

Ю.Н.Тюрин, А.А.Макаров

АНАЛИЗ ДАННЫХ НА КОМПЬЮТЕРЕ

Рекомендовано Учебно-методическим советом
по классическому университетскому образованию
в качестве учебного пособия по направлениям
«Математика», «Математика. Прикладная математика»

Москва
Издательство МЦНМО
2016

УДК 519.2, 681.3(075.8)

ББК 517.8, 32.973(я73)

Т98

Рецензенты:

Д. ф.-м. н., профессор *С. А. Айвазян*;
Д. ф.-м. н., профессор *В. Н. Тутубалин*

Научный редактор:

В. Э. Фигурнов

Тюрин Ю.Н., Макаров А.А.

Т98

Анализ данных на компьютере: учебное пособие. — Новое изд. — М.: МЦНМО, 2016. — 368 с., ил.

ISBN 978-5-4439-1011-6

В учебном пособии без лишнего формализма излагаются основные идеи и понятия математической статистики, необходимые на практике для анализа данных. На примерах подробно рассмотрены важнейшие постановки статистических задач и методы их решения, включая расчеты на компьютере в пакете SPSS.

Рекомендуется для студентов широкого круга математических, естественно-научных и социально-экономических специальностей, а также для всех, кто сталкивается на практике с обработкой и анализом данных.

Предыдущее издание вышло в 2008 г. в ИД «ФОРУМ».

УДК 519.2, 681.3(075.8)

ББК 517.8, 32.973(я73)

ISBN 978-5-4439-1011-6

© Тюрин Ю.Н., Макаров А.А., 2008

© МЦНМО, 2016

Предисловие авторов

О чем написана эта книга. Эта книга – о теории и практике статистического анализа данных: о лежащем в его основе принципе случайного выбора и случайности, о статистических законах и статистических моделях, о статистическом прогнозе, о статистической точности и т.д. В этой книге мы старались как можно более просто и понятно передать *дух статистической науки*, объяснить читателю самые необходимые и употребительные методы обработки данных и рассказать, как из неточных, подверженных ошибкам и колебаниям данных можно извлекать твердые и обоснованные выводы. Мы также стремились показать, как в настоящее время реально осуществляются прикладные статистические расчеты на персональных компьютерах. Для этого мы используем один из ведущих в мире и популярных в России пакет программ SPSS.

Для кого предназначена книга. Эта книга — учебное пособие для всех, кто хочет освоить теоретические и практические основы статистических методов анализа данных и применять их в своей деятельности. Книга обобщает опыт преподавания авторами курсов теории вероятностей и математической статистики в Московском государственном университете им. М.В. Ломоносова, Высшей школе экономики, других вузах, крупных государственных и частных компаниях и практику решения прикладных задач. Нашими слушателями, собеседниками, заказчиками были самые разные люди: экономисты и социологи, бизнесмены и менеджеры, государственные служащие и инженеры, психологи и политологи, медики, биологи и т.д. и, разумеется, студенты, обучающиеся по соответствующим специальностям. При написании книги мы имели в виду всех этих потенциальных читателей и стремились к тому, чтобы книга была доступна и полезна для них всех.

От читателя мы не требуем наличия каких-либо предварительных знаний о теории вероятностей и математической статистике. Но минимальные знания по математике (в объеме первого курса вуза), а также начальные навыки работы с компьютером для чтения книги и применения изложенных в ней методов анализа данных все же нужны.

О математическом формализме. Хотя статистика в сколь угодно развитом виде невозможна без математики, мы стремились в нашем изложении свести к минимуму математический формализм. Так, мы почти нигде не приводим формальных математических доказательств.

Мы заменяем их обсуждениями и объяснениями. На наш взгляд, эта особенность делает книгу доступной самому широкому кругу читателей.

Особенности изложения. В этой книге мы всюду и на равных правах представляем как классические гауссовские методы, так и их «более молодые» непараметрические конкуренты, описываем условия применимости, преимущества и недостатки этих методов. Все излагаемые темы мы сопровождаем рассмотрением примеров, взятых из обычной статистической практики. На этих примерах мы стараемся показать, как в конкретных условиях следует строить статистические модели и как затем с их помощью можно делать выводы из имеющихся статистических данных. Иногда мы применяем к одному примеру разные методы анализа данных, что позволяет наглядно сравнить эти методы между собой.

О компьютерных разделах книги. Современный прикладной статистический анализ немислим без компьютера. Поэтому изложение статистической теории в книге соединено с рассказом о ее современном компьютерном осуществлении. Изложив статистическую тему, мы затем показываем, как типичные задачи по этой теме решаются в статистическом пакете SPSS. Мы считаем, что владение компьютерными методами анализа данных — это один из базовых элементов подготовки современного специалиста.

Предыдущие издания книги. Первое издание этой книги вышло в начале 1995 г. В то время внедрение компьютерных программ анализа данных в массовую деловую и учебную практику делало первые шаги. Два последующих издания в 1998 и 2003 гг. заметно отличались как от первого издания, так и друг от друга за счет расширения статистических тем и рассматриваемых пакетов анализа данных. Настоящее издание опирается на 4-е издание книги в 2008 г. В нем мы ограничились рассказом о наиболее важных и употребительных на практике основных статистических понятиях и моделях. В качестве программы статистического анализа данных мы ограничились пакетом SPSS, который широко распространен в социально-экономическом анализе данных в России и широко используется в учебном процессе во многих российских университетах, включая Национальный исследовательский университет «Высшая школа экономики», где его используют при подготовке специалистов по многим социальным, управленческим и гуманитарным специальностям.

За прошедшие годы объем доступных статистических данных в научной и деловой практике рос лавинообразно, породив новую популярную область анализа данных «Big Data». Это сделало еще более востребованным представление о базовых понятиях, методах и алгоритмах анализа данных и условиях их применения. Кроме того, в

России по-прежнему наблюдается дефицит учебной литературы по непараметрическому анализу данных. С ростом популярности различных социально-экономических, политических и других рейтингов и экспертных оценок востребованность этих методов возрастает. Думаем, что подробное обсуждение материала в книге будет полезно многим студентам и специалистам в различных областях вне зависимости от того, какими компьютерными средствами анализа данных они пользуются.

Таблицы. В книгу включен небольшой набор таблиц математической статистики. Таблицы очень полезны при изучении методов анализа данных — решение «вручную» нескольких задач по анализу данных (а это можно сделать только при наличии таблиц) позволяет наглядно ощутить, как работают соответствующие статистические методы. Кроме того, таблицы полезны и при использовании непараметрических методов анализа данных с малыми объемами наблюдений — в этих случаях приближения, используемые компьютерными программами, могут быть весьма не точны, использование таблиц позволяет сделать более обоснованные выводы.

О прикладной статистике. О роли прикладной статистики в современной жизни и о некоторых особенностях развития этой науки в нашей стране вы можете прочесть в предисловии редактора книги. А далее находится раздел «Как читать эту книгу», в котором описан порядок размещения материала в книге.

Благодарности. Мы рады выразить благодарность нашим коллегам, чья помощь и участие способствовали написанию этой книги. Мы особо признательны Д.С. Шмерлингу, который был инициатором создания этой книги. Его интерес и постоянное внимание нас всегда поддерживали. Мы благодарны В.Э. Фигурнову, который как редактор внес в текст много улучшений. Мы глубоко благодарны С.А. Айвазяну, М.В. Болдину, В.Н. Тутубалину за многие обсуждения, советы и поддержку. Общение с этими и другими учеными помогало нашему совершенствованию в статистической теории и практике.

Сведения об авторах. Юрий Николаевич Тюрин, д.ф.-м.н., заслуженный профессор Московского государственного университета им. М.В. Ломоносова.

Алексей Алексеевич Макаров, к.ф.-м.н., профессор, заведующий общеуниверситетской кафедрой высшей математики Национального исследовательского университета «Высшая школа экономики».

Предисловие редактора

Моторы реактивного самолета взрвели еще прежде, чем все восемь пассажиров взошли на борт, и они не успели пристегнуть ремни, как самолет уже катил по полю... Вице-президент выступил первым.

— Нас не удовлетворяют результаты этого месяца по Северо-Востоку. Цифры вам известны, как и мне. Я хочу знать, почему это происходит. И хочу, чтобы мне сказали, какие приняты меры.

Самолет к этому времени уже поднялся в воздух.

А. Хейли. «Колеса»

— Законы статистики везде одинаковы, — продолжал Николай Петрович солидно. — Утром, например, гостей бывает меньше, потому что публика еще исправна; но чем больше солнце поднимается к зениту, тем наплыв делается сильнее. И наконец, ночью, по выходе из театров — это почти целая оргия!

— И заметьте, — пояснил Семен Иванович, — каждый день, в одни и те же промежутки времени, цифры всегда одинаковые. Колебаний — никаких! Такова неизблемость законов статистики!

М.Е. Салтыков-Щедрин. «За рубежом»

В нашей повседневной жизни, бизнесе, иной профессиональной деятельности, а также в научных исследованиях мы постоянно сталкиваемся с событиями и явлениями с неопределенным исходом. Например, торговец не знает, сколько посетителей придет к нему в магазин, рабочий — сколько времени ему придется сегодня добираться до работы, бизнесмен — какой будет завтра или через месяц курс доллара, банкир — вернут или нет взятый у него заем, страховщик — когда и какое ему придется выплачивать страховое вознаграждение и т.д. При этом нам постоянно приходится принимать в подобных неопределенных, связанных со многими случайностями ситуациях свои решения, иногда очень важные. В быту или в несложном бизнесе мы можем принимать такие решения на основе здравого смысла, интуиции, предыдущего опыта. Здесь мы часто можем сделать некий «запас прочности» на действие случая: скажем, выходить из дома на десять минут раньше, чтобы уже почти наверняка не опаздывать на работу.

Однако в более серьезном бизнесе, в условиях жесткой конкуренции, решения должны приниматься на основе тщательного анализа имеющейся информации, быть обоснованными и доказуемыми. Например, вряд ли банк или совет директоров крупной корпорации примет решение

о вложении денег в некоторый проект только потому, что он кому-то «представляется выгодным». Здесь потребуются тщательный расчет, связанный с прогнозами состояния рынка и рентабельности вложений, оценками возможных рисков и их последствий и т.д. При этом уже вряд ли возможно делать большой запас прочности «на всякий случай», ибо тогда вас опередят конкуренты, умеющие считать лучше и, тем самым, принимать более правильные решения.

Для решения задач, связанных с анализом данных при наличии случайных и непредсказуемых воздействий, математиками и другими исследователями (биологами, психологами, экономистами и т.д.) за последние двести лет был выработан мощный и гибкий арсенал методов, называемых в совокупности математической статистикой (а также прикладной статистикой или анализом данных). Эти методы позволяют выявлять закономерности на фоне случайностей, делать обоснованные выводы и прогнозы, давать оценки вероятностей их выполнения или невыполнения. Введению в эти методы и посвящена данная книга.

Средства анализа данных на компьютерах. Широкому внедрению методов анализа данных в 60-х и 70-х годах XX века немало способствовало появление компьютеров, а начиная с 80-х годов XX века — персональных компьютеров. Статистические программные пакеты сделали методы анализа данных более доступными и наглядными: теперь уже не требовалось вручную выполнять трудоемкие расчеты по сложным формулам, строить таблицы и графики — всю эту черновую работу взял на себя компьютер, а человеку осталась главным образом творческая работа: постановка задач, выбор методов их решения и интерпретация результатов.

Результатом появления мощных и удобных пакетов для анализа данных на персональных компьютерах стало резкое расширение и изменение круга потребителей методов анализа данных. Если раньше эти методы рассматривались главным образом как инструмент научных исследований, то начиная с середины 80-х годов основными покупателями статистических пакетов (которые продаются в сотнях тысяч копий ежегодно) стали уже не научные, а коммерческие организации, а также правительственные и медицинские учреждения. Таким образом, методы анализа данных и статистические пакеты для компьютеров и других видов ЭВМ стали на Западе типичным и общепотребительным инструментом плановых, аналитических, маркетинговых отделов производственных и торговых корпораций, банков и страховых компаний, правительственных и медицинских учреждений. И даже представители мелкого бизнеса часто употребляют методы анализа данных либо самостоятельно, либо обращаясь к услугам консультационных компаний.

Примеры. Приведем несколько примеров применения методов статистического анализа данных в практических задачах.

1. Рассмотрим достаточно простую, но часто встречающуюся задачу. Предположим, что вы ввели важное нововведение: изменили систему оплаты труда, перешли на выпуск новой продукции, использовали новую технологию и т.п. Вам кажется, что это дало положительный эффект, но действительно ли это так? А может быть, этот кажущийся эффект определен вовсе не вашим нововведением, а естественной случайностью, и уже завтра вы можете получить прямо противоположенный, но столь же случайный эффект? Для решения этой задачи надо сформировать два набора чисел, каждый из которых содержит значения интересующего вас показателя эффективности до и после нововведения. Статистические критерии сравнения двух выборок покажут вам, случайны или неслучайны различия этих двух рядов чисел.

2. Другая важная задача — прогнозирование будущего поведения некоторого временного ряда: изменения курса доллара, цен и спроса на продукцию или сырье и т.д. Для такого временного ряда с помощью статистического пакета программ подбирают некоторое аналитическое уравнение — строят регрессионную модель.

Если мы предполагаем, что на интересующий нас показатель влияют некоторые другие факторы, их тоже можно включить в модель, предварительно (с помощью того же статистического пакета) проверив существенность (значимость) этого влияния. Затем на основе построенной модели можно сделать прогноз и указать его точность (см. рис. 1).

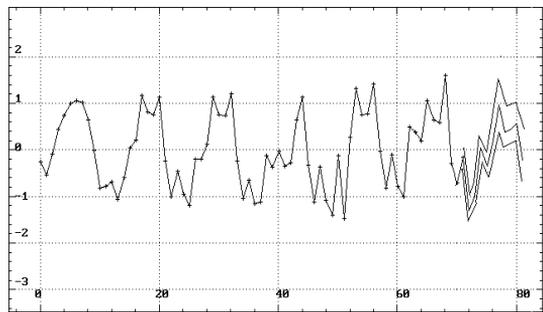


Рис. 1. График изменения объема транспортных перевозок и его прогноз

3. Во многих технологических процессах необходимо систематически контролировать состояние процесса, чтобы вовремя вмешаться при отклонениях его от нормального режима и предотвратить тем самым потери от выпуска некачественной продукции. Для этого используются статистические методы контроля качества, повсеместное и неукоснительное применение которых во многом определило поразительные успехи японской промышленности. Здесь мы наблюдаем замечательный пример внедрения статистических методов в широкую прак-

тику. Японскими специалистами были отобраны наиболее простые правила для оценивания динамики изменения качества продукции и его наглядного представления. Эти правила выражены самыми простейшими словами, и японские рабочие выучивают их наизусть, как молитву, после чего каждый простой рабочий знает, при каких обстоятельствах производственный процесс в порядке, когда надо быть настороже, а когда срочно вызывать бригаду наладчиков (рис. 2).

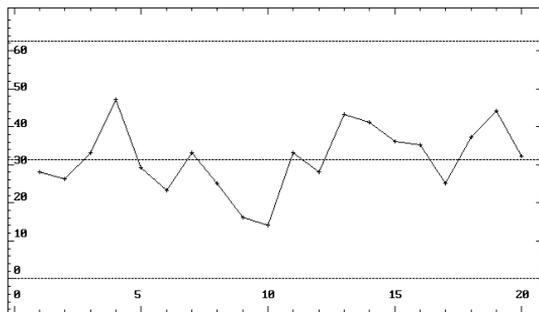


Рис. 2. Контрольная карта изменения показателя качества с зоной допустимых пределов изменения

4. Еще одна интересная и часто встречающаяся задача связана с классификацией объектов. Пусть, например, вы являетесь начальником кредитного отдела банка. Столкнувшись с невозвратом кредитов, вы решаете впредь выдавать кредиты лишь фирмам, которые «схожи» с теми, которые себя хорошо зарекомендовали, и не выдавать тем, которые «схожи» с неплательщиками или мошенниками.

Для классификации фирм можно собрать показатели их деятельности (например, размер основных фондов, валюту баланса, вид деятельности, объем реализации и т.д.) и провести кластерный анализ (в более сложных случаях — многомерное шкалирование, см. гл. 12) этих данных. Во многих случаях имеющиеся объекты удастся сгруппировать в несколько групп (кластеров), и вы сможете увидеть, не принадлежит ли запрашивающая кредит фирма к группе неплательщиков (рис. 3). Аналогичный пример: пусть у вас имеются данные о различных сортах пива, каждый из которых характеризуется множеством переменных: цвет, содержание алкоголя, других веществ, калорийность и т.п. Вы хотите закупать и продавать наиболее дешевое

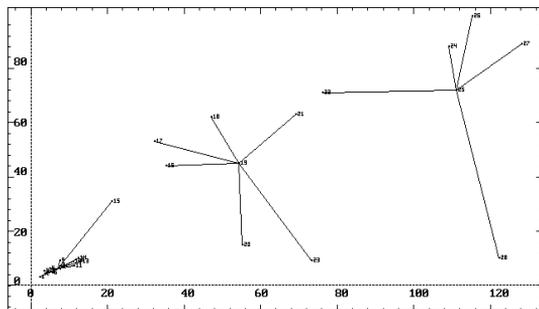


Рис. 3. С помощью кластерного анализа имеющаяся совокупность объектов разбита на три группы со схожими свойствами

пиво, но близкое по совокупности свойств к очень престижному и дорогому сорту. С помощью тех же методов вы сможете решить и эту задачу.

Можно было бы привести еще множество других интересных примеров применения методов анализа данных в самых разных областях: торговле и здравоохранении, образовании и управлении и т.д.

Универсальные и специальные методы. Следует подчеркнуть, что методы статистического анализа являются *универсальными* и могут применяться в самых разных областях человеческой деятельности. Скажем, предсказание курса доллара и прогноз спроса на автомобили делаются с помощью одних и тех же процедур. Поэтому требования неискушенных пользователей, чтобы им предоставили инструмент для анализа данных именно в банковском деле или именно в медицине, редко бывают обоснованными (примерно как просьба в канцелярском магазине продать чернила именно для третьего класса). В подавляющем большинстве случаев все нужные этим пользователям задачи могут быть решены с помощью универсальных пакетов.

Разумеется, есть и такие области человеческой деятельности, в которых возникают специфические, не встречающиеся в других областях задачи по анализу данных, и поэтому требуются специфические статистические средства. Однако таких областей очень мало. По-видимому, наиболее важная из них — это страховые (актуарные) расчеты, используемые в своей деятельности страховыми компаниями.

Методы визуализации данных. Чтобы решать, какие методы анализа надо применить к имеющимся данным и насколько удовлетворительны полученные результаты статистических процедур, нужно иметь возможность наглядно представлять себе эти данные и результаты. Поэтому практически все статистические пакеты обеспечивают широкий набор средств визуализации данных: построение графиков, двух- и трехмерных диаграмм, а часто и различные средства деловой графики. Все это помогает лучше представить обрабатываемые данные, получить общее представление об их особенностях и закономерностях. Результаты применения статистических процедур, как правило, представляются в наглядном графическом виде всегда, когда это возможно.

О предварительной подготовке для анализа данных. Хотя статистические пакеты для персональных компьютеров резко упростили применение методов статистического анализа данных, все же для осмысленного их употребления пользователи должны обладать определенной подготовкой: понимать, в каких ситуациях применимы различные статистические методы, знать, каковы их свойства, уметь интерпретировать результаты. На Западе такая подготовка обеспечивается обучением

основам анализа данных практически всех студентов и менеджеров: в программы университетов, школ бизнеса, технических и других колледжей входят систематические курсы прикладной статистики. Разработаны и широко используются курсы основ теории вероятностей и статистики и для старших классов средней школы. В достатке имеется специальная и популярная литература по анализу данных, ее всегда можно найти в книжных магазинах, торгующих научно-технической литературой. А при затруднениях можно лично или по телефону обратиться в одну из сотен консультационных фирм и получить там квалифицированную консультацию по постановкам задач, использованию статистических пакетов и т.д.

К сожалению, в нашей стране ситуация пока совершенно другая, хотя она постепенно изменяется к лучшему. В программу средней школы наконец введены основы теории вероятностей и статистики. В высшей школе, при подготовке специалистов все большее внимание уделяется навыкам практического анализа данных. В учебных программах многих университетов появились компьютерные практикумы по анализу данных, а преподавание математической статистики сопровождается компьютерными иллюстрациями расчетов. Все это со временем даст положительные результаты. А пока даже самые простейшие методы статистического анализа данных для многих отечественных руководителей и менеджеров остаются terra incognita. Для исправления положения (что абсолютно необходимо для конкуренции с западным бизнесом), по-видимому, потребуется время.

О современных методах анализа данных. При решении задач анализа данных в экономике, социологии, психологии и во многих других областях обычно необходимо использовать методы, разработанные в течение последних десятилетий — робастные (устойчивые) методы оценивания, методы, свободные от распределения, и т.д. Эти методы имеют гораздо более широкие границы применимости и успешно работают там, где классические методы, созданные в XIX и первой половине XX века, использоваться не могут и дают неверные результаты. Многие из этих методов включены в современные статистические пакеты.

К сожалению, у нас в стране во многих вузах изучают только классические методы анализа данных, часто не объясняя условия их применимости. Из-за этого многие специалисты, пытаясь анализировать свои данные, получают неверные выводы просто потому, что применили знакомый, но не подходящий в их ситуации метод анализа, хотя нужный им метод можно было вызвать из соседней строчки меню статистического пакета.

В этом отношении данная книга может быть очень полезна, так как в ней авторы описывают и классические, и современные (непараметрические) методы анализа данных, подробно и понятно объясняя читателям условия применимости, возможности, преимущества и недостатки этих методов.

Советы читателям. Что же можно посоветовать тем, кто собирается изучать методы анализа данных или применять в своей деятельности? Вот некоторые рекомендации.

1. Читать популярные (рассчитанные на прикладных специалистов, а не профессиональных математиков) книги по анализу данных. Кроме данной книги, из книг на русском языке стоит отметить книги [87], [10], [16], [39], [109], [23], [35], [61], [44], [21], [95], [122].

2. Читать документацию статистических пакетов. Очень часто она фактически является популярным учебником, наглядно и неформально объясняющим применение средств анализа данных, в том числе и самых мощных многомерных методов. Сказанное в полной мере относится к документации пакета SPSS, рассматриваемого в этой книге.

3. Практически применять в ходе изучения математической статистики и анализа данных статистические пакеты. Пользоваться их подсказками. Очень часто это помогает понять назначение метода и его свойства лучше и быстрее, чем что-либо другое.

Остается пожелать читателям этой (чрезвычайно, на мой взгляд, полезной и актуальной) книги успешно изучить изложенные в ней методы и научиться применять эти и другие методы анализа данных в своей практической деятельности.

Виктор Фигурнов

Как читать эту книгу

Структура книги. Материал, включенный в эту книгу, можно условно разбить на три части. Первую из них составляют главы 1 — 4 и 11. Здесь изложены основные *понятия теоретической и прикладной статистики*, владение которыми необходимо для осмысленного применения методов статистического анализа данных. Мы обсуждаем понятия случайной изменчивости, основные характеристики случайных величин, наиболее распространенные статистические распределения, основы проверки статистических гипотез и оценивания параметров, а также основы выборочных обследований и опросов. Все изложение ведется не в строго формальном математическом ключе, а на общепонятном уровне, с привлечением многочисленных примеров.

Вторая часть книги (главы 5—12) описывает *статистические модели*, наиболее часто используемые на практике для анализа данных. Сюда вошли анализ нормальных выборок, регрессионный и факторный (или дисперсионный) анализ, исследование связи признаков и таблицы сопряженности, методы проверки согласия статистической модели с данными опыта, а также краткий обзор других методов статистического анализа. При этом особое внимание мы уделили непараметрическим (свободным от распределения) методам, поскольку они имеют гораздо более широкие границы применимости (по сравнению с классическими гауссовскими), более устойчивы по отношению к отклонениям от моделей и лишь немного уступают в эффективности наилучшим параметрическим методам, когда эти последние можно применять.

Заметная часть книги (ее составляют последние параграфы глав 1 — 10) посвящена статистической обработке данных на компьютере в пакете SPSS. В этой части показано, как рассмотренные в книге задачи можно решать с помощью компьютера. Мы полагаем, что эти примеры будут полезны всем читателям, в том числе и пользователям других статистических пакетов. Ведь входные данные и результаты статистической обработки, как правило, мало зависят от конкретного пакета, поскольку определяются общепринятыми традициями.

Примеры. Все обсуждаемые в книге постановки задач мы старались иллюстрировать на примерах. При этом на одном и том же примере мы показывали работу как непараметрических методов, так и их параметрических (гауссовских) аналогов. Это позволило нам провести наглядное сравнение различных методов с точки зрения их примени-

мости, устойчивости и т.п. Кроме того, чтобы помочь читателю лучше понять алгоритмы обработки, мы разбирали применение статистических методов для одних и тех же данных как при ручных расчетах, так и при использовании компьютера. Данные для примеров взяты из известных монографий А. Хальда [107], Р. Готсданкера [33], М. Холлендера и Д.А. Вулфа [115] и др., а также из практической работы авторов.

Одной из особенностей этих примеров является сравнительно малый объем исходных данных. Это довольно характерная ситуация для большинства прикладных исследований, особенно в гуманитарных областях. А поскольку на подобных объемах выборки практически невозможна эффективная проверка гипотез об их распределении, а процедуры отбраковки грубых наблюдений бесполезны или малоэффективны, мы рассматривали в первую очередь непараметрические статистические методы, как более универсальные.

Таблицы. В приложении книги приведены статистические таблицы. Они кроме широко распространенных в учебной литературе таблиц нормального и связанного с ним распределений включают малодоступные таблицы для основных непараметрических критериев и позволяют решать небольшие учебные задачи вручную.

Порядок чтения книги. Читать эту книгу можно в различном порядке. Тем, кто только начинает знакомиться с теорией статистики, мы советуем прочитать сначала главы 1, 3, 4 и 11. Они содержат базовые понятия прикладной статистики. К главе 2, содержащей сведения об основных вероятностных распределениях, можно обращаться по мере необходимости. Те, кто уже знаком с такими понятиями, как случайная величина, распределение вероятностей, статистические гипотезы и оценки, статистические критерии, уровни значимости, доверительные интервалы и т.п., могут начинать чтение с любой из интересующих их глав.

Предварительные сведения. От читателя этой книги мы старались не требовать особой математической подготовки — сведений из программы первого курса вуза более чем достаточно. Для использования компьютерных разделов книги вполне достаточно общего знакомства с интерфейсом Windows.

В конце каждой главы приведен список дополнительной литературы по рассматриваемым в главе темам.

Основные понятия прикладной статистики

Цель этой главы — познакомить читателя с основными понятиями теории вероятностей и статистики, на которые опирается анализ данных изменчивой (случайной) природы. Не стремясь к строгому формальному изложению, мы расскажем о случайных событиях и случайных величинах, об их характеристиках: распределении вероятностей, математическом ожидании, дисперсии и т.д. Будут введены наиболее распространенные понятия описательной статистики, используемые при обработке данных, такие как генеральная совокупность, выборка, выборочная функция распределения, медиана, квантили, гистограмма и др. В конце главы мы опишем, как можно вычислить соответствующие характеристики на компьютере.

1.1. Случайная изменчивость

Статистика изучает числа, чтобы обнаружить в них закономерности. Все мы хорошо знакомы с закономерными явлениями и закономерными изменениями, они составляют главный объект научных исследований. Например, исследователя могут интересовать вопросы типа: как изменяется давление в жидкости с изменением глубины? С какой скоростью движутся падающие тела? Как будет проходить химическая реакция, если мы определенным образом изменим температуру, давление и концентрации участвующих в реакции веществ и т.п. Знание законов природы позволяют нам ответить на подобные вопросы, не производя реальных опытов, т.е. заранее. Например, мы можем точно вычислить, какие вещества и в какой пропорции образуются при той или иной химической реакции, или предсказать, когда в данной местности произойдет следующее солнечное затмение.

Но отнюдь не во всех ситуациях интересующий нас результат полностью и жестко определяется влияющими на него факторами. Например, мы не можем указать, сколько часов будет светить электрическая лампочка или как долго будет служить телевизионный приемник. Невозможно предвидеть число посетителей магазина и количество товаров, которое они купят, каков будет результат бросания игральных костей и т.д. Ответы на подобные вопросы можно получить, только проведя

соответствующие испытания. Часто явления (ситуации), в которых результат полностью определяется влияющими на него факторами, называются *детерминированными* или *закономерными*, а те, в которых это не выполняется — *недетерминированными* или *стохастическими*.

Идея случайности. Для описания явлений с неопределенным исходом (как в повседневной жизни, так и в науке) используется *идея случайности*. Согласно этой идее результат явления с неопределенным исходом как бы определяется неким случайным испытанием, случайным экспериментом, случайным выбором. Иначе говоря, считается, что для выбора исхода в неопределенной ситуации природа словно бы бросает кости. Вопрос о том, насколько применим такой подход к явлениям окружающего мира, решается не путем его логического обоснования, а по результатам практического применения.

Замечание. Вопросы о том, существует ли случайность «на самом деле», о происхождении случайного и соотношении закономерного и случайного являются дискуссионными философскими темами. Действительно, закономерные изменения, как подчеркивает само их название, порождены определенными причинами, которые могут быть названы, указаны и изучены. Отыскивая эти причины, мы исходим из убеждения, что если нечто изменилось, так это потому, что изменилось что-то другое, и это другое служит причиной первому. Когда же изменения происходят при полной неизменности условий, в которых протекает явление, мы объясняем это случайностью. Но поскольку полной неизменности условий на практике достичь невозможно, сохраняется логическая возможность отрицать наличие в природе случайности и объяснять неопределенность результатов эксперимента воздействием неизвестных нам и неучтенных факторов. Мы не будем входить в эти философские споры и будем рассматривать проблемы случайности чисто технически, принимая этот подход лишь как модель для описания непредсказуемой изменчивости, дабы на его основе получать количественные выводы и рекомендации для практики.

Случайная изменчивость. Мы все хорошо знаем, что такое закономерность. Например, при формулировке законов природы мы говорим, что если одна величина принимает такое-то значение, то другая примет такое-то. Случайная изменчивость нам знакома в меньшей степени, а потому о ней надо поговорить подробнее. Для начала лучше взять такой пример, где случайная изменчивость действует отдельно от закономерной, так сказать, «в чистом виде».

Рассмотрим пример, заимствованный из книги А. Хальда. В табл. 1.1 приведены размеры головок 200 заклепок, изготовленных станком (который делает их тысячами). Все контролируемые условия, в которых работал станок, оставались неизменными. В то же время диаметры головок раз от разу несколько изменялись. Характерная черта случайных колебаний — эти изменения выглядят бессистемными, хаотичными. Действительно, если бы в этих изменениях мы смогли обнару-

жить какую-либо закономерность, у нас появились бы основания, чтобы искать ответственную за эту закономерность причину, тем самым изменчивость не была бы чисто случайной. Если бы, скажем, с течением времени размер головки заклепки проявил тенденцию к увеличению, мы могли бы попытаться связать это, например, с износом инструмента.

Таблица 1.1

Диаметры 200 головок заклепок, мм											
13.39	13.43	13.54	13.64	13.40	13.55	13.40	13.26	13.42	13.50	13.32	13.31
13.28	13.52	13.46	13.63	13.38	13.44	13.52	13.53	13.37	13.33	13.24	13.13
13.53	13.53	13.39	13.57	13.51	13.34	13.39	13.47	13.51	13.48	13.62	13.58
13.57	13.33	13.51	13.40	13.30	13.48	13.40	13.57	13.51	13.40	13.52	14.56
13.40	13.34	13.23	13.37	13.48	13.48	13.62	13.35	13.40	13.36	13.45	13.48
13.29	13.58	13.44	13.56	13.28	13.59	13.47	13.46	13.62	13.54	13.20	13.38
13.43	13.36	13.56	13.51	13.47	13.40	13.29	13.20	13.46	13.44	13.42	13.29
13.41	13.39	13.50	13.48	13.53	13.34	13.45	13.42	13.29	13.38	13.45	13.50
13.55	13.33	13.32	13.69	13.46	13.32	13.32	13.48	13.29	13.25	13.44	13.60
13.43	13.51	13.43	13.38	13.24	13.28	13.58	13.31	13.31	13.45	13.43	13.44
13.34	13.49	13.50	13.38	13.48	13.43	13.37	13.29	13.54	13.33	13.36	13.46
13.23	13.44	13.38	13.27	13.66	13.26	13.40	13.52	13.59	13.48	13.46	13.40
13.43	13.26	13.50	13.38	13.43	13.34	13.41	13.24	13.42	13.55	13.37	13.41
13.38	13.14	13.42	13.52	13.38	13.54	13.30	13.18	13.32	13.46	13.39	13.35
13.34	13.37	13.50	13.61	13.42	13.32	13.35	13.40	13.57	13.31	13.40	13.36
13.28	13.58	13.58	13.38	13.26	13.37	13.28	13.39	13.32	13.20	13.43	13.34
13.33	13.33	13.31	13.45	13.39	13.45	13.41	13.45				

Обсуждение случайной изменчивости не обязательно начинать с такого специального примера. Каждому известны более простые опыты, в которых результат определяется случаем: раздача игральных карт или костей домино, бросание игральные костей, монет и т.д. У всех этих примеров есть общая черта — непредсказуемость результатов для действий, проводящихся в неизменных условиях.

Закономерность и случайность. В большинстве явлений присутствуют оба вида изменчивости — и закономерная, и случайная, и для нахождения закономерностей нам приходится «отсеивать» мешающие случайные факторы. Например, при внесении удобрений на пшеничное поле мы не можем точно предсказать, какова будет урожайность на этом поле, поскольку она зависит от множества причин, которые мы считаем случайными (от погодных условий, нашествий вредителей, болезней растений и т.д.). Однако с помощью методов статистического анализа мы все же можем определить степень влияния на урожайность внесения удобрений и применения других агротехнических приемов. Для этого могут потребоваться многолетние тщательно спланированные эксперименты, с помощью которых влияние агротехнических приемов оценивается на фоне мешающих факторов.

Итак, статистический подход к изучению явлений природы состоит в мысленном разделении наблюдаемой изменчивости на две части — обусловленные закономерными и случайными причинами, и выявлению закономерной изменчивости на фоне случайной. Например, в табл. 1.2 и на рис. 1.1 отображено изменение урожайности зерновых (в центнерах с гектара) в СССР за 45 лет, с 1945 по 1989 г. Данные предоставлены А.И. Манелля, которому авторы выражают глубокую признательность.

