7788	0,7823	0,7858	0,7893
6	0,1614	0,1634	0,1655	0,1676	0,1696	6	0,7928	0,7964	0,7999	0,3035	0,8071
7	0,1717	0,1737	0,1758	0,1779	0,1799	7	0,8107	0,8144	0,8180	0,8217	0,8254
8	0,1820	0,1841	0,1861	0,1882	0,1903	8	0,8291	0,8328	0,8366	0,3404	0,3441
9	0,1923	0,1944	0,1965	0,1986	0,2007	9	0,8480	0,8518	0,8556	0,8595	0,3634
0,20	0,2027	0,2048	0,2069	0,2090	0,2111	0,70	0,8673	0,8712	0,8752	0,8792	0,8832
1	0,2132	0,2153	0,2174	0,2196	0,2216	1	0,8872	0,8912	0,8953	0,8994	0,9035
2	0,2237	0,2258	0,2279	0,2300	0,2321	2	0,9076	0,9118	0,9160	0,9202	0,9245
3	0,2342	0,2363	0,2384	0,2405	0,2427	3	0,9287	0,9330	0,9373	0,9417	0,9461
4	0,2443	0,2469	0,2490	0,2512	0,2533	4	0,9505	0,9549	0,9594	0,9639	0,9584
0,25	0,2554	0,2575	0,2597	0,2618	0,2640	0,75	0,973	0,978	0,982	0,987	0,991
6	0,2661	0,2683	0,2704	0,2726	0,2747	6	0,996	1,001	1,006	1,011	1,015
7	0,2769	0,2790	0,2812	0,2833	0,2855	7	1,020	1,025	1,030	1,035	1,040
8	0,2877	0,2899	0,2920	0,2942	0,2964	8	1,045	1,050	1,056	1,061	1,066
9	0,2986	0,3008	0,3029	0,3051	0,3073	9	1,071	1,077	1,082	1,088	1,093
0,30	0,3095	0,3117	0,3139	0,3161	0,3183	0,80	1,099	1,104	1,110	1,116	1,121
1	0,3205	0,3228	0,3250	0,3272	0,3294	1	1,127	1,133	1,139	1,145	1,151
2	0,3316	0,3339	0,3361	0,3383	0,3406	2	1,157	1,163	1,169	1,175	1,162
3	0,3428	0,3451	0,3473	0,3496	0,3518	3	1,188	1,195	1,201	1,208	1,214
4	0,3541	0,3564	0,3586	0,3609	0,3632	4	1,221	1,228	1,235	1,242	1,249
0,35	0,3654	0,3677	0,3700	0,3723	0,3746	0,85	1,256	1,263	1,271	1,278	1,286
6	0,3769	0,3792	0,3815	0,3838	0,3861	6	1,293	1,301	1,309	1,317	1,325
7	0,3884	0,3907	0,3931	0,3954	0,3977	7	1,333	1,341	1,350	1,353	1,367
8	0,4001	0,4024	0,4047	0,4071	0,4094	8	1,376	1,385	1,394	1,403	1,412
9	0,4118	0,4142	0,4165	0,4189	0,4213	9	1,422	1,432	1,442	1,452	1,463
0,40	0,4236	0,4260	0,4284	0,4308	0,4332	0,90	1,472	1,483	1,494	1,505	1,516
1	0,4356	0,4380	0,4404	0,4428	0,4453	1	1,528	1,539	1,551	1,564	1,576
2	0,4477	0,4501	0,4526	0,4550	0,4574	2	1,589	1,602	1,616	1,630	1,644
3	0,4599	0,4624	0,4648	0,4673	0,4698	3	1,658	1,673	1,689	1,705	1,721
4	0,4722	0,4747	0,4772	0,4797	0,4822	4	1,738	1,756	1,774	1,792	1,812
0,45	0,4847	0,4872	0,4897	0,4922	0,4948	0,95	1,832	1,853	1,874	1,897	1,921
6	0,4973	0,4999	0,5024	0,5049	0,5075	6	1,946	1,972	2,000	2,029	2,060
7	0,5101	0,5126	0,5152	0,5178	0,5204	7	2,092	2,127	2,165	2,205	2,249
8	0,5230	0,5256	0,5282	0,5308	0,5334	8	2,298	2,351	2,410	2,477	2,555
9	0,5361	0,5387	0,5413	0,5440	0,5466	9	2,647	2,759	2,903	3,106	3,453

Таблица XIV

Критические значения статистики правостороннего биномиального критерия при $p = 0{,}05$ и $p = 0{,}1$

$p = 0{,}05$									
$S_{кр}$ \ n	2	3	4	5	6	7	8	9	10
0	1	1	1	1	1	1	1	1	1
1	0,097	0,143	0,186	0,226	0,265	0,302	0,337	0,370	0,401
2	0,003	0,007	0,014	0,023	0,033	0,044	0,057	0,071	0,086
3		0	0,001	0,001	0,002	0,004	0,006	0,008	0,011
4			0	0	0	0	0	0,001	0,001
5				0	0	0	0	0	0
6					0	0	0	0	0
7						0	0	0	0
8							0	0	0
9								0	0
10									0
$p = 0{,}1$									
$S_{кр}$ \ n	2	3	4	5	6	7	8	9	10
0	1	1	1	1	1	1	1	1	1
1	0,190	0,271	0,344	0,410	0,469	0,522	0,570	0,613	0,651
2	0,010	0,028	0,052	0,081	0,114	0,150	0,187	0,225	0,264
3		0,001	0,004	0,009	0,016	0,026	0,038	0,053	0,070
4			0	0	0,001	0,003	0,005	0,008	0,013
5				0	0	0	0	0,001	0,002
6					0	0	0	0	0
7						0	0	0	0
8							0	0	0
9								0	0
10									0

Таблица XV

Критические значения статистики U критерия Манна—Уитни

$\alpha = 0{,}05$

m \ n	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	34	41	48	55	62	69	76	83	90	98	105	112	119	127

$\alpha = 0{,}01$

m \ n	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3			0	0	0	1	1	1	2	2	2	2	3	3
4	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	1	2	3	4	4	6	7	7	8	9	10	11	12	13
6	3	4	5	6	6	9	10	11	12	13	15	16	17	18
7	4	6	7	9	9	12	13	15	16	18	19	21	22	24
8	6	7	9	11	11	15	17	18	20	22	24	26	28	30
9	7	9	11	13	13	18	20	22	24	27	29	31	33	36
10	9	11	13	16	16	21	24	26	29	31	34	37	39	42
11	10	13	16	18	18	24	27	30	33	36	39	42	45	48
12	12	15	18	21	21	27	31	34	37	41	44	47	51	54
13	13	17	20	24	24	31	34	38	42	45	49	53	56	60
14	15	18	22	26	26	34	38	42	46	50	54	58	63	67
15	16	20	24	29	29	37	42	46	51	55	60	64	69	73
16	18	22	27	31	31	41	45	50	55	60	65	70	74	79
17	19	24	29	34	34	44	49	54	60	65	70	75	81	86
18	21	26	31	37	37	47	53	58	64	70	75	81	87	92
19	22	28	33	39	39	51	56	63	69	74	81	87	93	99
20	24	30	36	42	42	54	60	67	73	79	86	92	99	105

Замечания к использованию таблиц

1. Приведенные таблицы взяты из [1], [26], [29], [30].

2. С помощью табл. II можно определять значения $\varphi(x)$ и для отрицательных значений x, если воспользоваться свойствами четности функции $\varphi(x)$: $\varphi(-x) = \varphi(x)$.

3. С помощью табл. III можно определять значения $\Phi(x)$ и для отрицательных значений x, если воспользоваться свойством нечетности функции $\Phi(x)$: $\Phi(-x) = -\Phi(x)$.

4. При нахождении q-квантилей U_q для значений $q < 0{,}5$ с помощью табл. IV следует пользоваться равенством $U_q = -U_{1-q}$.

5. В табл. VI по заданной вероятности Q значение $\chi^2(v, Q)$ случайной величины $\chi^2(v)$, имеющей χ^2-распределение с v степенями свободы, определяется уравнением

$$\mathsf{P}\{\chi^2(v) \geqslant \chi^2(v, Q)\} = Q.$$

6. В табл. VII по заданной вероятности Q значение $t(v, Q)$ случайной величины $t(v)$, имеющей распределение Стьюдента с v степенями свободы, определяется уравнением

$$\mathsf{P}\{|t(v)| \geqslant t(v, Q)\} = Q.$$

7. В табл. VIII по заданной вероятности Q значение $F(v_1, v_2, Q)$ случайной величины $F(v_1 v_2)$, имеющей F-распределение с числом степеней свободы числителя v_1 и знаменателя v_2, определяется уравнением

$$\mathsf{P}\{F(v_1, v_2) \geqslant F(v_1, v_2, Q)\} = Q.$$

Эта таблица составлена лишь для случая, когда в статистике числитель больше знаменателя.

8. В табл. IX по заданным уровню значимости α и числу испытаний n значение $S(n, \alpha)$ «числа успехов» $S(n)$, имеющего биномиальное распределение с параметрами n и $p = 1/2$ определяется уравнением

$$\mathsf{P}\{S(n) \geqslant S(n, \alpha)\} = \alpha.$$

Числа в таблице расположены в три столбца, соответствующие близким значениям к уровням значимости $\alpha = 0{,}10;\ 0{,}05;\ 0{,}01$. Внутри каждого столбца слева в строке указаны значения $S(n, \alpha)$, где α близко к назначенному уровню, а справа даны значения $S(n, \alpha)$ при назначенном уровне α.

9. В табл. X для заданных объемов m и n двух выборок ($m \geqslant n$) и уровню значимости α при выполнении гипотезы H_0 критическое значение $W_{\text{кр}}$ определяется уравнением

$$\mathsf{P}\{W \geqslant W_{\text{кр}}\} = \alpha.$$

Числа в таблице расположены в три столбца, соответствующие близким значениям к уровням значимости $\alpha = 0{,}10;\ 0{,}05;\ 0{,}01$. Внутри каждого столбца слева в строке указаны значения $W_{\text{кр}}$, где α близко к назначенному уровню, а справа даны значения $W_{\text{кр}}$ при назначенном уровне α.

10. В табл. XI для заданных объема выборки n и уровня значимости α при выполнении гипотезы H_0 критическое значение $S_{\text{кр}} = x$ определяется уравнением $\mathsf{P}\{S \geqslant x\} = \alpha$.

11. В табл. XII для заданных числа способов обработки k, числа наблюдений для каждого из способов обработки n_i, $i = 1,\ 2,\ \ldots,\ k$ и уровня значимости α при выполнении гипотезы H_0 критическое значение $H_{\text{кр}}$ определяется уравнением

$$\mathsf{P}\{H \geqslant H_{\text{кр}}\} = \alpha.$$

В таблице приведены решения этого уравнения для тех значений α, которые близки к уровням $\alpha = 0{,}10;\ 0{,}05;\ 0{,}025;\ 0{,}01$. Таким образом, проверка гипотезы H_0 с помощью этой таблицы проводится лишь на приближенных к вышеуказанным уровнях значимости.

12. В табл. XIII по заданному значению $\hat{r}$ можно найти значение $z = \operatorname{arth} \hat{r}$ и наоборот, по заданному значению z можно найти значение $\hat{r} = \operatorname{th} z$. Для работы с отрицательными значениями $\hat{r}$ и z используется свойство нечетности функций $\operatorname{arth} \hat{r}$ и $\operatorname{th} z$.

13. В табл. XIV по заданным уровню значимости α, числу испытаний n, вероятности $p = 0{,}05$ или $p = 0{,}1$ критическое значение $S_{\text{кр}}$ определяется уравнением

$$\mathsf{P}\{S > S_{\text{кр}}\} = \alpha.$$

14. В табл. XV по заданным объемам m и n двух выборок и по уровню значимости $\alpha = 0{,}05$, $\alpha = 0{,}01$ находятся критические значения статистики U критерия Манна—Уитни.

15. В табл. IX, X, XI приняты следующие обозначения десятичных дробей. Ноль целых опускается; целая часть отделяется от дробной точкой, а не запятой; таким образом, например, запись .1719 означает 0,1719.

О СТАТИСТИЧЕСКИХ ПАКЕТАХ ПРОГРАММ ДЛЯ АНАЛИЗА ДАННЫХ НА ПЕРСОНАЛЬНЫХ КОМПЬЮТЕРАХ

Общая характеристика статистических пакетов программ

Реально изучаемые объекты и явления на практике обладают не одним признаком, а целым множеством признаков. Для совместного исследования значений этих признаков необходимо применение методов многомерного статистического анализа. Поскольку и число признаков, и число таких методов могут быть достаточно большими и приходится обрабатывать большие объемы информации, это приводит к весьма трудоемким вычислительным процедурам.

Вычислительная трудоемкость статистических методов анализа данных, в частности методов многомерного статистического анализа, потребовала создания статистических пакетов программ для анализа данных на персональных компьютерах. Статистические пакеты программ сделали методы анализа данных более доступными и наглядными. Всю трудоемкую работу расчетов по сложным формулам, построению таблиц и графиков взял на себя компьютер. Пользователю осталась творческая работа: постановка задач, выбор методов их решения, интерпретация результатов. Для осмысления применения методов статистического анализа данных пользователь должен обладать профессионализмом и хорошей интуицией. Он должен иметь определенную подготовку, чтобы знать, каковы условия применения различных статистических методов, каковы их свойства, возможности, преимущества и недостатки, как интерпретировать их результаты. Например, противоречия в математических и содержательных выводах свидетельствуют о некорректности решения задачи или некорректности интерпретации аналитических результатов. Таким образом, владение компьютерными методами анализа данных — это, с одной стороны, веление времени, а с другой стороны — необходимый элемент для широкого применения на практике методов статистического анализа данных.

Выбор типа программ статистического анализа зависит, прежде всего, от поставленной задачи, от уровня профессиональной подготовки пользователя и возможностей используемых компьютеров. Не

существует общего критерия оценки статистических пакетов программ. Статистические пакеты программ подразделяются на универсальные, позволяющие осуществлять комплексный статистический анализ и содержащие достаточно полный набор стандартных статистических методов, и специализированные, ориентируемые на конкретную предметную область и содержащие метод из одного-двух разделов математической статистики. Из статистических пакетов программ наиболее распространенными в России являются отечественные пакеты STADIA и ЭВРИСТА и зарубежные пакеты SPSS, STATGRAPHICS, STATISTICA. Все перечисленные пакеты программ, за исключением пакета ЭВРИСТА, являются универсальными. Статистический пакет ЭВРИСТА специализируется только на анализе временных рядов. Именно перечисленные пакеты программ рекомендуется применять на практике. Нецелесообразно использовать неполные пакеты, содержащие лишь, как правило, простейшие методы описательной статистики, и некоторые методы одного-двух других разделов статистики. Не следует также доверять методам обработки статистических данных, включенным в табличные процессоры общего назначения, в Excel, и в некоторые базы данных. Иногда в пакетах общего назначения, кроме традиционных методов анализа данных, имеются и малоизвестные или оригинальные методы, созданные разработчиками пакета. Такие методы нужно использовать очень осторожно и лишь после тщательного изучения документации пакета по таким методам.

Все рекомендуемые для использования статистические пакеты программ соответствуют первоочередным требованиям с точки зрения пользователя. Они просты для быстрого освоения, имеют удобные средства для ввода, преобразования и хранения данных, предоставляют удобные возможности для графического и табличного представления промежуточных и окончательных результатов обработки. Наконец, они имеют подробную документацию, играющую роль понятного учебника по использованию статистических методов. Очень удобно, что использование статистических процедур в документации иллюстрируется примерами. Наряду с документацией, осваивать пакет программ помогает встроенный справочник пакета и экспертная система, помогающая в выборе методов анализа данных и комментирующая полученные результаты.

Таким образом, работа со статистическими пакетами программ требует от пользователя определенной квалификации в прикладной статистике, изучения документации пакетов или специального обучения и владения основами программирования.

В каждом универсальном статистическом пакете программ достаточно полно представлены общепринятые статистические методы обработки из всех разделов анализа. Перечисление этих методов вместе с модификациями содержит более сотни наименований. Ограничимся перечислением лишь основных программ, представленных в универсальных статистических пакетах. Эти программы представляют методы описательной статистики, методы исследования функций распределения вероятностей, методы генерирования случайных выборок из разных распределений, методы оценивания параметров различных распределений, методы проверки различных статистических гипотез, методы однофакторного и двухфакторного анализов, в частности, дисперсионного анализа, методы регрессионного анализа (простая линейная регрессия, линейная множественная регрессия, нелинейная регрессия), методы корреляционного анализа (таблицы сопряженности, коэффициенты корреляции Спирмена, Кендалла, Пирсона), многомерные методы (дискриминантный, факторный, кластерный, шкалирования), методы анализа временных рядов и др.

Пояснения по назначению и процедуре применения методов, представленных в статистическом пакете программ, прежде всего, следует искать в документации пакета.

Методические указания по проведению статистического анализа в пакете STATISTICA

В. Т. Бордукова, Т. И. Бордукова

1. Организация исходных данных

Исходные данные в системе STATISTICA организованы в виде электронной таблицы. Таблица состоит из строк и столбцов. Столбцы называются Variable (Переменные), а строки Cases (Наблюдения). В качестве переменных обычно выступают исследуемые величины, а наблюдения — это значения, которые принимают переменные в отдельных измерениях. Система может работать как с численными, так и с текстовыми данными.

Файлы с исходными данными имеют расширение **.sta*. В системе **STATISTICA** в каталоге *stat\examples* имеется много файлов

с исходными данными для проведения практически любого анализа, предусмотренного в **STATISTICA**. Их можно использовать в качестве образца при подготовке своих данных для проведения конкретного анализа.

Пакет **STATISTICA** организован по модульному принципу. Это означает, что все методы статистической обработки, реализованные в системе, разбиты на несколько групп — модулей — в соответствии с разделами статистического анализа. Каждый модуль может работать независимо от остальных модулей системы. **Переключатель модулей (Module Switcher)** позволяет осуществить оперативный доступ ко всем модулям.