Таблица 1.2

Урожайность зерновых культур в СССР с 1945 по 1989 г.
(в центнерах с гектара в первоначально оприходованном весе)

Год	Урожайность	Год	Урожайность	Год	Урожайность
1945	5.6	1960	10.9	1975	10.9
1946	4.6	1961	10.7	1976	17.5
1947	7.3	1962	10.9	1977	15.0
1948	6.7	1963	8.3	1978	18.5
1949	6.9	1964	11.4	1979	14.2
1950	7.9	1965	9.5	1980	14.9
1951	7.4	1966	13.7	1981	12.6
1952	8.6	1967	12.1	1982	15.2
1953	7.8	1968	14.0	1983	15.9
1954	7.7	1969	13.2	1984	14.4
1955	8.4	1970	15.6	1985	16.2
1956	9.9	1971	15.4	1986	18.0
1957	8.4	1972	14.0	1987	18.3
1958	11.1	1973	17.6	1988	17.0
1959	10.4	1974	15.4	1989	18.8

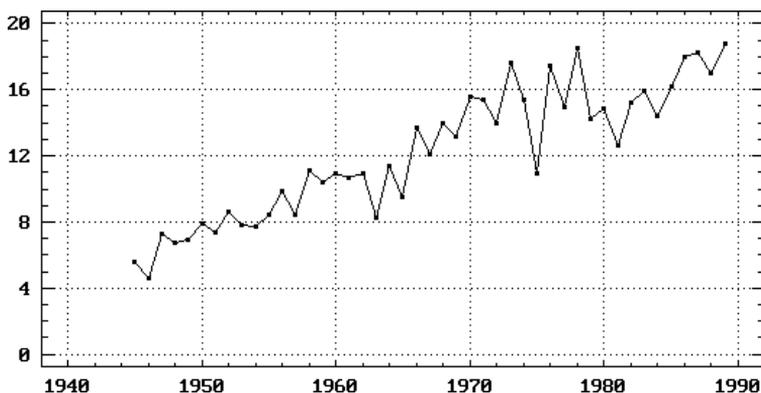


Рис. 1.1. Урожайность всех зерновых культур в СССР с 1945 по 1989 г. (ц/га)

Хорошо видно, что урожайность в целом возрастала (по-видимому, за счет улучшения агротехники и внесения минеральных удобрений). Ее рост и составляет закономерную часть картины. В то же время видны

и значительные колебания урожайности в разные годы, по-видимому, за счет погодных условий и иных факторов, изменения которых мы считаем случайными. Методы *математической статистики* позволяют в подобных ситуациях оценивать параметры имеющихся закономерностей, проверять те или иные гипотезы об этих закономерностях и т.д. В последующих главах этой книги мы рассмотрим, как решаются многие из подобных задач.

Однако случайности могут не только мешать нам постигать закономерности — они способны и сами порождать их. Рассмотрим, например, газ в некотором сосуде (скажем, воздух в комнате). Поведение каждой молекулы газа носит случайный характер, но вся совокупность этих молекул ведет себя вполне закономерно, подчиняясь хорошо известным законам физики. Так, давление газа на каждую единицу площади поверхности сосуда строго постоянно (колебания проявляются только для очень сильно разреженных газов), а объем газа, его давление и температура связаны друг с другом уравнением Менделеева—Клапейрона. Аналогично выбор времени для телефонных звонков каждый человек осуществляет сам, но нагрузка на телефонную станцию (АТС), распределение интервалов между звонками различных абонентов и т.д. подчиняются вполне определенным закономерностям. Изучением закономерностей, которые порождаются случайными событиями, занимается наука *теория вероятностей*.

1.2. События и их вероятности

Хотя результаты эксперимента (наблюдений, опыта), зависящего от случайных факторов, нельзя предсказать, все же разные возможные его исходы и связанные с ними события имеют неодинаковые шансы на появление. Количественное описание правдоподобия отдельных исходов и событий основывается на понятии вероятности. Предполагается, что каждому событию, возможному в данном случайном испытании, может быть приписана числовая мера его правдоподобия, называемая его *вероятностью*. Если, скажем, A есть случайное событие, то его вероятность обычно обозначается через $P(A)$. (Буква P — начальная в латинском слове «вероятность».) Вероятность *невозможного* события (которое никогда не происходит) принимается равной 0, а вероятность *достоверного* события (которое происходит всегда) принимается равной 1. Поэтому для любого события A : $0 \leq P(A) \leq 1$.

Свойства вероятности просты, естественны и в общем известны каждому. Однако перед тем, как рассказывать о них, необходимо дать некоторые определения, касающиеся случайных событий.

Случайные события.

Объединением, или суммой, событий A и B называют событие C , которое состоит в том, что происходит хотя бы одно из событий A и B . (C происходит тогда и только тогда, когда происходит либо A , либо B , либо оба вместе.) Обозначение:

$$C = A \cup B \quad \text{или} \quad C = A + B.$$

Пересечением, или произведением событий A и B называют событие C , которое состоит в том, что происходят оба события A и B . Обозначение:

$$C = A \cap B \quad \text{или} \quad C = AB.$$

Отрицанием события A называют такое событие, которое состоит в том, что A не происходит. Обозначение для него \bar{A} .

Событие, которое при нашем случайном испытании обязательно происходит, называют *достоверным*; которое не может произойти — *невозможным*. Вероятность достоверного события равна 1; вероятность невозможного события равна 0.

Если события A и B не могут произойти одновременно (т.е. если AB — невозможное событие), их называют *несовместимыми*. Несовместимы, например, события A и \bar{A} . В то же время $A + \bar{A}$ — событие достоверное.

Например, при бросании игральной кости:

- событие, состоящее в том, что в результате бросания кости выпадет 1, 2, 3, 4, 5 или 6 очков, является достоверным;
- событие, состоящее в том, что результате бросания кости выпадет 7 очков, является невозможным;
- объединением события A , состоящего в том, что в результате бросания кости выпадет меньше 4 очков, и события B , состоящего в том, что в результате бросания кости выпадет 3 или 6 очков, будет событие $A + B$, состоящее в том, что в результате бросания кости выпадет 1, 2, 3 или 6 очков;
- пересечение AB событий A и B состоит в том, что в результате бросания кости выпадет 3 очка;
- отрицание события A , обозначаемое \bar{A} , состоит в том, что в результате бросания кости выпадет 4, 5 или 6 очков.

Свойства вероятности. Теперь свойства вероятности перечислить просто:

1. $0 \leq P(A) \leq 1$ для любого события A .
2. $P(A + B) = P(A) + P(B)$, если события A и B несовместимы, а в общем случае $P(A + B) = P(A) + P(B) - P(AB)$.

3. Вероятность достоверного события равна 1, а невозможного события — нулю.

Для полного описания случайного эксперимента нужно указать все его возможные исходы и их вероятности. Например, бросание игральной кости, имеющей форму куба, приводит к выпадению одной из ее шести граней. Это шесть элементарных исходов, т.е. неразложимых на более простые. Если кость, как говорят, правильная, то эти шесть исходов равноправны и поэтому должны иметь равные вероятности. Следовательно, вероятность каждого из них равна $1/6$. Вероятности остальных (составных) событий может быть вычислена из приведенных выше свойств вероятности. Например, вероятность $P(B)$ события B , состоящего в том, что в результате бросания кости выпадет 3 или 6 очков, равна $1/3$. Действительно, это событие является объединением двух несовместимых событий: «выпало 3 очка» и «выпало 6 очков», вероятность каждого из которых равна $1/6$. Аналогично вероятность $P(A)$ события A , состоящего в том, что в результате бросания кости выпадет меньше 4 очков, равна $1/2$.

Не будем далее развивать эту тему, оставив ее теории вероятностей. Но все же нам придется ввести еще два важных понятия — независимости событий и условной вероятности.

Независимость событий.

Определение 1. События A и B называются независимыми, если

$$P(AB) = P(A)P(B).$$

На практике независимость событий обычно устанавливается не с помощью проверки этого равенства, а из условий опыта и других содержательных соображений. При этом указанное соотношение можно использовать для вычисления вероятности событий AB через вероятности событий A и B . Понятие независимости очень существенно для теории вероятностей. То, насколько в своей математической форме понятие независимости соответствует нашим интуитивным представлениям, лучше всего разобрать с помощью понятия *условной вероятности*.

Условная вероятность. Для простоты мы рассмотрим, как можно определить понятие условной вероятности в случайном испытании с конечным числом исходов. Пусть Ω — совокупность всех таких исходов, ω обозначает произвольный элементарный исход, $P(\omega)$ — его вероятность. Любые события A и B в этом опыте представляют собой некоторые подмножества Ω , поскольку они состоят из элементарных исходов. Обозначим через $P(A|B)$ условную вероятность события A

при условии, что произошло событие B . Достаточно определить условную вероятность для элементарных исходов ω . Те исходы ω , которые не входят в B , невозможны при наступлении события B , поэтому для них следует положить условную вероятность равной нулю:

$$P(\omega | B) = 0, \quad \text{если } \omega \notin B.$$

Для исходов ω , входящих в B , сумма их вероятностей $\sum_{\omega \in B} P(\omega)$ равна $P(B)$, а сумма их условных вероятностей должна быть равна единице. Действительно, $\sum_{\omega \in B} P(\omega | B)$ равна $P(B | B)$. Но при наступлении B событие B является достоверным, поэтому согласно свойству 3 вероятностей $P(B | B)$ равно 1. Чтобы это условие выполнялось, естественно положить для $\omega \in B$:

$$P(\omega | B) = P(\omega) / P(B).$$

Теперь мы можем определить условную вероятность для любого события A .

Определение. Условная вероятность события A при условии B есть

$$P(A | B) = \sum_{\omega \in A} P(\omega | B).$$

Из этого определения легко вывести, что:

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

Это соотношение в общем случае (когда число элементарных исходов не обязательно конечно) и принимают за определение условной вероятности. Из него легко следует известная формула умножения вероятностей:

$$P(AB) = P(A | B)P(B).$$

Заметим, что равноправие событий A и B позволяет написать также, что $P(AB) = P(B | A)P(A)$.

С помощью понятия условной вероятности мы можем дать другое определение независимости событий.

Определение 2. Событие A не зависит от события B , если

$$P(A | B) = P(A).$$

Иначе говоря, событие A не зависит от события B , если вероятность события A не зависит от того, произошло или нет событие B . Нетрудно показать, что два определения независимости события A от B , данные выше, эквивалентны. Так же можно показать, что если A не зависит

от B , то и B не зависит от A . Единственная оговорка, которую надо добавить к сказанному, — что условную вероятность можно определять таким образом, лишь если $P(B) > 0$.

1.3. Измерения вероятности

Раз мы ввели понятие вероятности как количественное выражение для правдоподобия случайного события, нам необходим метод ее численного выражения. Здесь возможны два пути — умозрения и прямого измерения.

Умозрительный способ определения численного значения вероятности зиждется в основном на понятии равновозможности тех или иных исходов эксперимента. Мы уже прибегали к помощи этого соображения при обсуждении бросания игральной кости. Основная область приложения этого принципа — случайный выбор и азартные игры. Поэтому принцип равновозможности исходов эксперимента имеет ограниченное применение. Кроме того, выводы из этого принципа всегда относятся к некоему идеальному случайному опыту, и то, насколько им подчиняется реальный эксперимент, само зачастую нуждается в проверке.

Измерение вероятности события отличается от измерения других физических величин. Для массы, скорости, температуры и большинства других физических величин есть специальные приборы, позволяющие выразить их числом (что и означает измерить). К сожалению, для вероятности такого прибора нет. Все же прямое измерение вероятности возможно, оно основано на независимых повторениях случайного эксперимента.

Пусть в определенном случайном эксперименте нас интересует вероятность некоторого события A . Допустим, что мы можем многократно осуществлять этот эксперимент в неизменных условиях, так что от опыта к опыту $P(A)$ не меняется. Проведем N таких повторений (иногда говорят — *реализаций*) этого опыта. Число N не должно зависеть от исходов отдельных опытов; например, оно может быть назначено заранее. Подсчитаем число тех опытов из N , в которых событие A произошло. Обозначим это число через $N(A)$. Рассмотрим отношение $N(A)/N$ — частоту события A в N повторениях опыта. *Оказывается, частота $N(A)/N$ приблизительно равна $P(A)$, если число повторений N велико.*

Указанная связь между частотой события и его вероятностью составляет содержание теоремы Бернулли, о которой подробнее мы будем говорить в главе 4. Там будет дана ее точная формулировка и доказательство. Кроме того, важен и вопрос о достигаемой точности приближения

частоты к вероятности, в частности, о числе опытов, необходимых для получения заранее указанной точности. Этому второму вопросу должно предшествовать прояснение содержания статистической точности, которое реализуется через посредство *доверительных интервалов*. Об этом речь пойдет в гл. 5.

Итак, задав вопрос об измерении вероятностей, мы столкнулись с неприятной неожиданностью — это измерение оказалось, во-первых, непростым с чисто физической точки зрения (многократное повторение опыта), а во-вторых, сопряженным с довольно сложными и новыми понятиями.

Особо надо подчеркнуть, что описанные выше опыты должны происходить независимо друг от друга в неизменных условиях, чтобы вероятность события сохранялась постоянной. При большом числе повторений опытов соблюсти это требование зачастую оказывается нелегко. Даже небольшие отклонения от статистической устойчивости могут оказать воздействие на результаты, особенно при высоких требованиях к точности выводов. Не говоря уже о том, что повторения опытов, да еще многократные, далеко не всегда возможны.

1.4. Случайные величины. Функции распределения

В случайных экспериментах нас часто интересуют такие величины, которые имеют числовое выражение. Например, у каждого человека имеется много числовых характеристик: рост, возраст, вес и т.д. Если мы выбираем человека случайно (например, из группы или из толпы), то случайными будут и значения указанных характеристик. Чтобы подчеркнуть это обстоятельство, что измеряемая по ходу опыта численная характеристика зависит от его случайного исхода и потому сама является случайной, ее называют *случайной величиной*.

Случайной величиной, в частности, является упомянутое выше число очков, выпадающее при бросании игральной кости. Случайна сумма очков, выпавших при бросании двух игральных костей (а также их разность, произведение и т.д.). Случайной величиной надо считать диаметр головки заклепки, изготавливаемой станком (см. табл. 1.1, где приведены значения, которые приняла эта случайная величина в 200 опытах).

Часто говорят, что случайная величина реализуется во время опыта. Если употребить это слово, то можно также сказать, что табл. 1.1 дает 200 *реализаций* этой случайной величины.

Каждая случайная величина задает *распределение вероятностей* на множестве своих значений. Если ξ — случайная величина, принимающая значения из X , то мы можем задать распределение вероятностей P_ξ на X следующим образом:

$$P_\xi(A) = P(\xi \in A).$$

Чтобы дать полное математическое описание случайной величины, надо указать множество ее значений и соответствующее случайной величине распределение вероятностей на этом множестве.

Виды случайных величин. В практических задачах обычно используются два вида случайных величин — *дискретные* и *непрерывные*, хотя бывают и такие случайные величины, которые не являются ни дискретными, ни непрерывными. Рассмотрим сначала дискретные случайные величины.

Дискретные случайные величины обладают тем свойством, что мы можем перечислить (перенумеровать) все их возможные значения. Таким образом, для задания распределения вероятностей, порожденных дискретными случайными величинами, надо только указать вероятности каждого возможного значения этой случайной величины. Например, число очков, выпавших при бросании игральной кости, — это дискретная случайная величина, так как она может принимать только 6 значений: 1, 2, 3, 4, 5 или 6. Для определения вероятностей любых событий, связанных с этой случайной величиной, нам надо только указать вероятности каждого из этих значений.

Определение. *Случайную величину называют дискретной, если множество ее возможных значений конечно либо счетно.*

Напомним, что множество называется счетным, если его элементы можно перенумеровать натуральными числами.

Каждое возможное значение дискретной случайной величины имеет положительную вероятность (иногда, впрочем, допускают, что некоторые значения могут иметь нулевые вероятности, особенно когда рассматривают не одно, а несколько дискретных распределений одновременно). Чтобы полностью описать дискретное распределение вероятностей, надо указать все значения, вероятности которых положительны (точнее, могут быть положительны), и вероятности этих значений.

Пример. При бросании двух игральных костей сумма выпавших очков может принимать значения от 2 до 12. При этом для правильных костей, бросааемых независимо, вероятность получить в сумме 2 очка равна $1/6 \times 1/6 = 1/36$, получить 3 очка — равна $2/36$ и т.д. Распределение вероятностей суммы выпавших очков определяется табл. 1.3.

Таблица 1.3

Значения	2	3	4	5	6	7	8	9	10	11	12
Вероятности	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Однако не все случайные величины могут быть описаны так просто, как дискретные случайные величины. Например, время службы электрической лампочки может, в принципе, принимать любое значение от нуля до бесконечности (как хорошо известно, это множество не является счетным). И если предполагается, что лампочка была вначале исправна, то вероятность того, что время ее службы будет в точности равно некоторому значению, будет равна нулю. Ненулевыми будут вероятности только сложных событий: например, что время службы лампочки — от одного до двух месяцев. Для подобных (так называемых *непрерывных*) случайных величин мы не можем задать их распределение путем указания вероятностей каждого возможного значения, так как все эти вероятности равны нулю. При описании таких случайных величин используются другие средства. В частности, если значениями случайной величины являются вещественные числа, то распределение случайной величины полностью определяется ее *функцией распределения*.

Функция распределения. Пусть ξ обозначает случайную величину, принимающую вещественные значения, x — вещественное число.

Определение. *Функцией распределения $F(x)$ случайной величины ξ называют $F(x) = P(\xi \leq x)$.*

Ясно, что функция $F(x)$ монотонно возрастает с ростом x (точнее сказать, не убывает, потому что могут существовать участки, на которых она постоянна). У дискретной случайной величины функция распределения ступенчатая, она возрастает скачком в тех точках, вероятности которых положительны. Это точки разрыва $F(x)$. На рис. 1.2 приведен график функции распределения для описанной выше случайной величины — суммы очков, выпавшей при бросании двух игральных костей.

Непрерывные случайные величины. Для случайной величины, принимающей вещественные значения, то свойство, что вероятность любого отдельного ее значения равна нулю, может легко быть выражено через функцию распределения.

Определение. *Случайную величину, принимающую вещественные значения, называют непрерывной, если непрерывна ее функция распределения.*

Непрерывным в этом случае называют и соответствующее распределение вероятностей. Для непрерывного распределения вероятность каждого отдельного значения случайной величины равна нулю. На этом

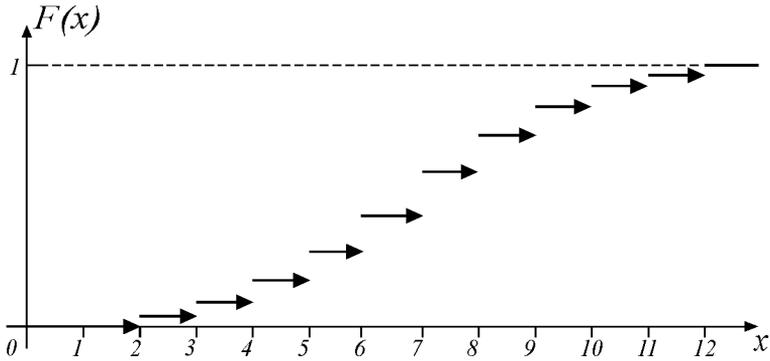


Рис. 1.2. График функции распределения суммы очков, выпавших на двух игральных костях

и основано противопоставление непрерывных и дискретных распределений — ведь для последних вся единичная вероятность распределена конечными положительными порциями. Для непрерывных же она как бы «разлита» по области определения случайной величины (в данном случае — по прямой).

Плотность вероятности. Нагляднее всего непрерывную случайную величину можно представить тогда, когда ее функция распределения не только непрерывна, но и дифференцируема (за исключением, может быть, конечного числа точек). В этом случае вероятности связанных с данной случайной величиной событий можно выразить через посредство так называемой функции *плотности вероятности*. Есть две эквивалентные формы определения плотности: интегральная и дифференциальная. Определение плотности вероятности в интегральной форме таково.

Определение. *Функция $p(t)$ называется плотностью вероятности в точке t (иногда — плотностью случайной величины ξ), если для любых чисел a, b (пусть $a < b$)*

$$P(a < \xi < b) = \int_a^b p(x) dx.$$

В дифференциальной форме определения плотности данное условие заменяется на следующее: для любого $\Delta > 0$ и любого¹ действительного t

$$P(t < \xi < t + \Delta) = p(t) \Delta + o(\Delta),$$

¹ Если говорить точно — любого, за исключением множества меры нуль. Предыдущее (интегральное) определение показывает, что функция плотности может быть произвольно изменена на любом множестве нулевой меры, все равно удовлетворяя определению. Практически, разумеется, используют наиболее регулярную и простую из возможных функций плотности.

где $o(\Delta)$ — малая (точнее, бесконечно малая) по сравнению с Δ величина.

Наглядное содержание второго из этих определений состоит в том, что вероятность, приходящаяся на малый отрезок, оказывается приблизительно пропорциональной длине этого отрезка, причем коэффициент пропорциональности равен значению функции плотности вероятности в некоторой точке этого отрезка.

Функция распределения и плотность связаны соотношениями:

$$F(x) = \int_{-\infty}^x p(t) dt, \quad p(x) = F'(x).$$

(для почти всех x — с теми же оговорками, что были сделаны выше).

Как правило, для приложений достаточно двух вышеописанных типов распределений — дискретного и непрерывного, точнее, имеющего плотность. Хотя можно встретиться с распределениями, представляющими собой смесь двух этих типов, и даже с более сложными. В главе 2 мы подробнее познакомимся с некоторыми важными для приложений законами вероятностей на числовой прямой.

Примеры. Покажем на примерах различные типы функций распределения и их свойства. Пусть случайная величина ξ может принимать только значения 0 и 1 с вероятностями соответственно p и $1 - p$ (причем $0 \leq p \leq 1$). В этом случае функция распределения имеет вид:

$$F(x) = \begin{cases} 0, & \text{если } x < 0; \\ p, & \text{если } 0 \leq x < 1; \\ 1, & \text{если } x \geq 1. \end{cases}$$

График этой функции изображен на рис. 1.3.

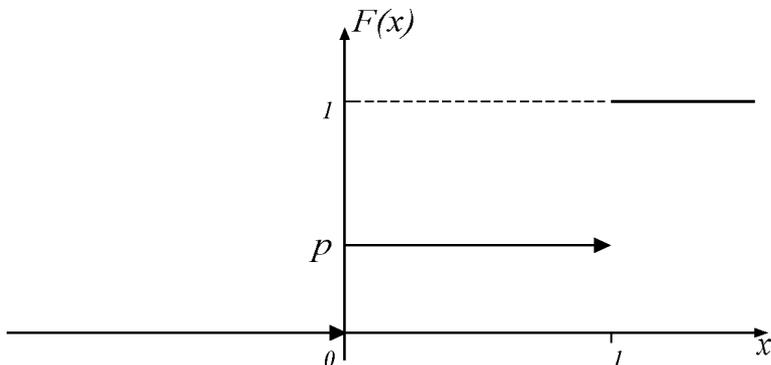


Рис. 1.3. График функции распределения, сосредоточенного в двух точках

Рассмотрим функцию распределения случайной величины более общего вида. Пусть случайная величина ξ принимает конечное число значений a_1, \dots, a_n ,

причем $P(\xi = a_k) = p_k \geq 0$, $(\sum_{k=1}^n p_k = 1)$. График функции этого дискретного распределения изображен на рис. 1.4. (Для удобства предположим, что возможные значения занумерованы в порядке возрастания.)

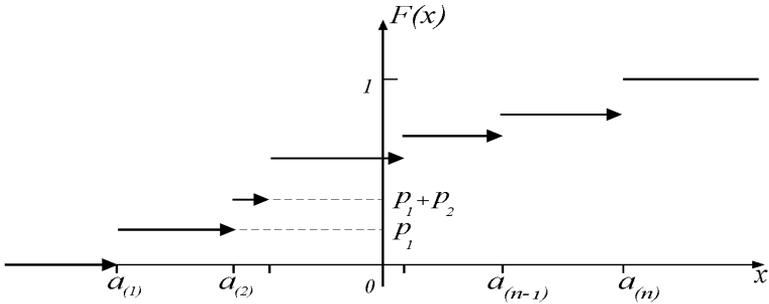


Рис. 1.4. График функции дискретного распределения

Рассмотрим пример непрерывного распределения вероятностей. Пусть функция плотности $p(t)$ равна

$$p(t) = \begin{cases} 0, & \text{если } t < 0; \\ 6t(1-t), & \text{если } 0 \leq t < 1; \\ 0, & \text{если } t \geq 1. \end{cases}$$

(Легко проверить, что в данном случае $\int_{-\infty}^{+\infty} p(t) dt = 1$, $p(t) \geq 0$, так что функция $p(t)$ может быть плотностью случайной величины). Функция распределения в этом примере равна

$$F(x) = \begin{cases} 0, & \text{для } x \leq 0; \\ -2x^3 + 3x^2, & \text{для } 0 \leq x \leq 1; \\ 1, & \text{для } x \geq 1. \end{cases}$$

График этой функции приведен на рис. 1.5.

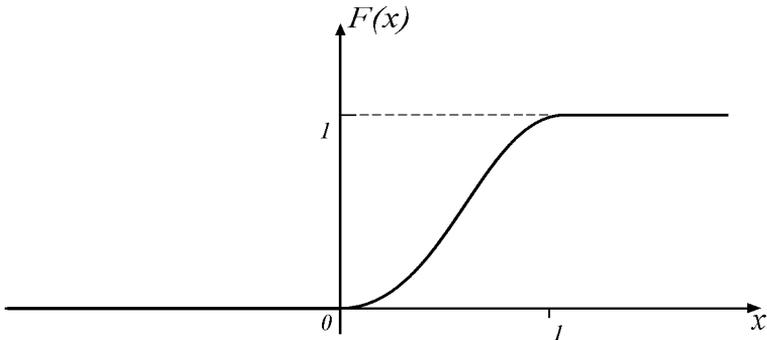


Рис. 1.5. Пример непрерывной функции распределения

В приведенных примерах можно заметить, что $F(x) \rightarrow 0$ при $x \rightarrow -\infty$ и $F(x) \rightarrow 1$, при $x \rightarrow +\infty$, и что $F(x)$ — неубывающая функция. Это общие свойства всех функций распределения.

Если в точке x функция распределения $y = F(x)$ имеет скачок, то величина этого скачка равна вероятности, сосредоточенной в точке x , т.е. вероятности события $\xi = x$. Если же точка x — точка непрерывности функции $y = F(x)$ и, более того, $F(x)$ имеет производную в этой точке, то график $F(x)$ в точке x имеет касательную, тангенс угла наклона которой равен плотности $p(x)$ в этой точке.

1.5. Числовые характеристики распределения вероятностей

Числовые характеристики распределения вероятностей полезны тем, что помогают составить наглядное представление об этом распределении. Наиболее часто употребляемыми характеристиками случайной величины (и соответствующего распределения вероятностей) служат *моменты* и *квантили*. Ниже мы их определим, но надо сделать оговорку: универсальные (пригодные для любых случайных величин) определения этих характеристик требуют весьма сложного математического аппарата (они основаны на теории меры, интеграла Лебега—Стилтьеса и т.д.), поэтому мы приводить их не будем. Вместо этого мы дадим более простые определения для дискретных и для непрерывных случайных величин.

Начнем с так называемого первого момента случайной величины ξ , называемого также *математическим ожиданием*, или *средним значением* ξ . Его обозначают через $M\xi$ или $E\xi$.

Определение. Для дискретной случайной величины ξ со значениями x_1, x_2, \dots , имеющих вероятности p_1, p_2, \dots

$$M\xi = \sum_k x_k p_k.$$

Если число возможных значений ξ конечно, то $M\xi$ всегда существует и не зависит от способа нумерации этих значений. В том случае, если число возможных значений ξ счетно, необходимо, чтобы сумма ряда $\sum_k x_k p_k$ не зависела от нумерации значений x , т.е. чтобы этот ряд сходилась абсолютно ($\sum_k |x_k| p_k < \infty$).

Определение. Для непрерывной случайной величины ξ с плотностью $p(x)$

$$M\xi = \int_{-\infty}^{\infty} x p(x) dx,$$

причем интеграл должен сходиться абсолютно.

Как говорилось выше, приведенные определения $M\xi$ не являются исчерпывающими, поскольку пригодны не для всех видов случайных величин. Общее определение математического ожидания выглядит следующим образом:

$$M\xi = \int x dP_{\xi}(x),$$

где $P_{\xi}(x)$ — распределение вероятностей, порожденное случайной величиной ξ . Приведенные выше формулы для дискретного и непрерывного распределений являются частными случаями этого выражения. Мы не будем пользоваться общим определением, так как это потребует множества математических знаний (о том, что такое $dP(x)$, в каком смысле понимается интеграл и т.д.).

Заметим, что существуют распределения вероятностей без математического ожидания и с такими случайными величинами иногда приходится сталкиваться на практике. Простой пример: пусть случайная величина ξ принимает значения $1^1, 2^2, \dots, n^n, \dots$ с вероятностями $2^{-1}, 2^{-2}$ и т.д. Тогда эта случайная величина не имеет математического ожидания.

Свойства математического ожидания. Перечислим без доказательства основные свойства математического ожидания.

1. Математическое ожидание постоянной равно этой постоянной.
2. Математическое ожидание суммы случайных величин равно сумме их математических ожиданий, т.е.

$$M(\xi + \eta) = M\xi + M\eta.$$

3. Математическое ожидание произведения случайной величины на константу равно произведению этой константы на математическое ожидание случайной величины, т.е.

$$Ma\xi = aM\xi.$$

(другими словами, постоянный множитель можно выносить за знак математического ожидания).

Полезно иметь в виду следующее геометрическое толкование математического ожидания. Пусть $F(x)$ — функция распределения случайной величины ξ . Тогда $M\xi$ равно разности площадей, заключенных ме-

жду осью ординат, прямой $y = 1$ и кривой $y = F(x)$ в интервале $(0, +\infty)$ и между осью абсцисс, кривой $y = F(x)$ и осью ординат в промежутке $(-\infty, 0)$ (см. рис. 1.6). Это правило позволяет во многих случаях находить математическое ожидание почти без вычислений, используя различные свойства функции распределения.

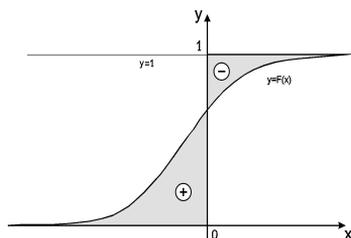


Рис. 1.6. Геометрическая интерпретация математического ожидания

Кроме среднего значения случайной величины, которое в определенном смысле характеризует центр распределения вероятностей, представляет интерес и разброс случайной величины относительно этого центра. Для характеристики (количественного описания) данного разброса в теории вероятностей используют *второй центральный момент* случайной величины. В русскоязычной литературе его называют *дисперсией* и обычно обозначают через $D\xi$.

Определение. *Дисперсией $D\xi$ случайной величины ξ называется величина*

$$D\xi = M(\xi - M\xi)^2 \quad \text{или} \quad D\xi = M\xi^2 - (M\xi)^2.$$

Дисперсия, так же как и математическое ожидание, существует не для всех случайных величин (не для всех распределений вероятностей).

Если необходимо, чтобы показатель разброса случайной величины выражался в тех же единицах, что и значение этой случайной величины, то вместо $D\xi$ используют величину $\sqrt{D\xi}$, которая называется *средним квадратическим отклонением*, или стандартным отклонением случайной величины ξ .

Свойства дисперсии. Из свойств дисперсии отметим следующие:

1. Дисперсия постоянной равна нулю.
2. Для любой неслучайной постоянной a

$$D(\xi + a) = D(\xi), \quad D(a\xi) = a^2 D(\xi).$$

Моменты. Кроме первого и второго моментов, при описании случайных величин иногда используются и другие моменты: третий, четвертый и т.д. Мы дадим их определения отдельно для дискретных и для непрерывных случайных величин.

Определение. Для дискретной случайной величины ξ со значениями x_1, x_2, \dots , имеющих вероятности p_1, p_2, \dots , k -м моментом $M\xi^k$ называется величина $M\xi^k = \sum_i x_i^k p_i$, а k -м центральным моментом называется величина $\sum_i (x_i - M\xi)^k p_i$. Для непрерывной случайной величины с плотностью $p(x)$, k -м моментом называется величина $\int_{-\infty}^{\infty} x^k p(x) dx$, а k -м центральным моментом называется величина $M(\xi - M\xi)^k = \int_{-\infty}^{\infty} (x - M\xi)^k p(x) dx$.

Чтобы приведенные формулы имели смысл, требуется, чтобы суммы и интегралы сходились абсолютно. Так же как математическое ожидание и дисперсия, моменты существуют не для всех случайных величин.

Асимметрия и эксцесс. В отличие от обычных моментов, центральные моменты не меняются при прибавлении к случайной величине постоянного слагаемого, т.е. они не зависят от выбора начала отсчета в шкале измерения случайной величины. Но от выбранной единицы измерения зависимость остается: если, скажем, случайную величину начать измерять не в метрах, а в сантиметрах, то значения центральных моментов также изменятся. Иногда это бывает неудобно. В таких случаях, чтобы устранить подобное влияние, моменты тем или иным способом *нормируют*, например, деля их на соответствующую степень среднего квадратического отклонения. В результате получается безразмерная величина, не зависящая от выбора начала отсчета и единиц измерения исходной случайной величины.

Чаще всего из нормированных моментов используются *асимметрия* и *эксцесс* — соответственно третий и четвертый нормированные центральные моменты. Для случайной величины ξ :

$$\text{асимметрия} = \frac{M(\xi - M\xi)^3}{(D\xi)^{3/2}}, \quad \text{эксцесс} = \frac{M(\xi - M\xi)^4}{(D\xi)^2}.$$

Принято считать, что асимметрия в какой-то степени характеризует несимметричность распределения случайной величины, а эксцесс — степень выраженности «хвостов» распределения, т.е. частоту появления удаленных от среднего значений. Иногда значения асимметрии и эксцесса используют для проверки гипотезы о том, что наблюдаемые данные (выборка) принадлежат заданному семейству распределений, например нормальному (см. п. 2.4). Так, для любого нормального распределения асимметрия равна нулю, а эксцесс — трем.

Квантили. Для случайных величин, принимающих вещественные значения, часто используются такие характеристики, как *квантили*.

Определение. Квантилью x_p случайной величины, имеющей функцию распределения $F(x)$, называется решение x_p уравнения $F(x) = p$.

Величину x_p часто называется p -квантилью или квантилью уровня p распределения $F(x)$. Среди квантилей чаще всего используются *медиана* и *квартили* распределения.

Медианой называется квантиль, соответствующая значению $p = 0.5$. **Верхней квартилью** называется квантиль, соответствующая значению $p = 0.75$. **Нижней квартилью** называется квантиль, соответствующая значению $p = 0.25$.

В описательной статистике (см. ниже) нередко используют *децили*, т.е. квантили уровней $0.1, 0.2, \dots, 0.9$. Знание децилей позволяет неплохо представлять поведение графика $y = F(x)$ в целом.

Отметим, что уравнение $F(x) = p$, определяющее p -квантили, для некоторых значений p , $0 < p < 1$ может не иметь решений либо иметь не единственное решение. Для соответствующей случайной величины ξ это означает, что некоторые p -квантили не существуют, а некоторые определены неоднозначно.