Структура диалога в каждом модуле имеет общие черты:

- после выбора в Переключателе модулей открывается **Стартовая панель** выбранного модуля;
- далее открываете файл данных, задавая, при необходимости, условия выбора наблюдений — кнопка **Select cases (Выбрать наблюдения)** — и веса переменных — кнопка **Weight (Вес)**;
- выбираете переменные для анализа из открытого файла данных;
- выбираете метод анализа из меню в **Стартовой панели** модуля;
- выбираете конкретную вычислительную процедуру и задаете ее параметры;
- производите запуск вычислительной процедуры;
- если процедура итерационная, то просматриваете результаты вычисления на каждом шаге в появившемся на экране окне (при этом вы можете добавить необходимое число итераций для увеличения точности оценок);
- используя графические возможности и специальные таблицы вывода с вычисленными разнообразными статистиками, осуществляете всесторонний просмотр и анализ результатов в окне **Results (Результаты)**;
- выбираете следующий шаг анализа.

В сложной задаче требуется работать с различными модулями, последовательно переключаясь между ними.

В системе **STATISTICA** реализован принцип постоянной логической подсказки. Если вы не знаете, что делать на данном этапе, то просто нажмите кнопку **Enter** и система сама отправит вас к нужному диалоговому окну.

Численные результаты анализа выводятся в форме электронных таблиц, которые можно сохранить в файле или распечатать.

2. Анализ данных в модуле Basic Statistic/Tables (Основные статистики/таблицы)

В этом модуле есть **Probability calculator** (**Вероятностный калькулятор**), заменяющий многие таблицы вероятностных распределений. С его помощью можно решать многие статистические задачи. Вероятностный калькулятор позволяет посмотреть на графики наиболее употребляемых функций распределения и их плотностей, определять квантили распределений. Вы можете воочию убедиться, например, что при больших степенях свободы (больших 30) χ^2-распределение и t-распределение практически совпадают с нормальным распределением.

Используя функцию Rnd(x), можно генерировать случайные числа, равномерно распределенные на интервале (0, x).

Словарь используемых в модуле статистических терминов

Valid N — истинное число наблюдений переменной;

Mean — выборочное среднее;

Confid −95% — нижняя граница 95% доверительного интервала для параметра;

Confid +95% — верхняя граница 95% доверительного интервала для параметра;

Sum — сумма значений переменной;

Minimum — минимальное значение переменной;

Maximum — максимальное значение переменной;

Range — размах (т. е. разность между максимумом и минимумом);

Variance — выборочная дисперсия;

Std. Dev — выборочное стандартное отклонение;

Std. Err — стандартная ошибка;

Skewness — выборочный коэффициент асимметрии;

Std. Err. Skewness — стандартная ошибка выборочного коэффициента асимметрии;

Kurtosis — выборочный коэффициент эксцесса;

Std. Err. Kurtosis — стандартная ошибка выборочного коэффициента эксцесса;

3. Модуль «ANOVA/MANOVA» (Одно- и многофакторный анализы)

Рассмотрим только случай однофакторного анализа.

При исследовании зависимостей одной из наиболее простых является ситуация, когда можно указать только один фактор, влияющий

на конечный результат, и этот фактор может принимать лишь конечное число значений (уровней). Такие задачи называются задачами *однофакторного анализа*. Типичный пример — сравнение по достигаемым результатам нескольких различных способов действий, направленных на достижение одной цели, скажем, нескольких школьных учебников или нескольких лекарств.

То, что, как мы считаем, должно оказывать влияние на конечный результат, называют фактором (в приведенных выше примерах факторами являются понятия «школьный учебник» и «лекарство»). Конкретную реализацию фактора (например, определенный школьный учебник или выбранное лекарство) называют *уровнем фактора* или *способом обработки*. Значения измеряемого признака (т. е. величину результата) часто называют *откликом*.

Для сравнения влияния фактора на результат необходим определенный статистический материал. Обычно его получают следующим образом: проводят несколько экспериментов для каждого уровня фактора. Итогом этих экспериментов являются k выборок (где k — количество уровней фактора), вообще говоря, разных объемов. Наиболее распространенным и удобным способом представления подобных данных является таблица.

Таблица 1

Обработки (соответствуют уровням фактора)	1	2	...	k
Результаты измерений	x_{11} x_{21} ... $x_{n_1,1}$	x_{12} x_{22} ... $x_{n_2,2}$	... ...	x_{1k} x_{2k} ... $x_{n_k,k}$

Опыт показывает, что при изменении уровня фактора наибольшей изменчивости в первую очередь, как правило, подвержено положение случайной величины, которую можно характеризовать медианой или математическим ожиданием, т. е. закон распределения не меняется, а меняется только математическое ожидание. Про распределения таких случайных величин говорят, что они относятся к семейству *сдвиговых* распределений. Часто в качестве такого семейства рассматривается семейство нормальных распределений, и для обработки данных применяются методы *дисперсионного анализа*.

Итак, мы предполагаем, что у нас имеется k выборок из нормальных распределений $N(\mu+\theta_1, \sigma^2)$, ..., $N(\mu+\theta_k, \sigma^2)$. Здесь μ — общее математическое ожидание, $\theta_1, \ldots, \theta_k$ — эффекты воздействия

различных уровней фактора на математическое ожидание, т. е. любое наблюдение можно записать в виде $x_{ij} = \mu + \theta_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$. Проверяется гипотеза об отсутствии эффекта воздействия фактора — H_0: $\theta_i = 0$, $i = 1, \ldots, k$ (или H_0: $\mu_i = \mu$, $i = 1, \ldots, k$). Для проверки этой гипотезы применяется F-отношение (см. ниже таблицу однофакторного дисперсионного анализа). Если гипотеза не отклоняется, то анализ заканчивается (это самый тривиальный случай, ибо обычно обработка проводится с целью воздействия на изучаемый признак). В противном случае, так как F-критерий не дает информации о том, какие именно из μ_i не совпадают с μ, возникает задача оценки величины эффектов обработки и выяснения качества полученных оценок, необходимо провести дополнительные исследования, анализируя так называемые контрасты. *Контрастом* называется линейная комбинация математических ожиданий

$$\gamma = \sum_{i=1}^{k} c_i \mu_i,$$

коэффициенты которой удовлетворяют условию

$$\sum_{i=1}^{k} c_i = 0.$$

Каждый контраст — это разность между взвешенными средними от математических ожиданий. Например,

$$\mu_1 - \mu_2, \quad \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4 + \mu_5}{3}$$

и т. д.

В других случаях предположение о нормальности распределений не является правомерным, и тогда используют различные непараметрические методы анализа, из которых наиболее разработаны ранговые методы.

При проведении однофакторного дисперсионного анализа в модуле **ANOVA/MANOVA** исходные данные представляются в виде двух столбцов. В первом столбце записываются значения уровня фактора, а во втором — соответствующие значения измеряемого признака. В качестве примера рассмотрим файл *Aggressn.sta* из папки *Example*. Здесь фактором является пол ребенка, а измеряемым признаком является уровень агрессии. Предположим, что выполнены условия проведения дисперсионного анализа, т. е. уровни агрессии у мальчиков и у девочек распределены по нормальным законам с неизвестными, но одинаковыми дисперсиями σ^2 и с неизвестными

математическими ожиданиями, соответственно равными μ_1 и μ_2. Проверим гипотезу H_0: $\mu_1 = \mu_2$ (гипотеза, что уровень агрессии не зависит от пола) против альтернативы, H_1: $\mu_1 \neq \mu_2$.

Находясь в любом модуле, можно открыть файл *Aggressn.sta*. Затем на стартовой панели щелкнуть мышью на **Analyses** и выбрать модуль **ANOVA/MANOVA**. Появится окно для выбора фактора и зависимой от него переменной. Чтобы сделать этот выбор, надо в первом списке щелкнуть мышью по имени **Gender (пол)**, а во втором списке — по имени **Aggression** и затем щелкнуть мышью по **OK** (рис. 15).

Снова появится диалоговое окно для того, чтобы указать некоторые дополнительные сведения о данных, если это требуется. Для нашей задачи никакой дополнительной информации не требуется, поэтому снова надо щелкнуть мышью по **OK** (рис. 16).

Появилось окно **ANOVA Results** — панель со множеством кнопок, щелкая мышью по которым, будем получать различные аспекты результатов дисперсионного анализа (рис. 17).

Первые два окна представляют собой так называемую таблицу дисперсионного анализа, а в третьем окне даны оценки математических ожиданий уровней агрессии для мальчиков и девочек.

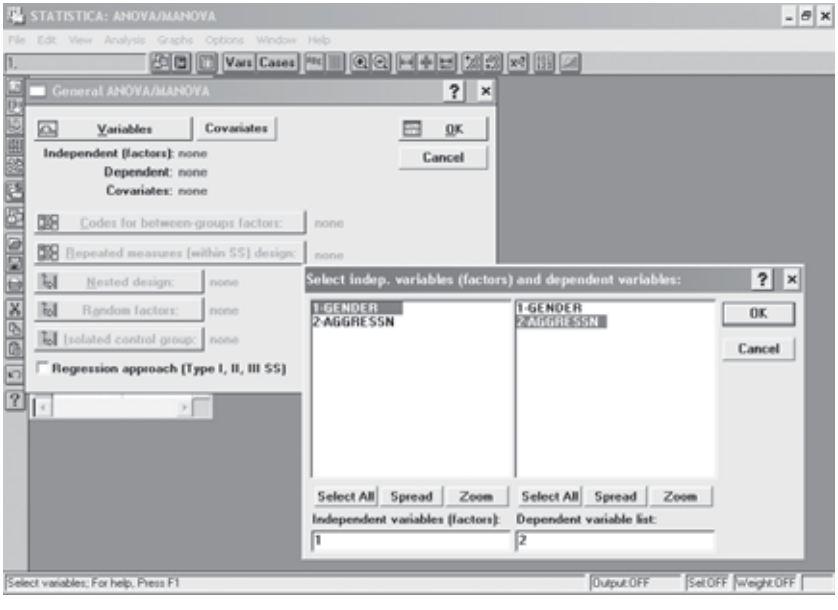


Рис. 15

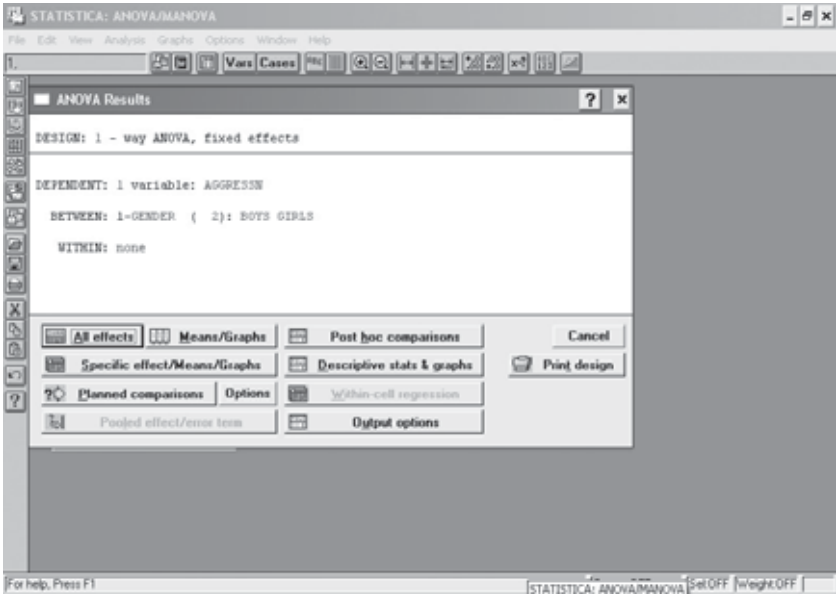

STATISTICA: ANOVA/MANOVA
File Edit View Analysis Graphs Options Window Help
Vars Cases
ANOVA Results
DESIGN: 1 - way ANOVA, fixed effects
DEPENDENT: 1 variable: AGGRESSN
BETWEEN: 1-GENDER (2): BOYS GIRLS
WITHIN: none
All effects
Means/Graphs
Post hoc comparisons
Cancel
Specific effect/Means/Graphs
Descriptive stats & graphs
Print design
Planned comparisons
Options
Within-cell regression
Pooled effect/error term
Output options
For help, Press F1
STATISTICA: ANOVA/MANOVA
Sel:OFF
Weight:OFF

Рис. 16

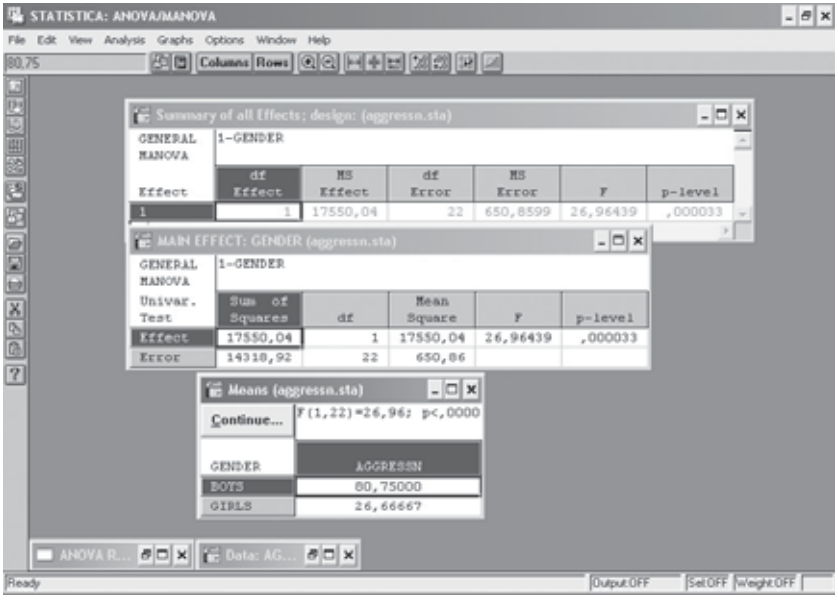

STATISTICA: ANOVA/MANOVA
File Edit View Analysis Graphs Options Window Help
Columns Rows
Summary of all Effects; design: (aggressn.sta)
GENERAL MANOVA
1-GENDER
Effect
df Effect
MS Effect
df Error
MS Error
F
p-level
1
1
17550,04
22
650,8599
26,96439
,000033
MAIN EFFECT: GENDER (aggressn.sta)
GENERAL MANOVA
1-GENDER
Univar. Test
Sum of Squares
df
Mean Square
F
p-level
Effect
17550,04
1
17550,04
26,96439
,000033
Error
14310,92
22
650,86
Means (aggressn.sta)
Continue...
F(1,22)=26,96; p<,0000
GENDER
AGGRESSN
BOYS
60,75000
GIRLS
26,66667
ANOVA R...
Data: AG...
Ready
Output:OFF
Sel:OFF
Weight:OFF

Рис. 17

Вычисленная статистика критерия Фишера и соответствующее ей p-значение подтверждают вывод о значимом различии средних значений изучаемого признака. Другими словами, пол ребенка существенно влияет на уровень его агрессии (рис. 18).

В окне **ANOVA Results** щелкнув мышью по кнопке, получим возможность более подробно ознакомиться с различными аспектами результатов дисперсионного анализа (рис. 19).

В частности, есть возможность протестировать, насколько хорошо выполнены предпосылки дисперсионного анализа для предложенных данных. Щелкнув мышью по кнопке с нарисованной на ней гистограммой (**Distribution of dep. Var.**), получим два окна. Первое окно содержит необходимые численные результаты для проверки с помощью критерия χ^2 того, что уровень агрессии для девочек является нормально распределенной случайной величиной. Вычисленная статистика критерия χ^2 и соответствующее ей p-значение подтверждают наше предположение, ибо p-значение меньше, чем $\alpha = 0{,}05$. Во втором окне показана соответствующая гистограмма распределения.