1.6. Независимые и зависимые случайные величины

Введем очень важное понятие *независимости* случайных величин. Это понятие не менее важно, чем понятие независимости событий, и тесно с ним связано. Говоря описательно, случайные величины ξ и η независимы, если независимы любые два события, которые выражаются по отдельности через ξ и η .

Для случайных величин, принимающих вещественные значения, мы можем дать следующее определение.

Определение. *Случайные величины ξ и η независимы, если*

$$P(AB) = P(A)P(B),$$

для любых событий $A = (a_1 < \xi < a_2)$ и $B = (b_1 < \eta < b_2)$, где числа a_1, a_2, b_1 и b_2 могут быть произвольными.

Нам незачем стремиться к большей математической аккуратности в определении независимости случайных величин, поскольку на практике им пользоваться приходится редко. Дело в том, что независимость случайных величин обеспечивается скорее схемой постановки опытов, нежели проверкой математических соотношений. В этом вновь проглядывает аналогия с независимостью событий.

Для независимых случайных величин можно пополнить список свойств математического ожидания и дисперсии:

$$\begin{aligned}M\xi\eta &= M\xi M\eta, \\D(\xi + \eta) &= D\xi + D\eta,\end{aligned}$$

если случайные величины ξ и η независимы и указанные моменты существуют.

Ковариация. Для зависимых случайных величин часто желательно знать степень их зависимости, связи друг с другом. Таких характеристик можно придумать много, но наиболее употребительны из них *ковариация* и *корреляция*.

Определение. Ковариацией $\text{cov}(\xi, \eta)$ случайных величин ξ и η называют

$$\text{cov}(\xi, \eta) = M(\xi - M\xi)(\eta - M\eta),$$

если указанное математическое ожидание существует.

Легко видеть, что верна и другая формула:

$$\text{cov}(\xi, \eta) = M\xi\eta - M\xi M\eta.$$

Поэтому для независимых случайных величин ковариация равна нулю. Обратное, естественно, неверно: равенство нулю ковариации не означает независимости случайных величин (придумайте пример!). Кроме того, ковариация вообще может не существовать (так же как и математические ожидания). Так что обращение в нуль ковариации признаков не является достаточным для их независимости, а только необходимым (и то лишь если ковариация существует).

Из других свойств ковариации отметим, что

$$\text{cov}(A\xi + a, B\eta + b) = AB \text{cov}(\xi, \eta),$$

если A, B, a, b — постоянные (неслучайные) величины.

Корреляция. Использование ковариации в качестве меры связи случайных переменных неудобно, так как величина ковариации зависит от единиц измерения, в которых измерены случайные величины. При переходе к другим единицам измерения (например, от метров к сантиметрам) ковариация тоже изменяется, хотя степень связи случайных переменных, естественно, остается прежней. Поэтому в качестве меры связи признаков обычно используют другую числовую величину, называемую *коэффициентом корреляции*.

Определение. Коэффициентом корреляции случайных величин ξ и η (обозначение $\text{corr}(\xi, \eta)$, либо $\rho(\xi, \eta)$, либо просто ρ) называют

$$\rho = \frac{\text{cov}(\xi, \eta)}{\sqrt{D\xi}\sqrt{D\eta}}.$$

Заметим, что для существования коэффициента корреляции необходимо (и достаточно) существование дисперсий $D\xi > 0$, $D\eta > 0$.

Отметим следующие свойства коэффициента корреляции:

1. Модуль коэффициента корреляции не меняется при линейных преобразованиях случайных переменных: $|\rho(\xi, \eta)| = |\rho(\xi', \eta')|$, где $\xi' = a_1 + b_1\xi$, $\eta' = a_2 + b_2\eta$, a_1, b_1, a_2, b_2 — произвольные числа.
2. $|\rho(\xi, \eta)| \leq 1$.
3. $|\rho(\xi, \eta)| = 1$ тогда и только тогда, когда случайные величины ξ и η линейно связаны, т.е. существуют такие числа a, b , что

$$P(\eta = a\xi + b) = 1.$$

4. Если ξ и η статистически независимы, то $\rho(\xi, \eta) = 0$. Уже отмечалось, что обратное заключение, вообще говоря, неверно. Об этом мы еще будем говорить.

Свойства 1 и 4 проверяются непосредственно. Докажем свойства 2 и 3 (при желании читатель может эти доказательства пропустить). Пусть t — переменная величина в смысле математического анализа. Рассмотрим дисперсию случайной величины $D(\eta - t\xi)$ как функцию переменной t . По свойствам дисперсии $D(\eta - t\xi) = t^2 D\xi - 2t \text{cov}(\xi, \eta) + D\eta$, т.е. она представляется квадратным трехчленом от t . Этот квадратный трехчлен неотрицателен, поскольку дисперсия всегда неотрицательна. Поэтому его дискриминант $[\text{cov}(\xi, \eta)]^2 - D\xi D\eta \leq 0$, а это и означает, что $|\rho(\xi, \eta)| \leq 1$ (свойство 2).

Для доказательства свойства 3 заметим, что при $|\rho(\xi, \eta)| = 1$ дискриминант приведенного выше квадратного трехчлена обращается в 0, а поэтому при некотором t_0 значение $D(\eta - t_0\xi)$ равно нулю. Равенство нулю дисперсии означает, что эта случайная величина постоянна, т.е. для некоторого c вероятность $P(\eta - t_0\xi = c)$ равна единице, что и требовалось доказать.

Итак, корреляция случайных величин принимает значения от -1 до 1 и может быть равна ± 1 , только если эти величины линейно зависят друг от друга. Значения корреляции, близкие к -1 или 1 , указывают, что зависимость случайных величин друг от друга почти линейная. Значения ковариации, близкие к нулю, означают, что связь между случайными величинами либо слаба, либо не носит линейного характера. Подробнее о связи между случайными величинами мы расскажем в главе 9.

1.7. Случайный выбор

Значительная часть статистики связана с описанием больших совокупностей объектов. Если интересующая нас совокупность слишком многочисленна, либо ее элементы малодоступны, либо имеются другие причины, не позволяющие изучать сразу все ее элементы, прибегают к изучению какой-то части этой совокупности. Эта выбранная для полного исследования группа элементов называется *выборкой* или *выборочной совокупностью*, а все множество изучаемых элементов — *генеральной совокупностью*. Естественно стремиться сделать выборку так, чтобы она наилучшим образом представляла всю генеральную совокупность, т.е. была бы, как говорят, *репрезентативной*. Как этого добиться? Если генеральная совокупность нам мало известна или совсем неизвестна, не удастся предложить ничего лучшего, чем чисто случайный выбор. Дадим его определение, начав со случайного выбора одного объекта.

Определение. *Выбор одного объекта называют чисто случайным, если все объекты имеют равные вероятности оказаться выбранными.*

Если речь идет о выборе одного объекта из N , это означает, что для каждого элемента вероятность выбора равна $1/N$.

Определение. *Выбор n объектов из N называют чисто случайным, если все наборы из n объектов имеют одинаковые вероятности быть выбранными.*

Чисто случайный выбор n объектов (иногда говорят — *случайную выборку объема n*) можно получить, извлекая из генеральной совокупности по одному объекту последовательно и чисто случайно.

Нарушение принципов случайного выбора порой приводило к серьезным ошибкам. Стал знаменитым своей неудачей опрос, проведенный американским журналом «Литературное обозрение» относительно исхода президентских выборов в США в 1936 г.

Кандидатами на этих выборах были Ф.Д. Рузвельт и А.М. Ландон. В качестве генеральной совокупности редакция журнала использовала телефонные книги. Отобрав случайно 4 миллиона адресов, она разослала по всей стране открытки с вопросом об отношении к кандидатам в президенты. Затратив большую сумму на рассылку и обработку открыток, журнал объявил, что на предстоящих выборах президентом США с большим перевесом будет избран Ландон. Результат выборов оказался противоположным этому прогнозу.

Здесь были совершены сразу две ошибки — во-первых, телефонные книги сами по себе дают не репрезентативную выборку из населения страны, хотя бы потому, что абоненты — в основном зажиточные главы семейств. Во-вторых, прислали ответы не все, а люди, не только достаточно уверенные в своем мнении, но и привыкшие отвечать на письма, т.е. в значительной части

представители делового мира, которые и поддерживали Ландона. Если бы редакция критически подошла к своей работе, она поняла бы, что методика опроса страдает изъянами.

Явление, подобное только что описанному, когда выборка представляет не всю генеральную совокупность, а лишь какой-то ее слой, какую-то ее часть, называется *смещением выборки*. Смещение — один из основных источников ошибок при использовании выборочного метода.

Однако для тех же самых президентских выборов социологи Дж. Гэллуп и Э. Роупер правильно предсказали победу Рузвельта, основываясь только на 4 тысячах анкет. Причиной этого успеха, прославившего его авторов, было не только правильное составление выборки. Они учли, что общество распадается на социальные группы, которые более однородны, в том числе по своим политическим взглядам. Поэтому выборка из слоя может быть относительно малочисленной с тем же результатом точности. Имея результаты обследования по слоям, можно характеризовать общество в целом. Сейчас такая методика является общепринятой.

Мы не станем обсуждать, как следует организовывать случайный выбор на практике, если генеральная совокупность — это реальные объекты. Но отметим, что при этом возникают свои проблемы и соответственно средства их разрешения. Подробно с этим кругом вопросов можно познакомиться в [54].

1.8. Выборки и их описание

1.8.1. Что такое выборка

В предыдущем параграфе мы использовали слово «выборка» для описания результата случайного выбора нескольких объектов из некоторой заданной генеральной совокупности. В этом смысле слово «выборка» используется, когда мы говорим «социологический опрос произведен на выборке из 2000 человек (респондентов)». Но в математической литературе слово «выборка» гораздо чаще используется в другом смысле. Дадим его определение.

Определение. *Выборкой называют последовательность независимых одинаково распределенных случайных величин.*

Именно в этом значении слово «выборка» употребляется в статистических задачах естествознания, и в этом значении оно будет встречаться далее, в этой книге.

Замечание. Происхождение данного значения слова «выборка» связано с давними ассоциациями всякого случайного испытания со случайным выбором из некоей совокупности. Если эта совокупность является конечной (как это и бывает на практике), то последовательные результаты случайных выборов из

нее не являются независимыми, поскольку каждое изъятие элемента из совокупности изменяет эту совокупность. Конечно, для обширных совокупностей извлечение одного или нескольких элементов мало изменяет вероятности выбора, но все же они не остаются постоянными в процессе выбора. В связи с этим иногда говорят о *бесконечных генеральных совокупностях* (популяциях) и о случайном выборе из них. Это образное выражение может сделать более наглядным представление о независимых случайных величинах.

1.8.2. Выборочные характеристики

Перечисленные в параграфе 1.4 характеристики случайной величины существенно опираются на знание закона ее распределения $F(x)$. Для практических задач такое знание — редкость. Здесь закон распределения обычно неизвестен, в лучшем случае он известен с точностью до некоторых неизвестных параметров. Как же тогда получить сведения о распределении случайной величины и его характеристиках? Это становится возможным, когда имеются независимые многократные повторения опыта, в котором мы измеряем значения интересующей нас случайной величины.

Предположим, что наблюдения над случайной величиной ξ можно повторять независимо и в неизменных условиях, получая ее независимые реализации x_1, x_2, \dots, x_n . Тогда x_1, x_2, \dots, x_n будут независимыми одинаково распределенными случайными величинами, т.е. *выборкой*. Зная величины x_1, x_2, \dots, x_n , мы можем построить приблизительные значения для функции распределения и других характеристик случайной величины ξ . Это и позволяет нам изучать свойства случайных величин, не зная их законов распределения.

Замечание. Мы уже встречались с идеей независимых повторений случайного опыта в неизменных условиях, когда обсуждали измерения вероятностей событий. Возвращение к этой идее не удивительно, поскольку для описания распределения случайной величины ξ мы как раз и должны уметь указывать вероятности всех событий, выражаемых через ξ .

Расскажем о том, как по имеющейся выборке можно получить приближенные значения для характеристик случайных величин. Начнем с функции распределения случайной величины.

Эмпирическая функция распределения.

Определение. *Выборочной (эмпирической) функцией распределения случайной величины ξ , построенной по выборке x_1, x_2, \dots, x_n , называется функция $F_n(x)$, равная доле таких значений x_i , что $x_i \leq x$, $i = 1, \dots, n$.*

Иначе говоря, $F_n(x)$ есть частота события $x_i \leq x$ в ряду x_1, x_2, \dots, x_n .

Для построения выборочной функции распределения удобно от выборки x_1, \dots, x_n перейти к вариационному ряду $x_{(1)}, \dots, x_{(n)}$.

Определение. *Вариационным рядом называют выборку, переименованную в порядке возрастания.*

Так, $x_{(1)}$ обозначает наименьшее из чисел x_1, \dots, x_n , $x_{(2)}$ — наименьшее из оставшихся после удаления $x_{(1)}$ и т.д. В частности, $x_{(n)}$ обозначает наибольшее из x_1, \dots, x_n . При $x < x_{(1)}$, по определению, $F_n(x) = 0$, в точке $x_{(1)}$ функция $F_n(x)$ совершает скачок, равный $1/n$, и остается постоянной до значения $x_{(2)}$, и т.д. Таким образом, выборочная функция распределения является ступенчатой с точками скачков $x_{(1)}, \dots, x_{(n)}$, причем величина каждого скачка равна $1/n$ (рис. 1.7).

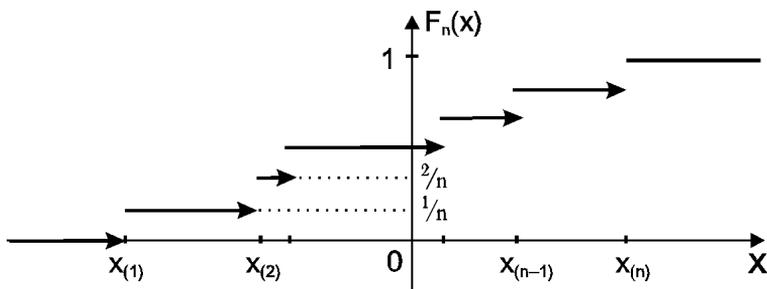


Рис. 1.7. Общий вид эмпирической функции распределения

Видно, что график эмпирической функции распределения напоминает график дискретного распределения вероятностей. Это не случайно: эмпирическую функцию выборки x_1, \dots, x_n можно рассматривать как функцию распределения вероятностей, где каждому значению x_i , $i = 1, \dots, n$ приписана вероятность $1/n$. Иногда поэтому вместо эмпирической (или выборочной) функции распределения употребляют название «функция распределения выборки».

Связь между эмпирической функцией распределения и функцией распределения (иногда, чтобы подчеркнуть разницу, говорят о теоретической функции распределения, что не вполне правильно, ибо никакой теории здесь нет) основана на уже упомянутой теореме Бернулли. Она такая же, как связь между частотой события и его вероятностью. Для любого числа x значение $F_n(x)$ представляет собой частоту события ($\xi \leq x$) в ряду из n независимых повторений. Поэтому $F_n(x) \rightarrow F(x)$ при $n \rightarrow \infty$.

Установлено, что выборочная функция распределения с ростом объема выборки n равномерно по x аппроксимирует теоретическую

функцию распределения $F(x)$ случайной величины ξ , т.е. величина $\sup_x |F_n(x) - F(x)|$ стремится к нулю при $n \rightarrow \infty$ с вероятностью 1.

Выборочные характеристики. На указанном выше свойстве выборочной функции распределения основаны многие методы математической статистики. Замена функции распределения $F(x)$ на ее выборочный аналог $F_n(x)$ в определении математического ожидания, дисперсии, медианы и т.п. приводит к *выборочному среднему, выборочной дисперсии, выборочной медиане* и т.д. Покажем, как действует это правило и чему равны соответствующие выборочные характеристики.

В случае математического ожидания, используя в качестве функции распределения случайной величины ξ выборочную функцию $F_n(x)$, мы подразумеваем, что некая случайная величина может принять значения $x_{(1)}, \dots, x_{(n)}$, каждое с вероятностью $1/n$. Воспользовавшись формулой для определения математического ожидания для дискретной случайной величины, приходим к следующему определению.

Средним значением выборки (выборочным средним), или *выборочным аналогом математического ожидания, называется величина*

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

Аналогично

Дисперсией выборки (выборочной дисперсией), или *выборочным аналогом дисперсии, называется величина*

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

Однако в статистике чаще в качестве выборочной дисперсии используют

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

поскольку математическое ожидание величины s^2 равно дисперсии ξ , т.е. $Ms^2 = D\xi$.

Выборочной квантилью называется решение уравнения

$$F_n(x) = p.$$

В частности, *выборочная медиана* есть решение уравнения

$$F_n(x) = 0.5.$$

Замечание. Решение уравнения $F_n(x) = 0.5$ при четном $n = 2k$ определено не однозначно. Действительно, для каждого x из промежутка $x_{(k)} \leq x < x_{(k+1)}$ $F(x) = 0.5$. В этом случае условились определить выборочную медиану как $\frac{x_{(k)} + x_{(k+1)}}{2}$. При нечетном $n = 2k + 1$ решение уравнения $F_n(x) = 0.5$ не существует, так как выборочная функция распределения принимает только значения из множества $\left\{ \frac{i}{2k+1}, i = 0, 1, \dots, 2k + 1 \right\}$. В связи с этим выборочную медиану определяют как $x_{(k+1)}$, ибо в этой точке $F_n(x)$ переходит через $1/2$. Выборочная медиана разбивает выборку пополам: слева и справа от нее оказывается одинаковое число элементов выборки. Заметим, что при больших значениях n : $F_n(x_{(k+1)}) = \frac{(k+1)}{2k+1} \rightarrow \frac{1}{2}$.

Важным свойством выборочных характеристик является то, что все они сходятся к соответствующим теоретическим характеристикам при растущих объемах выборки n . Характер этой сходимости будет рассмотрен в гл. 4 и 5, когда речь пойдет о законе больших чисел и о построении статистических оценок различных параметров распределения.

Выборочные ковариация и корреляция. Если в каждом наблюдении мы регистрируем значения не одной, а двух (или нескольких) случайных величин одновременно, мы получаем в результате двумерную (или многомерную) выборку. Для таких выборок тоже можно говорить о числовых характеристиках, например, о ковариации или корреляции компонент этой выборки.

Коэффициентом корреляции двумерной выборки $(x_1, y_1), \dots, (x_n, y_n)$, или **выборочным коэффициентом корреляции**, называют величину

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

(Иногда ее называют коэффициентом корреляции К. Пирсона.) Аналогично определяется **выборочная ковариация**, она равна $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$.

1.8.3. Ранги и ранжирование

Ранги. Во многих случаях имеющиеся в нашем распоряжении числовые данные (например, значения элементов выборки) носят в той или иной мере условный характер. Например, эти данные могут быть тестовыми баллами, экспертными оценками, данными о вкусовых или политических предпочтениях опрошенных людей и т.д. Анализ таких данных требует особой осторожности, поскольку многие предпосылки классических статистических методов (например, предположения о каком-либо конкретном, скажем нормальном, законе распределения)

для них не выполняются. Твердую основу для выводов здесь дают только соотношения между наблюдениями типа «больше-меньше», так как они не меняются при изменении шкалы измерений. Например, при анализе анкет с данными о симпатиях избирателей к политическим деятелям мы можем сказать, что политик, получивший больший балл в анкете, более симпатичен отвечавшему на вопросы человеку (респонденту), чем политик, получивший меньший балл. Но на сколько (или во сколько раз) он более симпатичен, сказать нельзя, так как для предпочтений нет объективной единицы измерения.

В подобных случаях (которые мы будем более подробно рассматривать в последующих главах) имеет смысл вообще отказаться от анализа конкретных значений данных, а исследовать только информацию об их взаимной упорядоченности. Для этого от исходных числовых данных осуществляют переход к их *рангам*.

Определение. *Рангом наблюдения называют тот номер, который получит это наблюдение в упорядоченной совокупности всех данных — после их упорядочения по определенному правилу (например, от меньших значений к большим или наоборот).*

Чаще всего упорядочение чисел (набор которых составляют упомянутые выше данные) производят по величине — от меньших к большим. Именно такое упорядочение и связанное с ним ранжирование (присвоение рангов) мы будем иметь в виду в дальнейшем.

Пример. Пусть выборка состоит из чисел 6, 17, 14, 5, 12. Тогда рангом числа 6 оказывается 2, рангом 17 будет 5 и т.д.

Определение. *Процедура перехода от совокупности наблюдений к последовательности их рангов называется ранжированием. Результат ранжирования называется ранжировкой.*

Статистические методы, в которых мы делаем выводы о данных на основании их рангов, называются ранговыми. Они получили широкое распространение, так как надежно работают при очень слабых предположениях об исходных данных (не требуя, например, чтобы эти данные имели какой-либо конкретный закон распределения). В последующих главах этой книги мы рассмотрим применение ранговых методов в наиболее распространенных практических задачах.

Средние ранги. Трудности в назначении рангов возникают, если среди элементов выборки встречаются совпадающие. (Так часто бывает, когда данные регистрируются с округлением.) В этом случае обычно используют *средние ранги*.

Средние ранги вводятся так. Предположим, что наблюдение x_i имеет ту же величину, что и некоторые другие из общего числа n наблюдений. (Эту совокупность одинаковых наблюдений из набора x_1, \dots, x_n называют *связкой*; количество таких одинаковых наблюдений в данной связке называют ее размером.) Средний ранг x_i в ранжировке наблюдений x_1, \dots, x_n есть среднее арифметическое тех рангов, которые были бы назначены x_i и всем остальным элементам связки, если бы одинаковые наблюдения оказались различны.

В качестве примера рассмотрим выборку 6, 17, 12, 6, 12. Ее ранжировка равна $1\frac{1}{2}$, 5, $3\frac{1}{2}$, $1\frac{1}{2}$, $3\frac{1}{2}$.

1.8.4. Методы описательной статистики

В практических задачах мы обычно имеем совокупность наблюдений x_1, x_2, \dots, x_n , на основе которых требуется сделать те или иные выводы. Часто этих наблюдений много — несколько десятков, сотен или тысяч, так что возникает задача компактного описания имеющихся наблюдений. В идеале таким описанием могло бы быть утверждение, что x_1, x_2, \dots, x_n являются выборкой, т.е. независимыми реализациями случайной величины ξ с известным законом распределения $F(x)$. Это позволило бы теоретически провести расчеты всех необходимых исследователю характеристик наблюдаемого явления.

Однако далеко не всегда мы можем утверждать, что x_1, x_2, \dots, x_n являются независимыми и одинаково распределенными случайными величинами. Во-первых, это не так-то просто проверить (для подтверждения этого требуются значительные объемы наблюдений и специальные, порой многочисленные, тесты). А во-вторых, часто заведомо известно, что это не так. Поэтому для компактного описания совокупности наблюдений x_1, x_2, \dots, x_n используют другие методы — методы описательной статистики.

Определение. *Методами описательной статистики принято называть методы описания выборок x_1, x_2, \dots, x_n с помощью различных показателей и графиков.*

Полезность методов описательной статистики состоит в том, что несколько простых и довольно информативных статистических показателей способны избавить нас от просмотра сотен, а порой и тысяч значений выборки.

Показатели описательной статистики. Описывающие выборку показатели можно разбить на несколько групп.

1. *Показатели положения* описывают положение данных на числовой оси. Примеры таких показателей — минимальный и максимальный элементы выборки (первый и последний члены вариационного ряда), верхний и нижний квартили (они ограничивают зону, в которую попадают 50% центральных элементов выборки). Наконец, сведения о середине совокупности могут дать выборочное среднее значение, выборочная медиана и другие аналогичные характеристики.
2. *Показатели разброса* описывают степень разброса данных относительно своего центра. К ним в первую очередь относятся: дисперсия выборки, стандартное отклонение, размах выборки (разность между максимальным и минимальным элементами), межквартильный размах (разность между верхней и нижней квартилью), коэффициент эксцесса и т.п. По сути дела, эти показатели говорят, насколько кучно основная масса данных группируется около центра.
3. *Показатели асимметрии*. Третья группа показателей отвечает на вопрос о симметрии распределения данных около своего центра. К ней можно отнести: коэффициент асимметрии, положение выборочной медианы относительно выборочного среднего и относительно выборочных квартилей, гистограмму и т.д.
4. *Показатели, описывающие закон распределения*. Наконец, четвертая группа показателей описательной статистики дает представление собственно о законе распределения данных. Сюда относятся графики гистограммы и эмпирической функции распределения, таблицы частот.

Применение показателей описательной статистики. Из перечисленных выше характеристик на практике по традиции чаще всего используются выборочное среднее, медиана и дисперсия (или стандартное отклонение). Однако для получения более точных и достоверных выводов мы настоятельно рекомендуем внимательно изучать и другие из перечисленных выше характеристик, а также обращать внимание на условия получения выборочных совокупностей.

Особое внимание следует обратить на наличие в выборке *выбросов* — грубых (ошибочных), сильно отличающихся от основной массы наблюдений. Дело в том, что даже одно или несколько грубых наблюдений способны сильно исказить такие выборочные характеристики, как среднее, дисперсия, стандартное отклонение, коэффициенты асимметрии и эксцесса. Проще всего обнаружить такие наблюдения с помощью перехода от выборки к ее вариационному ряду или гистограммы с достаточно большим числом интервалов группировки (см. ниже).

Подозрение о присутствии таких наблюдений может возникнуть, если выборочная медиана заметно отличается от выборочного среднего, хотя в целом совокупность симметрична; если положение медианы сильно несимметрично относительно минимального и максимального элементов выборки, и т.д.

Замечание. Наличие выбросов, т.е. грубых (ошибочных) наблюдений, может не только сильно исказить значения выборочных показателей — выборочного среднего, дисперсии, стандартного отклонения и т.д., — но и привести к многим другим ошибочным выводам. Дело в том, что большинство традиционных статистических методов весьма чувствительно к отклонениям от условий применимости метода. К сожалению, интенсивно развивающиеся в последние два десятилетия статистические методы, устойчивые к выбросам и другим отклонениям, еще не получили широкого распространения на практике, за исключением ранговых процедур для наиболее стандартных задач. Отчасти причиной здесь является значительная вычислительная сложность этих методов, из-за чего их применение невозможно без использования специальных компьютерных программ.

1.8.5. Наглядные методы описательной статистики

Рассмотренные выше вопросы и понятия дают первое представление о теоретических и выборочных характеристиках случайных величин. С различной степенью подробности и строгости этот материал изложен во многих учебниках по теории вероятностей и математической статистике, выбор которых должен определяться направленностью интересов и уровнем математической подготовки читателя.

Группировки. Нередко (для облегчения регистрации или при невысокой точности измерений) данные группируют, т.е. числовую ось разбивают на промежутки и для каждого промежутка указывают число n_j элементов выборки x_1, \dots, x_n , которые в него попали (здесь j — номер промежутка). Ясно, что $\sum_j n_j = n$.

В этом случае в качестве выборочного среднего и дисперсии используют следующие величины. Пусть t_1, t_2, \dots — центры (середины) выбранных промежутков. Тогда вместо выборочного среднего \bar{x} используют величину \bar{t} :

$$\bar{x} \simeq \bar{t} = \frac{\sum_j t_j n_j}{n} = \sum_j t_j \frac{n_j}{n},$$

а в качестве выборочной дисперсии s^2 :

$$s^2 \simeq \frac{1}{n-1} \sum_j (t_j - \bar{t})^2 n_j.$$

Приведем еще несколько полезных приемов описательной статистики для работы с выборкой. В качестве примера рассмотрим данные из табл. 1.1, в которой приведены результаты измерения диаметров 200 головок заклепок. Здесь случайная величина — диаметр изготавливаемой заклепки, приведенные 200 значений — ее независимые реализации.

Точечная диаграмма. Данные, собранные в таблицу, трудно обозреть. Они нуждаются в наглядном представлении. Одной из форм такого наглядного представления служит *точечная диаграмма*: табличные данные отмечаются точками на числовой шкале. Если некоторое число встречается в таблице несколько раз, его представляют соответствующим количеством точек. Точечная диаграмма для данных табл. 1.1 приведена на рис. 1.8.

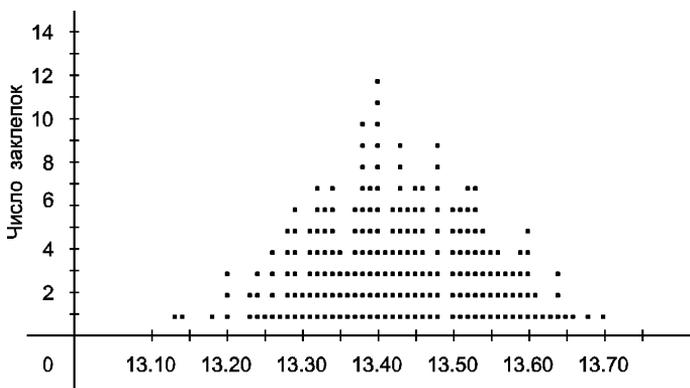


Рис. 1.8. Точечная диаграмма. Распределение диаметров 200 головок заклепок, выраженных в мм

Эта диаграмма удобна в том случае, когда одно и то же значение случайной величины повторяется в выборке несколько раз. В противном случае точечная диаграмма сводится к последовательности точек на оси абсцисс. Во всех случаях точечная диаграмма помогает построить график выборочной функции распределения.

Гистограмма. Более наглядное описание данных достигается путем группировки наблюдений в классы. Под группировкой, или классификацией, мы будем понимать некоторое разбиение интервала, содержащего все n наблюдаемых результатов x_1, \dots, x_n на m интервалов, которые будем называть *интервалами группировки*. Длины интервалов обозначим через $\Delta_1, \dots, \Delta_m$, а середины интервалов группировки — через t_1, \dots, t_m .

Число наблюдений n_{ij} в j -м интервале группировки равно количеству x_i , $i = 1, \dots, n$, удовлетворяющих неравенству

$$|x_i - t_j| < \frac{1}{2}\Delta_j.$$

Определим величину $h_j = n_j/n$, которая означает частоту попадания наблюдений в j -й интервал группировки. Для того чтобы избавиться от влияния размера интервала группировки на h_j , вводится величина $f_j = h_j/\Delta_j$.

Определение. *Графическое изображение зависимости частоты попадания элементов выборки от соответствующего интервала группировки называется гистограммой выборки.*

Подчеркнем, что в качестве ординаты здесь берется не сама частота, а частота, деленная на длину интервала группировки. Если все интервалы группировки имеют одинаковую длину, деление на Δ обычно опускают и n_j или h_j используют как ординаты, как это показано на нескольких рисунках ниже. На рис. 1.9 приведена гистограмма выборки при длине интервала группировки, равной 0.01 мм. Ординатой на этом рисунке является число заклепок в каждом интервале группировки.

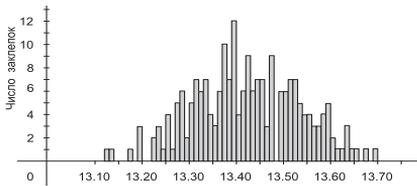


Рис. 1.9. Гистограмма. Длина интервала группировки равна 0.01 мм

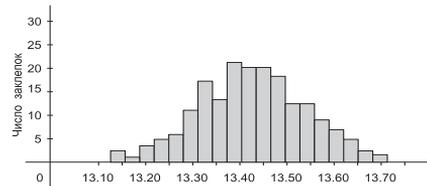


Рис. 1.10. Гистограмма. Длина интервала группировки равна 0.03 мм

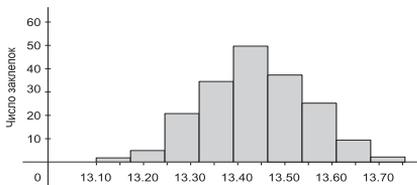


Рис. 1.11. Гистограмма. Длина интервала группировки равна 0.07 мм

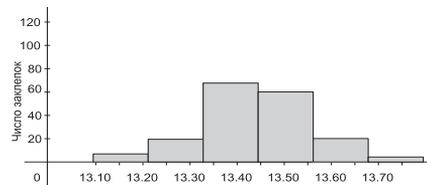


Рис. 1.12. Гистограмма. Длина интервала группировки равна 0.11 мм

Отметим, что согласно определению площадь каждого столбца гистограммы равна (точнее, пропорциональна) частоте попадания наблюдений в данный интервал группировки.

Ясно, что величина интервала группировки существенно влияет на общий вид гистограммы. Если длина интервала группировки мала, то

влияние случайных колебаний начинает преобладать, так как каждый интервал содержит при этом лишь небольшое число наблюдений. Этот эффект хорошо виден на рис. 1.9. На рис. 1.10—1.12 приведены гистограммы выборки при длине интервала группировки, равной 0.03, 0.07 и 0.11 мм соответственно. Из приведенных рисунков видно, что чем больше величина интервала группировки, тем более скрадываются характерные черты распределения.

Если группированное распределение должно являться основой для последующих вычислений, то, как правило, все интервалы группировки должны быть небольшими и иметь одну и ту же длину.

Пример. О пользе наглядных приемов описательной статистики красноречиво говорит следующий пример, относящийся еще к началу нашего века. Мы изложим его, следуя Р. Фишеру (одному из создателей современной математической статистики).

... Иоханес Шмидт из Карлсбергской лаборатории в Копенгагене был не только ихтиологом, но и неутомимым биостатистиком. Он развивал идею, что рыбы одного вида распадаются на относительно изолированные сообщества. Между этими группами он находил статистические различия по числу позвонков или лучей плавников. Для доказательства этого он строил гистограммы распределений числа позвонков (лучей плавников) для каждой из групп и сравнивал их между собой. Причиной различий сообществ рыб служит то, что эти сообщества не смешиваются при размножении: каждая группа мечет икру в своем месте. Часто такие различия были заметны даже между стаями рыб одного вида, обитавшими в одном фьорде.

Однако для угрей Шмидт не смог найти никаких статистических различий между выборками, выловленными даже в очень далеких друг от друга местах — будь то различные части европейского материка, Азорские острова, Нил или Исландия. Шмидт решил, что угри всех различных речных систем составляют одно сообщество, а значит, они должны иметь общее место размножения.

Через некоторое время это предположение подтвердилось в ходе экспедиции исследовательского судна «Диана». Одним из главных успехов этого плавания была поимка личинок угря в некотором ограниченном районе Западной Атлантики — Саргассовом море. Выяснилось, что все угри, независимо от своего «места жительства», отправляются выводить потомство только в Саргассово море.

1.9. Методы описательной статистики в пакете SPSS

В базовом модуле пакета SPSS широко представлены численные и графические методы описательной статистики. Процедуры, целенаправленно вычисляющие только описательные статистики, сгруппированы в пункте «Descriptive Statistics» (описательные статистики) меню статистических процедур пакета «Analyze» (анализ) (рис. 1.13).

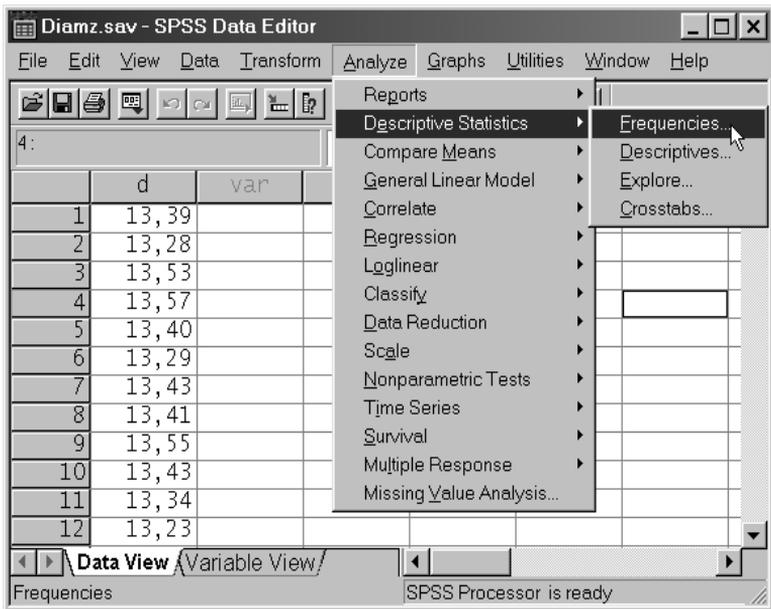


Рис. 1.13. Пакет SPSS. Редактор данных с меню «Analyze» и «Descriptive Statistics»

Рассмотрим рис. 1.13, кратко опишем назначения процедур, входящих в меню «Descriptive Statistics», и разберем на примерах наиболее употребительные из них.

Frequencies (частоты) — позволяет получить широкий набор числовых характеристик, включая частоты, проценты, накопленные (кумулятивные) проценты, среднее, дисперсию, стандартное отклонение, медиану, моду, сумму значений, минимальное и максимальное значения переменных, асимметрию, эксцесс, стандартные ошибки оценок асимметрии и эксцесса, квартили, процентиля, столбиковые диаграммы, гистограммы и др. для одной или нескольких выборок. Работа этой процедуры разобрана в примере 1.1к.

Descriptives (описательные статистики) — вычисляет основные описательные статистики: среднее значение, его стандартную ошибку, минимальное и максимальное значения, дисперсию, стандартное отклонение, размах, асимметрию и эксцесс и другие характеристики одной или нескольких выборок. Основное отличие этой процедуры от **Frequencies** в том, что она вычисляет нормированные значения выборок и сохраняет их в отдельных переменных редактора данных. Эту процедуру лучше всего применять для данных, закон распределения которых близок к нормальному.

Explore (разведочный анализ) — позволяет вычислить те же описательные статистики, что и две предыдущие процедуры, не только для нескольких выборок, но и для их подгрупп. Кроме того, эта процедура может выдавать различные устойчивые оценки среднего значения выборок, включая усеченные оценки, М-оценки Хубера и другие робастные оценки (см. [116], [108]), а также строить

диаграммы «ящик с усами» и «ствол-лист» (см. [98]), графики на нормальной вероятностной бумаге (см. п. 5.2) и многое другое.

Crosstabs (таблицы сопряженности) — используется для выяснения связи двух или нескольких переменных, измеренных в номинальных или порядковых шкалах (см. п. 9.3). Эта процедура строит двумерные и многомерные таблицы сопряженности и проверяет гипотезу о независимости переменных с помощью критерия хи-квадрат Пирсона и вычисляет различные меры связи между двумя переменными.

Различные элементы методов описательной статистики входят и во многие другие статистические процедуры пакета. В довольно полном объеме они представлены также в процедуре «Means» блока «Compare means» (см. рис. 1.13).

Рассмотрим несколько примеров.

Пример 1.1к. Для выборки диаметров головок заклепок (табл. 1.1) вычислить среднее значение, медиану, дисперсию, нижнюю и верхнюю квартили, а также минимальный и максимальный элементы.

Подготовка данных. В окне редактора базы данных пакета создать числовую переменную с именем **d** и ввести в нее (для определенности — по столбцам) значения из табл. 1.1. Если уже есть готовый файл данных, например **DIAMZ.sav**, загрузить его из пункта **Open** меню **File** панели управления пакета. (Данные в SPSS хранятся в собственном «экономном» формате. Файлы этого формата имеют расширение **sav**. Пакет обладает большими возможностями загрузки файлов других форматов.)

Выбор процедуры. В меню **Analyze** выбрать блок процедур **Descriptive Statistics**. В окне этого блока выбрать процедуру **Frequencies**, как это показано на рис. 1.13.

Выполнение процедуры **Frequencies**, как и большинства других статистических и графических процедур пакета, начинается с заполнения полей окна ввода данных и настройки параметров процедуры. Это окно представлено на рис. 1.14.

Заполнение полей окна ввода данных и параметров. Выделить щелчком мыши переменную **d** в левой части окна. (В этой части автоматически отображаются все переменные, загруженные в редактор пакета.) Перенести ее в окно **Variable(s)** — анализируемых переменных, щелкнув мышкой на стрелке переноса в центре окна. Затем перейти в меню настройки параметров выдачи числовых результатов (рис. 1.15), нажав кнопку **Statistics** в нижней части окна (см. рис. 1.14). (Кнопка **Charts** задает настройки вывода графиков и диаграмм.)

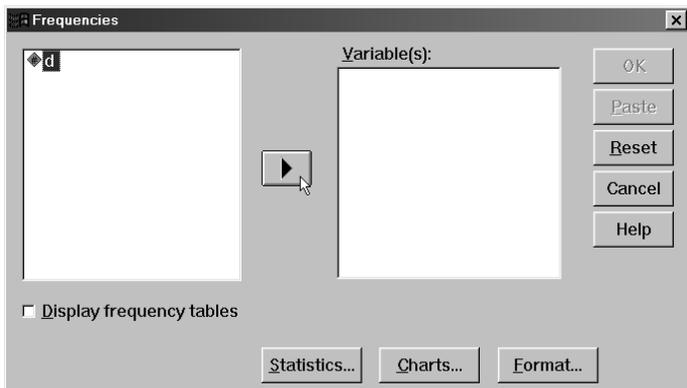


Рис. 1.14. Пакет SPSS. Окно ввода данных и параметров процедуры «Frequencies»

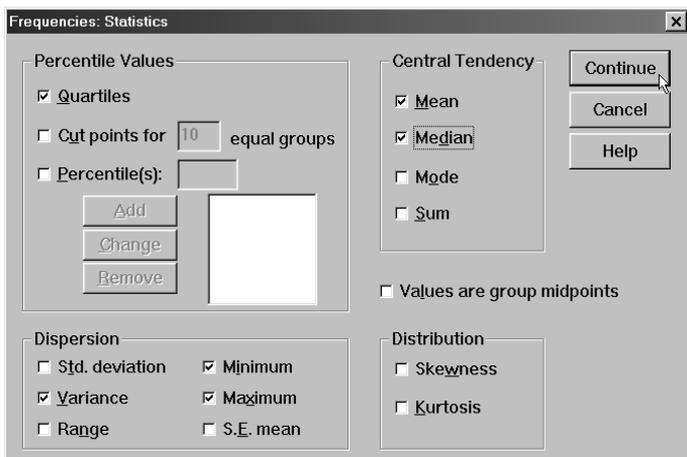


Рис. 1.15. Пакет SPSS. Окно задания вывода статистических характеристик данных процедуры «Frequencies»

Окно задания вывода статистических характеристик разбито на четыре крупных блока, каждый из которых отвечает за ту или иную группу описательных статистик выборки.

Перечислим эти группы и дадим перевод входящих в них характеристик.

Группа **Percentile Values**

Quartiles — квантили

Cut points for n equal groups — точки разбиения выборки на n равных групп

Percentile(s) — проценты

Группа **Central Tendency** (положение центра) характеризует положение центра выборки и включает следующие описательные статистики:

Mean — среднее значение
Median — медиана
Mode — мода
Sum — сумма

Группа **Dispersion** (разброс) включает различные показатели разброса выборки:

Std. deviation — стандартное отклонение
Variance — дисперсия
Range — размах
Minimum — минимум
Maximum — максимум
S.E. mean — стандартная ошибка среднего значения

Группа **Distribution** (распределение) включает набор статистик, характеризующих форму распределения выборки.

Skewness — коэффициент асимметрии
Kurtosis — коэффициент эксцесса

В окне задания вывода статистических характеристик отметить мышью требуемые в задаче характеристики, как это показано на рис. 1.15, и нажать кнопку **(Continue)**. Осуществится возврат в окно ввода данных и параметров процедуры (рис. 1.14), в котором следует нажать кнопку **(Ok)**. (Пассивность этой кнопки свидетельствует, что в процедуре либо не заданы данные для анализа, либо не до конца определены необходимые для работы процедуры параметры.)

Результаты. В окне навигатора вывода результатов (рис. 1.16) появится таблица результатов вычислений. Она, кроме заказанных описательных статистик, указывает, что анализируемая переменная включает 200 значений (**N Valid**) и число пропущенных наблюдений (**Missing**) равно 0.

Комментарий. В окно ввода данных **Variable(s)** процедуры **Frequencies** можно ввести сразу несколько переменных. Заданные описательные статистики будут рассчитаны отдельно для каждой из них.

Пример 1.2к. Сгруппировать данные примера 1.1к в диапазоне от 13 мм до 14.8 мм с шагом группировки 0.15 мм и вычислить частоты попадания в полученные интервалы группировки.

Авторам не известно простое прямое решение этой задачи в пакете SPSS. Один из косвенных путей — использование **frequency table** (таблицы частот), которая является табличным аналогом точечной диаграммы, т.е. для каждого наблюдения выборки указывает, сколько раз оно в ней встречается. Для получения **frequency table** необходимо вызвать процедуру **Frequencies** (см. пример 1.1к). В окне ввода данных и настройки параметров процедуры указать в качестве анализируемой переменной

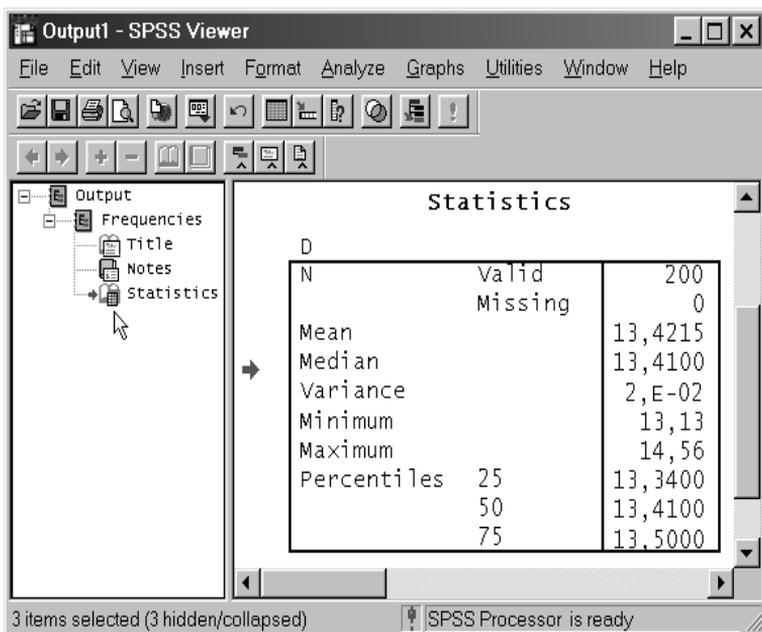


Рис. 1.16. Пакет SPSS. Окно навигатора вывода результатов процедуры «Frequencies» с таблицей результатов

d и задать выдачу **frequency table**, щелкнув мышкой в окне **Display frequency table** (см. рис. 1.14). В результате работы процедуры в окне навигатора вывода результатов пакета появится заказанная таблица, верхняя часть которой представлена на рис. 1.17.

В таблице все несовпадающие значения выборки упорядочены по возрастанию в первой колонке, и указано: число совпадающих значений в колонке **frequency**, процент этих значений в выборке в колонке **Percent** (процент в этой колонке вычисляется от всех наблюдений, включая пропущенные), процент значений от только имеющихся наблюдений в колонке **Valid Percent** и накопленный процент в колонке **Cumulative Percent**. Так как в нашей выборке нет пропущенных наблюдений, то значения колонок **Percent** и **Valid Percent** совпадают. Из этой таблицы видно, что минимальное значение 13.13 встретилось в выборке лишь один раз, что составляет 0.5% от объема выборки $n = 200$, а значение 13.20 присутствует в выборке 3 раза, что составляет 1.5%.

Из таблицы на рис. 1.17 путем простого, но утомительного расчета получаем частоты попадания наблюдений в каждый из требуемых интервалов группировки. Так, в интервал от 13 до 13.15 мм попало всего 2 значения, или 1% от всех наблюдений. В интервал от 13.15 до 13.30 мм

	Frequency	Percent	Valid Percent	Cumulative Percent
valid 13,13	1	,5	,5	,5
13,14	1	,5	,5	1,0
13,18	1	,5	,5	1,5
13,20	3	1,5	1,5	3,0
13,23	2	1,0	1,0	4,0
13,24	3	1,5	1,5	5,5
13,25	1	,5	,5	6,0
13,26	4	2,0	2,0	8,0
13,27	1	,5	,5	8,5
13,28	5	2,5	2,5	11,0
13,29	6	3,0	3,0	14,0
13,30	2	1,0	1,0	15,0
13,31	5	2,5	2,5	17,5
13,32	7	3,5	3,5	21,0
13,33	6	3,0	3,0	24,0

Рис. 1.17. Пакет SPSS. Таблица частот процедуры «Frequencies»

попало 14% наблюдений. Для получения этого значения достаточно вычесть из накопленного процента для значения 13.30 (15%) накопленный процент предыдущих интервалов (в данном случае 1%).

Комментарий. Изучение таблицы табуляции частот показывает, что в выборке находится одно сильно выделяющееся наблюдение. Оно могло заметно повлиять на вычисленные значения некоторых описательных статистик. Влияние этого наблюдения на различные выборочные статистики будет рассмотрено в гл. 5 и 10. В примере 1.3к мы проведем приведенные выше расчеты без учета сильно выделяющегося наблюдения.

Есть и другие не прямые пути решения этой задачи в SPSS. Так, можно использовать процедуру **Record** (перекодировки) из меню **Transform** панели управления редактора данных. Но и этой процедуре приходится вручную задавать границы всех интервалов группировки. Не будем более подробно останавливаться на этом способе.

Комментарий. Этот пример показывает, что даже в очень хорошем статистическом пакете может не быть удобных инструментов для решения очень простой задачи. Следует помнить, что часть возможностей пакета SPSS не доступна непосредственно из меню пакета, а требует обращения к командному языку пакета со всеми вытекающими отсюда преимуществами и неудобствами.

Пример 1.3к. Для выборки диаметров головок заклепок построить гистограмму частот с шагом группировки 0.075 мм на интервале от 13 до 13.75 мм (т.е. без учета сильно выделяющегося наблюдения).

Подготовка данных выполняется так же, как в примере 1.1к.

Выбор процедуры. В меню пункта **Graphs** (графики) панели управления редактора данных следует выбрать процедуру **Histogram**. При этом появится окно ввода данных и параметров процедуры (рис. 1.18).

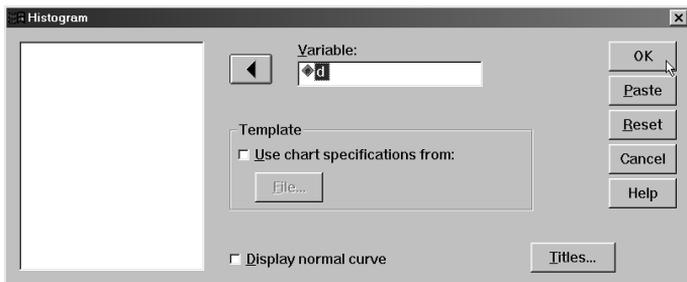


Рис. 1.18. Пакет SPSS. Окно ввода данных и параметров процедуры «Histogram»

Заполнение полей ввода данных. Следует выделить переменную **d** в левой части окна и перенести ее в поле **Variable**. Затем нажмите кнопку **OK**. Окно ввода параметров процедуры также позволяет предварительно оформить заголовки и подзаголовки графика (кнопка **Titles**) и использовать подготовленные ранее различные элементы оформления графика (блок **Template** (шаблоны)). Кроме того, на график гистограммы может быть наложена подобранная кривая плотности нормального распределения (опция **Display normal curve**).

Пакет SPSS в автоматическом режиме выбирает диапазон построения гистограммы и число интервалов группировки и помещает построенный график в окно навигатора вывода. Чтобы задать указанные выше параметры гистограммы вручную, необходимо дважды щелкнуть мышкой на полученном графике. При этом произойдет переход в окно редактора графиков пакета, как это показано на рис. 1.19а.

Для задания параметров группировки данных вручную в меню **Chart** (диаграммы) редактора графиков следует выбрать процедуру **Axis** (оси). В появившемся окне **Axis Selection** указать режим **Interval**. На экране появится окно **Interval Axis** (рис. 1.20). В блоке **Interval** выбрать режим **Custom** (заказ пользователя) и нажать кнопку **Define**.

В появившемся окне **Interval Axis: Define Custom Intervals** (рис. 1.21) задать требуемые границы диапазона группировки данных и число интервалов, как это показано на рис. 1.21. Нажать кнопку **Continue**, а затем кнопку **OK** в окне **Interval Axis**.

Результаты работы процедуры в окне редактора графиков приведены на рис. 1.19b. Одновременно происходит коррекция графика и в окне навигатора вывода результатов.

Комментарий. Гистограммы строят процедуры **Frequencies** и **Explore** из меню **Descriptive Statistics** и ряд других. При этом все они используют автоматический режим выбора параметров группировки данных. Все эти процедуры помещают построенный график гистограммы в окно навигатора вывода как самостоятель-

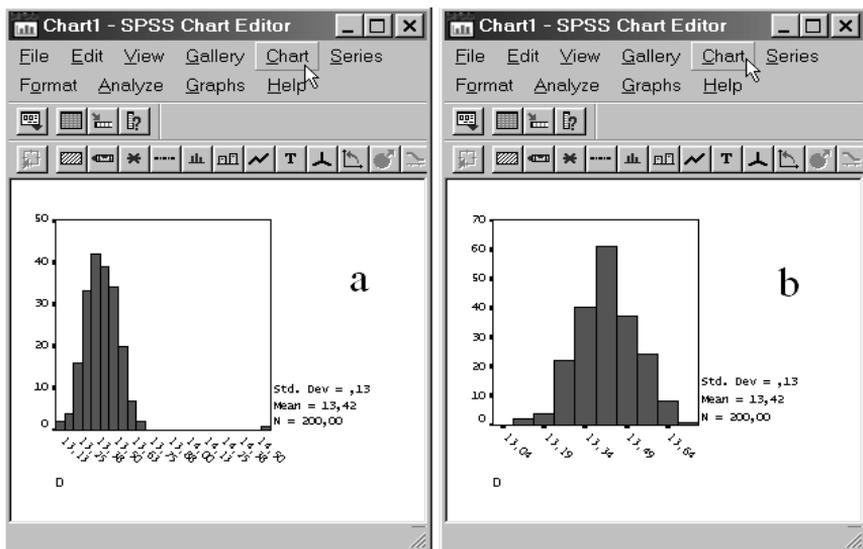


Рис. 1.19. Пакет SPSS. Гистограмма диаметров головок заклепок в окне графического редактора. а — построенная в автоматическом режиме, б — после задания параметров группировки вручную

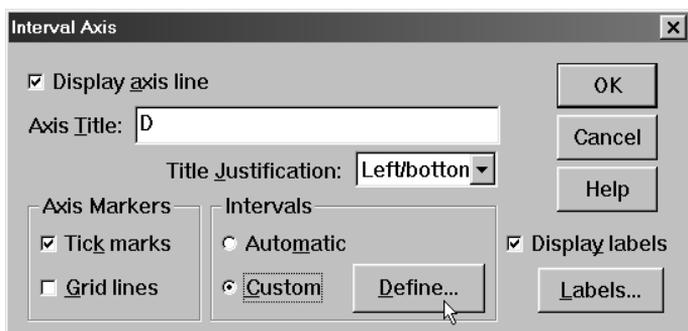


Рис. 1.20. Пакет SPSS. Окно настройки параметров осей графика

ный объект. Для задания требуемых параметров группировки гистограммы необходимо каждый раз проделать процедуру, описанную выше.

Другие возможности. На наш взгляд, наиболее полно различные описательные характеристики выборок представляет в пакете SPSS процедура **Explore**. Она очень хорошо позволяет судить о наличии в выборке нехарактерных значений, показывая несколько минимальных и максимальных элементов выборки и выделяя их на графике «ящик с усами». Кроме того эта процедура позволяет судить об устойчивости оценки среднего значения, вычисляя усеченные оценки и различные

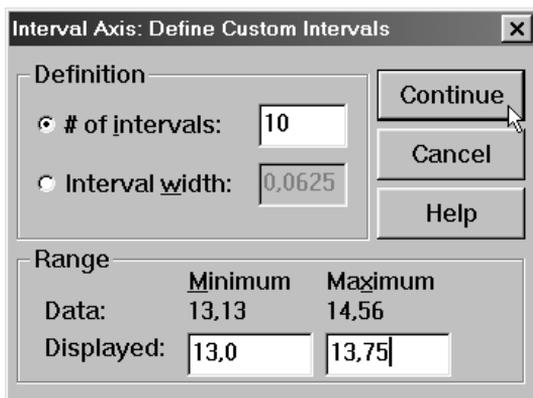


Рис. 1.21. Пакет SPSS. Окно настройки диапазона вывода и числа интервалов группировки гистограммы

робастные оценки среднего, а также строя доверительные интервалы для среднего значения. Мы частично расскажем о работе этой процедуры в гл. 10, обсуждая критерии согласия.

Дополнительная литература

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика. Основы эконометрики. Т. 1. Теория вероятностей и прикладная статистика. — М.: Юнити-Дана, 2001. — 656 с.
2. Борodin А.Н. Элементарный курс теории вероятностей и математической статистики. — СПб.: Лань, 2005. — 256 с.
3. Гнеденко Б.В., Хинчин А.Я. Элементарное введение в теорию вероятностей. 10-е изд., испр. — М.: Едиториал УРСС, 2003. — 208 с.
4. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. — М.: Мир. Т. 1, 1980. — 610 с., Т. 2, 1981. — 520 с.
5. Королев В.Ю. Теория вероятностей и математическая статистика: учеб. — М.: ТК Велби, Проспект, 2006. — 165 с.
6. Румшицкий Л.З. Элементарная теория вероятностей. 5-е изд., перераб. М.: Наука, 1976. — 240 с.
7. Томас Р. Количественные методы анализа хозяйственной деятельности: пер. с англ. — М.: Изд-во «Дело и Сервис», 1999. — 432 с.
8. Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. — М.: Мир, 1981. — 693 с.
9. Феллер В. Введение в теорию вероятностей и ее приложения: в 2 т. Т. 1; пер. с англ. — М.: Мир, 1984. — 528 с.

Важные законы распределения вероятностей

Подчиняются ли каким-то законам явления, носящие случайный характер? Да, но эти законы отличаются от привычных нам физических законов. Значения случайных величин невозможно предугадать даже при полностью известных условиях эксперимента, в котором они измеряются. Мы можем лишь указать *вероятности* того, что случайная величина принимает то или иное значение или попадает в то или иное множество. Зато зная *распределения вероятностей* интересующих нас случайных величин, мы можем делать выводы о событиях, в которых участвуют эти случайные величины. Правда, эти выводы будут также носить вероятностный характер. В последующих главах этой книги мы расскажем о том, как при знании распределений вероятностей или при некоторых предположениях относительно этих распределений делаются статистические выводы: как проверяются гипотезы, оцениваются параметры, определяются допустимые отклонения или вероятности ошибок этих оценок и т.д.

Но среди всех вероятностных распределений есть такие, которые используются на практике особенно часто. Эти распределения детально изучены и свойства их хорошо известны. Многие из этих распределений лежат в основе целых областей знания — таких, как теория массового обслуживания, теория надежности, контроль качества, теория измерений, теория игр и т.п. В этой главе мы расскажем о некоторых из таких распределений, покажем типичные ситуации, в которых они необходимы, дадим описания наиболее распространенных таблиц распределений и правил их использования. Материал главы имеет справочный характер и постоянно используется в дальнейшем тексте. При первом знакомстве его достаточно лишь бегло просмотреть, возвращаясь к нему в дальнейшем по необходимости.

Большинство применяемых на практике распределений являются дискретными или непрерывными. Среди дискретных распределений будут рассмотрены биномиальное и пуассоновское, среди непрерывных — показательное, нормальное и связанные с ним распределения: Стьюдента, хи-квадрат и F -распределение Фишера. Последние особенно часто используются при построении доверительных интервалов и проверке гипотез.

Более подробное изложение свойств этих и многих других распределений можно найти в [19], [65], [77], [87] и [111].

2.1. Биномиальное распределение

Область применения. Биномиальное распределение — это одно из самых распространенных дискретных распределений, оно служит вероятностной моделью для многих явлений. Оно возникает в тех случаях, когда нас интересует, сколько раз происходит некоторое событие в серии из определенного числа независимых наблюдений (опытов), выполняемых в одинаковых условиях. Поясним сказанное на примере.

Рассмотрим какое-либо массовое производство. Даже во время его нормальной работы иногда изготавливаются изделия, не соответствующие стандарту, т.е. дефектные. Обозначим долю дефектных изделий через p , $0 < p < 1$. Какое именно произведенное изделие окажется негодным, сказать заранее (до его изготовления) невозможно. Для описания подобной ситуации обычно используется следующая математическая модель:

- а) каждое изделие с вероятностью p может оказаться дефектным (с вероятностью $q = 1 - p$ оно соответствует стандарту); эта вероятность для всех изделий одинакова;
- б) появление как дефектных, так и стандартных изделий происходит независимо друг от друга. Это значит, что в нормальном процессе производства появление бракованного изделия не влияет на возможность появления брака в дальнейшем. Нарушение этого условия означает сбой нормального технологического режима.

Последовательность независимых испытаний, в которых результатом каждого из испытаний может быть один из двух исходов (например, успех и неудача), и вероятность «успеха» (или «неудачи») в каждом из испытаний одна и та же, называется *схемой испытаний Бернулли*. Поэтому мы можем перефразировать вышесказанное так: в нормальных условиях технологический процесс производства математически представляется схемой испытаний Бернулли.

Для чего же на производстве требуется подсчитывать число дефектных изделий? Как правило, это делается для контроля технологического процесса. При массовом производстве сплошная проверка качества изготовленных изделий обычно неоправданна. Поэтому для контроля качества из произведенной продукции наудачу отбирают определенное количество изделий (в дальнейшем — n) и проверяют их, регистрируют

найденное число бракованных изделий (в дальнейшем — X) и в зависимости от значения X принимают то или иное решение о состоянии производственного процесса. Теоретически X может принимать любые целые значения от 0 до n включительно, но, конечно, вероятности этих значений различны. Для того чтобы делаемые по значению X выводы были обоснованными, требуется знать распределение случайной величины X . Если выполняются приведенные выше условия схемы испытаний Бернулли, то распределение X является *биномиальным распределением*, и вероятности значений X можно получить очень просто.

Пронумеруем в произвольном порядке n проверяемых изделий (например, в порядке их поступления на контроль). Будем обозначать исход испытания каждого изделия нулем или единицей (ноль — нормальное изделие, единица — дефектное) и будем записывать итоги проверки партии из n изделий в виде последовательности из n нулей и единиц. Событие ($X = k$), или, другими словами, «среди n испытаний изделий оказалось k бракованных, а остальные $(n - k)$ — годные» — это совокупность всех последовательностей, содержащих в любом порядке k единиц и $(n - k)$ нулей. Вероятность того, что в результате проверки будет получена любая из таких последовательностей, равна $p^k(1-p)^{n-k}$, а число таких последовательностей — $C_n^k = \frac{n!}{k!(n-k)!}$. Поэтому, согласно свойствам вероятностей, описанным в п. 1.2, вероятность события ($X = k$) равна:

$$P(X = k) = C_n^k p^k (1-p)^{n-k} = \left(\frac{n!}{k!(n-k)!} \right) p^k q^{n-k}.$$

Определение. Случайная величина X имеет биномиальное распределение с параметрами n и p , если она принимает значения $0, 1, \dots, n$ с вероятностями:

$$P(X = k) = C_n^k p^k (1-p)^{n-k} \quad k = 0, 1, \dots, n.$$

Параметр p обычно называют вероятностью «успеха» в испытании Бернулли. В приведенном выше примере «успех» соответствует обнаружению бракованной детали. Распределение называется биномиальным, потому что вероятности $P(X = k)$ являются слагаемыми бинома Ньютона:

$$1^n = [p + (1-p)]^n = \sum_{k=0}^n C_n^k p^k (1-p)^{n-k} = \sum_{k=0}^n P(X = k).$$

Чтобы подчеркнуть зависимость $P(X = k)$ от p и n , вероятность $P(X = k)$ обычно записывают в виде:

$$P(X = k \mid n, p).$$

Свойства. Математическое ожидание и дисперсия случайной величины, имеющей биномиальное распределение, равны:

$$MX = np, \quad DX = np(1 - p).$$

Эти выражения легко получить с помощью следующего полезного приема. Введем для каждого отдельного испытания Бернулли случайную величину ξ , которая может принимать только два значения: 1, если испытание закончилось успехом, и 0, если неудачей. Если дать номера 1, 2, ... отдельным испытаниям, то те же номера надо присвоить и соответствующим им случайным величинам $\xi : \xi_1, \xi_2, \dots$. Тогда X можно представить в виде: $X = \xi_1 + \xi_2 + \dots + \xi_n$, причем случайные слагаемые в данной формуле статистически независимы и одинаково распределены. Для любого k от 1 до n выполняется $M\xi_k = p$, $D\xi_k = p(1 - p)$, поэтому, согласно свойствам математического ожидания и дисперсии из п. 1.5: $MX = nM\xi$, $DX = nD\xi$, что и приводит к указанным выше выражениям.

На рис. 2.1 показаны вероятности $P(X = k)$ при $n = 10$ для различных значений p ($p = 0.1, 0.2, 0.4$ и 0.5).

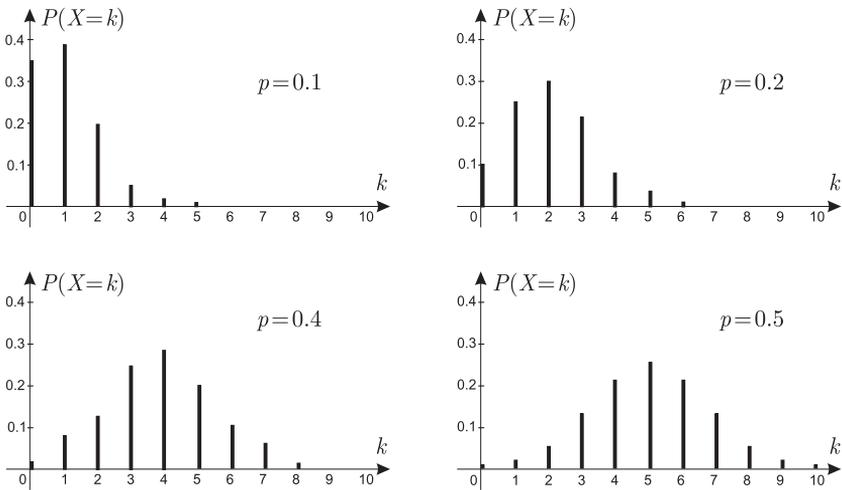


Рис. 2.1. Вид биномиального распределения для различных значений p при $n = 10$

Связь с другими распределениями. Биномиальное распределение тесно связано со многими другими распределениями. Ниже мы укажем наиболее часто используемые из этих связей. Описание других можно найти в [19], [111].

1. Биномиальное распределение с параметрами n и p может быть аппроксимировано нормальным распределением со средним np и стандартным отклонением $(np(1 - p))^{1/2}$, если только выполняются условия $np(1 - p) > 5$ и $0.1 \leq p \leq 0.9$. При условии $np(1 - p) > 25$ эту аппроксимацию можно применять независимо от значения p .

2. Биномиальное распределение с параметрами n и p может быть аппроксимировано распределением Пуассона со средним np при условии, что $p < 0.1$ и n достаточно велико.

Таблицы. Для биномиального распределения, как и для других распределений вероятностей, есть два типа таблиц.

В таблицах первого типа приводятся вероятности $P(X = k)$ при различных значениях p и n . Например, в [19] приведены таблицы $P(X = k | n, p)$ (с пятью десятичными знаками) для n от 5 до 30, с шагом по n , равным 5 (краткое обозначение: $n = 5(5)30$), и $p = 0.01; 0.02(0.02); 0.10(0.10); 0.50$. Последнее выражение для p означает, что в таблицах есть значения для $p = 0.01$, для $p = 0.02$, далее p изменяется с шагом 0.02 до 0.10 и со значения $p = 0.1$ оно изменяется с шагом 0.1 до 0.5.

В таблицах второго типа даны значения накопленных вероятностей биномиального распределения, т.е. значения

$$P(X \leq k | n, p) = \sum_{m=0}^k P(X = m | n, p).$$

Например, в [77], $P(X \leq k | n, p)$ даны для $n = 1(1)25$, $p = 0.005(0.005); 0.02(0.01); 0.10(0.05); 0.30(0.10); 0.50$, для $k = 0(1)n$.

В описаниях таблиц обычно можно найти указания, как поступать, если интересующие нас значения n и/или p в данных таблицах отсутствуют (см., например, [19]).

Замечание. Значения вероятностей $P(X = k)$ биномиального распределения с параметром $p > 0.5$ легко получить, зная соответствующие вероятности при $p < 0.5$. Действительно, если вероятность «успеха» $p > 0.5$, то вероятность «неудачи» $q = 1 - p < 0.5$. Поменяв названия «успех» и «неудача» одно на другое, мы сведем случай $p > 0.5$ к $p < 0.5$. Другими словами:

$$P(X = k | n, p) = P(X = n - k | n, 1 - p).$$

Это свойство учитывается при составлении статистических таблиц биномиального распределения.

2.2. Распределение Пуассона

Область применения. Распределение Пуассона играет важную роль в ряде вопросов физики, теории связи, теории надежности, теории массового обслуживания и т.д. — словом, всюду, где в течение определенного времени может происходить случайное число каких-то событий (радиоактивных распадов, телефонных вызовов, отказов оборудования, несчастных случаев и т.п.).

Рассмотрим наиболее типичную ситуацию, в которой возникает распределение Пуассона. Пусть некоторые события могут происходить в случайные моменты времени, а нас интересует число появлений таких событий в промежутке времени от 0 до T . (Например, это могут быть помехи в канале связи, появления метеоритов, дорожные происшествия и т.п.) Сделаем следующие предположения.