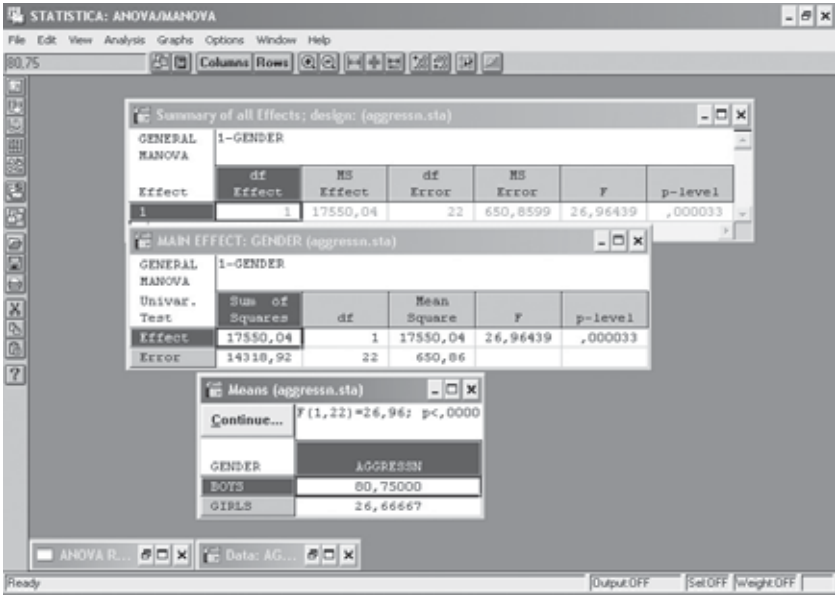


Рис. 18

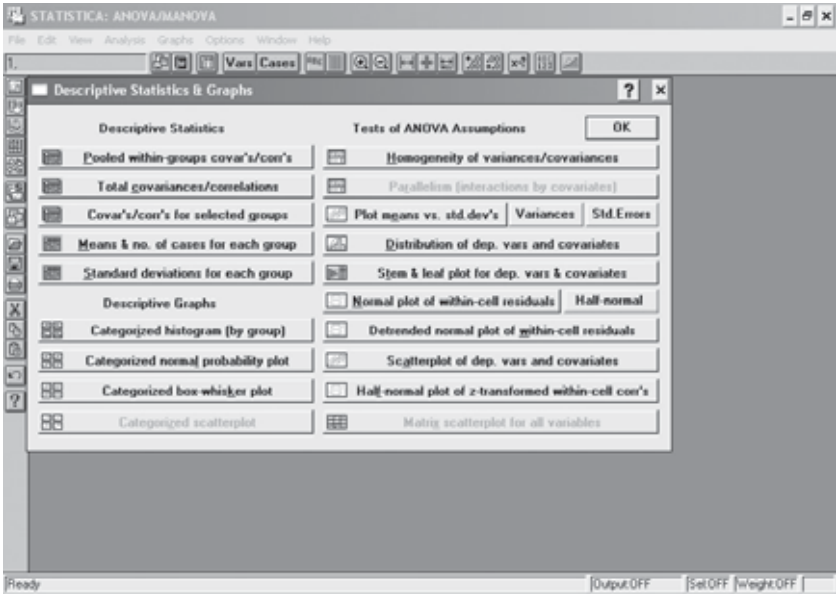

STATISTICA: ANOVA/MANOVA
Descriptive Statistics & Graphs
Descriptive Statistics
Tests of ANOVA Assumptions
OK
Pooled within-groups covar's/cor's
Total covariances/correlations
Covar's/cor's for selected groups
Means & no. of cases for each group
Standard deviations for each group
Descriptive Graphs
Categorized histogram (by group)
Categorized normal probability plot
Categorized box-whisker plot
Categorized scatterplot
Homogeneity of variances/covariances
Parallelism (interactions by covariates)
Plot means vs. std.dev's
Variances
Std.Errors
Distribution of dep. vars and covariates
Stem & leaf plot for dep. vars & covariates
Normal plot of within-cell residuals
Half-normal
Detrended normal plot of within-cell residuals
Scatterplot of dep. vars and covariates
Half-normal plot of z-transformed within-cell cor's
Matrix scatterplot for all variables
Ready
Output:OFF
Sel:OFF
Weight:OFF

Рис. 19

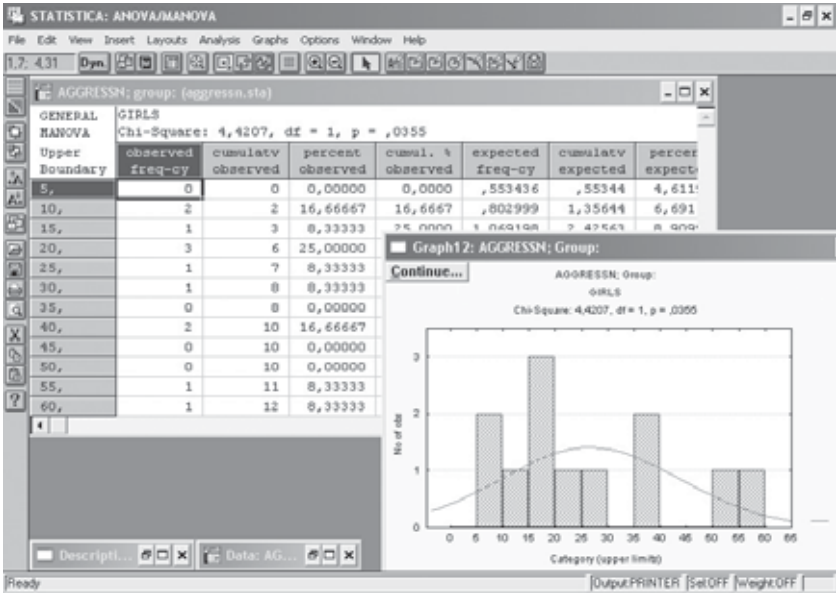

STATISTICA: ANOVA/MANOVA
File Edit View Insert Layouts Analysis Graphs Options Window Help
AGGRESSN; group: (aggressn.sta)
GENERAL
MANOVA
GIRLS
Chi-Square: 4,4207, df = 1, p = ,0355
Upper Boundary
observed freq-cy
cumulatv observed
percent observed
cumul. % observed
expected freq-cy
cumulatv expected
5, 0 0 0,00000 0,0000 ,553436 ,55344
10, 2 2 16,66667 16,6667 ,802999 1,35644
15, 1 3 8,33333
20, 3 6 25,00000
25, 1 7 8,33333
30, 1 8 8,33333
35, 0 8 0,00000
40, 2 10 16,66667
45, 0 10 0,00000
50, 0 10 0,00000
55, 1 11 8,33333
60, 1 12 8,33333
Graph12: AGGRESSN; Group:
Continue...
AGGRESSN; Group:
GIRLS
Chi-Square: 4,4207, df = 1, p = ,0355
No of obs
Category (upper limits)
Ready
Output:PRINTER
Sel:OFF
Weight:OFF

Рис. 20

4. Непараметрические методы проведения однофакторного анализа

Ранговый однофакторный анализ. Если мы ничего не знаем о распределении наблюдений, то непосредственно использовать для проверки нулевой гипотезы количественные значения наблюдений x_{ij} становится затруднительно. В этом случае проще всего опираться в своих выводах только на отношениях «больше—меньше» между наблюдениями, так как они не зависят от распределения наблюдений. При этом вся информация, которую мы используем из таблицы, содержится в тех рангах, что получают числа x_{ij} при упорядочении всей совокупности. Соответствующие критерии для проверки нулевой гипотезы называются *ранговыми*, они пригодны для любых непрерывных распределений наблюдений. Более того, они годятся и тогда, когда измерения x_{ij} сделаны в порядковой шкале, например, являются тестовыми баллами или экспертными оценками. Здесь конкретные численные значения величин x_{ij} вообще являются условностью, а содержательный смысл имеют лишь отношения «больше—меньше» между ними.

Мы будем в основном рассматривать наиболее ясный и простой случай, когда среди чисел x_{ij} нет совпадающих. При наличии совпадений (и использовании средних рангов) теоретическая схема действует как приближенная, а надежность ее выводов снижается тем более, чем больше совпадений. Ниже мы укажем, какие поправки делаются при наличии совпадений.

Упорядочим величины x_{ij} (все равно как — от большего к меньшему или от меньшего к большему). Обозначим через r_{ij} ранг числа x_{ij} во всей совокупности. Тогда табл. 1 преобразуется в табл. 2.

Таблица 2

Обработки (соответствуют уровням фактора)	1	2	...	k
Ранги результатов измерений	r_{11} r_{21} ... $r_{n_1,1}$	r_{12} r_{22} ... $r_{n_2,2}$	... ...	r_{1k} r_{2k} ... $r_{n_k,k}$

Важно отметить, что при выполнении гипотезы H_0 любые возможные расположения рангов по местам таблицы равновероятны.

Согласно сформулированной стратегии анализа возникает вопрос: нельзя ли объяснить наблюденное в опыте расположение рангов в таблице действием чистой случайности? Это можно сформу-

лировать в виде статистической гипотезы о том, что все k представленных выборок однородны, т. е. являются выборками из одного и того же распределения.

Если мы не можем сказать что либо определенное об альтернативах к H_0, можно воспользоваться для ее проверки свободным от распределения ранговым критерием Краскела—Уоллиса. Для каждого уровня фактора j (т. е. для каждого столбца исходной таблицы) надо вычислить

$$R_j = \sum_{i=1}^{n_i} r_{ij}$$

и

$$R_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_i} r_{ij},$$

где $R_{\cdot j}$ — это средний ранг, рассчитанный по столбцу. Если между столбцами нет систематических различий, средние ранги $R_{\cdot j}$, $j = 1, \ldots, k$, не должны значительно отличаться от среднего ранга, рассчитанного по всей совокупности. Ясно, что последний равен $(n+1)/2$. Поэтому в качестве меры отступления от чистой случайности можно взять величину

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} n_j \left(R_{\cdot k} - \frac{n+1}{2} \right)^2,$$

которая называется *статистикой Краскела—Уоллиса*.

Более «грубой» версией критерия Краскела—Уоллиса является *медианный* тест. Для его применения в каждом столбце таблицы вычисляют количество наблюдений, меньших и, соответственно, больших общей медианы. Для полученной таблицы, размером $2 \times k$ по критерию χ^2 с $k-1$ степенями свободы проверяется гипотеза об одинаковом распределении наблюдений, меньших общей медианы, и наблюдений больших ее.

В качестве примера выберите файл *Kruskal.sta* из папки *Example*. Обратите внимание, что все измеряемые значения проранжированы. В модуле **Nonparametric Statistics** выберите **Kruskal—Wallis ANOVA, median test**. В открывшемся окне, щелкнув мышью по кнопке **Variable**, укажите фактор CONDITN и зависимую переменную PERFRMNC. Вы увидите результаты анализа с помощью критерия Краскела—Уоллиса, затем, щелкнув на **Continue** — результаты **median test**. В обоих случаях p-значение меньше уровня значимости

$\alpha = 0{,}05$, т. е. гипотеза об однородности отвергается. Это подтверждают и гистограммы зависимой переменной для трех уровней фактора и, так называемые, «ящики с усами». Чтобы эту информацию увидеть, надо нажать соответственно на **Categorised histogram** и **Box & Whisker**. В данной задаче при нажатии на **Box & Whisker** вы получите для трех уровней фактора диаграмму, на которой указаны положения наименьшего и наибольшего значения измеряемого признака, а выделенный прямоугольником диапазон содержит 50% средних по величине ранга значений.

Словарь используемых статистических терминов в модуле ANOVA/MANOVA

Таблица 3

Таблица однофакторного дисперсионного анализа

Источник дисперсии	Сумма квадратов	Степени свободы	Средний квадрат	F-отношение
Между группами	$SS_B = \sum_{i=1}^{k} n_i(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2$	$v_B = k - 1$	$MS_B = \frac{SS_B}{v_B}$	$F = \frac{MS_B}{MS_W}$
Внутри групп	$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i\cdot})^2$	$v_W = n - k$	$MS_W = \frac{SS_W}{v_W}$	
Полная	$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{\cdot\cdot})^2$	$v_T = n - 1$		

Смысл используемых индексов следующий: B — Between (между), W — Within (внутри), T — Total (полная). В первом столбце таблицы перечисляются три источника дисперсии — между группами, внутри групп и полная, во втором — суммы квадратов для этих трех источников. Каждый средний квадрат вычисляется путем деления суммы квадратов на число степеней свободы, причем средний квадрат для полной дисперсии в таблице не приводится. Величина SS_W называется *остаточной суммой квадратов* или *суммой квадратов ошибок*.

Данные, приведенные в таблице, используются для проверки гипотез и построения различных доверительных интервалов. Так, MS_W является несмещенной оценкой общей дисперсии σ^2, $100(1-\alpha)\%$-ным доверительным интервалом для μ_i будет

$$\overline{x}_{i\cdot} \pm t_{1-\frac{\alpha}{2}}(v_W)\sqrt{\frac{MS_W}{n_i}}, \quad i = 1, \ldots, k.$$

Для проверки гипотезы H_0: $\gamma = \sum_{i=1}^{k} c_i \mu_i = 0$ против альтернативы H_1: $\gamma \neq 0$ с уровнем значимости α образуем следующий $100(1-\alpha)\%$-ный доверительный интервал

$$\sum_{i=1}^{k} c_i \overline{x}_i \pm t_{1-\frac{\alpha}{2}}(v_W) \sqrt{MS_W \sum_{i=1}^{k} \frac{c_i^2}{n_i}}.$$

Если этот доверительный интервал содержит ноль, принимается гипотеза H_0, в противном случае она отвергается с уровнем значимости α. В частности, для разности $\left(\mu_i - \mu_j\right)$ имеем $100(1-\alpha)\%$-ный доверительный интервал

$$(\overline{x}_{i\cdot} - \overline{x}_{j\cdot}) \pm t_{1-\frac{\alpha}{2}}(v_W) \sqrt{MS_W \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}, \quad i, j = 1, \ldots, k, \quad i \neq j,$$

где $t_{1-\frac{\alpha}{2})}(v_W)$ есть $(1-\alpha)$-квантиль t-распределения Стьюдента с v_W степенями свободы.

Если необходимо проверить одновременно несколько аналогичных гипотез, то уровень значимости совокупности всех критериев обычно будет сильно отличаться от α. Чтобы обойти эту трудность, можно использовать одну из трех *процедур множественного сравнения* для всех критериев, которая позволяет сохранить α в качестве общего уровня значимости: метод Шеффе, метод Тьюки и множественный t-метод.

Для проверки гипотезы H_0: $\gamma = \sum_{i=1}^{k} c_i \mu_i = 0$ против альтернативы H_1: $\gamma \neq 0$ с уровнем значимости α образуем следующие $100(1-\alpha)\%$-ный доверительные интервалы для каждого представляющего интерес контраста, причем общим для всех критериев уровнем значимости остается α.

Метод Шеффе

$$\sum_{i=1}^{k} c_i \overline{x}_{i\cdot} \pm T,$$

где

$$T^2 = kMS_W F_{1-\alpha}(k, n-k) \sum_{i=1}^{k} \frac{c_i^2}{n_i}.$$

Метод Тьюки (применим только в случае равных объемов выборок $n_1 = n_2 = \ldots = n_k = m$)

$$\sum_{i=1}^{k} c_i \overline{x}_{i\cdot} \pm T,$$

где

$$T = \frac{1}{2}\sqrt{\frac{MS_W}{m}} q_{1-\alpha} \sum_{i=1}^{k} |c_i|,$$

а $q_{1-\alpha}$ есть $(1-\alpha)$-квантиль распределения стьюдентизированного размаха с k и $v = n - k$ степенями свободы.

Множественный *t*-метод

$$\sum_{i=1}^{k} c_i \overline{x}_{i\cdot} \pm t_{1-\frac{\alpha}{2k}}(v_W) \sqrt{MS_W \sum_{i=1}^{k} \frac{c_i^2}{n_i}},$$

где $t_{1-\frac{\alpha}{2k}}(v_W)$ есть $\left(1 - \frac{\alpha}{2k}\right)$-квантиль t-распределения Стьюдента с v_W степенями свободы. Здесь p — число *заранее* выбранных контрастов, а для первых двух методов множество проверяемых контрастов может быть любым.

Замечание. В среднем для простых контрастов, содержащих не более трех математических ожиданий, метод Тьюки дает более короткие доверительные интервалы, чем метод Шеффе, для анализа же контрастов, содержащих более трех математических ожиданий метод Шеффе лучше.

5. Модуль Multiple Regression (Множественная регрессия)

Если вы хотите построить зависимость между многомерными переменными, подобрать простую линейную модель и оценить ее адекватность, воспользуйтесь модулем **Multiple Regression** (**Множественная регрессия**). В качестве примера рассмотрим оптовые цены на вина в зависимости от года закладки вина. Цены указаны в долларах за одну бутылку. Файл с исходными данными назовем *vino.sta*. Естественно предположить, что между возрастом вина и его ценой есть некоторая зависимость. Интуитивно понятно, чем более выдержанное вино, тем оно дороже. Итак, мы хотим приближенно выразить зависимость между этими величинами. Зачем это нам нужно? Например, имея такую формулу, можно прогнозировать стоимость вин на следующем аукционе, или, если вы планируете продавать вино, год закладки которого не представлен в таблице, то это поможет вам грамотно назначить цену.

Регрессионный анализ предназначен для изучения связей между одной зависимой переменной (в нашем случае *ценой*) и несколькими (в нашем случае одной переменной — *возрастом вина*) независимыми переменными. В линейном регрессионном анализе эта связь предполагается линейной. В самом простом случае, в линейной регрессионной модели имеется две переменные X и Y и мы выражаем Y как линейную функцию от X:

$$Y_i = \beta_1 \cdot X_i + \beta_0 + \varepsilon_i; \quad i = 1, 2, \ldots, n.$$

Здесь β_1 и β_0 — неизвестные коэффициенты регрессии, ε_i — ошибки измерения.

Уравнение называется уравнением линейной регрессии. В результате исследований мы, конечно, не сможем точно определить коэффициенты регрессии, а лишь получим их оценки, которые обозначим соответственно b_1 и b_0. Итак, получили приближенную зависимость $Y = b_1 \cdot X + b_0$.

Перейдем от исходных переменных Год и Цена к переменным Возраст и Цена_Лог по соответствующим формулам:

$$\text{Возраст} = 1972 - \text{Год}; \quad \text{Цена\_Лог} = \ln(\text{Цена}).$$

После этого таблица будет содержать четыре переменные и иметь вид, показанный на рис. 21.