1. Пусть вероятность появления события за малый интервал времени длины Δ примерно пропорциональна Δ , т.е. равна $a\Delta + o(\Delta)$, здесь $a > 0$ — параметр задачи, отражающий среднюю частоту событий.
2. Если в интервале времени длины Δ уже произошло одно событие, то условная вероятность появления в этом же интервале другого события стремится к 0 при $\Delta \rightarrow 0$.
3. Количества событий, происшедших на непересекающихся интервалах времени, независимы как случайные величины.

В этих условиях можно показать, что случайное число событий, происшедших за время от 0 до T , распределено по закону Пуассона с параметром $\lambda = aT$.

Определение. Случайная величина ξ , которая принимает только целые, неотрицательные значения $0, 1, 2, \dots$, имеет закон распределения Пуассона с параметром $\lambda > 0$, если

$$P(\xi = k | \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{для } k = 0, 1, 2, \dots$$

Свойства. Математическое ожидание и дисперсия случайной величины, имеющей распределение Пуассона с параметром λ , равны:

$$M\xi = \lambda, \quad D\xi = \lambda.$$

Эти выражения несложно получить прямыми вычислениями. Имеем:

$$\begin{aligned} M\xi &= \sum_{k=0}^{\infty} kP(\xi = k | \lambda) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{(k-1)}}{(k-1)!} e^{-\lambda} \\ &= \lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} = \lambda. \end{aligned}$$

Здесь была осуществлена замена $n = k - 1$ и использован тот факт, что $\sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^\lambda$. Аналогично можно вычислить дисперсию случайной величины ξ .

На рис. 2.2 показаны значения вероятностей $P(\xi = k | \lambda)$ для различных значений k и λ .

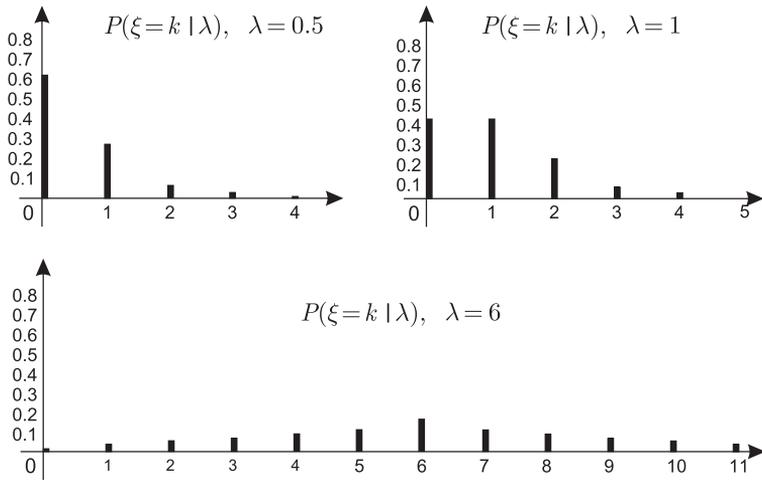


Рис. 2.2. Вид распределения Пуассона для различных значений k и λ

Связь с другими распределениями. 1. Выше уже указывалась связь между распределением Пуассона и биномиальным. Остановимся на этом вопросе более подробно.

При большом n и малом p действует приближенное соотношение:

$$C_n^k p^k (1-p)^{n-k} \simeq \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

где $\lambda = np$. Этот факт можно сформулировать в виде предельного утверждения: при всяком k , ($k = 0, 1, 2, \dots$)

$$\lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} C_n^k p^k (1-p)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{если существует} \quad \lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} np = \lambda > 0.$$

2. При $\lambda > 9$ распределение Пуассона может быть аппроксимировано нормальным распределением со средним λ и дисперсией λ .

3. Сумма n независимых случайных величин, имеющих пуассоновские распределения с параметрами $\lambda_1, \lambda_2, \dots, \lambda_n$ соответственно, имеет также распределение Пуассона с параметром

$$\lambda = \lambda_1 + \lambda_2 + \dots + \lambda_n.$$

Таблицы. Таблицы распределения Пуассона при различных значениях даны, например, в [19], [65], [77], а также в других сборниках таблиц и монографиях.

Дадим описание таблиц, приведенных в [19] для $P(\xi = k | \lambda)$. При этом значение λ изменяется от 0.1 (0.1) 15.0, а значение k изменяется с единичным шагом в таких пределах, где $P(\xi = k | \lambda) > 5 \cdot 10^{-7}$. Там

же указано, как вычислять значение $P(\xi = k | \lambda)$ с помощью таблиц функции распределения χ^2 , о которой речь пойдет ниже.

Более подробные таблицы распределения Пуассона даны в [65], где λ изменяется до 205. Отметим, что при больших значениях λ для вычисления $P(\xi = k | \lambda)$ можно использовать приближенную формулу

$$P(\xi = k | \lambda) \sim \frac{1}{\sqrt{\lambda}} \varphi\left(\frac{k - \lambda}{\sqrt{\lambda}}\right),$$

где φ — плотность нормального распределения с параметрами 0 и 1.

Наряду с таблицами для $P(\xi = k | \lambda)$ составлены и таблицы накопленной вероятности распределения Пуассона, т.е. таблицы для

$$P(\xi \leq k | \lambda) = \sum_{m=0}^k P(\xi = m | \lambda).$$

В [77] приведены таблицы $P(\xi \leq k | \lambda)$ для $\lambda = 0.01$ (0.01); 1 (0.05); 5 (0.1); 10 (0.5); 20 (1); 30 (5); 50 с точностью до $0.5 \cdot 10^{-4}$.

2.3. Показательное распределение

Область применения. Укажем две области применения статистических методов, в которых показательное распределение играет базовую роль.

К первой из них относятся задачи, связанные с данными типа «времени жизни». Понимать этот термин следует достаточно широко. В медико-биологических исследованиях под ним может подразумеваться продолжительность жизни больных при клинических исследованиях, в технике — продолжительности безотказной работы устройств, в психологии — время, затраченное испытуемым на выполнение тестовых задач, и т.д. Подробное изложение обработки подобных данных дано в [55].

Второй областью активного использования показательного распределения являются задачи массового обслуживания. Здесь речь может идти об интервалах времени между вызовами «скорой помощи», телефонными звонками или обращениями клиентов и т.д. В условиях модели п. 2.2, в которой речь шла о появлении в случайные моменты неких событий и которую мы использовали для иллюстрации распределения Пуассона, длина интервала времени между появлениями последовательных событий имеет показательное распределение.

Определение. Положительная случайная величина X имеет показательное распределение с параметром $\theta > 0$, если ее плотность задана формулой

$$p(x, \theta) = \theta e^{-\theta x} \quad (x \geq 0).$$

Показательное распределение часто называют еще экспоненциальным. Параметр θ в ряде прикладных областей именуют «отношением риска». Иногда вместо параметра θ используют параметр $b = 1/\theta$, тогда функция плотности записывается в виде:

$$p(x, b) = \frac{1}{b} e^{-x/b} \quad (x \geq 0).$$

Свойства. Математическое ожидание и дисперсия случайной величины X , распределенной по показательному закону с параметром θ , равны

$$MX = 1/\theta, \quad DX = 1/\theta^2.$$

Первое из этих соотношений придает параметру θ ясный вероятностный смысл: $1/\theta$ — это среднее время службы изделия, среднее время между вызовами и т.д.

На рис. 2.3. приведен графический вид плотности показательного распределения с параметром θ .

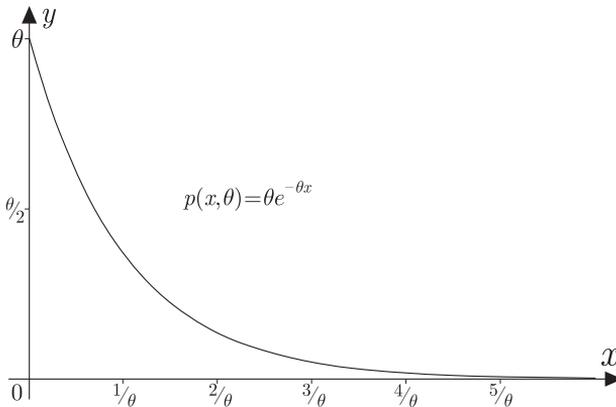


Рис. 2.3. Плотность показательного распределения с параметром θ

Функция показательного распределения, т.е. $P(X < x)$, равна

$$F(x, \theta) = \begin{cases} 1 - e^{-\theta x}, & \text{для } x \geq 0; \\ 0, & \text{для } x < 0. \end{cases}$$

Показательное распределение среди всех других выделяется, как иногда говорят, отсутствием «памяти», т.е. отсутствием последействия.

Это подразумевает следующее: для показательно распределенной случайной величины X (и только для такой)

$$P(X \geq s + t \mid X \geq t) = P(X \geq s)$$

для любых $s, t \geq 0$. Поясним смысл этой формулы на примере. Пусть X — время службы некоего изделия, и оно подчиняется экспоненциальному распределению. Тогда для изделия, прослужившего время t , вероятность прослужить дополнительное время s совпадает с вероятностью прослужить то же время s для нового (только начавшего работу) изделия. Как видим, это соотношение как бы исключает износ и старение. Поэтому в статистических моделях срока службы, если мы хотим учесть старение, приходится привлекать различного рода обобщения показательного распределения.

Связь с другими распределениями. Показательное распределение является частным случаем гамма-распределения, распределения Вейбулла и некоторых других. Подробную информацию на эту тему можно получить в [111].

Таблицы. Функция показательного распределения достаточно проста, поэтому специальные таблицы для этого распределения не нужны. Значения функции показательного распределения можно вычислить с помощью калькулятора.

2.4. Нормальное распределение

Область применения. Нормальное распределение относится к числу наиболее распространенных и важных, оно часто используется для приближенного описания многих случайных явлений, например, для случайного отступления фактического размера изделия от номинального, рассеяния снарядов при артиллерийской стрельбе и во многих других ситуациях, в которых на интересующий нас результат воздействует большое количество независимых случайных факторов, среди которых нет сильно выделяющихся.

Замечание. Использованию нормального распределения для приближенного описания распределений случайных величин не препятствует то обстоятельство, что эти величины обычно могут принимать значения только из какого-то ограниченного интервала (скажем, размер изделия должен быть больше нуля и меньше километра), а нормальное распределение не сосредоточено целиком ни на каком интервале. Дело в том, что вероятность больших отклонений нормальной случайной величины от центра распределения настолько мала, что ее практически можно считать равной нулю.

Определение. 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 задается формулой

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

Смысл параметров нормального распределения наглядно показан на рис. 2.4.

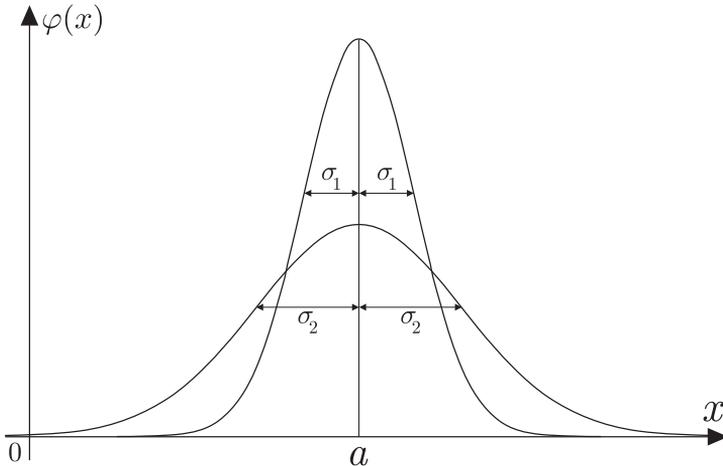


Рис. 2.4. Плотность нормального распределения со средним a и различными значениями дисперсии σ^2

Отметим, что $\varphi(x)$ стремится к нулю при $x \rightarrow -\infty$ и $x \rightarrow +\infty$. График функции $\varphi(x)$ симметричен относительно точки a . При этом в точке a функция $\varphi(x)$ достигает своего максимума, который равен $1/(\sqrt{2\pi}\sigma)$.

Параметр a характеризует положение графика функции на числовой оси (параметр положения). Параметр σ ($\sigma > 0$) характеризует степень сжатия или растяжения графика плотности (параметр масштаба). Как видим, вся совокупность нормальных распределений представляет собой двухпараметрическое семейство.

Свойства. Математическое ожидание и дисперсия случайной величины ξ , распределенной как $N(a, \sigma^2)$, равны

$$M\xi = a, \quad D\xi = \sigma^2.$$

Медиана нормального распределения равна a , так как плотность распределения симметрична относительно точки $x = a$.

Особую роль играет нормальное распределение с параметрами $a = 0$ и $\sigma = 1$, т.е. распределение $N(0, 1)$, которое часто называют *стандартным* нормальным распределением. Плотность стандартного нормального распределения есть

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

Функция распределения стандартного нормального распределения равна

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy.$$

Функцию $\Phi(\cdot)$ часто называют функцией Лапласа. Отметим, что $\Phi(x) = 1 - \Phi(-x)$, поэтому достаточно знать значения функции $\Phi(x)$ для $x \geq 0$. Это свойство функции $\Phi(x)$ используется при составлении таблиц.

Функцию произвольного нормального распределения $N(a, \sigma^2)$ можно легко выразить через $\Phi(\cdot)$. Для этого следует заметить, что если ξ распределена по закону $N(a, \sigma^2)$, то ее линейная функция $X = (\xi - a)/\sigma$ подчиняется стандартному нормальному распределению. Поэтому

$$P(\xi < x) = P\left(X < \frac{x - a}{\sigma}\right) = \Phi\left(\frac{x - a}{\sigma}\right).$$

Эта формула позволяет вычислять вероятности событий, связанных с произвольными нормальными случайными величинами, с помощью таблиц стандартного нормального распределения.

Аналогичным образом, легко показать, что если ξ распределена по нормальному закону, скажем, $N(a, \sigma^2)$, то случайная величина $k\xi + b$ (линейная функция ξ) имеет нормальное распределение $N(a + b, k^2\sigma^2)$.

Напомним, что площадь фигуры, ограниченная графиком функции плотности распределения, осью абсцисс и отрезками двух вертикальных прямых, $x = b$, $x = c$, есть вероятность попадания случайной величины в интервал (b, c) . В связи с этим полезно представить, как распределяются доли площадей между кривой $\varphi(x)$ и осью абсцисс (см. рис. 2.5). Более подробный анализ показывает, что случайная величина $N(0, 1)$ с вероятностью, примерно равной 0.94, попадает в интервал $(-2, 2)$, и с вероятностью, примерно равной 0.9973, — в интервал $(-3, 3)$. Отсюда для произвольной нормально распределенной случайной величины можно сформулировать правило, именуемое в литературе *правилом трех сигм*. А именно: нормальная случайная величина $N(a, \sigma^2)$ с вероятностью 0.9973 попадает в интервал $(a - 3\sigma, a + 3\sigma)$.

Таблицы. Для функции $\Phi(x)$ и ее производной, т.е. для плотности стандартного нормального распределения, существуют многочисленные

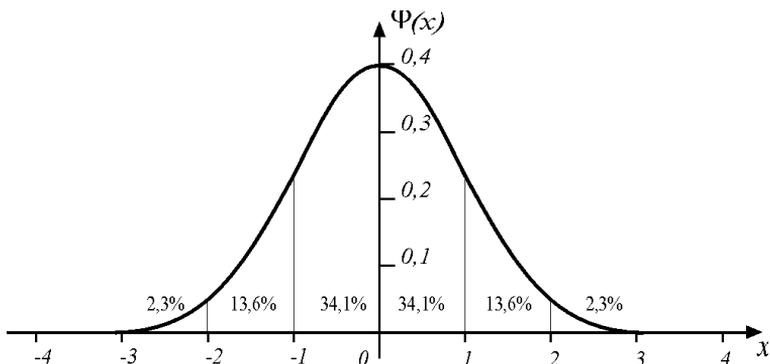


Рис. 2.5. Примерное распределение площадей под кривой функции плотности стандартного нормального распределения

таблицы разной степени подробности. Так, в [19] указаны значения $\Phi(x)$ с шестью значащими цифрами для $x = 0.000$ (0.001) 3.000 и с пятью значащими цифрами для $x = 3.00$ (0.01) 5.00 (в данном случае значащими называются все разряды десятичной дроби начиная с первого, отличного от девятки, например, если $\Phi(x) = 0.99976737$, то значащими цифрами считаются 76737).

Для статистических применений часто оказываются полезными таблицы, представляющие накопленную нормальную вероятность, отсчитываемую справа, т.е. таблицы, в которых в зависимости от x указаны значения $P(\xi \geq x) = 1 - \Phi(x)$. Например, в [115] дана таблица $P(\xi \geq x)$ для $x = 0.00$ (0.01) 3,5 с четырьмя значащими цифрами. Как будет показано в гл. 5, таблицы подобного вида более удобны в статистической практике, чем таблицы для $\Phi(x)$.

В большинстве сборников также приводятся таблицы квантилей стандартного нормального распределения. Они позволяют по заданному значению вероятности p , $0 < p < 1$ находить точку x , такую, что $P(\xi < x) = p$. Последнее бывает часто необходимо при проверке статистических гипотез.

2.5. Двумерное нормальное распределение

Область применения. Двумерное нормальное распределение используется при описании совместного распределения двух случайных переменных (двух признаков). В этой ситуации двумерное нормальное распределение является столь же важным, как одномерное нормальное распределение для описания одного случайного признака. Обсуждение

двумерного нормального распределения начнем с обсуждения многомерных распределений вообще.

Многомерные распределения. В гл. 1 мы установили, что для непрерывной одномерной случайной величины ξ ее функция плотности вероятности, скажем, $p(x)$ полностью задает распределение случайной величины: для любых чисел a, b ($a < b$)

$$P(a < \xi < b) = \int_a^b p(x) dx.$$

Аналогичным образом можно задать закон распределения случайной величины, принимающей значения не на числовой прямой, а на плоскости, в трехмерном пространстве, на сфере и т.д. Надо только иметь соответствующую функцию плотности $p(x)$. Тогда для любого множества X его вероятность $P(X)$ равна

$$P(X) = \int_X p(x) dx,$$

где интегрирование производится соответственно по области X в плоскости, трехмерном пространстве, сфере и т.д.

Двумерное нормальное распределение. В качестве примера определим двумерное нормальное распределение на плоскости. Пусть η_1 и η_2 — независимые случайные величины, имеющие стандартное нормальное распределение. Тогда двумерная случайная величина $\eta = (\eta_1, \eta_2)$ имеет *стандартное двумерное нормальное распределение*. Его плотность $p(x, y)$ равна:

$$p(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right).$$

Для одномерного случая все нормальные распределения могут быть получены как линейные преобразования стандартного нормального распределения: если $\xi \sim N(a, \sigma^2)$, то ξ можно представить в виде $\xi = a + \sigma\eta$, где случайная величина η имеет стандартное нормальное распределение. Аналогичным образом можно определить двумерные нормальные распределения — это те распределения, которые можно получить из стандартного двумерного распределения линейным преобразованием. По определению, случайная величина $\xi = (\xi_1, \xi_2)$ имеет двумерное нормальное распределение, если ее можно представить в виде

$$\begin{cases} \xi_1 = a_1 + b_1\eta_1 + c_1\eta_2 \\ \xi_2 = a_2 + b_2\eta_1 + c_2\eta_2 \end{cases},$$

где $a_1, b_1, c_1, a_2, b_2, c_2$ — некоторые вещественные числа. Заметим, что согласно свойствам нормальных случайных величин, компоненты двумерной нормальной случайной величины, т.е. ξ_1 и ξ_2 , являются нормальными (одномерными) случайными величинами. Разумеется, случайные величины ξ_1 и ξ_2 могут быть зависимыми. Ниже мы покажем, что ξ_1 и ξ_2 зависимы тогда и только тогда, когда их ковариация (или корреляция) не равна нулю.

Аналогичным образом можно определить и многомерные нормальные распределения.

Частные (маргинальные) плотности. Если $\xi = (\xi_1, \xi_2)$ — двумерная случайная величина, то ее компоненты ξ_1 и ξ_2 — тоже случайные величины. Можно показать, что если ξ имеет плотность $p(x, y)$, то ξ_1 и ξ_2 тоже непрерывные случайные величины, имеющие плотности $p_1(x)$ и $p_2(y)$ (называемые *частными плотностями*), и эти плотности выражаются формулами:

$$p_1(x) = \int_{-\infty}^{\infty} p(x, y) dy, \quad p_2(y) = \int_{-\infty}^{\infty} p(x, y) dx.$$

Характеристики многомерных распределений. Чаще всего в качестве характеристик многомерных распределений используются те или иные функции от компонент (координат) многомерных случайных величин, имеющих данное распределение. Например, для двумерной случайной величины $\xi = (\xi_1, \xi_2)$ мы можем рассматривать ее математическое ожидание $M\xi = (M\xi_1, M\xi_2)$ и вторые центральные моменты:

$$\sigma_{11} = D\xi_1, \quad \sigma_{22} = D\xi_2, \quad \sigma_{12} = \sigma_{21} = \text{cov}(\xi_1, \xi_2).$$

Если $\xi = (\xi_1, \xi_2)$ имеет плотность $p(x, y)$, то эти моменты, естественно, выражаются в виде интегралов от плотности. Например,

$$M\xi_1 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x p(x, y) dx dy.$$

Двумерная нормальная плотность. Укажем формулы для плотности двумерного нормального распределения. Пусть $\xi = (\xi_1, \xi_2)$ — двумерная нормальная случайная величина. Формула для плотности будет выглядеть проще, если мы от ξ перейдем к случайной величине $\eta = (\eta_1, \eta_2)$, где $\eta_1 = (\xi_1 - a_1)/\sqrt{\sigma_{11}}$, $\eta_2 = (\xi_2 - a_2)/\sqrt{\sigma_{11}}$, где $a_1 = M\xi_1$, $a_2 = M\xi_2$, а σ_{11} и σ_{22} были определены выше. Тогда η_1 и η_2 — случайные величины, имеющие стандартное нормальное распределение. Пусть их корреляция (она же ковариация) равна ρ . Легко видеть, что $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ — то же самое, что величина корреляции исходных случайных величин ξ_1 и ξ_2 . Тогда можно показать, что функция плотности $p(x_1, x_2)$ двумерной случайной величины $\eta = (\eta_1, \eta_2)$ равна:

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)}\right\}.$$

Для исходной двумерной случайной величины $\xi = (\xi_1, \xi_2)$ плотность вероятности в точке (x_1, x_2) равна

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}} \times \exp\left\{-\frac{1}{2(1-\rho^2)} \left[\frac{(x_1 - a_1)^2}{\sigma_{11}} - 2\rho \frac{(x_1 - a_1)(x_2 - a_2)}{\sqrt{\sigma_{11}\sigma_{22}}} + \frac{(x_2 - a_2)^2}{\sigma_{22}} \right]\right\}.$$

Практически это выражение используют редко.

2.6. Распределения, связанные с нормальным

Область применения. При операциях с нормальными случайными величинами, которые приходится проводить при анализе данных, возникает несколько новых видов распределений (и соответствующих им случайных величин). В первую очередь это распределение Стьюдента, χ^2 и F -распределения. Эти распределения играют очень важную роль в прикладном и теоретическом анализе. Так, при выяснении точности и достоверности статистических оценок используются процентные точки распределений Стьюдента и хи-квадрат. Распределение статистик многих критериев, использующихся для проверки различных предположений, хорошо приближается этими распределениями.

2.6.1. Распределение хи-квадрат

Определение. Пусть случайные величины $\xi_1, \xi_2, \dots, \xi_n$ — независимы, и каждая из них имеет стандартное нормальное распределение $N(0, 1)$. Говорят, что случайная величина χ_n^2 , определенная как:

$$\chi_n^2 = \xi_1^2 + \dots + \xi_n^2,$$

имеет распределение хи-квадрат с n степенями свободы. Для обозначения этого распределения также обычно используется выражение χ_n^2 .

Ясно, что χ_n^2 (для любого $n \geq 1$) с вероятностью 1 принимает положительные значения. Функция плотности χ_n^2 в точке $x (x > 0)$ равна

$$\frac{1}{2^{n/2}} \frac{1}{\Gamma(n/2)} x^{n/2-1} e^{-x/2}.$$

где $\Gamma(\cdot)$ есть гамма-функция. На практике эта плотность распределения непосредственно используется редко.

Заметим, что показательное распределение с параметром $\theta = 1/2$ из параграфа 2.3 — это распределение χ^2 с двумя степенями свободы.

На рис. 2.6 изображены функции плотности распределения хи-квадрат с различным числом степеней свободы.

Свойства. Нетрудно убедиться, что математическое ожидание и дисперсия случайной величины χ_n^2 равны:

$$M\chi_n^2 = n, \quad D\chi_n^2 = 2n.$$

Таблицы. Для случайной величины χ_n^2 составлены разнообразные таблицы (см. [19], [65], [77]). Чаще всего они содержат значения p -квантилей случайных величин χ_n^2 , $n = 1, 2, \dots, m$ (если вероятность

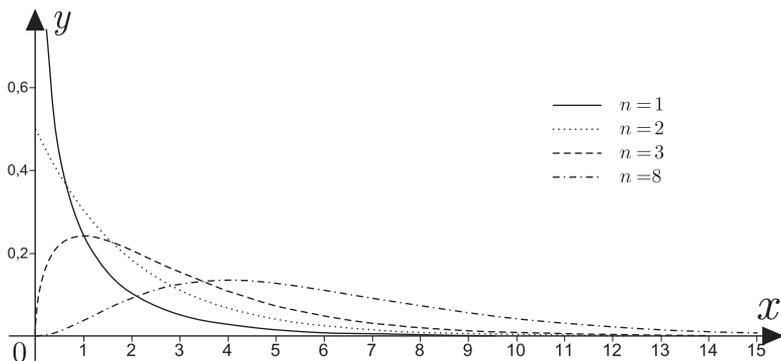


Рис. 2.6. Функции плотности распределения хи-квадрат с различным числом степеней свободы n

выражена в процентах, их называют процентными точками и, соответственно, говорят о таблицах процентных точек). Аргумент p , $0 < p < 1$, при этом пробегает тот или иной набор значений.

2.6.2. Распределение Стьюдента

Определение. Пусть случайные величины $\xi_0, \xi_1, \dots, \xi_n$ — независимы, и каждая из них имеет стандартное нормальное распределение $N(0, 1)$. Введем случайную величину

$$t_n = \frac{\xi_0}{\sqrt{\frac{1}{n} \sum_{i=1}^n \xi_i^2}}.$$

Ее распределение называют распределением Стьюдента. Саму случайную величину часто называют стьюдентовской дробью, стьюдентовым отношением и т.п. Число n , $n = 1, 2, \dots$ называют числом степеней свободы распределения Стьюдента.

Плотность распределения Стьюдента в точке x равна

$$\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}.$$

Из определения видно, что плотность симметрична относительно $x = 0$. Это обстоятельство используют при составлении таблиц.

На рис. 2.7 изображены функции плотности распределения Стьюдента с различным числом степеней свободы.

Свойства. Можно показать, что:

$$Mt_n = 0, \quad Dt_n = \frac{n}{n-2}.$$

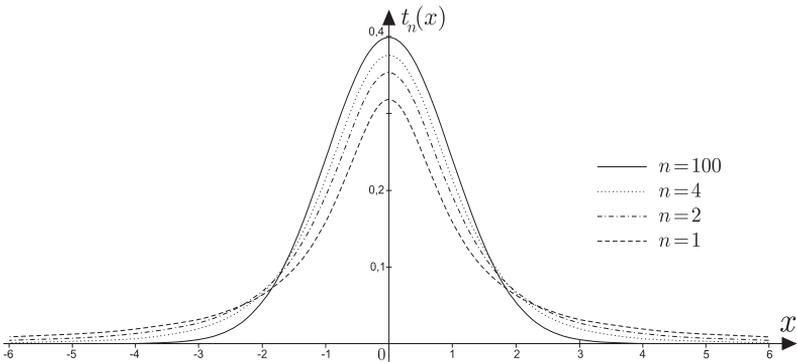


Рис. 2.7. Функции плотности распределения Стьюдента с различным числом степеней свободы n

Таблицы. В сборниках обычно приводятся таблицы процентных точек для последовательных $n = 1, 2, \dots$ вплоть до некоторого значения. При больших n обычно рекомендуют использовать таблицы стандартного нормального распределения, иногда с поправками.

2.6.3. F-распределение

Определение. Пусть $\eta_1, \dots, \eta_m; \xi_1, \dots, \xi_n$ (где m, n — натуральные числа) обозначают независимые случайные величины, каждая из которых распределена по стандартному нормальному закону $N(0, 1)$. Говорят, что случайная величина $F_{m,n}$, определенная как

$$F_{m,n} = \frac{\frac{1}{m} (\eta_1^2 + \dots + \eta_m^2)}{\frac{1}{n} (\xi_1^2 + \dots + \xi_n^2)},$$

имеет F -распределение с параметрами m и n . Натуральные числа m, n называют числами степеней свободы. F -распределение иногда называют еще распределением дисперсионного отношения (смысл этого названия станет ясен в гл. 6).

Плотность $F_{m,n}$ выражается довольно сложной формулой, которая редко непосредственно используется на практике, поэтому мы ее приводить не будем.

На рис. 2.8 изображены функции плотности F -распределения с различным числом степеней свободы.

Свойства. Математическое ожидание и дисперсия случайной величины $F_{m,n}$ равны:

$$MF_{m,n} = \frac{n}{n-2} \quad \text{для } n > 2, \quad DF_{m,n} = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad \text{для } n > 4.$$

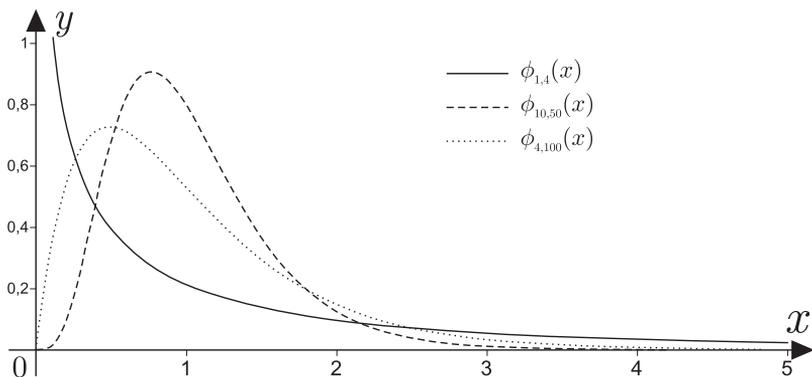


Рис. 2.8. Функции плотности F -распределения с различным числом степеней свободы

Таблицы. Семейство F -распределений зависит от двух натуральных параметров m и n , в связи с чем даже таблицы процентных точек занимают большой объем. Ради экономии места они часто публикуются в сжатом виде, поэтому при их практическом использовании приходится прибегать к дополнительным вычислениям и интерполяции.

2.7. Законы распределения вероятностей в пакете SPSS

Современные статистические пакеты, как правило, предоставляют обширную справочную информацию по различным семействам вероятностных распределений, заменяя различные статистические таблицы. Кроме того, они позволяют моделировать выборки случайных (или более строго псевдослучайных) величин с заданными распределениями вероятностей. Это позволяет использовать эти пакеты не только в курсах математической статистики, но и в курсах теории вероятностей.

Однако пакет SPSS отчасти неудобен для первого знакомства с вероятностными распределениями. В пакете отсутствуют удобные возможности для построения графиков плотности и функции распределения. Также нет отдельного меню, предоставляющего прямой доступ к различным распределениям. Все это ни в коей мере не препятствует эффективному статистическому анализу данных. В большинстве статистических процедур пакета для вычисляемых статистик указываются их минимальные уровни значимости (см. п. 3.4), а этого вполне достаточно для обоснованного принятия решений.

Вычислить значение функций распределения вероятностей или квантиля в пакете можно для широкого класса дискретных и непрерыв-

ных распределений, включая: биномиальное, Пуассона, геометрическое, гипергеометрическое, нормальное, t -распределение Стьюдента, хи-квадрат, экспоненциальное, F -распределение, гамма, бета, Коши, Лапласа, Парето, логистическое, Вейбулла и др. Для этого следует вызвать процедуру **Compute** (вычисления) из меню **Transform** (преобразования) панели управления редактора данных пакета. Эта процедура позволяет пользователю задать функциональное выражение, используя обширную библиотеку функций и знаки арифметических и логических операций. При этом в качестве аргументов функционального выражения могут выступать переменные из редактора пакета, а результат также является переменной, создаваемой процедурой в редакторе пакета. Разберем работу этой важной процедуры на примерах.

Пример 2.1к. Найдём p -квантили экспоненциального распределения со средним значением 4 для $p = 0.95; 0.975; 0.99$.

Подготовка данных. В редакторе данных пакета создадим переменную **prob** и введем в нее значения вероятностей $p = 0.95; 0.975; 0.99$, как это показано на рис. 2.9.

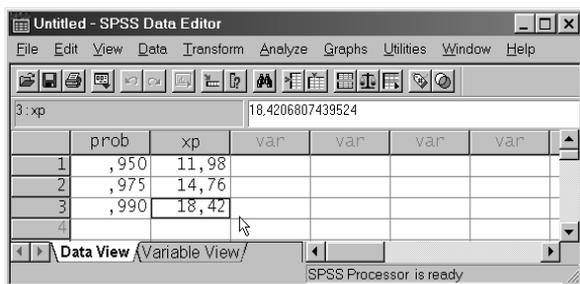


Рис. 2.9. Пакет SPSS. Окно редактора данных с исходными данными и результатами работы процедуры

Выбор процедуры. В пункте **Transform** панели управления редактора данных пакета вызвать процедуру **Compute**.

Комментарий. Вызов этой процедуры в SPSS возможен только при наличии в редакторе данных пакета хотя бы одного значения одной переменной.

Заполнение полей ввода данных. Окно ввода данных процедуры **Compute** представлено на рис. 2.10.

В поле **Target Variable** необходимо указать имя переменной, например **xp**, куда будет помещен результат вычислений. В списке функций выделить функцию **IDF.EXP(p, scale)** и щелчком мыши на стрелке переноса перенести эту функцию в окно **Numeric Expression**. При этом параметры процедуры p и $scale$ заменятся знаками «?».

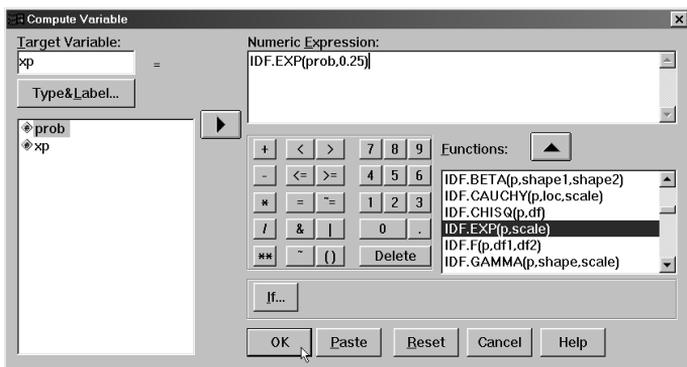


Рис. 2.10. Пакет SPSS. Окно ввода данных и задания параметров процедуры «Compute» для вычисления квантилей экспоненциального распределения

Комментарии. 1. Процедуры вычисления квантилей распределения имеют общий префикс **IDF** (сокращение от Inverse Density Function (обратная функция распределения)). За префиксом следует сокращенное имя распределения (**EXP** для экспоненциального распределения).

2. Назначение каждой процедуры и общий смысл параметров процедуры можно узнать, щелкнув правой кнопкой мыши на имени процедуры.

3. Процедуры вычисления значений функции распределения вероятностей имеют общий префикс **CDF** (сокращение от Cumulative Density Function (функция распределения)).

Вместо знаков «?» параметров процедуры следует указать требуемые значения. Первый параметр обозначает вероятность, для которой будет вычисляться квантиль. На его место следует ввести имя переменной из редактора данных, в котором хранятся требуемые в задаче вероятности — **prob** (см. рис. 2.10). (Все имена переменных из редактора данных показаны в левом нижнем поле окна ввода данных и параметров процедуры.) Вторым параметром **scale** (масштаб) процедуры **IDF.EXP** является отношение риска — θ . Его связь с математическим ожиданием указана в п. 2.3. А именно: $\theta = 1/MX$. Поэтому в качестве второго параметра следует ввести $0.25 = 1/4$, как это показано на рис. 2.10.

Комментарий. Числовые константы в качестве параметров процедур в SPSS должны использовать разделитель точку между целой и дробной частью, вне зависимости от того, какой разделитель целой и дробной части используется в редакторе пакета.

Закончив задание функционального выражения и его параметров, нажать **OK**.

Результаты. Результаты работы процедуры отобразятся в заданной переменной **xp** в редакторе данных (правый столбец на рис. 2.9).

Следующий пример посвящен моделированию случайных (псевдослучайных) выборок из заданного распределения. Такие выборки используются, например, для организации случайного выбора из заданной генеральной совокупности (см. гл. 1), для имитации реальных физических процессов при проведении различных расчетов и т.д.

Пример 2.2к. Создадим выборку размера 10 из равномерного распределения на отрезке $[0, 5]$.

Подготовка данных. В пакете SPSS присутствует обширный набор датчиков «псевдослучайных» чисел, но ни в одном из них нет параметра объем выборки. Все эти процедуры будут формировать выборки, равные по объему переменной, уже введенной (загруженной) в редактор пакета. Таким образом, для получения случайной выборки из 10 значений необходимо ввести в редактор произвольную переменную с 10 наблюдениями. Скажем, набор чисел от 1 до 10 или набор из 10 единиц (см. рис. 2.11).

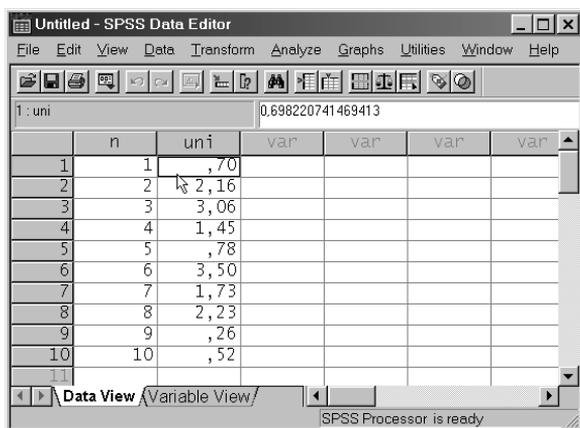


Рис. 2.11. Пакет SPSS. Редактор данных с исходными данными и результатами генерации выборки из равномерного распределения

Выбор процедуры. Такой же, как и в предыдущем примере.

Заполнение полей ввода данных и параметров процедуры.

В целом аналогично предыдущему примеру. В качестве переменной для результатов процедуры указать переменную *uni*. В окне **Numeric Expression** ввести функцию `RV.UNIFORM(min, max)`. Параметры *min* и *max* означают границы для равномерного распределения. То есть выражение в окне **Numeric Expression** должно иметь окончательный вид `RV.UNIFORM(0,5)`.

Комментарии. 1. Почти все процедуры, генерирующие псевдослучайные выборки из различных законов распределения, имеют префикс **RV** (сокращение

от Random Variable — случайная величина). После префикса следует имя распределения. (Исключением являются специализированные процедуры **NORMAL** и **UNIFORM**.)

2. Генерируемая псевдослучайная последовательность в SPSS определяется некоторым начальным, очень большим целым числом. Его значение можно задать в процедуре **Random Number Seed** (инициализация датчика случайных чисел) из меню **Transform** панели управления редактора данных.

Результаты. После задания указанного выше функционального выражения нажать . Результаты генерации представлены в переменной `uni` редактора данных (рис. 2.11).

Дополнительная литература

1. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. — М.: Наука, 1983. — 416 с.
2. *Вадзинский Р.И.* Справочник по вероятностным распределениям. — СПб.: Наука, 2001. — 295 с.
3. *Гнеденко Б.В., Хинчин А.Я.* Элементарное введение в теорию вероятностей. 10-е изд., испр. — М.: Едиториал УРСС, 2003. — 208 с.
4. *Кендэл М., Стьюарт А.* Теория распределений. — М.: Наука, 1966.
5. *Ликеш И., Ляга И.* Основные таблицы математической статистики. — М.: Финансы и статистика, 1985. — 356 с.
6. *Мюллер П., Нойман П., Шторм Р.* Таблицы по математической статистике. — М.: Финансы и статистика, 1982. — 278 с.
7. *Хастингс Н., Пикок Дж.* Справочник по статистическим распределениям. — М.: Статистика, 1980. — 95 с.

Основы проверки статистических гипотез

Во многих случаях нам требуется на основе тех или иных данных решить, справедливо ли некоторое суждение. Например, верно ли, что два набора данных исходят из одного и того же источника? Что А — лучший стрелок, чем В? Что от дома до работы быстрее доехать на метро, а не на автобусе, и т.д. Если мы считаем, что исходные данные для таких суждений в той или иной мере носят случайный характер, то и ответы можно дать лишь с определенной степенью уверенности, и имеется некоторая вероятность ошибиться. Например, предложив двум персонам А и В выстрелить по три раза в мишень и осмотрев результаты стрельбы, мы лишь предположительно можем сказать, кто из них лучший стрелок: ведь возможно, что победителю просто повезло и он по чистой случайности стрелял намного точнее, чем обычно, либо, наоборот, проигравшему не повезло, так как он стрелял намного хуже, чем обычно. Поэтому при ответе на подобные вопросы хотелось бы не только уметь принимать наиболее обоснованные решения, но и оценивать вероятность ошибочности принятого решения.

Рассмотрение таких задач в строгой математической постановке приводит к понятию *статистической гипотезы*. В этой главе мы обсуждаем, что такое статистические гипотезы, какие существуют способы их проверки, каковы наилучшие методы действий и с какими понятиями они связаны. Мы проиллюстрируем эти понятия на примере нескольких важных и часто встречающихся ситуаций и на этих же примерах покажем, как естественные проблемы надо переводить на математико-статистический язык, чтобы они могли стать предметом статистического исследования. Среди задач, рассматриваемых в этой главе, — проверка гипотез в схеме испытаний Бернулли, гипотез о положении одной выборки и о взаимном смещении двух выборок. Проверка гипотез в более сложных ситуациях рассматривается в последующих главах этой книги.

3.1. Статистические модели

Идея случайного выбора. Прежде чем приступить к описанию статистических гипотез, обсудим еще раз понятие случайного выбора, которое уже рассматривалось в гл. 1.

Если опустить детали и некоторые (хотя и важные) исключения, можно сказать, что весь статистический анализ основан на *идеи случайного выбора*. Мы принимаем тезис, что имеющиеся данные появились как результат случайного выбора из некоторой генеральной совокупности, нередко — воображаемой. Обычно мы полагаем, что этот случайный выбор произведен природой. Впрочем, во многих задачах эта генеральная совокупность вполне реальна, и выбор из нее произведен активным наблюдателем.

Для краткости будем говорить, что все данные, которые мы собираемся изучить как единое целое, представляют собой *одно наблюдение*. Природа этого собирательного наблюдения может быть самой разнообразной. Это может быть одно число, последовательность чисел, последовательность символов, числовая таблица и т.д. Обозначим на время это собирательное наблюдение через x . Раз мы считаем x результатом случайного выбора, мы должны указать и ту генеральную совокупность, из которой x был выбран. Это значит, что мы должны указать те значения, которые могли бы появиться вместо реального x . Обозначим эту совокупность через X . Множество X называют также *выборочным пространством*, или пространством выборов.

Мы предполагаем далее, что указанный выбор произошел в соответствии с неким распределением вероятностей на множестве X , согласно которому каждый элемент из X имеет определенные шансы быть выбранным. Если X — конечное множество, то у каждого его элемента x есть положительная вероятность $p(x)$ быть выбранным. Случайный выбор по такому вероятностному закону легко понимать буквально. Для более сложно устроенных бесконечных множеств X приходится определять вероятность не для отдельных его точек, а для подмножеств. Случайный выбор одной из бесконечного множества возможностей вообразить труднее, он похож на выбор точки x из отрезка или пространственной области X .

Соотношение между наблюдением x и выборочным пространством X , между элементами которого распределена вероятность, — в точности такое же, как между элементарными исходами и пространством элементарных исходов, с которым имеет дело теория вероятностей (и которые мы обсуждали в гл. 1). Благодаря этому теория вероятностей становится основой математической статистики, и поэтому, в частности, мы можем применять вероятностные соображения к задаче проверки статистических гипотез.

Прагматическое правило. Ясно, что раз мы приняли вероятностную точку зрения на происхождение наших данных (т.е. считаем, что они получены путем случайного выбора), то все дальнейшие сужде-

ния, основанные на этих данных, будут иметь вероятностный характер. Всякое утверждение будет верным лишь с некоторой вероятностью, а с некоторой тоже положительной вероятностью оно может оказаться неверным. Будут ли полезными такие выводы и можно ли вообще на таком пути получить достоверные результаты?

На оба эти вопроса следует ответить положительно. Во-первых, знание вероятностей событий полезно, так как у исследователя быстро вырабатывается вероятностная интуиция, позволяющая ему оперировать вероятностями, распределениями, математическими ожиданиями и т.п., извлекая из этого пользу. Во-вторых, и чисто вероятностные результаты могут быть вполне убедительными: вывод можно считать практически достоверным, если его вероятность близка к единице.

Можно высказать следующее *прагматическое правило*, которым руководствуются люди и которое соединяет теорию вероятностей с нашей деятельностью.

- *Мы считаем практически достоверным событие, вероятность которого близка к 1.*
- *Мы считаем практически невозможным событие, вероятность которого близка к 0.*

И мы не только так думаем, но и поступаем в соответствии с этим!

Изложенное прагматическое правило, в строгом смысле, конечно, неверно, поскольку оно не защищает полностью от ошибок. Но ошибки при его использовании будут редки. Правило полезно тем, что дает возможность практически применять вероятностные выводы.

Иногда то же правило высказывают чуть по-другому: *в однократном испытании маловероятное событие не происходит (и наоборот — обязательно происходит событие, вероятность которого близка к 1)*. Слово «однократный» вставлено ради уточнения, ибо в достаточно длинной последовательности независимых повторений опыта упомянутое маловероятное (в одном опыте!) событие встретится почти обязательно. Но это уже совсем другая ситуация.

Остается еще неразъясненным, какую вероятность следует считать малой. На этот вопрос нельзя дать количественного ответа, пригодного во всех случаях. Ответ зависит от того, какой опасностью грозит нам ошибка. Довольно часто — при проверке статистических гипотез, например, о чем см. ниже — полагают малыми вероятности, начиная с $0.01 \div 0.05$. Другое дело — надежность технических устройств, например тормозов автомобиля. Здесь недопустимо большой будет вероятность отказа, скажем, 0.001 , так как выход из строя тормозов один раз на тысячу торможений повлечет большое число аварий. Поэтому при расчетах

надежности нередко требуют, чтобы вероятность безотказной работы была бы порядка $1 - 10^{-6}$. Мы не будем обсуждать здесь, насколько реалистичны подобные требования: может ли обеспечить такую точность в расчете вероятности неизбежно приближенная математическая модель и как затем сопоставить расчетные и реальные результаты.

Предупреждения. 1. Следует дать несколько советов, как надо строить статистические модели, притом зачастую в задачах, не имеющих явного статистического характера. Для этого надо присущие обсуждаемой проблеме черты выразить в терминах, относящихся к выборочному пространству и распределению вероятностей. К сожалению, в общих словах этот процесс описать невозможно. Более того, этот процесс является творческим, и его невозможно *заучить*, как, скажем, таблицу умножения. Но ему можно *научиться*, изучая образцы и примеры и следуя их духу. Мы разберем несколько таких примеров в параграфе 3.3. В дальнейшем мы также будем уделять особое внимание этой стадии статистических исследований.

2. При формализации реальных задач могут возникать весьма разнообразные статистические модели. Однако математической теорией подготовлены средства для исследования лишь ограниченного числа моделей. Для ряда типовых моделей теория разработана очень подробно, и там можно получить ответы на основные вопросы, интересующие исследователя. Некоторую часть таких стандартных моделей, с которыми на практике приходится иметь дело чаще всего, мы обсудим в данной книге. Другие можно найти в более специальных и подробных руководствах и справочниках.

3. Об ограниченности математических средств стоит помнить и при математической формализации эксперимента. Если возможно, надо свести дело к типовой статистической задаче. Эти соображения особенно важны при *планировании* эксперимента или исследования; при сборе информации, если речь идет о статистическом обследовании; при постановке опытов, если мы говорим об активном эксперименте.

3.2. Проверка статистических гипотез (общие положения)

В этом параграфе мы рассмотрим основные теоретические понятия и подходы, используемые при проверке статистических гипотез. Этот материал весьма важен, но непрост в освоении. Поэтому при каких-либо затруднениях при чтении данного параграфа целесообразно заглянуть чуть вперед в п. 3.3 — там показано, как описываемые понятия и подходы возникают в практических задачах.

Статистические гипотезы. В обычном языке слово «гипотеза» означает предположение. В том же смысле оно употребляется и в научном языке, используя в основном для предположений, вызывающих сомнения. В математической статистике термин «гипотеза» означает

предположение, которое не только вызывает сомнения, но и которое мы собираемся в данный момент проверить.

При построении статистической модели приходится делать много различных допущений и предположений, и далеко не все из них мы собираемся или можем проверить. Эти предположения относятся как к выборочному пространству, так и к распределению вероятностей $P(\cdot)$ на нем.

Вопросов о выборочном пространстве обычно не возникает. Вопросы и сомнения относятся к распределению вероятностей. Среди них бывают и такие: обладает ли $P(\cdot)$ определенным свойством? (Это свойство $P(\cdot)$ выражает в статистической форме вопрос, интересующий исследователя с содержательных позиций.) Вопрос можно поставить в форме проверки предположения: сначала высказать гипотезу «Распределение вероятностей обладает таким-то свойством», а затем спросить, верно ли это. Предположение может быть как о конкретном законе распределения (например: «данные являются выборкой из нормального закона с заданными параметрами»), так и о частных характеристиках распределения, таких как симметрия, принадлежность к определенному типу, о значениях параметров и т.д. Соответственно различают простые и составные (сложные) гипотезы:

- *простая гипотеза* полностью задает распределение вероятностей;
- *сложная гипотеза* указывает не одно распределение, а некоторое множество распределений. Обычно это множество распределений, обладающих определенным свойством (свойствами).

Статистическая проверка гипотезы состоит в выяснении того, насколько совместима эта гипотеза с имеющимся (наблюдаемым) результатом случайного выбора. Надо, следовательно, решить, совместимо ли с наблюдением x определенное множество распределений вероятностей $P(\cdot)$, соответствующих данной гипотезе.

Как итог обсуждения можно высказать следующее определение.

Определение. *Статистическая гипотеза — это предположение о распределении вероятностей, которое мы хотим проверить по имеющимся данным.*

Остается выяснить, как это можно сделать.

Проверка гипотез. Поговорим прежде о проверке гипотез вообще. Лучше всего, если гипотезу можно проверить непосредственно, — тогда не возникает никаких методических проблем. Но если прямого способа проверки у нас нет, приходится прибегать к проверкам косвенным. Это

значит, что приходится довольствоваться проверкой некоторых следствий, которые логически вытекают из содержания гипотезы. Если некоторое явление логически неизбежно следует из гипотезы, но в природе не наблюдается, то это значит, что гипотеза неверна. С другой стороны, если происходит то, что при гипотезе происходить не должно, это тоже означает ложность гипотезы. Заметим, что подтверждение следствия еще не означает справедливости гипотезы, поскольку правильное заключение может вытекать и из неверной предпосылки. Поэтому, строго говоря, косвенным образом *доказать* гипотезу нельзя, хотя *опровергнуть* — можно.

Впрочем, когда косвенных подтверждений накапливается много, общество зачастую расценивает их как убедительное доказательство в пользу гипотезы. В языке это отражается так, что бывшую гипотезу начинают именовать законом.

Скажем, когда Ньютон выдвинул для объяснения движения небесных тел свой закон всемирного тяготения, он выглядел как некое предположение. По отношению к планетам он давал не больше сведений, чем законы Кеплера. Ньютону нужны были новые объекты, на которых он мог бы проверить действие своего открытия. Таким небесным телом могла бы быть Луна. Мы знаем сейчас, что на ее движение оказывают влияние своим притяжением не только Земля, но и Солнце, а также другие планеты. Поэтому ее движение не является в точности эллиптическим, а из-за близости Луны к Земле мы можем наблюдать эти отклонения. Ньютону удалось объяснить многие особенности движения Луны, но полностью удовлетворен он не был. Может быть, именно поэтому он так долго медлил с опубликованием своего открытия. Для решения этой и других задач небесной механики понадобились усилия лучших ученых следующего, восемнадцатого века¹.

Однако впоследствии на основании формулы Ньютона были объяснены не только движение Луны, но и траектории комет, открыты планеты Уран, Нептун и Плутон. Поэтому предположение Ньютона стало считаться уже не гипотезой, а законом природы, в справедливости которого никто не сомневается. Лишь во второй половине XX века, когда стало возможным измерять координаты небесных тел (в частности, искусственных спутников Земли) с точностью до сантиметров, их траектории стало необходимо рассчитывать не по закону Ньютона, а по более точным формулам общей теории относительности Эйнштейна.

Для проверки естественно-научных гипотез часто применяется такой принцип: гипотезу отвергают, если происходит то, что при ее справедливости происходить не должно. Проверка статистических гипотез происходит так же, но с оговоркой: место невозможных событий

¹ К слову сказать, теория движения Луны должна быть очень точной, ибо у нас (у человечества) есть очень мощные возможности ее проверки — лунные и солнечные затмения, сведения о которых сохранились в истории за многие тысячелетия. Теория должна не только достаточно точно предсказывать даты близящихся затмений (что относительно нетрудно), но и рассчитывать эти даты на много веков назад и получать при этом верные результаты. Такой точности добиться нелегко.

занимают события практически невозможные. Причина этого проста: пригодных способов для проверки невозможных событий, как правило, просто нет.

Альтернативы. Повторим вышесказанное чуть более формально и точно. Итак, пусть H — статистическая гипотеза, т.е. предположение о распределении вероятностей на выборочном пространстве. Будем далее говорить о вероятностях событий, вычисленных в предположении, что H справедлива, или, коротко — о вероятностях при H , обозначая их $P(\cdot|H)$. Если H — простая гипотеза, то для всякого события A (A — множество в выборочном пространстве) его вероятность $P(A|H)$ определена однозначно. Если гипотеза H сложная (состоит из многих простых), то $P(A|H)$ обозначает все возможные при H значения вероятности события A .

Выберем уровень вероятности ε , $\varepsilon > 0$. Условимся считать событие практически невозможным, если его вероятность меньше ε . Когда речь идет о проверке гипотез, число ε называют *уровнем значимости*.

Выберем событие A , вероятность которого при гипотезе меньше ε , т.е. $P(A|H) < \varepsilon$. (Если H — сложная гипотеза, то меньше ε должны быть все возможные при H значения вероятности A .) Правило проверки H теперь таково:

На основании эксперимента мы отвергаем гипотезу H на уровне значимости ε , если в этом эксперименте произошло событие A .

Таким образом, уровень значимости есть вероятность ошибочно отвергнуть гипотезу, когда она верна.

Определение. *Событие A называется критическим для гипотезы H , или критерием для H . Если $P(A|H) \leq \varepsilon$, то ε называют гарантированным уровнем значимости критерия A для H .*

Теперь обсудим вопрос о том, как следует выбирать критическое событие. Далеко не всякое маловероятное при гипотезе событие целесообразно использовать для ее проверки. Например, если это событие имеет одну и ту же вероятность и при соблюдении, и при несоблюдении гипотезы, то информация о том, произошло событие или нет, не даст нам ровно никаких сведений о гипотезе. Поэтому при выборе события A следует принимать во внимание вероятность этого события не только при соблюдении гипотезы, но и при ее несоблюдении!

На практике нас, однако, обычно интересуют не все возможные «несоблюдения» гипотезы H , а лишь некоторые. Во-первых, обычно у наблюдаемого явления x имеются или предполагаются некоторые свойства, которые выполняются и при соблюдении, и при несоблюдении

H , что ограничивает круг возможных распределений при несоблюдении H . Во-вторых, нас могут интересовать некоторые специфические (например, наиболее часто встречающиеся) нарушения H , и мы можем захотеть построить правило проверки H , «чувствительное» именно к этим видам отклонений. Поэтому при проверке статистических гипотез рассматривают не только множество распределений на X , допустимых при выполнении H , но и указывают множество H' распределений на X , которые мы рассматриваем в качестве «альтернативы» гипотезе H .

Определение. *Распределения, с которыми мы можем встретиться в случае нарушения H , называют альтернативными распределениями, или альтернативами. (Иногда говорят также о конкурирующих распределениях и о конкурирующих гипотезах.)*

Ниже мы увидим, что обычно «специализированные», т.е. рассчитанные на более узкий круг альтернатив, способы проверки статистических гипотез являются (для этих альтернатив!) более «мощными», чем «универсальные», т.е. рассчитанные на широкий круг альтернатив.

Выбор критического события. Теперь вернемся к вопросу выбора критического события A . Идеальным было бы найти для проверки H такое событие, которое не может произойти при гипотезе и обязательно происходит при альтернативе: появление (непоявление) такого события было бы наилучшим индикатором для H . Прекрасно подошло бы и такое критическое событие, вероятность которого близка к 0 при гипотезе и близка к 1 при альтернативе. Однако существование такого события возможно не всегда. Например, при проверке гипотезы о том, что некоторый параметр распределения равен a , против альтернативы о том, что он не равен a , такого события указать нельзя, поскольку при приближении параметра распределения к a вероятность любого события будет приближаться к тому значению, которое она имела бы при параметре, равном a . В подобных случаях приходится довольствоваться меньшим: в качестве критического выбирают событие, вероятность (вероятности — если гипотеза сложная) которого (малая при гипотезе) *увеличивается* по мере удаления распределения от гипотетического (гипотетических).

В некоторых случаях эту мысль удастся осуществить в виде выбора оптимального критического множества заданного уровня значимости. Именно так обстоит дело для многих широко используемых статистических моделей. Например, в схеме Бернулли для некоторых практически важных гипотез и альтернатив существуют наилучшие (наиболее мощные) критерии. Но в целом такие удачи редки. Теоретиками предлага-

лись многие идеи, как рационально выбирать критические множества. Но удовлетворительного общего решения этой проблемы нет.

Статистики критериев. Обычно для построения критического множества используется следующий подход. Пусть T — некоторая функция на множестве X , принимающая числовые значения. Мы будем называть T *статистикой критерия*. Как правило, статистику T выбирают таким образом, чтобы ее распределения при гипотезе и при альтернативе как можно более различались (в случае, если множества распределений H и H' «касаются» друг друга — чтобы различие в распределениях T было как можно большим по мере удаления истинного распределения наблюдений от гипотетического). При таком выборе статистики T обычно некоторые значения T (например, слишком большие или слишком малые) являются нетипичными при гипотезе и типичными при альтернативе. Поэтому для построения критического множества A выбирают некоторое множество вещественных чисел A' (множество «нетипичных» при гипотезе значений статистики T), и полагают множество A как

$$A = \{x \mid T(x) \in A'\}.$$

Это множество будет критическим для гипотезы на уровне $\max_{P \in H} P(A)$. Поскольку множество A полностью определяется по A' , множество A' тоже называют *критическим*.

Читатель может подумать, что мы не продвинулись ни на шаг вперед: вместо выбора критического множества A надо выбирать критическое множество A' . Но дело в том, что обычно множество A' устроено очень просто. Например, если статистика критерия T выбрана так, что она принимает небольшие значения при гипотезе и большие — при альтернативе, то множество A' следует выбирать как $\{y \mid y \geq a\}$, где a — некоторое число. При другом поведении статистики T множество A' может быть устроено по-другому, например $\{y \mid y \leq a\}$ или $\{y \mid y \leq a \text{ или } y \geq b\}$. Разумеется, следует выбирать множество A' так, чтобы $\max_{P \in H} P(A) \leq \varepsilon$, где ε — уровень значимости критерия. С конкретными примерами применения данного подхода можно познакомиться ниже в этой главе.

Ошибки первого и второго рода. При проверке статистических гипотез возможны ошибочные заключения двух типов:

- отвержение гипотезы в случае, когда она на самом деле верна;
- неотвержение (принятие) гипотезы, если она на самом деле неверна.

Эти возможности называются соответственно *ошибками первого рода* и *ошибками второго рода*.

Из-за различного подхода к гипотезе и альтернативе наше отношение к ошибкам первого и второго рода также неодинаково. При построении статистических критериев мы фиксируем максимальную допустимую вероятность ошибки первого рода (т.е. уровень значимости критерия) и стремимся выбрать критическое множество таким образом, чтобы минимизировать вероятность ошибки второго рода (или хотя бы сделать так, чтобы эта вероятность была как можно меньше по мере удаления истинного распределения от гипотетического или гипотетических).

Мощность критерия. Обозначим через β вероятность ошибки второго рода статистического критерия. Если альтернативная гипотеза является сложной, то эта вероятность, естественно, зависит от выбора конкретного альтернативного распределения. Если мы рассматриваем альтернативы из какого-либо параметрического семейства распределений P_θ , значение β также можно считать функцией от θ .

Величину $1 - \beta$ обычно называют *мощностью критерия*. Ясно, что мощность критерия может принимать любые значения от 0 до 1. Чем ближе мощность критерия к единице, тем более эффективен (более «мощен») критерий. Многие известные статистические критерии получены путем нахождения наиболее мощного критерия при заданных предположениях о гипотезе и альтернативе.

3.3. Примеры статистических моделей и гипотез

Покажем на примерах, как может проходить математическая формализация практических задач и как сформулированные на естественном языке вопросы превращаются в статистические гипотезы.

Тройной тест. Рассмотрим распространенный в психологии тройной тест (его другое название — тест дегустатора, см. [107]). Он состоит из серии одинаковых опытов, в каждом из которых испытуемому предъявляют одновременно три стимула. Два из них идентичны, а третий несколько отличается. Испытуемый, ориентируясь на свои ощущения, должен указать этот отличающийся стимул. Например, испытуемому могут быть предложены три стакана с жидкостью: два с чистой водой, а третий — со слабым раствором сахара, либо наоборот — два стакана подслащенных, а третий — с чистой водой. Задание для испытуемого — указать стакан, отличающийся от двух других.

Опыты стараются организовать так, чтобы они проходили в одинаковых условиях и чтобы в каждом из них испытуемый мог полагаться только на свои ощущения. В результате подобного однократного экс-

перимента можно получить как правильный, так и неправильный ответ. При слабой концентрации раствора, когда его трудно отличить от воды, из одного ответа нельзя сделать определенного заключения о способности испытуемого чувствовать данную концентрацию. Испытуемый может случайно ошибиться, даже если в целом он способен отличать данную концентрацию сахара от чистой воды. С другой стороны, правильный ответ не исключает того, что испытуемый его просто угадал, не отличая раствора от воды.

Эти свойства эксперимента мы можем перечислить в виде следующих допущений:

- в каждом испытании ответ испытуемого случаен;
- существует вероятность правильного ответа, которая неизменна во все время испытаний;
- результаты отдельных испытаний статистически независимы.

Коротко это выражается так: статистической моделью эксперимента служит схема Бернулли.

Сформулировав математическую модель явления, перейдем к выдвиганию статистических гипотез. Интересующая нас способность испытуемого характеризуется вероятностью правильного ответа, которую мы обозначим p . В этом опыте она нам неизвестна. Естественно, эта вероятность зависит от степени концентрации сахара. Если концентрация очень мала и не воспринимается, то у испытуемого нет оснований для выбора. Он «наудачу» будет указывать один из трех стаканов. В этих условиях вероятность правильного ответа $p = 1/3$.

Предположим, что экспериментатора интересует, начиная с каких концентраций испытуемый отличает раствор от воды. Тогда для данной концентрации экспериментатор может выдвинуть предположение, что испытуемый ее ощутит не состоянием. В изложенной модели это предположение превращается в статистическую гипотезу о том, что $p = 1/3$. Примем следующую форму записи статистической гипотезы: $H : p = 1/3$. Если же экспериментатор предполагает, что испытуемый может ощутить наличие сахара, то соответствующая статистическая гипотеза состоит в том, что $p > 1/3$, т.е. $H : p > 1/3$. Возможна и гипотеза о том, что $p < 1/3$, она соответствует тому, что испытуемый способен отличить раствор от воды, но принимает одно за другое.

Экспериментатор может выдвигать и другие гипотезы о способности испытуемого к различению концентраций. Например, возможна такая гипотеза: испытуемый способен ощутить присутствие сахара, ошибаясь один раз из десяти. В этом случае вероятность правильного ответа равна 0.9 и гипотеза примет вид: $H : p = 0.9$.

Заметим, что с чисто математической точки зрения гипотеза вида $H : p = 1/3$ проще, чем $p > 1/3$ или $p < 1/3$. Действительно, при $p = 1/3$ мы имеем дело с одним (полностью заданным) биномиальным распределением, а в других случаях перед нами семейство распределений. Ясно, что с одним распределением иметь дело проще.

Сейчас мы не будем рассматривать процесс проверки этих гипотез (он описан в п. 3.4), а вместо этого приведем еще один пример перевода естественно-научной задачи на статистический язык, т.е. построения статистической модели явления и выдвижения гипотезы для проверки.

Парные наблюдения. На практике часто бывает необходимо сравнить два способа действий по их результатам. Речь может идти о сравнении двух методик обучения, эффективности двух лекарств, производительности труда при двух технологиях и т.д. В качестве конкретного примера рассмотрим эксперимент, в котором выясняется, на какой из сигналов человек реагирует быстрее: на свет или на звук.

Эксперимент был организован следующим образом (см. [33]). Каждому из семнадцати испытуемых в случайном порядке поочередно подавались два сигнала: световой и звуковой. Интенсивность сигналов была неизменна в течение всего эксперимента. Увидев или услышав сигнал, испытуемый должен был нажать на кнопку. Время между сигналом и реакцией испытуемого регистрировал прибор. Результаты эксперимента приведены в табл. 3.1.

Таблица 3.1

Время реакции на свет и на звук, в миллисекундах

i	x_i	y_i	i	x_i	y_i
1	223	181	9	200	155
2	104	194	10	191	156
3	209	173	11	197	178
4	183	153	12	183	160
5	180	168	13	174	164
6	168	176	14	176	169
7	215	163	15	155	155
8	172	152	16	115	122
			17	163	144

i — номер испытуемого, $i = 1, \dots, 17$; x_i — время его реакции на звук, y_i — время его реакции на свет.

Вместо поставленного выше вопроса о том, на какой из сигналов человек отвечает быстрее, выдвинем другой: можно ли считать, что время реакции человека на свет и на звук одинаково? Логически эти вопросы тесно связаны: если мы отвечаем отрицательно на второй из них, мы тем самым признаем, что различия есть. После этого уже

нетрудно понять, когда время реакции меньше. Если же на второй вопрос мы отвечаем положительно, то первый после этого просто снимается. С математической же точки зрения второй вопрос проще, как мы увидим из дальнейшего обсуждения.

Итак, время реакции на звук, X , и время реакции на свет, Y , различно у разных людей, несмотря на то, что во время опыта они находились в одинаковых условиях. Ясно, что наблюдаемый разброс во времени реакции не связан с изучаемым явлением (различием двух действий). По-видимому, этот разброс можно объяснить различиями между испытуемыми и/или нестабильностью времени отклика на сигнал у каждого испытуемого. Как бы то ни было, эти колебания не имеют отношения к той закономерности, что нас интересует. *Поэтому мы объявляем их случайными.* Так сделан первый шаг к статистической модели: переменные x_i и y_i признаны реализациями случайных величин, скажем, X_i и Y_i . Поскольку каждый испытуемый решал свои задачи самостоятельно, не взаимодействуя с другими испытуемыми и не испытывая с их стороны влияния, мы будем считать случайные величины $X_1, Y_1, \dots, X_{17}, Y_{17}$ независимыми (в теоретико-вероятностном смысле).

Выбор статистической модели. Дальнейшее уточнение статистической модели в подобных задачах может идти различными путями, в зависимости от природы эксперимента и наших знаний о ней. Один путь связан с предположением о том, что случайные величины X_i и Y_i имеют некоторые конкретные законы распределения. Например, мы можем предположить, что X_i и Y_i — независимы и имеют нормальные распределения с одной и той же дисперсией (обозначим ее σ^2). Тогда, если ввести для средних значений обозначения: $MX_i = a_i, MY_i = b_i$ где $i = 1, \dots, 17$, то можно сформулировать наши допущения так: случайные величины X_i, Y_i подчиняются распределениям $N(a_i, \sigma^2), N(b_i, \sigma^2)$ соответственно, где параметры $a_1, b_1, \dots, a_{17}, b_{17}, \sigma^2$ нам неизвестны. При этих обозначениях выдвинутый вопрос о равном времени реакции на свет и на звук может быть сформулирован как статистическая гипотеза:

$$H : a_1 = b_1, a_2 = b_2, \dots, a_{17} = b_{17}.$$

Если экспериментатор уверен, что группа испытуемых достаточно однородна, он может дополнительно предположить, что $a_1 = \dots = a_{17}$ и $b_1 = \dots = b_{17}$. Если обозначить общие значения параметров через a и b соответственно, то статистическую модель в этом случае можно сформулировать так: случайные величины X_1, \dots, X_{17} независимы и распределены по закону $N(a, \sigma^2)$; случайные величины Y_1, \dots, Y_{17} тоже независимы, не зависят от X_1, \dots, X_{17} и распределены по закону

$N(b, \sigma^2)$. Параметры a, b и σ^2 неизвестны. Тогда гипотезу о равном времени реакции можно записать следующим образом:

$$H : a = b.$$

Ясно, что задача с меньшим числом неопределенных параметров, как во второй постановке, в принципе должна давать более точные ответы. При проверке гипотез это означает, что мы сможем принять или отвергнуть проверяемую гипотезу с большей степенью уверенности. Но следует помнить, что уменьшение количества параметров в модели является следствием принятия дополнительных предположений об имеющихся данных. Так, в приведенном выше примере мы предположили, что $MX_1 = \dots = MX_{17}$ и $MY_1 = \dots = MY_{17}$, что и дало нам возможность уменьшить количество параметров в модели с 35 до 3. Но если сделанные дополнительные предположения являются неправомерными, то использование полученной математической модели может привести к неверному заключению. Например, при обработке наших данных по однородной схеме можно получить неверный ответ, если фактически эти данные однородными не являются.