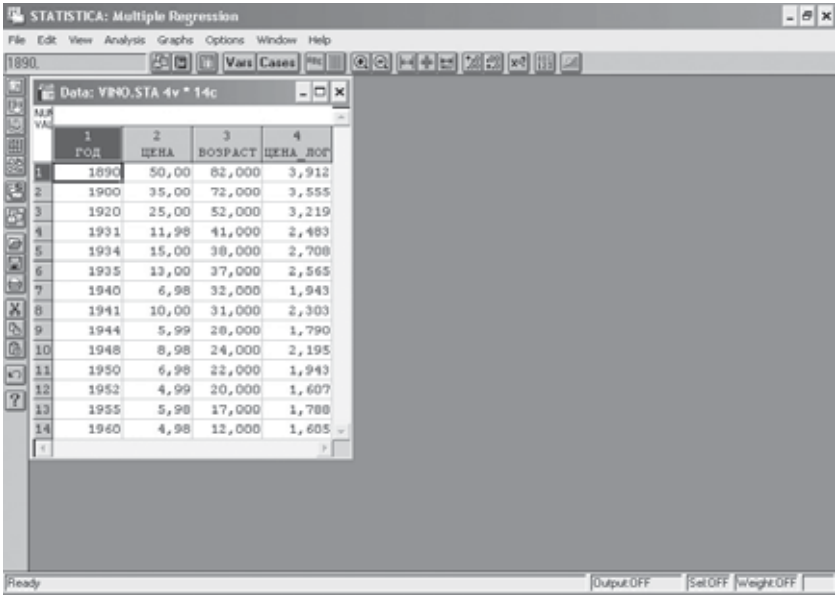


Рис. 21

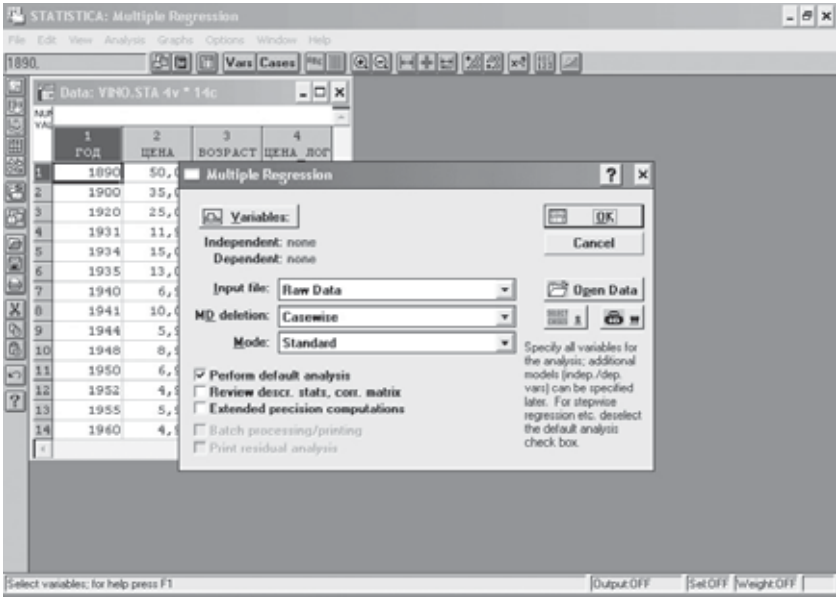

STATISTICA: Multiple Regression
Data: VINO.STA 4v * 14c
Multiple Regression
Variables:
Independent: none
Dependent: none
Input file: Raw Data
MD deletion: Casewise
Mode: Standard
Perform default analysis
Review descr. stats, corr. matrix
Extended precision computations
Batch processing/printing
Print residual analysis
OK
Cancel
Open Data
Specify all variables for the analysis; additional models (indep./dep. vars) can be specified later. For stepwise regression etc. deselect the default analysis check box.
Select variables; for help press F1

Рис. 22

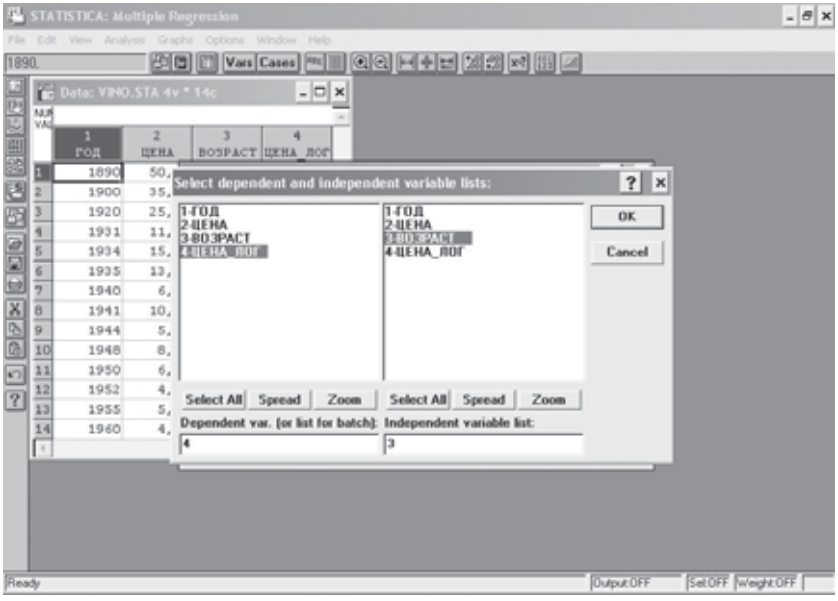

STATISTICA: Multiple Regression
Data: VINO.STA 4v * 14c
Select dependent and independent variable lists:
1-ГОД
2-ЦЕНА
3-ВОЗРАСТ
4-ЦЕНА_ЛОГ
OK
Cancel
Select All
Spread
Zoom
Dependent var. (or list for batch):
Independent variable list:
Ready

Рис. 23

Нажмите кнопку **Open Data** (**Открыть данные**) и откройте созданный файл данных *vino.sta*. Далее выберите переменные для анализа. Выбор переменных осуществляется с помощью кнопки **Variables** (**Переменные**), находящейся в левом верхнем углу панели (рис. 22).

Далее нужно выбрать переменные для анализа. **Зависимой** переменной у нас является Цена_Лог, а **независимой** — Возраст. Для их задания необходимо нажать кнопку **Variables** (**Переменные**) на Стартовой панели. В открывшемся окне **Select dependent and independent variable list** (**Выбрать зависимую переменную и список независимых переменных**) в качестве зависимой переменной указать Цена_Лог, а в качестве списка независимых переменных указать только переменную Возраст (рис. 23).

Нажмите кнопку **OK** в правом верхнем углу и вы снова окажетесь в Стартовой панели модуля **Множественная регрессия**. В **Стартовой панели** вы можете задать дополнительные опции и параметры анализа. Например, вы можете выбрать подмножество случаев для анализа, либо приписать вес переменным — эти опции относятся к исходным данным. Вы также можете задать опции, которые относятся непосредственно к статистической процедуре: задать правило обработки пропущенных данных, выбрать метод обработки по

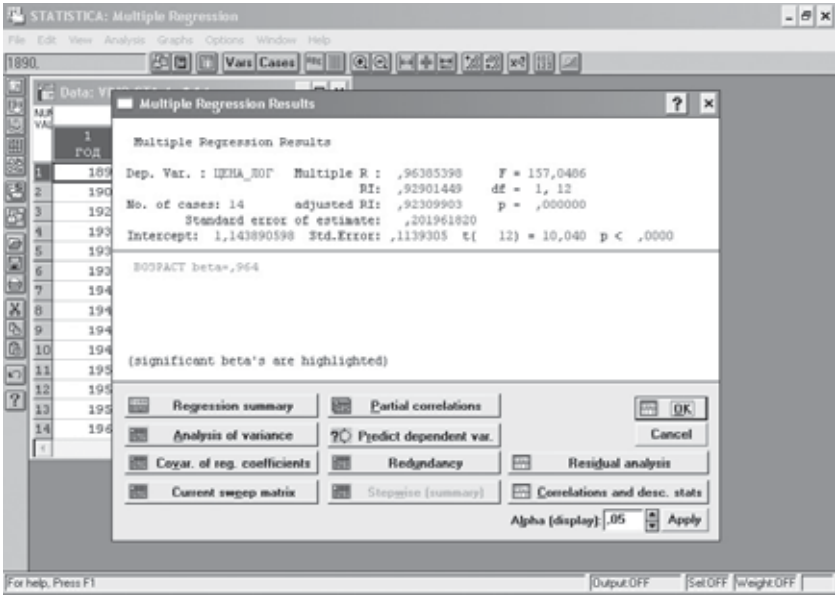


Рис. 24

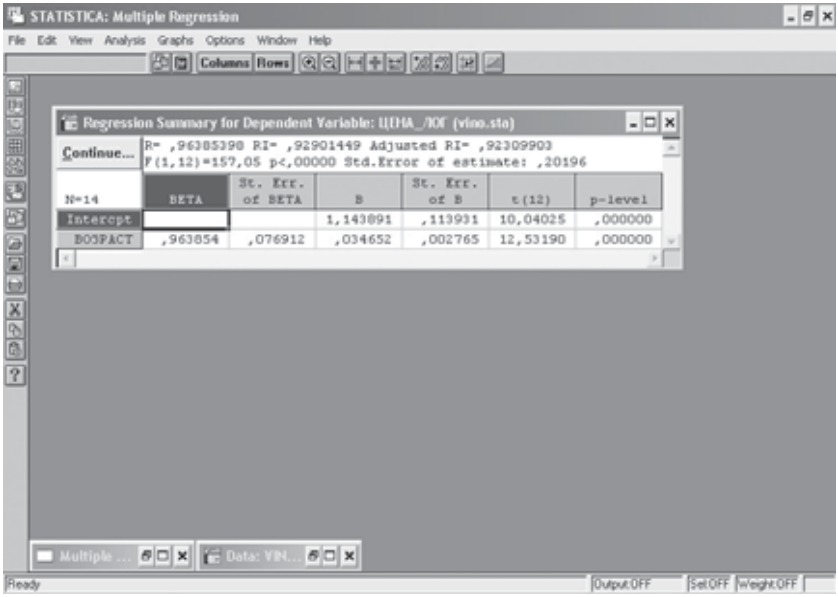


Рис. 25

умолчанию. Мы выбрали опцию **Расчет** с расширенной точностью и выбор метода обработки по умолчанию, но этот выбор не обязателен. В **Стартовой панели** нажмите на кнопку **OK** и через секунду появится окно с результатами анализа. Окно результатов имеет следующую структуру. Верхняя часть окна — информационная, а нижняя содержит функциональные кнопки, с помощью которых можно всесторонне рассмотреть результаты анализа.

В информационной части, прежде всего, смотрим на значение коэффициента детерминации R1. В данном случае $R1 = 0{,}929$. Это значит, что построенная регрессия объясняет 92,9% разбросов значений Цена_Лог относительно среднего. Это хороший результат (рис. 24, 25).

6. Исследование остатков

Поясним, прежде всего, что такое остатки. Задавая различные значения Возраста, мы по модели получим **Predicted values (Предсказанные значения)** переменной Цена_Лог. Разности между исходными и предсказанными значениями зависимой переменной называются остатками. Исследуя остатки, вы можете оценить адекватность выбранной модели. Для этого нажмите в окне **Ре-**

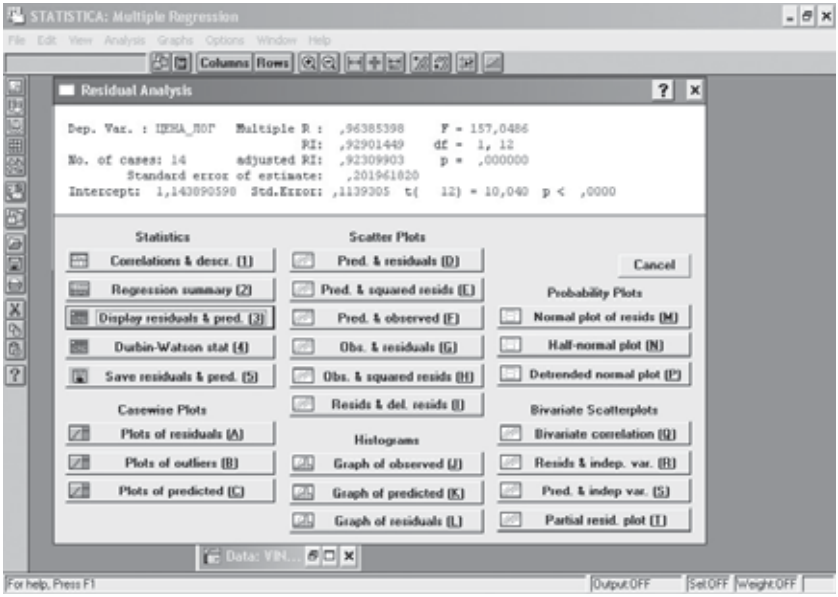


Рис. 26

зультаты множественной регрессии кнопку **Residual analysis** (**Анализ остатков**). С помощью функциональных кнопок можно всесторонне просмотреть остатки в графической и табличной формах (рис. 26).

В частности, из графика остатков на вероятностной бумаге видно, что они достаточно хорошо ложатся на прямую, которая соответствует нормальному закону. Поэтому предположение о нормальном распределении ошибок является верным.

Из графика остатков относительно линии предсказанных значений видно, что в поведении остатков нет закономерности. Нет оснований говорить об их коррелируемости или отдельных выбросах. Все это говорит об адекватности выбранной модели (рис. 27, 28).

Рассмотрим более сложный пример множественной регрессии, в которой список независимых переменных содержит более одной переменной. Создадим файл с данными о капитальных затратах на строительство атомных электростанций с реактором водяного охлаждения. Данные собраны о 32 электростанциях США. Требуется установить зависимость этих затрат от ряда параметров, приведенных в таблице, предсказать стоимость строительства новой элек-

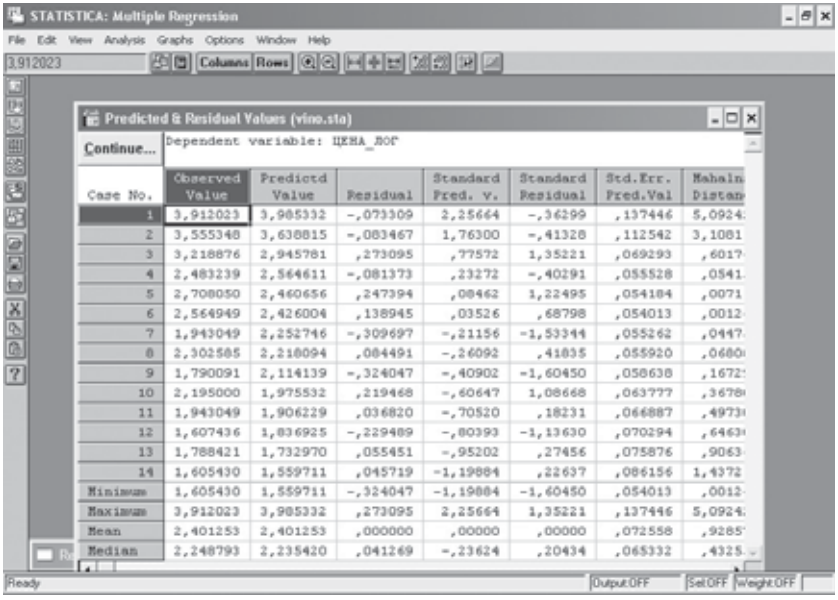

STATISTICA: Multiple Regression
File Edit View Analysis Graphs Options Window Help
Predicted & Residual Values (vino.sta)
Continue...
Dependent variable: ЦЕНА_БОТ
Case No. | Observed Value | Predictd Value | Residual | Standard Pred. v. | Standard Residual | Std.Err. Pred.Val | Mahaln Distan
1 | 3,912023 | 3,985332 | -,073309 | 2,25664 | -,36299 | ,137446 | 5,0924
2 | 3,555348 | 3,638815 | -,083467 | 1,76300 | -,41328 | ,112542 | 3,1081
3 | 3,218876 | 2,945781 | ,273095 | ,77572 | 1,35221 | ,069293 | ,6017
4 | 2,483239 | 2,564611 | -,081373 | ,23272 | -,40291 | ,055528 | ,0541
5 | 2,708050 | 2,460656 | ,247394 | ,08462 | 1,22495 | ,054184 | ,0071
6 | 2,564949 | 2,426004 | ,138945 | ,03526 | ,68798 | ,054013 | ,0012
7 | 1,943049 | 2,252746 | -,309697 | -,21156 | -1,53344 | ,055262 | ,0447
8 | 2,302585 | 2,218094 | ,084491 | -,26092 | ,41835 | ,055920 | ,0680
9 | 1,790091 | 2,114139 | -,324047 | -,40902 | -1,60450 | ,058638 | ,1672
10 | 2,195000 | 1,975532 | ,219468 | -,60647 | 1,08668 | ,063777 | ,3678
11 | 1,943049 | 1,906229 | ,036820 | -,70520 | ,18231 | ,066887 | ,4973
12 | 1,607436 | 1,836925 | -,229489 | -,80393 | -1,13630 | ,070294 | ,6463
13 | 1,788421 | 1,732970 | ,055451 | -,95202 | ,27456 | ,075876 | ,9063
14 | 1,605430 | 1,559711 | ,045719 | -1,19884 | ,22637 | ,086156 | 1,4372
Minimum | 1,605430 | 1,559711 | -,324047 | -1,19884 | -1,60450 | ,054013 | ,0012
Maximum | 3,912023 | 3,985332 | ,273095 | 2,25664 | 1,35221 | ,137446 | 5,0924
Mean | 2,401253 | 2,401253 | ,000000 | ,00000 | ,00000 | ,072558 | ,9285
Median | 2,248793 | 2,235420 | ,041269 | -,23624 | ,20434 | ,065332 | ,4325
Ready
Output:OFF
Set:OFF
Weight:OFF

Рис. 27

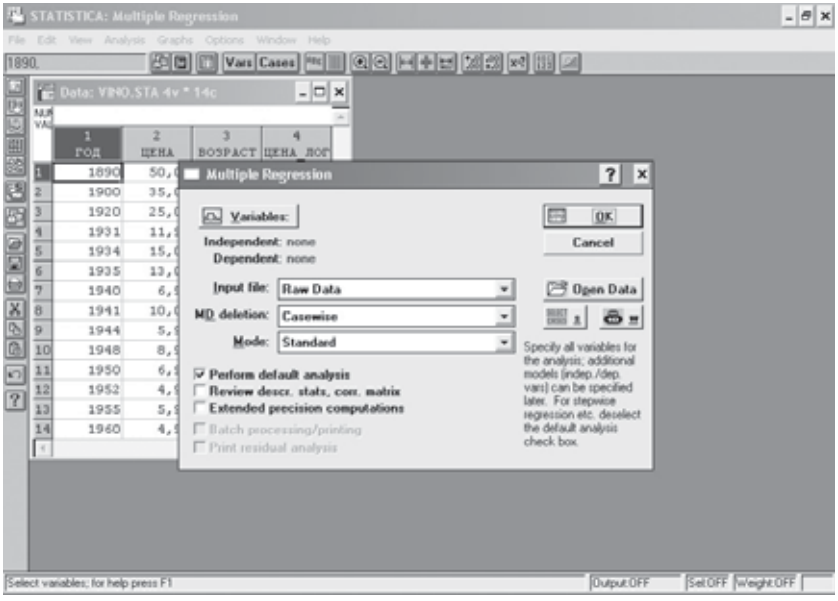

STATISTICA: Multiple Regression
File Edit View Analysis Graphs Options Window Help
Data: VINO.STA 4v * 14c
1 ГОД | 2 ЦЕНА | 3 ВОЗРАСТ | 4 ЦЕНА_БОТ
1890 | 50,
1900 | 35,
1920 | 25,
1931 | 11,
1934 | 15,
1935 | 13,
1940 | 6,
1941 | 10,
1944 | 5,
1948 | 8,
1950 | 6,
1952 | 4,
1955 | 5,
1960 | 4,
Multiple Regression
Variables:
Independent: none
Dependent: none
OK
Cancel
Input file: Raw Data
Open Data
MD deletion: Casewise
Mode: Standard
Perform default analysis
Review descr. stats, corr. matrix
Extended precision computations
Batch processing/printing
Print residual analysis
Specify all variables for the analysis; additional models (indep./dep. vars) can be specified later. For stepwise regression etc. deselect the default analysis check box.
Select variables; for help press F1
Output:OFF
Set:OFF
Weight:OFF

Рис. 28

тростанции, выделить наиболее значимые величины, влияющие на цену станции. Структура данных следующая (рис. 29):

C — цена в млрд. долларов, приведенная к курсу 1974 года;

D — срок разрешения на строительство;

T1 — промежуток времени между обращением за разрешением и получением разрешения на строительство;

T2 — промежуток времени между получением оперативной лицензии и получением разрешения на строительство;

S — номинальная мощность электростанции (МВт);

PR — наличие в той же самой местности ранее построенной электростанции на PBO, если есть, то PR = 1, иначе PR = 0;

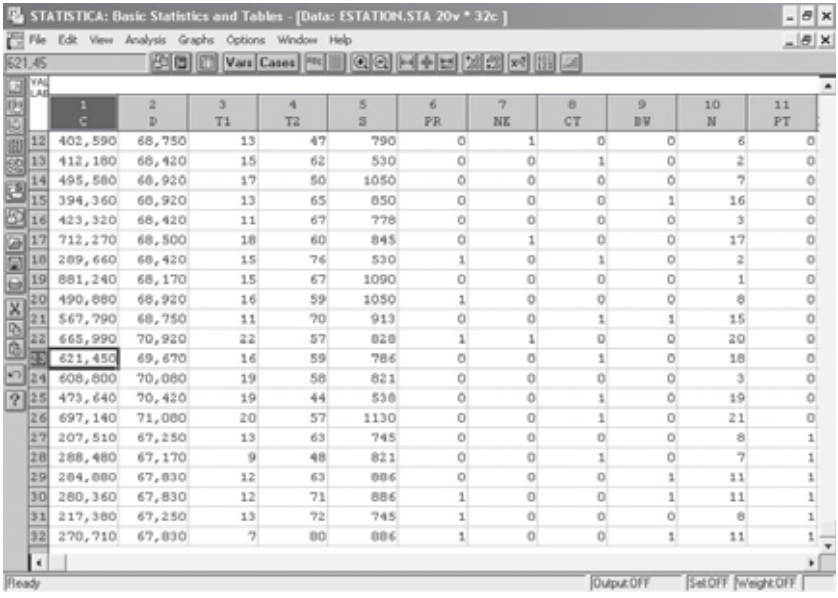
NE — характеристика района, в котором строится станция;

CT — использование нагревательной башни, если 1 — то используется, если 0, то нет;

BW — использование силовой установки фирмы Babcock—Wilcox, если 1 — то используется, если 0, то нет;

N — суммарное число электростанций, построенное архитектором-инженером станции;

PT — электростанция строится под частичным надзором, если такой надзор есть, то PT = 1, иначе PT = 0.



	1 C	2 D	3 T1	4 T2	5 S	6 PR	7 NE	8 CT	9 BW	10 N	11 PT
12	402,590	68,750	13	47	790	0	1	0	0	6	0
13	412,180	68,420	15	62	530	0	0	1	0	2	0
14	495,580	68,920	17	50	1050	0	0	0	0	7	0
15	394,360	68,920	13	65	850	0	0	0	1	16	0
16	423,320	68,420	11	67	778	0	0	0	0	3	0
17	712,270	68,500	18	60	845	0	1	0	0	17	0
18	289,660	68,420	15	76	530	1	0	1	0	2	0
19	881,240	68,170	15	67	1090	0	0	0	0	1	0
20	490,880	68,920	16	59	1050	1	0	0	0	8	0
21	567,790	68,750	11	70	913	0	0	1	1	15	0
22	665,990	70,920	22	57	828	1	1	0	0	20	0
23	621,450	69,670	16	59	786	0	0	1	0	18	0
24	608,800	70,080	19	58	821	0	0	0	0	3	0
25	473,640	70,420	19	44	538	0	0	1	0	19	0
26	697,140	71,080	20	57	1130	0	0	1	0	21	0
27	207,510	67,250	13	63	745	0	0	0	0	8	1
28	288,480	67,170	9	48	821	0	0	1	0	7	1
29	284,080	67,830	12	63	886	0	0	0	1	11	1
30	280,360	67,830	12	71	886	1	0	0	1	11	1
31	217,380	67,250	13	72	745	1	0	0	0	8	1
32	270,710	67,830	7	80	886	1	0	0	1	11	1

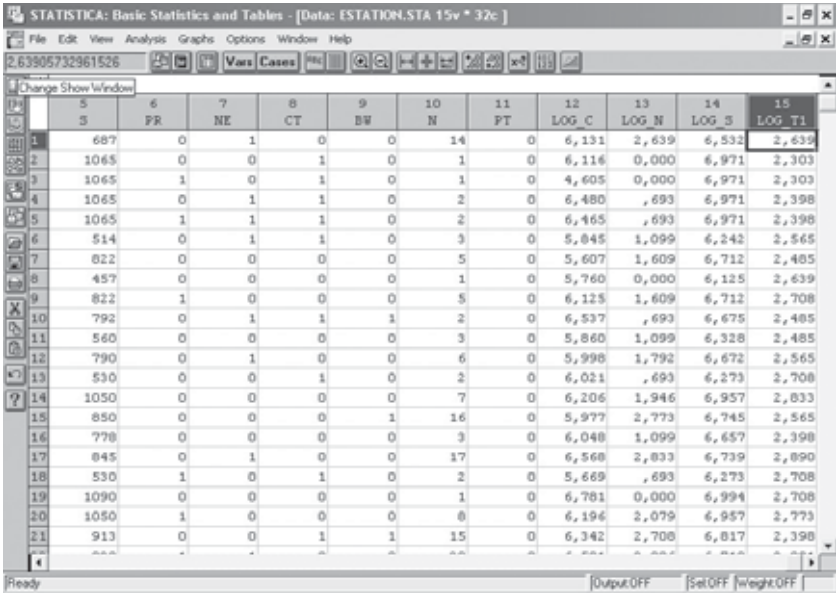
Рис. 29

Воспользуемся векторными обозначениями. Обозначим n-мерный вектор зависимой переменной через $\boldsymbol{Y}$, а через $\boldsymbol{X}$ — матрицу независимых переменных. Размер матрицы равен $n \times p$, где n — число наблюдений, а p — число независимых переменных.

Тогда можно записать $\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Здесь $\boldsymbol{\varepsilon}$ — вектор остатков, а $\boldsymbol{\beta}$ — вектор коэффициентов регрессии, для которого в результате получим вектор оценок $\boldsymbol{b}$.

В данном примере зависимая переменная C — цена, а в список независимых входят все остальные переменные этого примера. Предварительно выполним преобразования переменных. Вместо C, N, S, T1, T2 будем рассматривать их натуральные логарифмы (рис. 30).

Для начала статистического анализа необходимо вызвать *Стартовую панель* модуля. Для этого надо войти в меню **Analysis** и выбрать команду **Startup Panel**. Далее необходимо выбрать переменные для анализа. Для их задания необходимо нажать кнопку **Variables** (**Переменные**) на Стартовой панели. В открывшемся окне **Select dependent and independent variable list** (**Выбрать зависимую переменную и список независимых переменных**) в качестве зависимой переменной указать Log_C, а в качестве списка независимых переменных



	5 S	6 PR	7 NE	8 CT	9 BW	10 N	11 PT	12 LOG_C	13 LOG_N	14 LOG_S	15 LOG_T1
1	687	0	1	0	0	14	0	6,131	2,639	6,532	2,639
2	1065	0	0	1	0	1	0	6,116	0,000	6,971	2,303
3	1065	1	0	1	0	1	0	4,605	0,000	6,971	2,303
4	1065	0	1	1	0	2	0	6,480	,693	6,971	2,398
5	1065	1	1	1	0	2	0	6,465	,693	6,971	2,398
6	514	0	1	1	0	3	0	5,845	1,099	6,242	2,565
7	822	0	0	0	0	5	0	5,607	1,609	6,712	2,485
8	457	0	0	0	0	1	0	5,760	0,000	6,125	2,639
9	822	1	0	0	0	5	0	6,125	1,609	6,712	2,708
10	792	0	1	1	1	2	0	6,537	,693	6,675	2,485
11	560	0	0	0	0	3	0	5,860	1,099	6,328	2,485
12	790	0	1	0	0	6	0	5,998	1,792	6,672	2,565
13	530	0	0	1	0	2	0	6,021	,693	6,273	2,708
14	1050	0	0	0	0	7	0	6,206	1,946	6,957	2,833
15	850	0	0	0	1	16	0	5,977	2,773	6,745	2,565
16	778	0	0	0	0	3	0	6,048	1,099	6,657	2,398
17	845	0	1	0	0	17	0	6,568	2,833	6,739	2,890
18	530	1	0	1	0	2	0	5,669	,693	6,273	2,708
19	1090	0	0	0	0	1	0	6,781	0,000	6,994	2,708
20	1050	1	0	0	0	8	0	6,196	2,079	6,957	2,773
21	913	0	0	1	1	15	0	6,342	2,708	6,817	2,398

Рис. 30

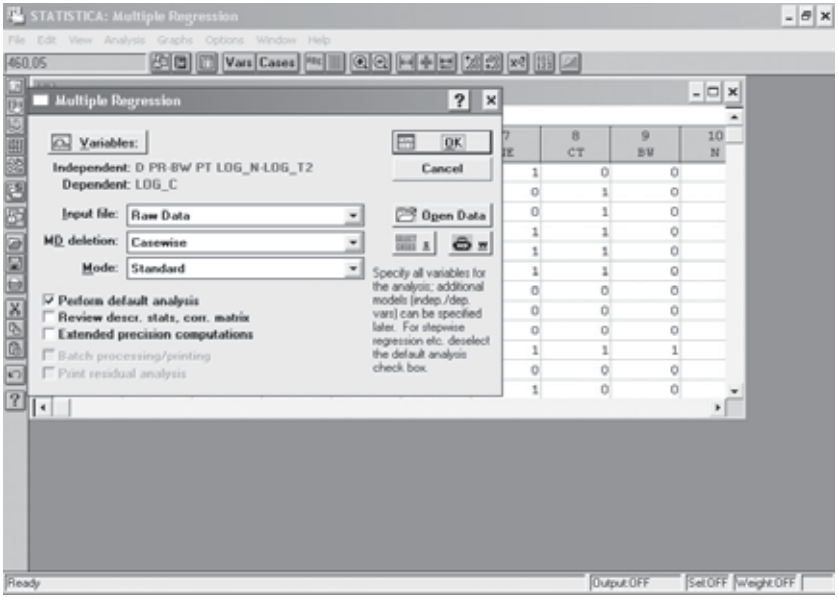

STATISTICA: Multiple Regression
Multiple Regression
Variables:
Independent: D PR-BW PT LOG_N-LOG_T2
Dependent: LOG_C
Input file: Raw Data
MD deletion: Casewise
Mode: Standard
Perform default analysis
Review descr. stats, corr. matrix
Extended precision computations
Batch processing/printing
Print residual analysis
OK
Cancel
Open Data
Specify all variables for the analysis; additional models (indep./dep. vars) can be specified later. For stepwise regression etc. deselect the default analysis check box.
Ready

Рис. 31

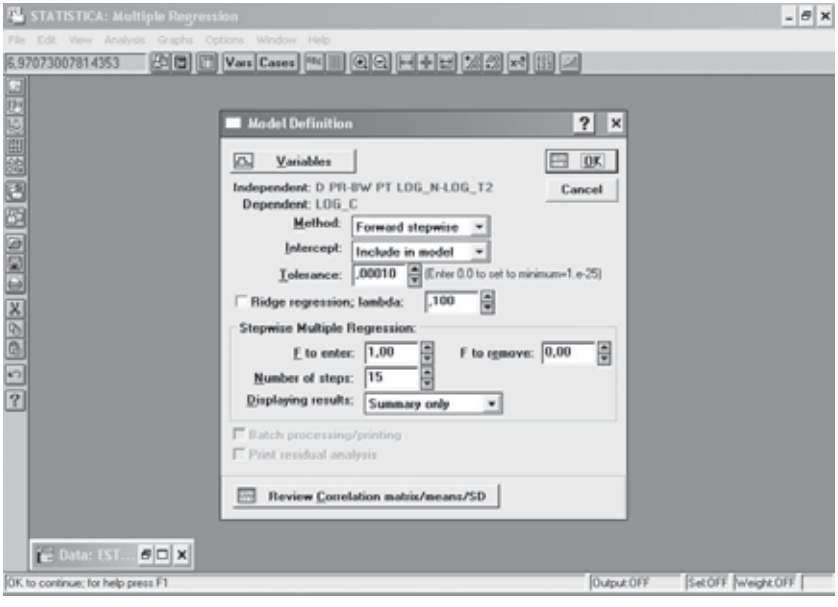

STATISTICA: Multiple Regression
Model Definition
Variables
Independent: D PR-BW PT LOG_N-LOG_T2
Dependent: LOG_C
Method: Forward stepwise
Intercept: Include in model
Tolerance: ,00010
(Enter 0.0 to set to minimum=1.e-25)
Ridge regression; lambda: ,100
Stepwise Multiple Regression:
F to enter: 1,00
F to remove: 0,00
Number of steps: 15
Displaying results: Summary only
Batch processing/printing
Print residual analysis
Review Correlation matrix/means/SD
OK
Cancel
OK to continue; for help press F1

Рис. 32

указать переменные D, PR, NE, CT, BW, PT, Log_N, Log_S, Log_T1, Log_T2 (рис. 31).

Нажмите кнопку **OK** в правом верхнем углу и вы снова окажетесь в Стартовой панели модуля **Множественная регрессия**. В Стартовой панели вы можете задать дополнительные опции и параметры анализа. Например, вы можете отменить выбор метода анализа по умолчанию. После нажатия на кнопку **OK** появится следующее диалоговое окно **Model Definition** для выбора метода анализа (рис. 32).

В списке методов выберите один из пошаговых регрессионных методов, например, **Forward stepwise** (**Включение**), значения остальных параметров оставьте неизменными. Нажмите **OK**. Система произведет вычисления, и на экране появится окно с результатами (рис. 33, 34, 35). (Метод пошаговой регрессии состоит в том, что на каждом шаге в анализ включается или исключается некоторая независимая переменная. Таким образом, формируется подмножество наиболее «значимых» переменных, которые удовлетворительно описывают зависимую переменную. Существует несколько разновидностей метода пошаговой регрессии.)

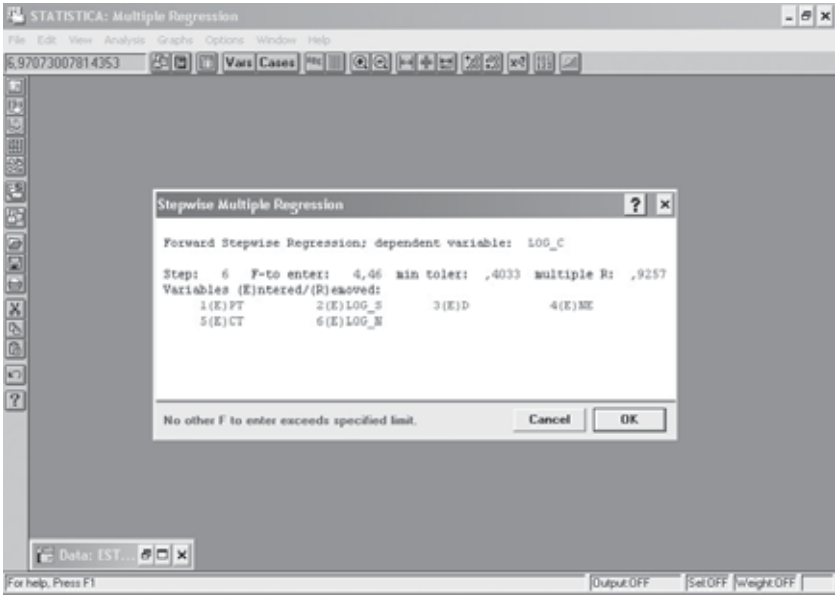


Рис. 33

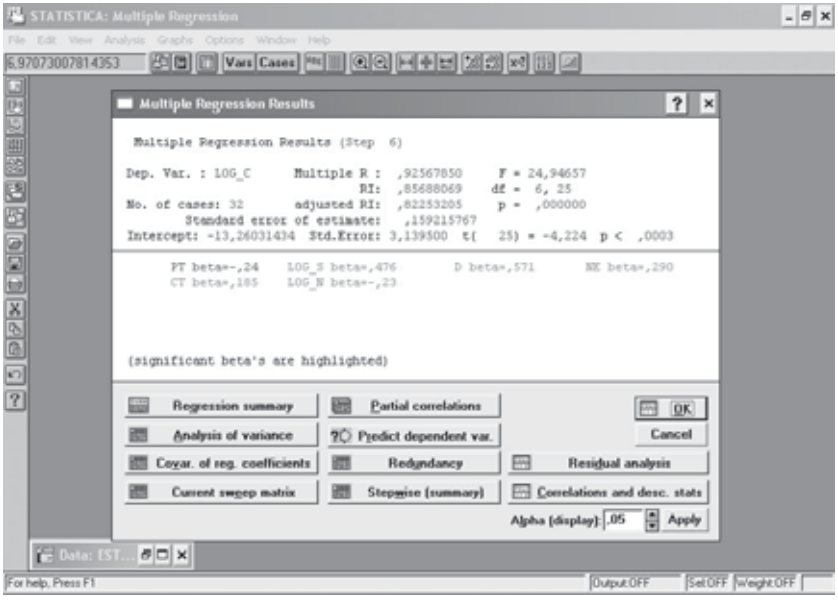

STATISTICA: Multiple Regression
Multiple Regression Results
Multiple Regression Results (Step 6)
Dep. Var. : LOG_C Multiple R : ,92567850 F = 24,94657
RI: ,85688069 df = 6, 25
No. of cases: 32 adjusted RI: ,82253205 p = ,000000
Standard error of estimate: ,15921576
Intercept: -13,26031434 Std.Error: 3,139500 t(25) = -4,224 p < ,0003
PT beta=-,24 LOG_S beta=,476 D beta=,571 NE beta=,290
CT beta=,185 LOG_N beta=-,23
(significant beta's are highlighted)
Regression summary
Partial correlations
OK
Analysis of variance
Predict dependent var.
Cancel
Covar. of reg. coefficients
Redundancy
Residual analysis
Current sweep matrix
Stepwise (summary)
Correlations and desc. stats
Alpha (display): .05
Apply
Data: EST...
For help, Press F1
Output:OFF
Set:OFF
Weight:OFF