Итак, при построении статистической модели постоянно приходится вводить упрощающие математические предположения и одновременно оценивать, насколько они приемлемы с содержательной точки зрения. И часто надо быть готовым к тому, чтобы отказаться от недопустимых предположений или заменить их чем-то другим.

Другой путь построения статистической модели — так называемый *непараметрический*. Здесь мы не делаем предположений о том, что наблюдаемые случайные переменные имеют какой-либо параметрический закон распределения. В этом случае мы делаем меньше математических допущений, а значит, здесь меньше опасности принять неоправданное предположение. Зато при этом мы используем не всю информацию об имеющихся данных, а только ту ее часть, которая не зависит от конкретного вида распределения исходных данных. Например, при проверке гипотезы о равном времени реакции на свет и звук мы должны будем использовать не сами значения времен реакций X_i и Y_i , а их *ранги* в объединенной выборке X_i и Y_i . По сравнению с параметрическим методом (если предположения о параметрическом характере случайных событий справедливы) мы получим при этом несколько менее точные выводы, но зато непараметрический метод имеет гораздо более широкую область применимости. Более подробно мы обсудим непараметрический подход к описанной задаче в п. 3.6.1.

Итак, при построении статистической модели приходится делать ряд предположений. Большую часть этих предположений мы не проверяем

(и часто даже и не можем проверить). Некоторые предположения мы выбираем для проверки их совместимости со статистическим материалом, и называем такие предположения статистическими гипотезами. Ниже мы расскажем, как осуществляется проверка статистических гипотез.

3.4. Проверка статистических гипотез (прикладные задачи)

3.4.1. Схема испытаний Бернулли

Вероятности событий при гипотезе. Обратимся к описанному выше тройному тесту. Мы выяснили, что статистической моделью этого теста является схема испытаний Бернулли, и выдвинули несколько статистических гипотез, которые были сформулированы так: $H : p = 1/3$, $H : p > 1/3$, $H : p = 0.9$, где p — вероятность правильного ответа в одном испытании.

Пусть для определенности число испытаний $n = 10$. (Вообще-то десяти испытаний для серьезных выводов недостаточно. Мы выбрали $n = 10$ только ради простоты изложения, чтобы сделать последующие расчеты легко обозримыми.) В качестве наблюдения x в этой схеме эксперимента должны выступать результаты этих 10 испытаний, т.е. последовательность длины 10 вида *успех, неудача, неудача, успех* и т.д. Соответственно пространство X состоит из $2^n = 2^{10}$ всевозможных таких последовательностей. Вероятность любой из них равна $p^S(1-p)^{n-S}$, где S — число правильных ответов. Можно показать, что статистические решения, основанные на S , не будут менее точными, чем решения, основанные на полной записи результатов. (Это очень интересная математическая особенность, на которой мы не можем останавливаться. Скажем лишь, что это означает, что вся информация, необходимая для принятия решений о величине p , заключена в числе успехов S , а сведения о конкретном чередовании успехов и неудач не важны.) Поэтому проверку гипотез мы будем проводить, основываясь на числе успехов S , которое имеет биномиальное распределение, подробно разобранный в гл. 2.

Для проверки первой гипотезы надо выбрать такое событие, вероятность которого, вычисленная согласно гипотетическому распределению вероятностей, была бы малой. Обозначим это событие через A . Выберем некоторое число ε , и все события, вероятность которых меньше ε , будем считать маловероятными. Пусть, например, $\varepsilon = 0.02$. Вероятность A , которую мы обыкновенно обозначаем через $P(A)$, сейчас удобно записать как $P(A | H)$, отмечая, что эта вероятность вычислена при гипотезе H .

Таблица 3.2

k	0	1	2	3	4	5
$P(S = k H)$	0.0173	0.0868	0.1950	0.2602	0.2276	0.1365
k	6	7	8	9	10	
$P(S = k H)$	0.0569	0.0163	0.0030	0.0004	0.0000	

Рассмотрим некоторые примеры событий и вычислим их вероятности. В табл. 3.2 приведены вероятности событий вида $\{S = k\}$ при $p = 1/3$.

Легко видеть, что половина этих событий маловероятна согласно выбранному нами критерию.

В табл. 3.3 приведены вероятности событий, заключающихся в том, что правильных ответов больше или равно заданному числу, т.е. событий вида $\{S \geq k\}$, $k = 0, 1, 2, \dots, 10$.

Таблица 3.3

k	0	1	2	3	4	5
$P(S \geq k H)$	1.000	0.9827	0.8959	0.7009	0.4407	0.2131
k	6	7	8	9	10	
$P(S \geq k H)$	0.0766	0.0197	0.0034	0.0004	0.0000	

Здесь тоже несколько событий имеют вероятность меньше 0.02. Как видим, для выбора маловероятного при H события A имеется довольно много возможностей. Как мы говорили в п. 3.2, надо выбрать A так, чтобы $P(A | H)$ была малой, но при нарушении H становилась бы большой. То есть выбрать такое A , которое неправдоподобно при H и естественно (обыкновенно, не удивительно) при рассматриваемой альтернативе к H . Как мы установили в п. 3.3, альтернативой к гипотезе $H : p = 1/3$ может быть совокупность распределений, для которых $p > 1/3$. Таким образом, с простой гипотезой H конкурирует сложная альтернатива $H_1 : p > 1/3$. Эту альтернативу называют односторонней (правосторонней), чтобы отличить от двусторонней альтернативы $H_2 : p \neq 1/3$.

Можно, разумеется, рассматривать и простые альтернативы к гипотезе H . Рассмотрим, например, альтернативу $H_3 : p = 0.9$, и разберем в этой ситуации, как осуществить выбор множества A , руководствуясь изложенным выше принципом.

Вероятности событий при альтернативе. Посмотрим, как изменяются вероятности событий, приведенных в табл. 3.3, когда они вычисляются при альтернативе $p = 0.9$. Соответствующие значения даны в табл. 3.4.

Таблица 3.4

k	0	1	2	3	4	5
$P(S \geq k H_3)$	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
k	6	7	8	9	10	
$P(S \geq k H_3)$	0.9984	0.9872	0.9298	0.7361	0.3487	

Анализируя табл. 3.2, видим, что события $S = 7$, $S = 8$, $S = 9$, $S = 10$ маловероятны как каждое в отдельности, так и все вместе взятые, т.е. объединение этих событий, которое можно записать в виде $S \geq 7$, имеет вероятность, равную 0.0197 (см. табл. 3.3). Из табл. 3.4 видно, что вероятность события $S \geq 7$, вычисленного при альтернативе, равна 0.9872, т.е. событие $S \geq 7$ при справедливости альтернативы практически достоверно. Поэтому в качестве критического для гипотезы $H : p = 1/3$ при ее проверке против конкурирующей гипотезы $H_3 : p = 0.9$ можно взять событие $\{S \geq 7\}$.

Может возникнуть следующий вопрос: почему мы не включили событие $S = 0$ в выбираемое нами маловероятное (при первой гипотезе) событие A , вместо, например, событий $S = 7$ и $S = 8$? Ответ дает расчет вероятности события $A = \{S = 0\} \cup \{S \geq 9\}$ при альтернативе. Действительно, $P(A | H_3) = 0.7361$, т.е. это событие менее вероятно при альтернативе, чем выбранное выше.

Разобранный пример характеризует в некотором смысле идеальную ситуацию, когда удастся найти такое событие A , которое практически невозможно при H и практически достоверно при альтернативе. В этом случае по результатам эксперимента, в зависимости от того, произошло или нет A , мы уверенно можем судить, имеем ли дело с H или с альтернативой.

Сложная альтернатива. С точки зрения экспериментатора, разумной альтернативой к гипотезе $H : p = 1/3$ является сложная альтернатива $H_1 : p > 1/3$. Эта альтернатива не задает конкретного распределения вероятностей в схеме Бернулли. Вероятности событий при альтернативе H_1 зависят от конкретного значения параметра p , $1/3 < p \leq 1$. Они изменяются вместе с изменением этого параметра, и мы можем судить о тенденции изменения этой вероятности. Очевидно, что чем больше значение p , тем больше вероятность появления большого числа успехов S . Это наглядно показывает сравнение табл. 3.2 и 3.4. Выше было установлено, что событие $A = \{S \geq 7\}$ при справедливости первой гипотезы маловероятно. В то же время, чем больше значение p , тем больше вероятность этого события. Так, при $p = 0.9$ $P(S \geq 7 | H_3) = 0.9872$.

Поэтому разумно именно с помощью этого события судить о справедливости H — если альтернативой к H выступает $H_1 : p > 1/3$.

Предположим, что мы провели обсуждаемый эксперимент и получили для S конкретное значение. Обозначим это наблюдаемое значение как $S_{\text{набл.}}$, чтобы отличать случайную величину S от ее реализации $S_{\text{набл.}}$. Пусть, к примеру, $S_{\text{набл.}} = 7$. Тем самым осуществилось событие $\{S \geq 7\}$. Поэтому мы отвергаем гипотезу $H : p = 1/3$ на уровне значимости 0.02 в пользу альтернативы $H_1 : p > 1/3$.

Упоминание уровня значимости в заключительном решении существенно — от его величины зависит, отвергаем мы гипотезу или нет. Пусть, например, $\varepsilon = 0.005$. Тогда критическое множество есть $\{S \geq 8\}$, и опыт, в котором $S_{\text{набл.}} = 7$, не отвергает гипотезу $H : p = 1/3$ на уровне значимости 0.005 против альтернативы $H_1 : p > 1/3$.

Выбор уровня значимости всегда произволен. Неприятно, что от этого произвола зависит решение — отвергнуть или нет гипотезу. В данном примере (и во многих других случаях) есть более гибкий способ действий — указать *минимальный уровень значимости*, на котором можно отвергнуть гипотезу.

Критическое событие в нашей задаче имеет вид $\{S \geq C\}$, где C — некоторое критическое значение. Чем больше число C , тем менее вероятно при гипотезе H событие $\{S \geq C\}$. Тем больше поэтому уверенность, что H надо отвергнуть, если $S_{\text{набл.}} \geq C$. Наибольшей достижимой уверенности соответствует наименьший возможный уровень значимости, который в нашей задаче есть $P(S \geq S_{\text{набл.}} | H)$.

Наименьший уровень значимости полезно вычислять во всех случаях, так как он характеризует, насколько сильно наблюдаемое значение $S_{\text{набл.}}$ противоречит гипотезе H .

Виды альтернатив. В примере испытаний Бернулли, которые обсуждались выше, разумный класс альтернатив к гипотезе $H : p = 1/3$ был определен как $p > 1/3$. Такие альтернативы называют *односторонними* (в данном случае правосторонними). Встречаются статистические задачи и с левосторонними альтернативами. Основную гипотезу в этом случае приходится отвергать, если успехов в опыте зарегистрировано неестественно мало с точки зрения гипотезы H . Иначе говоря, критическое множество A имеет вид $A = \{S \leq C\}$, а число C выбирается так, чтобы $P(S \leq C | H)$ была малой.

Наиболее общими альтернативами являются *двусторонние* альтернативы. Пусть основная (нулевая, как часто говорят) гипотеза имеет вид $H_0 : p = p_0$, где p — некоторое определенное число. Если невозможно заранее указать направление изменения p при отступлении от

H_0 , приходится рассматривать альтернативу вида $H : p \neq p_0$. Руководствуясь изложенными принципами проверки статистических гипотез и характером изменения распределения вероятностей между возможными значениями S (числом успехов) при разных p , мы заключаем, что в данном случае следует отвергнуть гипотезу H_0 и тогда, когда $S_{\text{набл.}}$ неправдоподобно велико, и тогда, когда оно неправдоподобно мало. Напомним, что все эти вероятности вычисляются так, как это предписывает проверяемая гипотеза H_0 .

Следовательно, надо выбрать два критических значения для S , а именно верхнее и нижнее, скажем, x и y . Выбрать их необходимо так, чтобы $P(S \leq y | H_0) + P(S \geq x | H_0)$ была малой. Гипотеза H_0 отвергается, если $S_{\text{набл.}} \leq y$, либо $S_{\text{набл.}} \geq x$. Уровень значимости этого критерия есть $P(S \leq y | H_0) + P(S \geq x | H_0)$.

Замечание. Обычно описанное правило оформляют несколько иначе, следя за отклонением наблюдаемого S от его ожидаемого значения np_0 . Напомним, что математическое ожидание числа успехов в схеме Бернулли равно $MS = np_0$. С помощью таблиц выбирают число z так, что вероятность $P(|Sn_{\text{набл.}}| \geq z | H_0)$ оказывается малой. Гипотезу $H_0 : p = p_0$ отвергают, если $|S_{\text{набл.}} - np_0| \geq z$. В этом случае статистикой критерия служит уже не S , а $|S - np_0|$. Здесь также можно вычислять минимальный уровень значимости, на котором можно отвергнуть $H_0 : p = p_0$ против двусторонней альтернативы $p \neq p_0$. Он равен $P(|S - np_0| \geq |S_{\text{набл.}} - np_0| | H)$.

3.4.2. Критерий знаков для одной выборки

На изложенном выше способе проверки статистических гипотез в схеме Бернулли основан широко распространенный *критерий знаков*. Для его применения достаточны очень слабые предположения о законе распределения данных, такие как независимость наблюдений и однозначная определенность медианы. Напомним, что медианой распределения случайной величины ξ называется такое число θ , для которого $P(\xi < \theta) = P(\xi > \theta) = 1/2$.

Предположим, что в результате многочисленных измерений артериального кровяного давления у пациентов некой поликлиники было установлено его медианное значение θ . Эти измерения возобновились после летних отпусков. У первых N пациентов были зарегистрированы значения давления крови x_1, \dots, x_N . Можно ли считать, что медианный уровень давления понизился после летнего отдыха?

Как обычно, проще проверить гипотезу о том, что значение медианы θ не изменилось. При этом надо рассматривать только односторонние альтернативы — в данном случае левосторонние (как будет описано

ниже). Если гипотеза будет отвергнута, это будет означать положительный ответ на поставленный выше вопрос.

Проверка этой гипотезы с помощью критерия знаков проводится следующим образом. Рассмотрим случайную величину $X - \theta$. Так как, согласно гипотезе, $\text{med } X = \theta$, то $P\{(X - \theta) > 0\} = P\{(X - \theta) < 0\} = 1/2$. В выборке $x_i - \theta$, $i = 1, \dots, N$ подсчитаем число положительных разностей и обозначим его через S . Для формализации этого алгоритма удобно ввести функцию

$$s(x) = \begin{cases} 1, & \text{при } x > 0, \\ 0, & \text{при } x < 0. \end{cases}$$

Тогда $S = \sum_{i=1}^N s(x_i - \theta)$. Случайная величина $s(x)$ принимает два значения: 0 и 1. Согласно выдвинутой гипотезе, вероятность каждого из этих значений равна $1/2$. Таким образом, видно, что задача сводится к схеме испытаний Бернулли, в которой через S обозначено число «успехов», и следует проверить гипотезу $H : p = 1/2$. В нашем примере надо рассматривать левосторонние альтернативы, но вообще альтернативы могут быть как односторонними, так и двусторонними, в зависимости от решаемой задачи.

Отметим важное обстоятельство в приведенном примере. Гипотеза о значении медианы случайной величины, выдвинутая нами первоначально, не определяла однозначно закон распределения X и тем самым не позволяла вычислить вероятность произвольных значений X . В связи с этим мы были вынуждены перейти к случайной величине $s(x - \theta)$, которая задает только знак разности $x - \theta$. При этом вероятностное распределение $s(x - \theta)$ определяется уже однозначно. Изложенный критерий получил название *критерия знаков*, так как он работает фактически только со знаками преобразованных некоторым образом случайных величин. Этот критерий хорош именно тем, что требует очень немногого от функции распределения случайной величины и очень прост в применении.

3.5. Проверка гипотез в двухвыборочных задачах

Область применения. Рассмотрим часто встречающуюся на практике задачу сравнения двух выборочных совокупностей. В духе основной статистической предпосылки мы будем рассматривать эти совокупности как случайные. Например, нас может интересовать сравнение двух методов обработки, т.е. двух разных действий, направленных к од-

ной цели: двух лекарств, двух рационов питания, двух методик обучения или профессиональной подготовки и т.д.

Данные. Для исследования нужны однородные объекты, разделенные на две группы. Взаимные влияния и взаимодействия объектов должны быть исключены. Для каждого объекта регистрируется некоторая его числовая характеристика. Возникающие при этом две группы чисел можно рассматривать как две независимые выборки.

Постановка задачи. Рассмотрим вопрос о том, какие задачи целесообразно рассматривать при сравнении двух выборок. Вспомним, что обычно две выборки получаются как характеристики двух обработок, т.е. как результаты применения различных условий эксперимента к двум группам однородных объектов. Опыт применения статистики показывает, что изменение условий эксперимента обычно сказывается прежде всего на изменении положения распределения измеряемой числовой характеристики на числовой прямой. Масштаб и форма распределения при малых изменениях условий эксперимента обычно остаются практически неизменными. При больших различиях в условиях эксперимента наряду с изменением положения распределения изменяется и его разброс (дисперсия). И совсем редко происходит изменение самой формы распределения. Поэтому при исследовании различий в двух выборках часто предполагают, что законы распределения двух анализируемых выборок отличаются только сдвигом, т.е. принадлежат *сдвиговому семейству распределений*.

Определение. *Распределение $G(x)$ принадлежит сдвиговому семейству распределений F , задаваемому распределением $F(x)$, если существует такая θ , что для любого x : $F(x) = G(x - \theta)$.*

Другими словами, если случайная величина ξ имеет распределение $F(x)$, то распределение $G(x)$ случайной величины η принадлежит сдвиговому семейству F тогда и только тогда, когда для некоторого неслучайного числа θ распределения случайных величин η и $\xi + \theta$ совпадают.

Для некоторых сдвиговых семейств (например, для семейства, порожденного нормальным распределением) построены весьма эффективные критерии для проверки гипотезы H против альтернатив сдвига $\theta \neq 0$ (см., например, гл. 5). Однако эти критерии предполагают, что F и G принадлежат определенному семейству, а поэтому могут давать неправильные результаты при невыполнении этого условия. Другой класс критериев — непараметрические критерии — не требует этого предположения. Такие критерии не зависят от распределений F и G (если эти распределения непрерывны) и эффективно работают при более широком классе альтернатив. В частности, с их помощью можно найти

различия в случайных величинах при альтернативах $F \leq G$ и $F \geq G$. Дадим определения этих понятий.

Определение. Мы говорим, что $F \geq G$, где F и G — функции распределения, если для любого числа x выполняется $F(x) \geq G(x)$. Мы говорим, что $F \leq G$, если для любого числа x выполняется $F(x) \leq G(x)$.

Смысл этого определения состоит в том, что при $F \geq G$ случайная величина X , имеющая закон распределения F , имеет тенденцию принимать меньшие значения, чем случайная величина Y с законом распределения G , т.е. для любого x выполняется $P(X < x) \geq P(Y < x)$.

Методы. Ниже мы расскажем, как проверить однородность двух выборок с помощью критерия Манна–Уитни или критерия Уилкоксона. Методы анализа двух выборок, имеющих нормальный закон распределения, будут рассмотрены отдельно в гл. 5.

3.5.1. Критерий Манна–Уитни

Область применения критерия Манна–Уитни — анализ двух независимых выборок. Размеры этих выборок могут различаться.

Назначение критерия — проверка гипотезы о статистической однородности двух выборок. Иногда эту гипотезу называют гипотезой об отсутствии эффекта обработки (имея в виду, что одна из выборок содержит характеристики объектов, подвергшихся некоему воздействию, а другая — характеристики контрольных объектов).

Данные. Рассматриваются две выборки x_1, \dots, x_m (выборка x) и y_1, \dots, y_n (выборка y) объемов m и n . Обозначим закон распределения первой выборки через F , а второй — через G .

Допущения. 1. Выборки x_1, \dots, x_m и y_1, \dots, y_n должны быть независимы.

2. Законы распределений F и G непрерывны. Отсюда следует, что с вероятностью 1 среди чисел x_1, \dots, x_m и y_1, \dots, y_n нет совпадающих.

Гипотеза. Утверждение об однородности выборок x_1, \dots, x_m и y_1, \dots, y_n , в введенных выше обозначениях можно записать в виде $H : F = G$.

Альтернативы. В качестве альтернатив к H могут выступать все возможности $F \neq G$. Однако критерий Манна–Уитни способен обнаруживать отнюдь не все возможные отступления от $H : F = G$. Этот критерий предназначен в первую очередь для проверки H против альтернативы $F \geq G$ (правосторонняя альтернатива, «перетекание» вероят-

ностей вправо) или альтернативы $F \leq G$ (левосторонняя альтернатива, т.е. уход вероятностей влево). Можно рассматривать и объединение обеих возможностей (двусторонняя альтернатива).

Метод. Критерий Манна–Уитни повторяет основные идеи критерия знаков и в определенном смысле является его продолжением. Он основан на попарном сравнении результатов из первой и второй выборок.

Условимся, что всякое событие $x_i < y_j$ обозначает «успех», а всякое событие $x_i > y_j$ — «неудачу». Смысл такой терминологии может быть связан с тем, что мы предполагаем, что вторая группа лучше первой, и рады подтверждению наших представлений. Изменяя i от 1 до m и j от 1 до n , получаем mn парных сравнений элементов выборок x и y . Обозначим число успехов в этих парных сравнениях через U . Ясно, что U может принимать любое целое значение от 0 до mn .

Определение. Введенная выше случайная величина U называется **статистикой Манна–Уитни**.

Вычислив значение $U_{\text{набл.}}$, мы можем приступить к проверке гипотезы H .

1. Зададим уровень значимости α или выберем метод, связанный с определением наименьшего уровня значимости статистики U , который описан ниже.

2. Для правосторонних альтернатив найдем по таблицам такое критическое значение $U_{\text{п.}}(\alpha, m, n)$, что

$$P\{U \geq U_{\text{п.}}(\alpha, m, n)\} = \alpha.$$

При этом критическая область для гипотезы H против правосторонних альтернатив будет иметь вид:

$$\{U \geq U_{\text{п.}}(\alpha, m, n)\}.$$

При проверке H против левосторонних альтернатив надо найти критическое значение $U_{\text{л.}}(\alpha, m, n)$, такое, что

$$P\{U \leq U_{\text{л.}}(\alpha, m, n)\} = \alpha.$$

Здесь критическая область примет вид:

$$\{U \leq U_{\text{л.}}(\alpha, m, n)\}.$$

В таблицах (см. [77], [83], [115]) обычно приводятся критические значения, соответствующие числам α из ряда 0.05, 0.025, 0.01, 0.005, 0.001. Ввиду дискретного характера распределения вероятностей между возможными значениями случайной величины U приведенные выше уравнения не всегда имеют точное решение, и в таблицах они приво-

дятся приближенно. Для вычисления по таблицам значений $U_{л.}(\alpha, m, n)$ можно воспользоваться соотношением

$$U_{л.}(\alpha, m, n) + U_{п.}(\alpha, m, n) = mn,$$

вытекающим из симметрии распределения статистики U относительно своего центра $mn/2$.

3. Отвергнем гипотезу H против правосторонних (левосторонних) альтернатив при попадании $U_{набл.}$ в соответствующую критическую область.

4. При проверке H против двусторонних альтернатив в качестве критического множества можно взять объединение

$$\{U \leq U_{л.}(\alpha, m, n)\} \cup \{U \geq U_{п.}(\alpha, m, n)\},$$

т.е. отвергнуть H , если происходит одно из двух ранее упомянутых критических событий. Ввиду уже отмеченной симметрии этому критерию можно дать вид

$$\left| U - \frac{mn}{2} \right| \geq \left| U_{п.}(\alpha, m, n) - \frac{mn}{2} \right|.$$

При таком выборе критического множества уровень значимости удваивается. Теперь он равен 2α (с теми же оговорками насчет дискретности распределения U , что были сделаны выше). Если мы желаем сохранить и здесь уровень значимости α , надо взять $U_{л.}(\frac{\alpha}{2}, m, n)$ и $U_{п.}(\frac{\alpha}{2}, m, n)$

Приближение для больших выборок. Смотри п. 3.5.2 и связь между статистикой Манна–Уитни и статистикой Уилкоксона, указанную там же в разделе «Обсуждение».

Обсуждение. Укажем некоторые свойства статистики U и соображения, приводящие к описанному выше методу проверки гипотезы.

Распределение вероятностей U при гипотезе H . Хотя статистика Манна–Уитни является суммой одинаково распределенных случайных величин, принимающих значения 0 и 1, она не имеет биномиального распределения, так как эти величины являются зависимыми (например, зависимы результаты сравнения x_1 с y_1 и x_1 с y_2). Поэтому распределение статистики U приходится рассчитывать, используя специальные таблицы или асимптотические приближения.

Однако расчет распределения статистики U значительно упрощается тем, что при выполнении гипотезы H это распределение не зависит от закона распределения выборок (если эти распределения непрерывны). Распределение U при гипотезе H зависит только от объемов выборок — m и n . В справочниках [77], [83], [115] приводятся таблицы, по которым можно найти вероятность $P(U \geq k)$ для различных k при небольших значениях m и n .

Заметим, что при справедливости гипотезы H (т.е. при совпадении законов распределения F и G) выполняется $P(x_i < y_j) = P(x_i > y_j) = 0.5$. Поэтому

при H количества успехов и неудач должны быть приблизительно равны, т.е. U не должно значительно отклоняться от $mn/2$.

Распределение статистики U при нарушении гипотезы. Рассмотрим, как может вести себя U при различных альтернативах. В отличие от поведения U при гипотезе, здесь распределение U зависит от F и G , поэтому мы можем описать его свойства лишь для отдельных типов альтернатив. Проще всего указать свойства U для односторонних альтернатив: правосторонних (если $F \geq G$) или левосторонних (если $F \leq G$). Легко видеть, что для правосторонних альтернатив выполняется $P(x_i < y_j) > 0.5$, поэтому значение U , т.е. общее число успехов $x_i < y_j$, скорее всего, должно превосходить $mn/2$ и тем значительнее, чем больше $P(x_i < y_j)$. Для левосторонних альтернатив ($F \leq G$) соотношение обратное: $P(x_i < y_i) < 0.5$, поэтому общее число успехов, как правило, должно быть меньше $mn/2$, и тем меньше, чем меньше $P(x_i < y_j)$.

Итак, для односторонних альтернатив статистика Манна–Уитни имеет ясные свойства, поэтому на ее основе можно построить критерий для проверки гипотезы H против таких альтернатив.

Метод проверки гипотезы. В связи с таким поведением статистики U для проверки гипотезы H против указанных выше возможных альтернатив разумно предложить следующее правило: отвергнуть H , если наблюдаемое U (в дальнейшем $U_{\text{набл.}}$) значительно отклоняется от $mn/2$ — значения, ожидаемого от U при гипотезе H (от математического ожидания U при гипотезе H). Чем больше отклоняется от $mn/2$ наблюдаемое значение U , т.е. $U_{\text{набл.}}$, тем сильнее мы сомневаемся в том, что H верна. Разумеется, U может значительно отклоняться от $M(U | H)$ и за счет действия случая, когда H выполняется, но чем больше отклонение, тем оно при H менее вероятно и тем труднее объяснить это отклонение случайностью. Скорее всего, если отклонение велико, оно вызвано не случаем, а закономерной причиной — тем, что распределения G и F не совпадают.

Силу таких доводов против $H : F = G$ в пользу, например, правосторонней альтернативы $F \geq G$ можно выразить количественно, вычислив $P(U \geq U_{\text{набл.}} | H)$. Это вероятность того, что при независимом повторении эксперимента мы получим такое же или еще более сильное свидетельство против H (в пользу правосторонней альтернативы), как уже имеющееся $U_{\text{набл.}}$. Если $U_{\text{набл.}}$ велико, то вышеназванная вероятность мала, и наоборот. Если эта вероятность столь мала, что подобное событие кажется практически невозможным при H , гипотезу H следует отвергнуть (по имеющемуся наблюдению $U_{\text{набл.}}$) в пользу правосторонней альтернативы.

Рекомендация изменяется очевидным образом, если с H конкурируют левосторонние альтернативы. Наконец, в случае двусторонних альтернатив надо вычислить вероятность

$$P \left\{ \left| U - \frac{mn}{2} \right| \geq \left| U_{\text{набл.}} - \frac{mn}{2} \right| \right\}$$

и в зависимости от того, насколько она мала, отвергнуть гипотезу.

Описанный способ действий имеет определенные преимущества перед стандартной процедурой проверки статистических гипотез, как она описана в п. 3.2. Главное то, что здесь не приходится заранее выбирать уровень значимости, что всегда выглядит несколько произвольно. Описанный подход автоматически доставляет нам тот наименьший уровень значимости, на котором (по имею-

щимся наблюдениям) можно отвергнуть гипотезу H в пользу соответствующей альтернативы. В данном случае есть и еще одно дополнительное преимущество: как мы уже отмечали выше, из-за дискретности распределения U традиционные номинальные уровни значимости типа 0.05, 0.025, 0.001 и т.д. могут быть достигнуты лишь приближенно. В обсуждаемом методе проверки приближение исчезает: мы получаем точное значение вероятности, если обращаемся к достаточно подробным таблицам распределений U .

Совпадения. Выше отмечалось, что из условия непрерывности распределений F и G следует отсутствие повторов в выборках. На практике же такие повторы встречаются часто. Во многих случаях причиной этого является не нарушение исходных предположений, а ограниченная точность при записи наблюдений.

Допустим, что некоторые элементы выборки икс совпали с некоторыми элементами из выборки игрек, т.е. $x_i = y_j$ для некоторых $i, j (i = 1, \dots, m; j = 1, \dots, n)$. В этом случае статистику U вычисляют так: к числу успехов прибавляют уменьшенное вдвое число событий вида $(x_i = y_j)$. Таким образом, каждое совпадение икса и игрека считается за половину успеха. Далее с так подсчитанным числом успехов поступают так, как описано выше.

При наличии совпадающих наблюдений получаемые при использовании описанных критериев выводы имеют приближенный характер, и эти приближения тем хуже (и выводы тем сомнительнее), чем больше среди наблюдений совпадающих, т.е. чем сильнее отступление от исходных математических предположений. В тех случаях, когда результаты (X и Y) могут принимать лишь ограниченное число значений (что влечет за собой большое количество совпадений), этот метод применять не следует. К сожалению, четкого разграничения в этом вопросе сделать нельзя.

3.5.2. Критерий Уилкоксона

Область применения. Критерий Уилкоксона применяется в той же ситуации, что и критерий Манна–Уитни. В отличие от этого критерия и критерия знаков, он имеет дело не со знаками некоторых случайных величин, а с их рангами. Исторически критерий Уилкоксона был одним из первых критериев, основанных на рангах (см. п. 1.8).

Рассмотрим ранги элементов объединения двух выборок x_1, \dots, x_m и y_1, \dots, y_n . Для получения рангов совокупность всех наблюдений следует упорядочить в порядке возрастания. (Напомним, что если функции распределения F и G выборок x и y непрерывны, то в их совокупности нет совпадающих значений и, следовательно, результат упорядочивания однозначен. Как поступать в противном случае, будет сказано ниже, в разделе «Совпадения».)

Пусть, например, первая выборка состоит из чисел 6, 17 и 14, вторая — из чисел 5 и 12. Тогда ранги величин первой группы есть 2, 5, 4, второй — 1, 3.

Нетрудно понять, что последовательность рангов совокупности объема $m + n$ является некоторой перестановкой чисел $1, \dots, m + n$. Верно и обратное: любая перестановка чисел $1, \dots, m + n$ может оказаться ранговой последовательностью. Так что множество возможных ранговых последовательностей — это совокупность перестановок чисел $1, 2, \dots, m + n$. Их общее число равно $(m + n)!$.

Зная распределения случайных величин X_1, \dots, X_m и Y_1, \dots, Y_n , мы можем (по крайней мере, теоретически) вычислить вероятность того, что результат их ранжирования будет заданной перестановкой. Поэтому каждое распределение случайных величин X_1, \dots, X_m и Y_1, \dots, Y_n порождает некоторое распределение вероятностей на указанном множестве перестановок. Ясно, что если исходные данные однородны (X_1, \dots, X_m и Y_1, \dots, Y_n в совокупности являются независимыми и одинаково распределенными случайными величинами), то в качестве последовательности рангов с равными шансами может появиться любая перестановка чисел от 1 до $m + n$. Число таких перестановок равно $(m + n)!$, поэтому вероятность каждой равна $1/(m + n)!$. Заметим, что этот результат никак не зависит от распределения самих наблюдений.

Посмотрим, как изменяется распределение вероятностей среди ранговых последовательностей (т.е. среди перестановок) при отступлениях от однородности выборок. В качестве нарушений однородности мы будем рассматривать те же ситуации, что и при обсуждении критерия Манна–Уитни в предыдущем пункте: левосторонние альтернативы $F \leq G$ и правосторонние альтернативы $F \geq G$. Для правосторонних альтернатив $P(x_i < y_j) > 0.5$, т.е. наблюдения из второй группы имеют тенденцию превосходить наблюдения из первой. Поэтому ранг наблюдений из второй группы чаще будет принимать значения из правой части ряда чисел $1, 2, \dots, m + n$. Если же отступление таково, что $P(x_i < y_j) < 0.5$, то ранги игроков чаще будут принимать значения из левой части ряда чисел $1, 2, \dots, m + n$. Переход от рангов игроков к их сумме позволяет резче отметить эти закономерности.

Таким образом, ранги в какой-то мере способны характеризовать, например, положение одной выборки по отношению к другой, и в то же время они не зависят от неизвестных нам распределений выборок x и y . Это обстоятельство и легло в основу ранговых методов, широко применяемых в настоящее время в различных задачах.

Вернемся к непосредственному обсуждению критерия Уилкоксона.

Назначение. Критерий Уилкоксона используется для проверки гипотезы об однородности двух выборок. Нередко одна из выборок пред-

ставляет характеристики объектов, подвергшихся перед тем какому-то воздействию (обработке). В этом случае гипотезу однородности можно назвать *гипотезой об отсутствии эффекта обработки*.

Данные. Рассматриваются две выборки x_1, \dots, x_m и y_1, \dots, y_n , объемов m и n . Обозначим закон распределения первой выборки через F , а второй — через G .

Допущения. 1. Выборки x_1, \dots, x_m и y_1, \dots, y_n независимы между собой.

2. Законы распределения выборок F и G непрерывны.

Гипотеза. Во введенных выше обозначениях гипотезу об однородности выборок можно записать в виде $H : F = G$.