Рис. 34

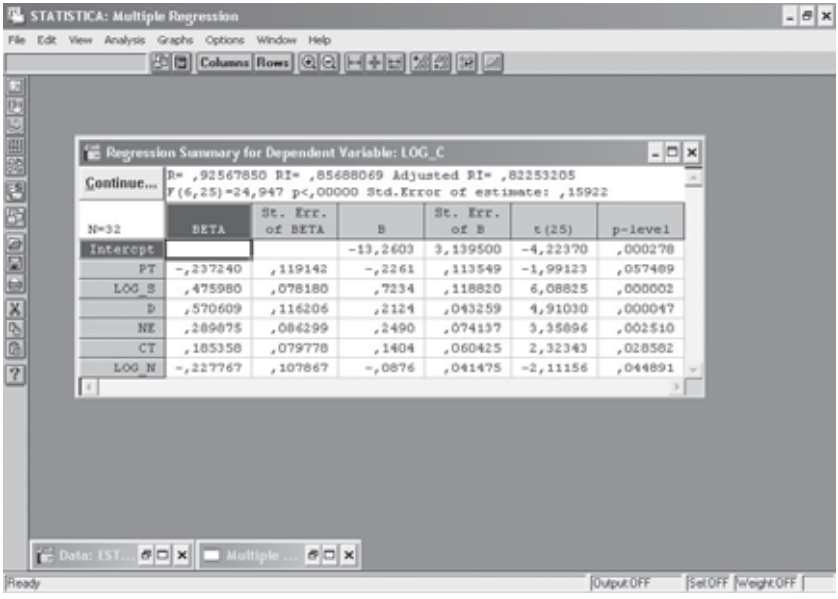

STATISTICA: Multiple Regression
File Edit View Analysis Graphs Options Window Help
Columns Rows
Regression Summary for Dependent Variable: LOG_C
Continue...
R= ,92567850 RI= ,85688069 Adjusted RI= ,82253205
F(6,25)=24,947 p<,00000 Std.Error of estimate: ,15922
N=32
BETA
St. Err. of BETA
B
St. Err. of B
t(25)
p-level
Intercept -13,2603 3,139500 -4,22370 ,000278
PT -,237240 ,119142 -,2261 ,113549 -1,99123 ,057489
LOG_S ,475980 ,078180 ,7234 ,118820 6,08825 ,000002
D ,570609 ,116206 ,2124 ,043259 4,91030 ,000047
NE ,289875 ,086299 ,2490 ,074137 3,35896 ,002510
CT ,185358 ,079778 ,1404 ,060425 2,32343 ,028582
LOG_N -,227767 ,107867 -,0876 ,041475 -2,11156 ,044891
Data: EST...
Multiple ...
Ready
Output:OFF
Set:OFF
Weight:OFF

Рис. 35

Словарь используемых статистических терминов в модуле «Множественная регрессия»

Таблица 4

Таблица дисперсионного анализа для модели множественной линейной регрессии

Источник дисперсии	Сумма квадратов	Степени свободы	Средний квадрат	F-отношение
Регрессия	$SS_d = \sum_{i=1}^{p} b_i \sum(x_{ij} - \overline{x}_i)y_i$	$v_D = p$	$MS_D = \frac{SS_D}{v_D}$	$F = \frac{MS_D}{MS_R}$
Отклонение от регрессии	$SS_R = SS_T - SS_D$	$v_R = n - p - 1$	$MS_R = s^2 = \frac{SS_R}{v_R}$	
Полная	$SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2$	$v_T = n - 1$		

Введем обозначение $S^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2$ для суммы квадратов отклонений значений $\boldsymbol{Y}$ от линии регрессии. То есть S^2 — это мера ошибки, связанной с «подгонкой» выборочных данных посредством модели линейной регрессии. Вектор МНК-оценок $\boldsymbol{b}$ минимизирует S^2. Величина SS_R (остаточная сумма квадратов) есть значение S^2, которое получается при подстановке вектора $\boldsymbol{b}$ вместо $\boldsymbol{\beta}$.

Если эту величину разделить на число степеней свободы остатков (или ошибок) $v_R = n - p - 1$, получится несмещенная оценка дисперсии ошибок Φ^2, называемая остаточным средним квадратом ошибки MS_R, иногда эту величину обозначают s^2.

Multiple R — выборочный коэффициент множественной корреляции. Множественный коэффициент корреляции есть максимальное значение парного коэффициента корреляции между $\boldsymbol{Y}$ и линейной комбинацией $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$. Более того, $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{X}_1 + \ldots + \boldsymbol{\beta}_p \boldsymbol{X}_p$ является линейной комбинацией, на которой этот максимум достигается.

R1 = (Multiple R)^2 = SS_D/SS_T — выборочный коэффициент детерминации, он показывает долю общей дисперсии (общего разброса) относительно выборочного среднего зависимой переменной, которая объясняется построенной регрессией. Итак, чем больше R1, тем лучше модель аппроксимирует $\boldsymbol{Y}$. Величина $(1 - \text{R1})^{1/2}$ есть доля стандартного отклонения $\boldsymbol{Y}$, оставшаяся «необъясненной» зависимостью от $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$.

Adjusted R1 = 1 – ((1 – R1) · *n*/(*n* – *p*)) — скорректированный выборочный коэффициент детерминации, где n — число наблюдений, а p — число независимых переменных плюс 1, так как в модель включен свободный член.

p — вероятность того, что при выполнении гипотезы H_0 статистика критерия g принимает значение g_0 или даже более экстремальное, чем g_0. (Экстремальные значения определяются критической областью.) Эта вероятность называется ***p*-значением**. Если p меньше, чем α, то гипотеза H_0 отвергается с уровнем значимости α, в противном случае гипотеза H_0 принимается. Эта процедура эквивалентна проверке того, попадает ли вычисленное значение критерия в критическую область для уровня значимости α. Если попадает, то гипотеза H_0 отвергается с уровнем значимости α, в противном случае гипотеза H_0 принимается.

$\boldsymbol{p = P\{g \leqslant g_0\}}$ — для левосторонней критической области;

$\boldsymbol{p = P\{g \geqslant g_0\}}$ — для правосторонней критической области;

$\boldsymbol{p = 2 \cdot \min[P\{g \leqslant g_0\}, P\{g \geqslant g_0\}]}$ — для двусторонней критической области.

Std. Err of estimate — стандартная ошибка оценки — мера рассеяния наблюдаемых значений относительно регрессионной прямой, т. е. s.

Std. Err — стандартная ошибка оценки b_k в уравнении регрессии, $k = 0, 1, \ldots, p$, т. е. оценка стандартного отклонения b_k от β_k. Обозначим ее $s(b_k)$. Так как каждая из этих величин является функцией от MS_R и имеет v_R степеней свободы, то $100(1-\alpha)\%$-ный доверительный интервал для β_k есть $b_{k\pm}s(b_k)t_{1-(\alpha/2)}(v_R)$, $k = 0, 1, \ldots, p$.

df — число степеней свободы критерия;

***t*(*df*)** — значение t-критерия (t-критерий используется для проверки гипотезы о равенстве нулю коэффициента регрессии β_k, $k = 0, 1, \ldots, p$). Соответствующее p-значение есть удвоенная площадь под кривой плотности распределения $t(v_R)$ справа от точки, соответствующей вычисленному значению $|t|$. Рассмотрим коэффициент корреляции между $\boldsymbol{Y}$ и $\boldsymbol{X}_m$ при фиксированных значениях остальных переменных $\boldsymbol{X}_k$, где $k = 1, \ldots, p$ и $k \neq m$. Он называется **частным коэффициентом корреляции** и обозначается $\boldsymbol{r_{gh \cdot c}}$, где $\boldsymbol{g} = \boldsymbol{Y}$, $\boldsymbol{h} = \boldsymbol{X}_m$, а $\boldsymbol{c}$ — множество всех остальных переменных $\boldsymbol{X}_k$, где $k = 1, \ldots, p$ и $k \neq m$. Оценки $r_{gh \cdot c}$ частных коэффициентов корреляции $\Delta_{gh \cdot c}$ отсутствуют в выходных данных модуля, но их легко получить, используя t-статистику для проверки гипотезы H_0: $\boldsymbol{\beta}_m = 0$, а именно,

$$r_{gh \cdot c} = t/(t^2 + n - p - 1)^{1/2}.$$

Для проверки гипотезы H_0: $r_{gh\cdot c} = 0$ можно использовать статистику

$$t = (r_{gh\cdot c}(n - p - 3)^{1/2})/(1 - r^2_{gh\cdot c})^{1/2}.$$

Если H_0 истинна, статистика имеет t-распределение Стьюдента с $(n - p - 3)$ степенями свободы.

Для проверки гипотезы H_0: $r_{gh\cdot c} = \rho_0$, где ρ_0 — заданное значение, можно воспользоваться преобразованием Фишера

$$v = \ln((1 + r_{gh\cdot c})/(1 - r_{gh\cdot c}))/2.$$

Статистикой критерия является $z = (v - \mu_v)/\sigma$.

Здесь $\mu_v = \ln((1 + \rho_0)/(1 - \rho_0))/2$, а $\sigma_v = 1/(n - 3)$.

Если H_0 истинна и n велико, то распределение z аппроксимируется посредством $N(0, 1)$, p-значение зависит от альтернативной гипотезы и H_0 отвергается, если $p < \alpha$.

F(df) — значение F-критерия (F-критерий используется для проверки гипотезы о значимости регрессии, $F = MS_D/MS_R$. Соответствующее p-значение есть площадь под кривой плотности распределения $F(v_D, v_R)$ справа от точки, соответствующей вычисленному значению F.

Если необходимо проверить гипотезу о равенстве нулю некоторого подмножества из m коэффициентов регрессии, то надо поступить следующим образом: переупорядочить независимые переменные так, чтобы они были первыми в списке. Тогда проверка гипотезы H_0: $\beta_1 = \ldots = \beta_m = 0$ эквивалентна проверке гипотезы о том, что m переменных $X_1, \ldots, X_m$ не улучшают предсказание Y относительно предсказания, получаемого с помощью регрессии Y по $X_{m+1}, \ldots, X_p$. Для проверки H_0 сначала вычислим регрессию Y по переменным $X_{m+1}, \ldots, X_p$ и из анализа соответствующей таблицы дисперсионного анализа получим остаточную сумму квадратов SS'_R. Затем вычислим регрессию Y по переменным $X_1, \ldots, X_p$. Остаточную сумму квадратов и средний квадрат для этого случая обозначим через SS_R и MR_R соответственно. Тогда статистика критерия для H_0 имеет вид:

$$F = (SS'_R - SS_R)/(m \cdot MS_R).$$

Для гипотезы H_0 она имеет F-распределение с m и v_R степенями свободы. Соответствующее p-значение есть площадь под кривой плотности распределения $F(m, v_R)$ справа от точки, соответствующей вычисленному значению F.

Замечание. В регрессионной модели коэффициент β_i измеряет степень изменения Y в зависимости от X_i, когда значения X_j, $j = 1, 2, \ldots, p$, $j \neq i$, фиксированы. Однако эти коэффициенты могут быть несравнимы по величине из-за различия в единицах измерения $X_1, \ldots, X_p$. Эта трудность может быть преодолена применением *стандартизованных независимых переменных*. Введем переменные $Z_j = X_j / s_j$ для $j = 1, 2, \ldots, p$, где s_j^2 — выборочная дисперсия X_j. Модель множественной линейной регрессии в терминах Z_j будет теперь определяться уравнениями

$$Y_i = \gamma_0 + \gamma_1 z_{1i} + \ldots + \gamma_p z_{pi} + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

(Обозначим МНК-оценку для γ_i через c_i.)

Преимущество стандартизации состоит в том, что новые коэффициенты $\gamma_0, \ldots, \gamma_p$ измеряют теперь степень изменения Y в одной и той же шкале. Это позволяет делать выводы о влиянии независимых переменных $Z_1, \ldots, Z_p$ на Y. Так, большое значение c_j указывает на высокую степень влияния Z_j на Y.

7. Модуль «Факторный анализ»

Этот модуль включает **анализ главных компонент** и **факторный анализ**.

Имеется p исходных переменных (признаков) $X_1, \ldots, X_p$, имеющих совместное нормальное распределение с вектором математических ожиданий $\mu = (\mu_1, \ldots, \mu_p)$ и ковариационной матрицей $\Sigma = (\sigma_{ij})$, $i, j = 1, \ldots, p$. Часто требуется определить взаимосвязь между переменными $X_1, \ldots, X_p$. Эта взаимосвязь называется *структурой зависимости* и может быть измерена ковариациями, или, что эквивалентно, корреляциями между $X_1, \ldots, X_p$. В некоторых случаях можно найти линейные комбинации $Y_1, \ldots, Y_q$ переменных $X_1, \ldots, X_p$ ($q < p$), по которым можно получить структуру зависимости между $X_1, \ldots, X_p$. Таким образом, получается сжатое описание структуры зависимости, несущее почти всю информацию, содержащуюся в самих переменных. Одним из методов анализа структуры зависимости является *анализ главных компонент*. Суть метода состоит в том, что ищутся такие линейные комбинации исходных переменных

$$Y_1 = \sum_{i=1}^{p} \alpha_{1j} X_j, \quad \ldots, \quad Y_p = \sum_{j=1}^{p} \alpha_{pj} X_j,$$

что

$$\operatorname{cov}(Y_i, Y_j) = 0, \quad i, j = 1, \ldots, p, \quad i \neq j,$$

$$V(Y_1) \geqslant V(Y_2) \geqslant \ldots \geqslant V(Y_p), \quad \sum_{i=1}^{p} V(Y_i) = \sum_{i=1}^{p} \sigma_{ii}.$$

Из этих формул видно, что переменные $Y_1, \ldots, Y_p$ некоррелированы и упорядочены по убыванию дисперсии. Более того, общая

дисперсия после перехода от $X_1, \ldots, X_p$ к $Y_1, \ldots, Y_p$ остается без изменений. Такое подмножество первых q переменных $Y_1, \ldots, Y_q$ будет объяснять большую часть общей дисперсии и, таким образом, получится сжатое описание структуры зависимости исходных переменных.

Рассмотрим теперь более общий подход к преобразованиям исходных переменных. Для этого введем *факторную модель*

$$X_1 = \sum_{j=1}^{m} \lambda_{1j} F_j + U_1, \quad \ldots, \quad X_p = \sum_{i=1}^{m} \lambda_{pj} F_j + U_p,$$

где λ_{ij} — постоянные и называются *факторными нагрузками*. Переменные $F_1, \ldots, F_m$ называются *общими (латентными) факторами*, поскольку они используются для представления всех p исходных переменных. Предполагается, что общие факторы некоррелированы и имеют единичные дисперсии. Переменные $U_1, \ldots, U_p$ называются *характерными факторами*, поскольку для каждой переменной X_i определяется своя переменная U_i, $i = 1, \ldots, p$. Предполагается, что характерные факторы некоррелированы и $V(U_i) = \tau_i$, $i = 1, \ldots, p$.

Таким образом, p линейных комбинаций модели главных компонент можно рассматривать как p общих факторов, описывающих структуру зависимости исходных переменных, в то время как $m \ll p$ общих факторов факторной модели описывают основную часть структуры зависимости, а специфические факторы — оставшуюся часть. Другими словами, в модели главных компонент вся дисперсия приписывается p общим факторам, тогда как в факторном анализе дисперсия каждой исходной переменной делится на две части: дисперсию, обусловленную наличием общих факторов (*общность*), и дисперсию, обусловленную вариацией каждой исходной переменной (*специфичность*). Техника факторного анализа направлена на оценку факторных нагрузок и специфических дисперсий, а также на вычисление так называемых *факторных значений*. После того как факторные нагрузки найдены, остается еще задача «наилучшей» интерпретации общих факторов. Для этого используется метод вращения факторов, который из-за субъективности является наиболее спорной частью факторного анализа. Казалось бы, для определения упомянутых оценок теоретически оправдано применение метода максимального правдоподобия. Однако, этот метод сложен для реализации на компьютере и поэтому не получил широкого распространения. Наиболее часто используются метод определения

главных факторов, центроидный метод (обычный и групповой), множественный групповой метод, метод сокращения ранга и т. д.

Основная задача факторного анализа состоит в том, чтобы обнаружить скрытые общие факторы, объясняющие взаимосвязи между наблюдаемыми признаками объекта. Число наблюдаемых признаков объекта может быть большим и связи между ними чрезвычайно сложными, однако, наблюдая объект, вы выдвигаете гипотезу, что существует небольшое число факторов, которые влияют на измеряемые признаки. Естественно стремление, с одной стороны, выявить как можно меньшее число скрытых общих факторов, а с другой стороны, желание, чтобы выделенные факторы как можно лучше приближали измеряемые признаки и описывали связи между ними.

Перейдем к постановке задачи факторного анализа.

В результате N-кратного измерения n признаков мы имеем матрицу измерений $X_{i,j}$, где $i = 1, \ldots, N$, $j = 1, \ldots, n$. Центрируем и нормируем случайные величины x_{ij} по формуле

$$z_{ij} = \frac{(x_{ij} - \overline{x}_{\cdot j})}{s_j}.$$

Это так называемая *стандартная* форма задания признаков Z_j, $j = 1, \ldots, n$. Задача состоит в том, чтобы представить Z_j в виде линейнойкомбинации небольшого числа скрытых общих факторов, т. е. в виде

$$Z_j = \sum_{i=1}^{m} a_{ij} F_j + d_j U_j, \quad j = 1, \ldots, n, \quad m \ll n.$$

Если пространство общих факторов найдено, то с помощью поворота координатных осей можно найти бесконечно много решений. Необходимо так подобрать оси, чтобы результаты можно было интерпретировать в терминах данной предметной области. Это подчас очень не простая задача. Задача интерпретации факторов значительно облегчается получением *простой структуры* факторных нагрузок, т. е. целью процедуры вращения является представление каждой исходной переменной одним фактором или небольшим числом факторов. Нагрузки остальных факторов при этом должны быть близки к нулю. В факторном анализе существует много графических и аналитических методов вращения для получения простой структуры. Минимизируется так называемая *целевая функция*, зависящая от факторных нагрузок и некоторого коэффициента $0 \leqslant \gamma \leqslant 1$. При $\gamma = 0$ вращение, получаемое в результате минимизации целевой функции называется *квартимаксом*. Метод квартимакс максимизирует дисперсию квадратов факторных нагрузок, т. е. выбираются

факторные нагрузки с достаточно большим диапазоном значений. При этом большие значения нагрузок увеличиваются, а маленькие становятся еще меньше, и в результате каждый фактор связывается с возможно меньшим числом исходных переменных. При $\gamma = 1$ вращение, получаемое в результате минимизации целевой функции называется *варимаксом*. Этот метод применяется наиболее часто. Метод варимакс максимизирует дисперсию квадратов нагрузок для каждого фактора в отдельности, что приводит к увеличению больших и уменьшению маленьких значений факторных нагрузок. Но в этом случае простая структура получается для каждого фактора в отдельности, тогда как в методе *квартимакс* простая структура получается для всех факторов одновременно.

В качестве примера рассмотрим следующую задачу. В файле *factor.sta* из папки *Example* собраны результаты опроса 100 взрослых людей относительно степени их удовлетворенности жизнью.

В файле даны значения следующих переменных:

1) Work_1 — удовлетворенность работой, 1-я компонента;
2) Work_2 — удовлетворенность работой, 2-я компонента;
3) Work_3 — удовлетворенность работой, 3-я компонента;
4) Hobby_1 — удовлетворенность свободным временем, 1-я компонента;
5) Hobby_2 — удовлетворенность свободным временем, 2-я компонента;
6) Home_1 — удовлетворенность домашней жизнью, 1-я компонента;
7) Home_2 — удовлетворенность домашней жизнью, 2-я компонента;
8) Home_3 — удовлетворенность домашней жизнью, 3-я компонента;
7) Miscel_1 — общая удовлетворенность, 1-я компонента;
8) Miscel_2 — общая удовлетворенность, 2-я компонента;

Многомерность каждой переменной обусловлена различными аспектами удовлетворенности, например, вы можете быть удовлетворены зарплатой, но не довольны рабочим коллективом и т. д.

Откройте модуль **«Факторный анализ»** с помощью переключателя модулей. На экране появится стартовая панель модуля **«Факторный анализ»**. В строке **Input File** укажите тип файла, с которым вы будете работать. В модуле можно работать с обычным файлом данных (**Raw Data**) и с корреляционной матрицей, ранее вычисленной для этих данных.

Выберите **Raw Data**. В строке **Missing Data** (**Пропущенные данные**) задайте способ обработки пропущенных значений, например, **Casewise** (**способ исключения пропущенных значений**). Щелкните мышью по клавише. Перед вами появится окно **Open Data File**. Выберите файл *factor.sta* и щелкните по кнопке **OK**. Выбрав файл, вы автоматически окажетесь в стартовой панели модуля **«Факторный анализ»** (рис. 36).

Инициировав кнопку **Variables**, выберите переменные для анализа. Кнопка **Select All** позволяет выбрать все переменные сразу (рис. 37).

Щелкнув мышью на кнопку **Spread** (**Распахнуть**), вы увидите окно выбора переменных для анализа с расширенным описанием переменных.

Файл открыт, и переменные для анализа выбраны. Теперь можно приступить к выявлению основных факторов, влияющих на удовлетворенность человека жизнью. Щелкнув на стартовой панели модуля кнопку **OK**, вы начнете анализ выбранных данных. Модуль обработает пропущенные значения тем способом, который вы указали, вычислит корреляционную матрицу и предложит на выбор несколько методов факторного анализа. Вычисление корреляционной матрицы, если она не задана изначально, является первым этапом

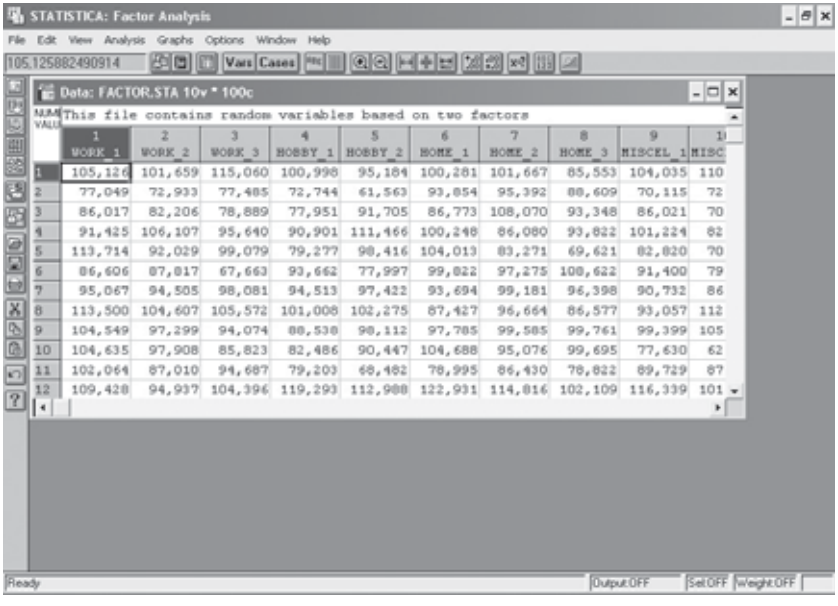


Рис. 36

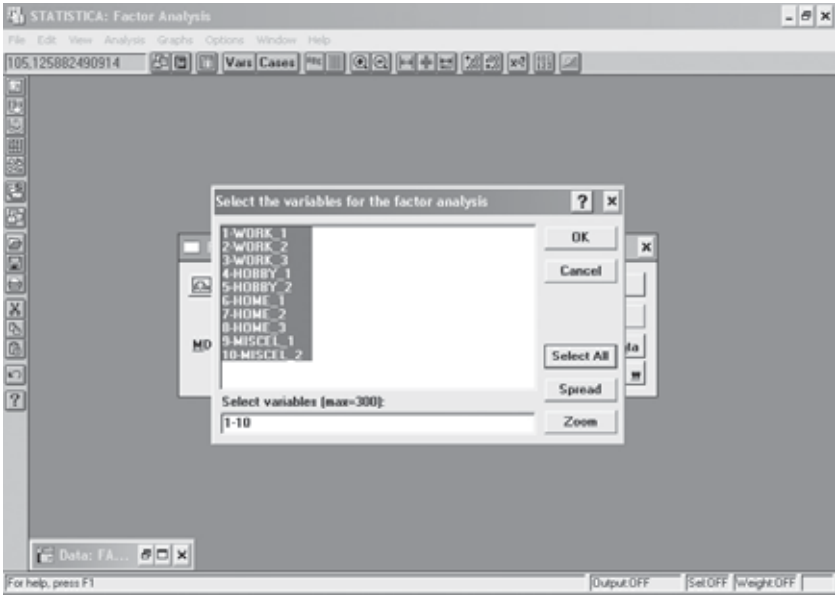

STATISTICA: Factor Analysis
File Edit View Analysis Graphs Options Window Help
105.125882490914
Vars Cases
Select the variables for the factor analysis
1-WORK_1
2-WORK_2
3-WORK_3
4-HOBBY_1
5-HOBBY_2
6-HOME_1
7-HOME_2
8-HOME_3
9-MISCEL_1
10-MISCEL_2
OK
Cancel
Select All
Spread
Zoom
Select variables (max=300):
1-10
Data: FA...
For help, press F1
Output:OFF
Set:OFF
Weight:OFF

Рис. 37

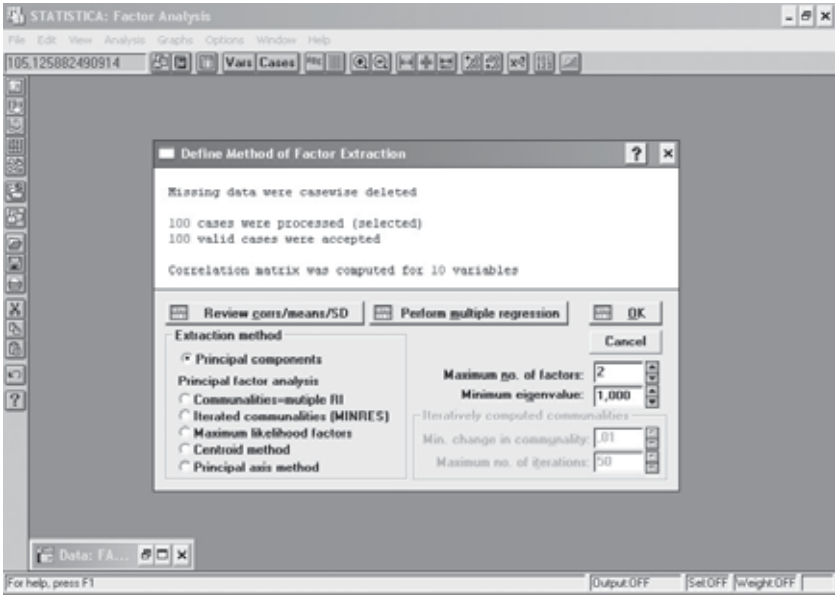

STATISTICA: Factor Analysis
File Edit View Analysis Graphs Options Window Help
105.125882490914
Vars Cases
Define Method of Factor Extraction
Missing data were casewise deleted
100 cases were processed (selected)
100 valid cases were accepted
Correlation matrix was computed for 10 variables
Review corrs/means/SD
Perform multiple regression
OK
Cancel
Extraction method
Principal components
Principal factor analysis
Communalities=multiple RI
Iterated communalities (MINRES)
Maximum likelihood factors
Centroid method
Principal axis method
Maximum no. of factors:
2
Minimum eigenvalue:
1,000
Iteratively computed communalities
Min. change in communality:
,01
Maximum no. of iterations:
50
Data: FA...
For help, press F1
Output:OFF
Set:OFF
Weight:OFF

Рис. 38

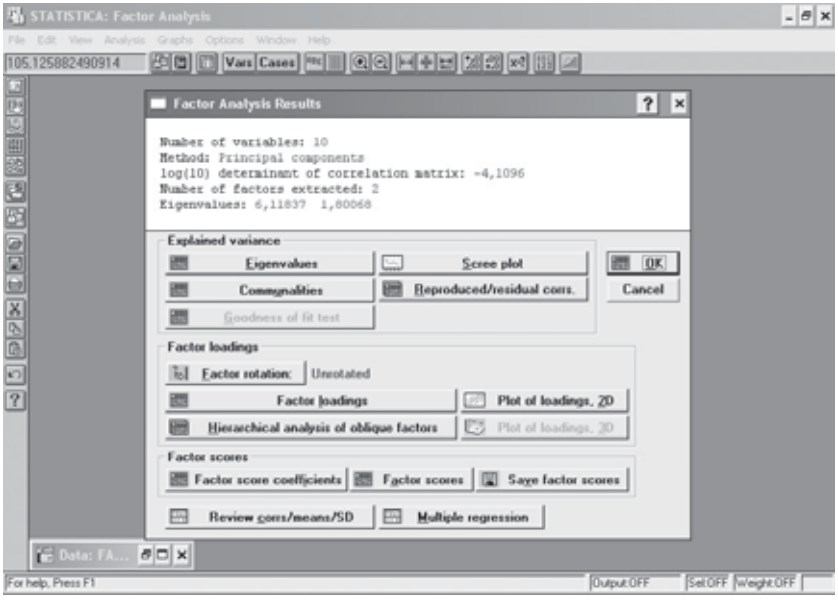

STATISTICA: Factor Analysis
File Edit View Analysis Graphs Options Window Help
105,125882490914
Vars Cases
Factor Analysis Results
Number of variables: 10
Method: Principal components
log(10) determinant of correlation matrix: -4,1096
Number of factors extracted: 2
Eigenvalues: 6,11837 1,80068
Explained variance
Eigenvalues
Scree plot
OK
Communalities
Reproduced/residual corrs.
Cancel
Goodness of fit test
Factor loadings
Factor rotation:
Unrotated
Factor loadings
Plot of loadings, 2D
Hierarchical analysis of oblique factors
Plot of loadings, 3D
Factor scores
Factor score coefficients
Factor scores
Save factor scores
Review corrs/means/SD
Multiple regression
Data: FA...
For help, Press F1
Output:OFF
Set:OFF
Weight:OFF

Рис. 39

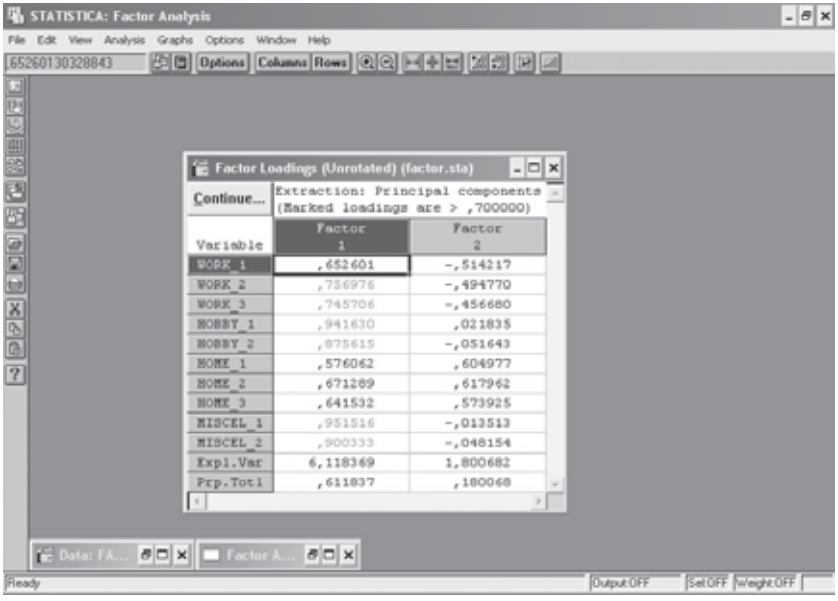

STATISTICA: Factor Analysis
File Edit View Analysis Graphs Options Window Help
,65260130328843
Options
Columns
Rows
Factor Loadings (Unrotated) (factor.sta)
Continue...
Extraction: Principal components
(Marked loadings are > ,700000)
Variable
Factor 1
Factor 2
WORK_1 ,652601 -,514217
WORK_2 ,756976 -,494770
WORK_3 ,745706 -,456680
HOBBY_1 ,941630 ,021835
HOBBY_2 ,875615 -,051643
HOME_1 ,576062 ,604977
HOME_2 ,671289 ,617962
HOME_3 ,641532 ,573925
MISCEL_1 ,951516 -,013513
MISCEL_2 ,900333 -,048154
Expl.Var 6,118369 1,800682
Prp.Totl ,611837 ,180068
Data: FA...
Factor A...
Ready
Output:OFF
Set:OFF
Weight:OFF

Рис. 40

факторного анализа. Щелкнув мышью на кнопку **OK**, вы увидите окно **Define Method of Factor Extraction** (рис. 38).

Нижняя часть окна имеет четыре функциональные кнопки и опции для выбора метода, а также поля, в которых проводятся установки для итерационного вычисления общностей. Обратите внимание на поля в правой части окна. Эти поля определяют число факторов выделяемых системой. Собственные значения, меньше указанного во втором поле, системой игнорируются. Инициировав кнопку **Reviews/ corrs/Means/SD (Просмотреть корреляции, средние, стандартные отклонения)**, вы откроете окно **Просмотреть описательные статистики**. С помощью кнопки **Perform Multiple regression (Выполнить множественную регрессию)** можно провести регрессионный анализ, не выходя из модуля. Группа опций, объединенная под заголовком **Extraction Method (Метод выделения)** позволяет выбрать метод обработки. В зависимости от критерия оптимальности возможен анализ либо методом **Principal Component (Методом главных компонент)**, либо одним из методов, объединенных в группу **Principal Factor Analysis (Анализ главных факторов)**. В эту группу входят следующие методы: **Communalities = Multiple R**2 — общности равны квадрату коэффициента множественной корреляции**; **Iterated Communalities (MINRES) — Итеративных общностей (Минимальных остатков)**; **Centroid Method — Центроидный метод**; **Principal axis Method — метод главных осей**.

Выберите опцию **Principal Component**, нажмите **OK**, и вы увидите окно с результатами факторного, точнее компонентного анализа (рис. 39, 40).

Обратите внимание на кнопку **Factor rotation — Вращение факторов**. Щелкнув по ней, вы можете выбрать различные повороты осей. Можно искать нужный поворот эмпирически, но в системе предложено несколько полезных процедур, которыми желательно пользоваться. Возможны следующие методы: **Varimax — Варимакс**, **Biquartimax — Биквартимакс**, **Quartimax — Квартимакс**, **Equamax — Эквимакс**. Слово **normalized (нормализованные)** в названии методов говорит о том, что факторные нагрузки в процедуре нормализуются, т. е. делятся на корень квадратный из общности. Слово **raw (Исходные)** говорит о том, что факторные нагрузки ненормализованные.

Проведем вращение методом **Varimax normalized** (рис. 41, 42).

Вы видите, что выделены два общих фактора Factor1 и Factor2. Factor1 определяется как отношение к работе (факторные нагрузки у переменных Work_1, Work_2 и Work_3 максимальны по этому фактору и минимальны по другому). Относительно Factor2 можно сказать, что он определяется как удовлетворенность домашней жизнью.

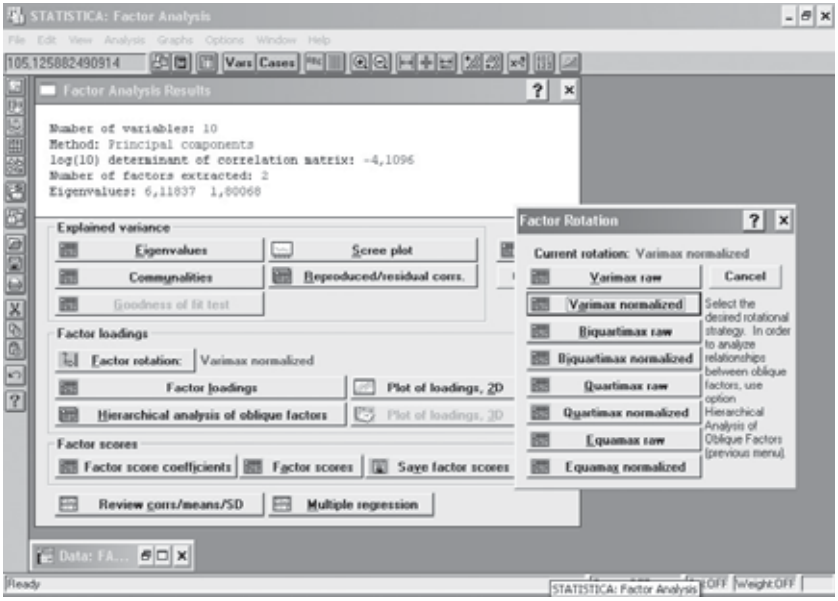

STATISTICA: Factor Analysis
Factor Analysis Results
Number of variables: 10
Method: Principal components
log(10) determinant of correlation matrix: -4,1096
Number of factors extracted: 2
Eigenvalues: 6,11837 1,80068
Explained variance
Eigenvalues
Scree plot
Communalities
Reproduced/residual corrs.
Goodness of fit test
Factor loadings
Factor rotation:
Varimax normalized
Factor loadings
Plot of loadings, 2D
Hierarchical analysis of oblique factors
Plot of loadings, 3D
Factor scores
Factor score coefficients
Factor scores
Save factor scores
Review corrs/means/SD
Multiple regression
Factor Rotation
Current rotation: Varimax normalized
Varimax raw
Cancel
Varimax normalized
Biquartimax raw
Biquartimax normalized
Quartimax raw
Quartimax normalized
Equamax raw
Equamax normalized
Select the desired rotational strategy. In order to analyze relationships between oblique factors, use option Hierarchical Analysis of Oblique Factors (previous menu).
Ready
STATISTICA: Factor Analysis
Weight:OFF

Рис. 41

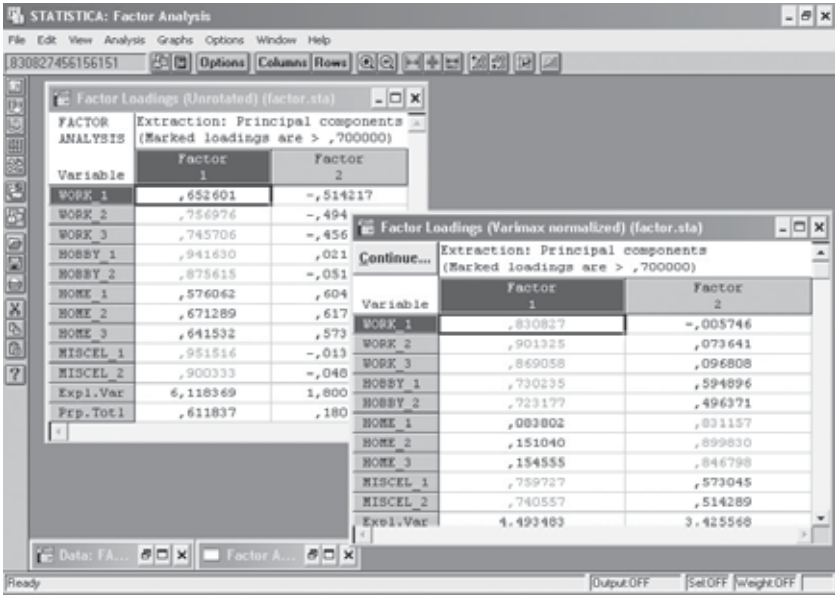

STATISTICA: Factor Analysis
File Edit View Analysis Graphs Options Window Help
Options
Columns
Rows
Factor Loadings (Unrotated) (factor.sta)
FACTOR ANALYSIS
Extraction: Principal components (Marked loadings are > ,700000)
Variable
Factor 1
Factor 2
WORK_1 ,652601 -,514217
WORK_2 ,756976 -,494
WORK_3 ,745706 -,456
HOBBY_1 ,941630 ,021
HOBBY_2 ,875615 -,051
HOME_1 ,576062 ,604
HOME_2 ,671289 ,617
HOME_3 ,641532 ,573
MISCEL_1 ,951516 -,013
MISCEL_2 ,900333 -,048
Expl.Var 6,118369 1,800
Prp.Totl ,611837 ,180
Factor Loadings (Varimax normalized) (factor.sta)
Continue...
Extraction: Principal components (Marked loadings are > ,700000)
Variable
Factor 1
Factor 2
WORK_1 ,830827 -,005746
WORK_2 ,901325 ,073641
WORK_3 ,869058 ,096808
HOBBY_1 ,730235 ,594896
HOBBY_2 ,723177 ,496371
HOME_1 ,003802 ,831157
HOME_2 ,151040 ,899830
HOME_3 ,154555 ,846798
MISCEL_1 ,759727 ,573045
MISCEL_2 ,740557 ,514289
Expl.Var 4,493483 3,425568
Data: FA...
Factor A...
Ready
Output:OFF
Set:OFF
Weight:OFF

Рис. 42

ЛИТЕРАТУРА

[1] *Айвазян С. Л.*, *Мхитарян В. С.* Прикладная статистика. Т. 1, 2. — М. : Юнити-дана, 2001.

[2] *Айвазян С. А.*, *Мхитарян В. С.* Прикладная статистика в задачах и упражнениях. — М. : Юнити-дана, 2001.

[3] *Артемьева Е. Ю.*, *Мартынов Е. М.* Вероятностные методы в психологии. — М. : МГУ, 1975.

[4] *Аткинсон Р.*, *Бауэр Г.*, *Кротерс Э.* Введение в математическую теорию обучения. — М. : Мир, 1969.

[5] *Афифи А.*, *Эйзен С.*, Статистический анализ. Подход с использованием ЭВМ. — М. : Мир, 1982.

[6] *Благуш П.* Факторный анализ с обобщениями. — М. : Финансы и статистика, 1989.

[7] *Боровиков В.* Программа Statistica для студентов и инженеров. — М. : Компьютер пресс, 2001.

[8] *Бочаров П. П.*, *Печинкин А. В.* Математическая статистика. — М. : РУДН, 1994.

[9] *Гласе Дж.*, *Стенли Дж.* Статистические методы в педагогике и психологии. — М.; Прогресс, 1976.

[10] *Готтсданкер Р.* Основы психологического эксперимента. — М. : МГУ, 1982.

[11] *Гусев А. Н.* Дисперсионный анализ в экспериментальной психологии. — М. : УМК «Психология», 2000.

[12] *Гусев А. Н.*, *Измайлов Ч. А.*, *Михалевская М. Б.* Измерение в психологии. — М. : УМК «Психология», 2005.

[13] *Джонсон Н.*, *Лион Ф.* Статистика и планирование эксперимента в технике и науке. — М. : Мир, 1980.

[14] *Дэйвисон М.* Многомерное шкалирование. — М. : Финансы и статистика, 1988.

[15] *Ермолаев О. Ю.* Математическая статистика для психологов. — М. : Флинта-МПСИ, 2002.

[16] *Кендалл М.*, *Стюарт А.* Многомерный статистический анализ и временные ряды. — М. : Наука, 1976.

[17] *Кендалл М.*, *Стюарт А.* Статистические выводы и связи. — М. : Наука, 1973.

[18] *Кендалл М.* Ранговые корреляции. — М. : Статистика, 1975.

[19] *Ким Дж.-О.*, *Мюллер Ч. У. и др*. Факторный, дискриминантами и кластерный анализ. — М. : Финансы и статистика, 1989.
[20] *Кричевец А. Н.*, *Шикин Е. В.*, *Дьячков А. Г.* Математика для психологов. — М. : Флинта-МПСИ, 2002.
[21] *Крылов В. Ю.* Методологические и теоретические проблемы математической психологии. — М. : Янус-К, 2000.
[22] *Купер К.* Индивидуальные различия. — М. : Аспект пресс, 2000.
[23] *Митина О. В.*, *Михайловская И. Б.* Факторный анализ для психологов. — М. : УМК «Психология», 2001.
[24] *Михеев В. И.* Методика получения и обработки экспериментальных данных в психолого-педагогических исследованиях. — М. : РУДН, 1986.
[25] *Наследов А. Д.* Математические методы психологического исследования. — СПб. : Речь, 2004.
[26] Психологическая диагностика детей и подростков / под ред. К. М. Гуревича, Е. М. Борисовой. — М. : Международная педагогическая академия, 1995.
[27] *Пфанцагль И.* Теория измерений. — М. : Мир, 1976.
[28] *Романко В. К.* Курс теории вероятностей и математической статистики для психологов. — М. : МГППИ, 2000.
[29] *Сидоренко Е. В.* Методы математической обработки в психологии. — СПб. : Социально-психологический центр, 1996.
[30] *Суходольский Г. В.* Основы математической статистики для психологов. — Л. : ЛГУ, 1972.
[31] *Терехина А. Ю.* Анализ данных методами многомерного шкалирования. — М. : Наука, 1986.
[32] *Тюрин Ю. Н.*, *Макаров А. Л.* Анализ данных на компьютере / под ред. В. Э. Фигурнова. — М. : Инфра-М, 2003.
[33] *Урбах В. Ю.* Статистический анализ в биологических и медицинских исследованиях. — М. : Медицина, 1975.
[34] *Феллер В.* Введение в теорию вероятностей и ее приложения. Т. 1. — М. : Мир, 1964.
[35] *Холлендер М.*, *Вулф Д.* Непараметрические методы статистики. — М. : Финансы и статистика, 1983.
[36] *Чистяков В. П.* Курс теории вероятностей. — М. : Агар, 1996.

ОТВЕТЫ

Глава 1

1. а) 1/2; б) 11/36.
2. 1/1190.
3. а) 228/1309; б) 91/1309.
4. а) 3/10; б) 5/6.
5. 1/21.
6. $12!/12^{12}$.
7. $C_8^2 \cdot 8!/8^8$.
8. 105/512.
9. 1/5.
10. $7!/7^7$.
11. $4^6 \cdot C_9^6/C_{36}^6$.
12. $C_{20}^4 \cdot C_{15}^6/C_{36}^6$.
13. C_{10}^3/C_{25}^3.
14. 1/60.
15. 1/2.
16. а) 0,72; б) 0,01.
17. 3/4.
18. 1/3.
19. 1/7.
20. а) 1/30; б) 19/30; в) 1/6.
21. Да.
22. а) 0,018; б) 0,648; в) 0,002; г) 0,044; д) 0,954.
23. 1/3.
24. 303/625.
25. 9/25.
26. 2/3.
27. 281/1296.
28. 37/72.
29. а) 0,008; б) 1/16.
30. а) 39/50; б) 81/195.
31. а) 19/70; б) 56/171.
32. 7/18.
33. а) 0,8; б) 7/16.
34. 0,41.
35. а) 5/16; б) 25/32.
36. 5/16.
37. а) 0,06; б) 0,1.
38. а) 0,972; б) 0,271.
39. Вероятнее выиграть две партии из четырех.

Глава 2

6. б) 1/4.
7. б) 1/2.
8. а) $1/\pi$; в) 3/4.
9. а) 0,34; б) 0,84.
10. а) 1/2; б) 1/2.
13. а) $F((x-4)/2)$; б) $F(\sqrt[3]{x})$; в) $F(\sqrt[5]{x+1})$; г) $F(\ln x)$ при $x > 0$ и 0 при $x \leqslant 0{,}14$.
14. а) $\mathsf{M}\xi = 3/4$, $\mathsf{D}\xi = 11/4$; б) $\mathsf{M}\xi = 0$, $\mathsf{D}\xi = 1/2$; в) $\mathsf{M}\xi = 5/3$, $\mathsf{D}\xi = 5/9$.
15. $\mathsf{M}\xi = \mathsf{D}\xi = 3/2$.
16. $\mathsf{M}\xi = 2/3$, $\mathsf{D}\xi = 1/18$.
17. $\mathsf{M}\xi = 5/3$, $\mathsf{D}\xi = 1/18$.
18. $\mathsf{M}\xi = \pi/2$, $\mathsf{D}\xi = \pi^2/4 - 2$.
19. а) $\mathsf{M}\eta = 6$, $\mathsf{D}\eta = 2$; б) $\mathsf{M}\eta = -4$, $\mathsf{D}\eta = 2$; в) $\mathsf{M}\eta = 16$, $\mathsf{D}\eta = 10$; г) $\mathsf{M}\eta = -13$, $\mathsf{D}\eta = 13$.
20. $p(x_1) = 0{,}6$; $p(x_2) = 0{,}4$.
21. $p(x_1) = p(x_2) = 0{,}15$; $p(x_3) = 0{,}7$.
22. 0.
23. г) 0.
24. г) 0.
25. 0,13
26. 1.
27. б) $\mathsf{M}\xi = \pi/4$; $\mathsf{D}\xi = \pi^2/16 + \pi/2 - 2$.

Глава 3

1. б) $\overline{x} = 1{,}4$; $s^2 = 1{,}14$; $\overline{v} = 77\%$.
2. б) $\overline{x} = 3{,}6$; $s^2 = 1{,}04$; $\overline{v} = 28\%$.
3. б) $\overline{x} = 14{,}35$; $s^2 = 2{,}63$; $\overline{v} = 11\%$.
4. б) $\overline{x} = 16{,}93$; $s^2 = 5{,}21$; $\overline{v} = 13\%$.
5. б) $\overline{x} = 1{,}335$; $s^2 = 0{,}26$; $\overline{v} = 38\%$.

Глава 4

1. а) $(\overline{x} - 1{,}33;\ \overline{x} + 1{,}33)$; б) $(\overline{x} - 0{,}98;\ \overline{x} + 0{,}98)$; в) $(\overline{x} - 0{,}785;\ \overline{x} + 0{,}785)$.
2. а) $(\overline{x} - 2{,}28;\ \overline{x} + 2{,}28)$; б) $(\overline{x} - 1{,}53;\ \overline{x} + 1{,}53)$; в) $(\overline{x} - 1{,}34;\ \overline{x} + 1{,}34)$.
3. а) 69; б) 273.
4. а) 0,88; б) 0,65; в) 0,99.
5. а) (1,37; 6,32); б) (1,65; 7,58).

Глава 6

1. H_1. **2.** H_0. **3.** H_0.
4. H_0. **5.** H_0. **6.** H_0.
7. H_1. **8.** H_1. **9.** H_0.
10. H_0. **11.** H_0. **12.** H_0.

Глава 7

1. H_0. **2.** H_0. **3.** H_0.
4. H_1.

Глава 8

1. H_0. **2.** H_1. **3.** H_1.

Глава 9

1. H_0 отклоняется; (0,608; 0,9).
2. H_0 отклоняется; (0,285; 0,798).
3. б) 0,956; в) H_0 отклоняется.
4. $\hat{r}^{(s)} = 0{,}945$; $\hat{r} = 0{,}778$.
5. $\widehat{W}(3) = 0{,}897$.
6. а) 0,672; в) (0,546; 0,798).
7. а) 0,039.

Глава 10

2. $\hat{a} = -40$, $\hat{b} = 0{,}6$.
3. $\hat{a} = 89{,}9$; $\hat{b} = -0{,}33$.

Глава 12

1. а) 1; б) 2; в) 2.
2. а) 2; б) 2; в) 1.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Минимальные системные требования определяются соответствующими требованиями программ Adobe Reader версии не ниже 11-й либо Adobe Digital Editions версии не ниже 4.5 для платформ Windows, Mac OS, Android и iOS; экран 10"

Учебное электронное издание

Романко Василий Кириллович

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ПСИХОЛОГИИ

Учебное пособие

Ведущий редактор *М. С. Стригунова*
Художники *Н. В. Зотова, Н. А. Новак*
Технический редактор *Е. В. Денюкова*
Оригинал-макет подготовлен *М. Ю. Пановым* в пакете LaTeX 2ε

Подписано к использованию 23.09.19.
Формат 125×200 мм

Издательство «Лаборатория знаний»
125167, Москва, проезд Аэропорта, д. 3
Телефон: (499) 157-5272
e-mail: info@pilotLZ.ru, http://www.pilotLZ.ru

Учебное пособие позволяет получить достаточный объем знаний и приобрести начальные практические навыки по применению математических методов анализа данных. Материал излагается в доступной форме, сопровождается множеством примеров; для его понимания не требуется математической подготовки. К каждой главе приведены задачи с ответами.

Книга написана на основе практических занятий и курса лекций, читавшихся автором в течение ряда лет студентам Московского городского психолого-педагогического университета.

Учебное пособие адресовано прежде всего студентам, преподавателям и специалистам в области психологии и педагогики. Может быть использовано студентами и исследователями в различных областях науки гуманитарных направлений – социологии, политологии и др.