Метод. 1. Рассмотрим ранги игроков в общей совокупности выборок x и y . Обозначим их через S_1, \dots, S_n .

2. Вычислим величину

$$W_{\text{набл.}} = S_1 + \dots + S_n,$$

называемую **статистикой Уилкоксона**. Таблицы распределения статистики W (при гипотезе однородности) можно найти в [19], [77], [115] и др.

3. Зададим уровень значимости α или выберем метод, связанный с определением наименьшего уровня значимости, приведенный ниже.

4. Для проверки H на уровне значимости α против правосторонних альтернатив $P(x_i < y_j) > 0.5$ найдем по таблице верхнее критическое значение $W(\alpha, m, n)$, т.е. такое значение, для которого

$$P(W \geq W(\alpha, m, n)) = \alpha.$$

Гипотезу следует отвергнуть против правосторонней альтернативы при уровне значимости α , если $W_{\text{набл.}} \geq W(\alpha, m, n)$.

5. Для проверки H на уровне значимости α против левосторонних альтернатив $P(x_i < y_j) < 0.5$, необходимо вычислить нижнее критическое значение статистики W . В силу симметричности распределения W нижнее критическое значение есть $n(m+n+1) - W(\alpha, m, n)$. Гипотеза H должна быть отвергнута на уровне значимости α против левосторонней альтернативы, если $W_{\text{набл.}} \leq n(m+n+1) - W(\alpha, m, n)$.

6. Гипотеза H отвергается на уровне 2α против двусторонней альтернативы $P(x_i < y_j) \neq 0.5$, если

$$W_{\text{набл.}} \geq W(\alpha, m, n) \text{ или } W_{\text{набл.}} \leq n(m+n+1) - W(\alpha, m, n).$$

Напомним, что альтернативы должны выбираться из содержательных соображений, связанных с условиями получения экспериментальных данных.

7. Более гибкое правило проверки H связано с вычислением наименьшего уровня значимости, на котором гипотеза H может быть отвергнута. Для разных альтернатив речь идет о вычислении вероятностей:

$$P(W \geq W_{\text{набл.}}),$$

$$P(W \leq W_{\text{набл.}}),$$

$$P(|W - n(m+n+1)/2| \geq |W_{\text{набл.}} - n(m+n+1)/2|).$$

Гипотеза отвергается, если соответствующая вероятность оказывается малой.

Приближение для больших выборок. На практике часто приходится сталкиваться с ситуацией, когда объемы выборок m и n выходят за пределы, приведенные в таблицах. В этом случае используют аппроксимацию распределения W предельным распределением статистики W при $m \rightarrow \infty$ и $n \rightarrow \infty$. Перейдем от величины W к $W^* = (W - MW)/\sqrt{DW}$. Ниже будет показано, что $MW = n(m+n+1)/2$. Также можно показать, что $DW = mn(m+n+1)/12$. Доказано, что в условиях H , при допущениях 1 и 2 и при больших m, n случайная величина W^* распределена приблизительно по нормальному закону с параметрами $(0, 1)$.

Обозначим через z_α верхнее критическое значение стандартного нормального распределения. Его можно найти с помощью таблицы квантилей нормального распределения для любого $0 < \alpha < 0.5$. Благодаря симметрии распределения нижнее критическое значение равно $-z_\alpha$. Правило проверки H перефразируем так:

- отвергнуть H на уровне α против альтернативы $P(x_i < y_j) > 0.5$, если $W_{\text{набл.}}^* \geq z_\alpha$;
- отвергнуть H на уровне α против альтернативы $P(x_i < y_j) < 0.5$, если $W_{\text{набл.}}^* \leq -z_\alpha$;
- отвергнуть H на уровне 2α против альтернативы $P(x_i < y_j) \neq 0.5$, если $|W_{\text{набл.}}^*| \geq z_\alpha$.

Правило, связанное с вычислением наименьшего уровня значимости, при использовании нормального приближения выглядит так: отвергнуть H (против соответствующих альтернатив), если оказывается малой вероятность $1 - \Phi(W_{\text{набл.}}^*)$ для альтернативы $P(x_i < y_j) > 0.5$, $\Phi(W_{\text{набл.}}^*)$ для альтернативы $P(x_i < y_j) < 0.5$, и $2\Phi(|W_{\text{набл.}}^*|) - 1$ для альтернативы $P(x_i < y_j) \neq 0.5$, где $\Phi(u)$ — функция нормального распределения (функция Лапласа), равная $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-x^2/2} dx$.

Функция нормального распределения и её обратная, которая называется функцией квантилей стандартного нормального распределения, подробно табулированы. Упомянутое ранее верхнее критическое значе-

ние z_α с помощью функции Φ можно определить как решение уравнения $1 - \Phi(z_\alpha) = \alpha$.

Замечание. Указанное выше нормальное приближение для вычисления критических значений статистики W хорошо действует даже для небольших значений m и n , если только α не слишком мало. (Так, для $m = n = 8$ приближенные квантили практически не отличаются от точных.)

Обсуждение. Рассмотрим подробнее свойства статистики W и соображения, положенные в основу критерия Уилкоксона.

Область определения. Случайная величина W может принимать все целые значения от минимального значения $\frac{n(n+1)}{2}$ до максимального $mn + \frac{n(n+1)}{2}$. Минимальное значение W мы получаем, когда рангами игроков служат (в той или иной последовательности) числа $1, 2, \dots, n$. Максимальное значение W возникает, когда этими рангами служат $m+1, m+2, \dots, m+n$.

Заметим, что W не изменится, если произвольно поменять порядок следования чисел, служащих рангами игроков (как не изменится и при перенумерации самих игроков). Чтобы упростить обсуждение, можно поэтому говорить далее о рангах игроков, упорядоченных по возрастанию. Пусть S_1, S_2, \dots, S_n обозначают именно упорядоченные ранги, так что $S_1 < S_2 < \dots < S_n$.

Распределение вероятностей. Статистика Уилкоксона была определена нами как сумма (упорядоченного) набора рангов игроков S_1, \dots, S_n . Вероятность каждого такого упорядоченного набора при выдвинутой гипотезе H — одна и та же и равна $(C_{m+n}^n)^{-1} = \frac{m!n!}{(m+n)!}$. Таким образом, при гипотезе H распределение W не зависит от закона распределения выборок x и y , так как от них не зависит распределение упорядоченной последовательности рангов. Для каждой пары (m, n) распределение W можно рассчитать. Покажем на примере, как это делается.

Пусть $m = 3$ и $n = 2$. Вычислим число всех возможных пар рангов игроков. Оно равно $C_{3+2}^2 = 10$. Следовательно, вероятность каждого упорядоченного набора рангов равна 0.1. Выпишем все возможные наборы рангов S_1, S_2 и соответствующую им сумму:

S_1, S_2	1,2	1,3	1,4	1,5	2,3	2,4	2,5	3,4	3,5	4,5
W	3	4	5	6	5	6	7	7	8	9

Таким образом, получаем следующее распределение W :

W	3	4	5	6	7	8	9
$P(W)$	0.1	0.1	0.2	0.2	0.2	0.1	0.1

Отметим, что распределение W симметрично относительно точки $n(m + n + 1)/2$ — середины отрезка $[n(n + 1)/2, nm + n(n + 1)/2]$. Из этого свойства легко вывести, что $M(W | H) = n(m + n + 1)/2$.

Рассмотрим случайную величину $W - n(m + n + 1)/2$. Согласно симметрии закона распределения относительно точки $n(m + n + 1)/2$, вероятность p_k , что эта величина примет некоторое значение k , равна вероятности p_{-k} ,

что она примет значение $-k$. Согласно определению математического ожидания, $M(W - n(m + n + 1)/2 | H) = \sum_{k=-nm/2}^{nm/2} k p_k = 0$. Учитывая, что математическое ожидание разности равно разности математических ожиданий, а математическое ожидание константы равно самой константе, получаем: $MW = n(m + n + 1)/2$.

Распределение статистики W при нарушении гипотезы. Чтобы оправдать сделанный выше выбор критических событий (критериев) для проверки H против рассмотренных альтернатив, надо изучить распределение статистик U и W при этих альтернативах. Когда F и G не одинаковы, распределения U и W уже не свободны от их влияния. Поэтому точно вычислить и указать распределения U и W можно (в принципе) только для каждой конкретной пары F и G . Тем не менее характер изменения распределений статистик U и W при переходе от гипотезы к альтернативам — не всем, но некоторым — установить можно. Это легко сделать для односторонних альтернатив. Например, когда $P(x_i < y_i) > 0.5$ (правосторонняя альтернатива), распределение вероятностей W «перетекает» от середины к правому концу того множества значений, которое может принимать W . Для левосторонних альтернатив аналогичное «перетекание» вероятности происходит влево — тем сильнее, чем больше $P(x_i < y_i)$ отличается от 0.5.

На рис. 3.1 мы попытались наглядно представить это положение, условно представляя распределение статистики W при гипотезе и при альтернативах с помощью плотностей — хотя искомые распределения дискретны и плотностей не имеют. Но так получается выразительнее. (При желании можно считать, что нарисованные непрерывные кривые изображают что-то вроде огибающих графиков дискретных вероятностей.)

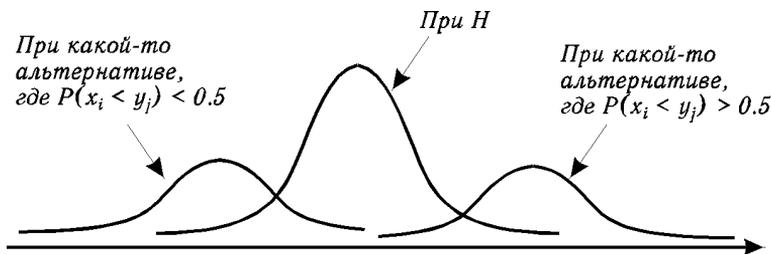


Рис. 3.1. Схематическое изображение распределений W

Из рис. 3.1 ясно, что гипотеза H должна отвергаться при слишком больших или при слишком малых значениях W в зависимости от того, какие альтернативы мы рассматриваем. При том выборе критериев, который был описан выше, их мощность возрастает при удалении $P(x_i < y_j)$ от 0.5. Это правило и лежит в основе описанного выше метода.

Связь со статистикой Манна–Уитни. Нетрудно проверить, что для всех m, n : $W = U + n(n + 1)/2$. Это соотношение показывает эквивалентность статистик U и W . Поэтому их применения приводят к одинаковым результатам.

Совпадения. Мы описали критерий Уилкоксона для проверки гипотезы об однородности двух выборок в условиях, когда функции распределений данных непрерывны и, тем самым, в выборках не должно быть совпадающих наблю-

дений. Однако на практике совпадающие наблюдения — не редкость. Чаще всего это происходит не потому, что нарушается условие непрерывности, а из-за ограниченной точности записи результатов измерений (например, рост человека обычно измеряется с точностью до 1 см). Применение критерия Уилкоксона к таким данным приводит к приближенным выводам, точность которых тем ниже, чем больше совпадающих значений.

Когда среди наблюдений встречаются одинаковые, им приписываются *средние ранги*. По определению, средний ранг числа z_i в совокупности чисел z_1, z_2, \dots, z_n есть среднее арифметическое из тех рангов, которые были бы назначены z_i и всем остальным значениям, совпадающим с z_i , если бы они оказались различными. После такого назначения рангов применяются описанные ранее процедуры.

Упомянутые группы одинаковых наблюдений называют *связками*. Количество элементов в связке называют ее размером. Наличие связей влияет на асимптотические распределения статистики Уилкоксона. Так, при использовании нормальной аппроксимации следует в формуле для вычисления W^* заменить DW на

$$\frac{mn}{12} \left[(m+n+1) - \frac{\sum_{k=1}^g t_k(t_k^2-1)}{(m+n)(m+n-1)} \right],$$

где t_1, t_2, \dots, t_g — размеры наблюдаемых связок среди игроков, g — общее число связок среди игроков. Наблюдение, не совпавшее с каким-либо другим наблюдением, рассматривается как связка размера 1 и в формуле, заменяющей DW , не учитывается.

При больших по размеру связках и (или) большом их числе применение критерия Уилкоксона сомнительно.

3.6. Парные наблюдения

Рассмотренное в предыдущем параграфе сравнение двух совокупностей наблюдений (двух выборок) часто проводится для обнаружения результата какого-либо воздействия (выявления эффекта обработки) либо, напротив, для подтверждения его отсутствия. Чем более однородными окажутся выбранные для эксперимента объекты (для контроля и воздействия), чем меньше их случайные различия, тем точнее (и по меньшему числу наблюдений) можно будет дать ответ на вопрос. Кстати, формирование однородной группы экспериментальных объектов составляет важную и не всегда простую задачу.

Ясно, что различие между объектами, выбранными для воздействия и для контроля (или для двух разных воздействий, если интерес представляет их сопоставление), будет наименьшим, если в обоих качествах выступает один и тот же объект. Если это возможно, то далее обычным порядком мы составляем группу экспериментальных объектов (по-прежнему стремясь к тому, чтобы они были однородны — значение этого выяснится в п. 3.6.2). Далее для каждого объекта мы измеряем два

значения интересующей нас характеристики (например, до воздействия и после или при двух разных воздействиях). Так возникают пары наблюдений и парные данные. Но, конечно, парные данные могут возникать и иначе (скажем, при наблюдениях над близнецами, которые во многих отношениях считаются идентичными).

3.6.1. Критерий знаков для анализа парных повторных наблюдений

Назначение. Критерий знаков используется для проверки гипотезы об однородности наблюдений внутри каждой пары (иногда говорят — для проверки гипотезы об отсутствии эффекта обработки).

Данные. Рассмотрим совокупность случайных пар $(x_1, y_1), \dots, (x_n, y_n)$ объема n . Введем величины $z_i = y_i - x_i, i = 1, \dots, n$.

Допущения. 1. Все z_i предполагаются взаимно независимыми. Заметим, что мы не требуем независимости между элементами x_i и y_i с одинаковым номером i . Это весьма важно на практике, когда наблюдения делаются для одного объекта и тем самым могут быть зависимы.

2. Все z_i имеют равные нулю медианы, т.е. $P(z_i < 0) = P(z_i > 0) = 1/2$. Подчеркнем, что законы распределения разных z_i могут не совпадать.

Гипотеза. Утверждение об отсутствии эффекта обработки для повторных парных наблюдений $(x_1, y_1), \dots, (x_n, y_n)$ можно записать в виде

$$H : P(x_i < y_i) = P(x_i > y_i) = 0.5 \quad \text{для всех } i = 1, \dots, n.$$

Метод. 1. Перейдем от повторных парных наблюдений $(x_1, y_1), \dots, (x_n, y_n)$ к величинам $z_i, i = 1, \dots, n$, введенным выше.

2. К совокупности $z_i, i = 1, \dots, n$ применим критерий знаков для проверки гипотезы о равенстве нулю медиан распределений величин $z_i, i = 1, \dots, n$ (см. п. 3.4.2).

Приближение для больших совокупностей. Следует воспользоваться нормальной аппроксимацией биномиального распределения. См. п. 2 раздела «Связь с другими распределениями» параграфа 2.1 гл. 2.

Связанные данные. Если среди значений z_i есть нулевые, то их следует отбросить и соответственно уменьшить n до числа ненулевых значений z_i .

Оценка эффекта обработки. Нередко для z_i рассматривают модель $z_i = \theta + e_i$, $i = 1, \dots, n$, где e_i — ненаблюдаемые случайные величины, θ — некоторая константа, характеризующая положение одного распределения относительно другого (скажем, до воздействия и после). Эту константу часто именуют эффектом обработки. Принятые выше допущения 1 и 2 переносятся на величины e_1, \dots, e_n . Гипотеза однородности формулируется в виде гипотезы о нулевом эффекте обработки $H : \theta = 0$.

Введенные величины θ и представления $z_i = \theta + e_i$ оказываются полезными, если в ходе проверки гипотезы выясняется, что $\theta \neq 0$ и что поэтому надо оценить количественно то различие, которое привносит обработка (воздействие).

Пример. Покажем, как использовать критерий знаков для анализа данных о времени реакции на звук и на свет. В этом примере рассматривается группа испытуемых, а целью исследования служит проверка гипотезы о равенстве времени реакций на звук и на свет. Порядок организации эксперимента позволяет предположить, что полученные данные на одном испытуемом независимы от аналогичных данных для остальных.

Осуществив переход от пар $(x_1, y_1), \dots, (x_n, y_n)$ к величинам z_i , $i = 1, \dots, n$ и запишем последние в виде: $z_i = \theta + e_i$, $i = 1, \dots, n$.

Выполняются ли для сформулированной задачи допущения, используемые в критерии знаков? Независимость e_i обеспечивается условиями организации эксперимента. Априорно предполагаемая непрерывность распределений рассматриваемых выборок обеспечивает непрерывность распределения e_i . В случае совпадения распределений времени реакции на звук и на свет справедливо следующее соотношение $P(x_i - y_i > 0) = P(x_i - y_i < 0) = 1/2$. Следовательно, $P(z_i > 0) = P(z_i < 0) = 1/2$, т.е. медиана распределения z_i равна нулю. Таким образом, предположение $\theta = 0$ обеспечивает выполнение допущения 2.

Одной из разумных альтернатив нулевой гипотезе в данном случае является предположение о том, что $\theta < 0$. Далее мы будем использовать критерий знаков против этой односторонней альтернативы.

В табл. 3.5 приведены соответствующие расчеты для данного примера.

Обозначим число положительных значений z_i через $S_{\text{набл}}$. Из табл. 3.5 видно, что $S_{\text{набл}}$ равно трем, а среди z_i есть одно значение, равное 0. В таких случаях необходимо уменьшить число наблюдений z_i на число значений z_i , равных 0, т.е. перейти от $n = 17$ к $n = 16$.

Вычислим вероятность $P(S \leq S_{\text{набл}} | H)$. Для этого воспользуемся таблицами биномиального распределения при $p = 1/2$, $n = 16$ (см. [19], [77]). Учитывая, что в силу симметрии при $p = 1/2$ $P(S \leq S_{\text{набл}} | H) = P(S \geq n - S_{\text{набл}} | H)$, получаем:

$$P(S \leq S_{\text{набл}} | H) = P(S \geq 16 - 3 | H) = P(S \geq 13 | H) = 0.0106.$$

То есть минимальный уровень значимости, на котором можно отвергнуть гипотезу о том, что $\theta = 0$ против односторонних альтернатив, равен 0.0106. Учитывая малость этого числа, заключаем, что гипотезу следует отвергнуть в пользу альтернативы $\theta < 0$.

i	x_i	y_i	z_i	$S(x_i)$
1	223	181	-42	-
2	104	194	90	+
3	209	173	-36	-
4	183	153	-30	-
5	180	168	-12	-
6	168	176	8	+
7	215	163	-52	-
8	172	152	-20	-
9	200	155	-45	-
10	191	156	-35	-
11	197	178	-19	-
12	183	160	-23	-
13	174	164	-10	-
14	176	169	-7	-
15	155	155	0	0
16	115	122	+7	+
17	163	144	-19	-

Обсуждение. Одно из главных достоинств критерия знаков — его простота. Другой важной особенностью этого критерия являются скромные требования к первоначальному статистическому материалу. Эти требования описываются с помощью модели парных наблюдений.

3.6.2. Анализ повторных парных наблюдений с помощью знаковых рангов (критерий знаковых ранговых сумм Уилкоксона)

Если можно дополнительно предположить, что случайные величины z_1, \dots, z_n из предыдущего пункта непрерывны и одинаково распределены, то для проверки гипотезы однородности можно применить более мощный критерий, основанный на статистике T знаковых ранговых сумм Уилкоксона.

Метод. 1. Вычислим абсолютные разности $|z_1|, \dots, |z_n|$. Пусть R_i обозначает ранг z_i в совместном упорядочении $|z_1|, \dots, |z_n|$ от меньшего к большему.

2. Определим переменные ψ_i , $i = 1, \dots, n$, где

$$\psi_i = \begin{cases} 1, & \text{если } z_i > 0; \\ 0, & \text{если } z_i < 0. \end{cases}$$

3. Вычислим наблюдаемое значение $T = \sum_{i=1}^n \psi_i R_i$, далее мы будем называть его $T_{\text{набл.}}$.

4. Для одностороннего критерия для проверки $H : P(z_i < 0) = P(z_i > 0)$ против правосторонней альтернативы $P(z_i < 0) < P(z_i > 0)$ на уровне значимости α :

- отклонить H , если $T_{\text{набл.}} \geq t(\alpha, n)$;
- принять H , если $T_{\text{набл.}} < t(\alpha, n)$,

где критическое значение $t(\alpha, n)$ удовлетворяет уравнению $P(T \geq t(\alpha, n) | H) = \alpha$. Таблицу критических значений можно найти в [115].

Для одностороннего критерия для проверки той же гипотезы против левосторонней альтернативы $P(z_i < 0) > P(z_i > 0)$ на уровне значимости α :

- отклонить H , если $T_{\text{набл.}} \leq \frac{n(n+1)}{2} - t(\alpha, n)$;
- принять H , если $T_{\text{набл.}} > \frac{n(n+1)}{2} - t(\alpha, n)$.

Для двустороннего критерия для проверки той же гипотезы H против двусторонних альтернатив $P(z_i < 0) \neq P(z_i > 0)$ на уровне значимости 2α :

- отклонить H , если $T_{\text{набл.}} \geq t(\alpha, n)$ или $T_{\text{набл.}} \leq \frac{n(n+1)}{2} - t(\alpha, n)$;
- принять H , если $\frac{n(n+1)}{2} - t(\alpha, n) < T_{\text{набл.}} < t(\alpha, n)$.

Замечание. Поскольку распределение статистики T дискретно, уравнение, определяющее $t(\alpha, n)$: $(P(T \geq t(\alpha, n)) = \alpha)$, имеет точное решение не для всех значений α (при фиксированном n). Поэтому либо в качестве $t(\alpha, n)$ придется взять приближенное решение, либо изменить α так, чтобы уравнение можно было решить точно.

Приближение для большой выборки. При выполнении гипотезы H статистика

$$T^* = \frac{T - MT}{\sqrt{DT}} = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

имеет асимптотическое (при $n \rightarrow \infty$) распределение $N(0, 1)$. Приведем приближение нормальной теории для проверки H , для определенности, против правосторонней альтернативы: H отклоняется, если $T_{\text{набл.}} \geq z_\alpha$, в противном случае H принимается. Здесь z_α — квантиль уровня $(1 - \alpha)$ стандартного нормального распределения $N(0, 1)$. Остальные правила трансформируются аналогично.

Совпадения. Если среди значений z_i есть нулевые, то их следует отбросить, соответственно уменьшив n до количества ненулевых значений z_i . Если среди ненулевых значений $|z_i|$ есть равные, то для вычисления T надо использовать средние ранги для величин $|z_1|, \dots, |z_n|$ и

далее использовать те же методы, что и без совпадений. Для приближения для больших выборок рекомендуется в формуле для вычисления T^* значение DT заменить на

$$\frac{1}{24} \left[n(n+1)(2n+1) - \frac{1}{2} \sum_{j=1}^g t_j(t_j-1)(t_j+1) \right],$$

где g — число связей, t_1, \dots, t_g — их размеры. Определение связей см. в разделе 3.5.2 при обсуждении статистики W^* .

3.7. Проверка статистических гипотез в пакете SPSS

Ниже будет рассмотрено, как в пакете SPSS реализуются методы проверки гипотез в схеме испытаний Бернулли, как можно проверять гипотезы о равенстве медианы выборки заданному значению, а также применять критерии знаков и знаковых рангов Уилкоксона для парных сравнений. Другие задачи проверки гипотез рассматриваются в последующих главах.

Схема испытаний Бернулли. В пакете SPSS есть несколько способов осуществить проверку нулевой гипотезы о значении вероятности успеха p в схеме испытаний Бернулли. В примере 3.1к будут разобраны два из них. Первый удобен, когда число испытаний невелико или нам заранее известно значение полученного числа «успехов» $S_{\text{набл.}}$. Второй следует применять, если число наблюдений велико и данные представляют последовательность результатов испытаний в схеме Бернулли. Подобный вид записи характерен для социологических или маркетинговых опросов, когда для каждого респондента фиксируется его принадлежность к одной из двух групп.

Пример 3.1к. Используя данные тройного теста, найдем минимальный уровень значимости критерия, основанного на значении числа «успехов» в схеме испытаний Бернулли, для проверки гипотезы о значении вероятности успеха против односторонних альтернатив.

Первый способ. Он заключается в непосредственном вычислении выражения $1 - P(X \leq S_{\text{набл.}} - 1)$. Для этого используется функция биномиального распределения с заданными параметрами (числом испытаний n и вероятностью успеха в случае нулевой гипотезы p)

Подготовка данных. В редакторе пакета создать переменную $s1$ и внести в нее значение $S_{\text{набл.}} - 1 = 6$.

Выбор процедуры. В меню Transform (преобразования) панели управления редактора пакета вызвать процедуру Compute. Работа этой процедуры подробно разобрана в примере 2.1к (см. п. 2.7). В окне ввода данных и параметров процедуры (см. рис. 2.10) в поле Target Variable ввести переменную sign_1, а в поле Numeric Expression — выражение: $1 - \text{CDF.BINOM}(s1,10,0.333)$, где $\text{CDF.BINOM}(x,n,p)$ — функция биномиального распределения вероятностей в точке x , n — число испытаний, p — вероятность успеха.

Результаты. Закончив ввод выражения, нажать [OK]. В редакторе данных пакета появится заданная выше переменная sign_1 со значением искомого уровня значимости 0.0195 (пакет округлит это значение в зависимости от заданного числа цифр после запятой в десятичной записи числа).

Второй способ.

Подготовка данных. Для этого способа данные должны представлять переменную пакета, например result, содержащую последовательность из 0 и 1 результатов испытаний в схеме Бернулли, где 1 означает «успех», а 0 — «неудачу», как это показано на рис. 3.2.

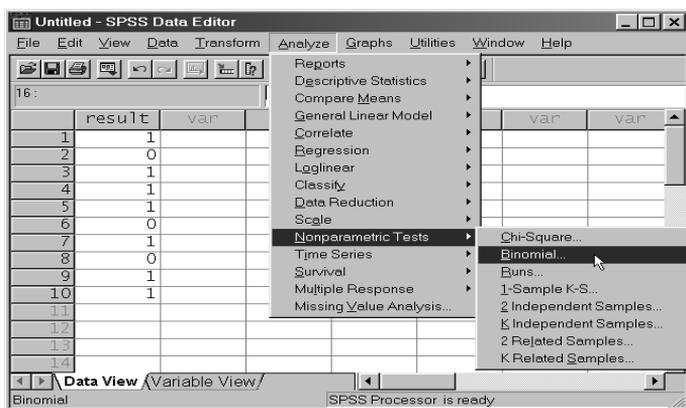


Рис. 3.2. Пакет SPSS. Редактор данных с исходными данными и меню выбора процедуры «Binomial»

Выбор процедуры. В меню Analyze редактора данных выбрать блок методов Nonparametric Tests (непараметрические тесты), а в нем процедуру Binomial, как это показано на рис. 3.2.

Заполнение полей ввода данных. Окно ввода данных этой процедуры приведено на рис. 3.3. В нем следует перенести переменную result из левой части окна в поле Test Variable List (список переменных для

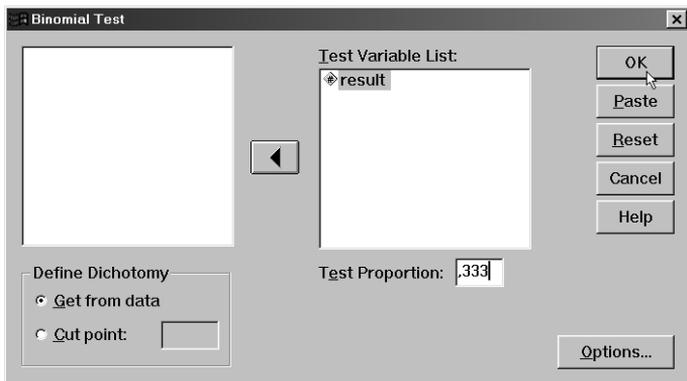


Рис. 3.3. Пакет SPSS. Окно ввода данных и параметров процедуры «Binomial»

теста). В поле **Test Proportion** (вероятность успеха) укажите вероятность успеха для нулевой гипотезы. Кнопка **Option** позволяет дополнительно задать выдачу описательной статистики выборки. Блок **Define Dichotomy** регулирует задание понятий «успех» и «неудача» в выборке.

Результаты. Таблица результатов работы процедуры представлена на рис. 3.4. Она включает число «успехов» и «неудач», их долю в общем объеме выборки (поле **Observed Prop**), значение вероятности «успеха» (поле **Test Prop**) и минимальный уровень значимости против односторонних альтернатив (поле **Exact Sig. (1-tailed)**). Это значение округлено до трех значащих цифр после запятой.

Binomial Test

	Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
RESULT	Group 1	1	,700	,333	,020
	Group 2	0	,300		
	Total	10	1,000		

Рис. 3.4. Пакет SPSS. Таблица результатов работы процедуры «Binomial»

Два следующих примера посвящены проверке гипотез с помощью непараметрических критериев знаков и знаковых рангов. В случае одной выборки критерий знаков используется для проверки гипотезы о равенстве медианы заданному значению (пример 3.2к). Для парных данных критерий знаковых рангов Уилкоксона используется при проверке гипотезы об отсутствии «эффекта обработки» (пример 3.3к).

Пример 3.2к. В задаче о скорости реакции на звук и на свет проверим гипотезу о том, что медиана распределения скорости реакции на свет равна 155 миллисекундам.

Подготовка данных. В редакторе пакета определить три переменные *sound*, *light* и *median*. В первые две из них ввести данные из табл. 3.1, как это показано на рис. 3.5. В переменную *median* ввести значения гипотетической медианы.

	light	sound	median	var	var	var
1	181	223	155			
2	194	104	155			
3	173	209	155			
4	153	183	155			
5	168	180	155			
6	176	168	155			
7	163	215	155			
8	152	172	155			
9	155	200	155			
10	156	191	155			
11	178	197	155			
12	160	183	155			
13	164	174	155			
14	169	176	155			
15	155	155	155			
16	122	115	155			
17	144	163	155			

Рис. 3.5. Пакет SPSS. Форма ввода наблюдений для процедуры парных сравнений

Выбор процедуры. В меню *Analyze* редактора данных выбрать блок методов *Nonparametric Tests* (непараметрические тесты), а в нем процедуру *2 Related Samples* (две связанные выборки) (см. рис. 3.2).

Заполнение полей ввода данных. На экране ввода данных выбранной процедуры (рис. 3.6) выделить сначала переменную *light* (ее имя при этом будет присвоено *Variable 1* блока *Current Selection* (текущий выбор)). Затем выделить переменную *median* (ее имя присвоится переменной *Variable 2*). Нажав кнопку со стрелкой в центре окна, перенести выбранные переменные в поле *Test Pair(s) List* (список пар для обработки), как это показано на рис. 3.6.

В блоке *Test Type* (тип теста) отметить критерий знаков *Sign* и нажать кнопку **OK**.

Результаты. В окне навигатора вывода результатов пакета появятся две таблицы (рис. 3.7) с результатами расчетов.

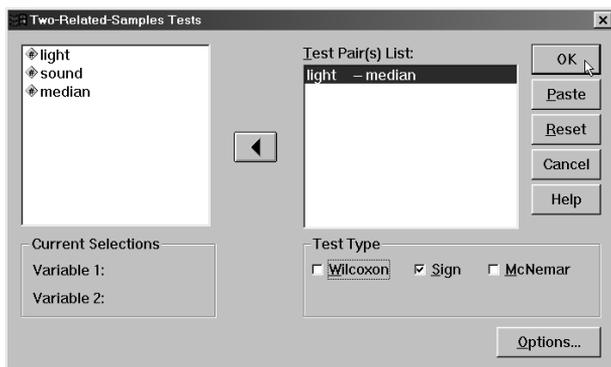


Рис. 3.6. Пакет SPSS. Окно ввода данных и параметров процедуры «2 Related Samples»

Frequencies

		N
MEDIAN - LIGHT	Negative Differences ^a	11
	Positive Differences ^b	4
	Ties ^c	2
	Total	17

a. MEDIAN < LIGHT

b. MEDIAN > LIGHT

c. LIGHT = MEDIAN

Test Statistics^b

	MEDIAN - LIGHT
Exact Sig. (2-tailed)	,118 ^a

a. Binomial distribution used.

b. Sign Test

Рис. 3.7. Пакет SPSS. Вывод результатов критерия знаков

В таблице **Frequencies** приведено число отрицательных **Negative Differences** и положительных **Positive Differences** разностей в переменной **MEDIAN - LIGHT**, а также число совпадающих значений **Ties** в анализируемых переменных. Таблица **Test Statistics** содержит значение минимального уровня значимости критерия знаков против двусторонних альтернатив. Это значение достаточно велико (более 10%), и поэтому нет оснований отвергнуть нулевую гипотезу.

Комментарий. Значение минимального уровня значимости, выдаваемое процедурой, базируется на использовании биномиального распределения вероятностей и является точным.

Пример 3.3к. С помощью критерия знаковых рангов Уилкоксона проверить гипотезу о совпадении распределений времени реакции на звук и на свет в предположении, что оба распределения непрерывны.

Подготовка данных. См. пример 3.2к.

Выбор процедуры. Такой же, как и в примере 3.2к.

Заполнение полей ввода данных. На рис. 3.8 приведено окно параметров ввода данных процедуры 2 Related Samples на стадии присвоения имен переменных Variable 1 и Variable 2 в блоке Current Selection. Выбранные переменные перенести в поле Test Pair(s) List, нажав стрелку в центре окна. Указав в блоке Test Type критерий знаковых рангов Wilcoxon, нажать кнопку **OK**.

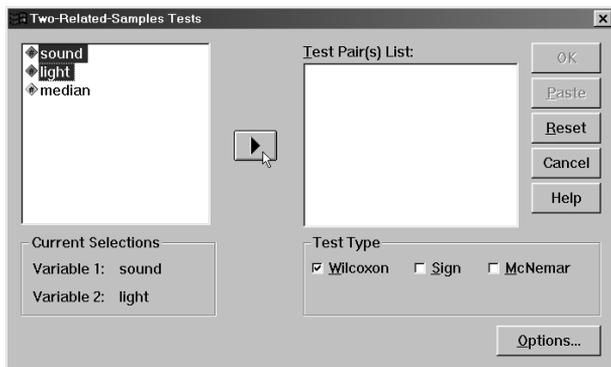


Рис. 3.8. Пакет SPSS. Окно ввода данных и параметров процедуры «2 Related Samples» на стадии выбора переменных

Ranks

		N	Mean Rank	Sum of Ranks
SOUND - LIGHT	Negative Ranks	3 ^a	6,83	20,50
	Positive Ranks	13 ^b	8,88	115,50
	Ties	1 ^c		
	Total	17		

a. SOUND < LIGHT

b. SOUND > LIGHT

c. LIGHT = SOUND

Test Statistics^b

	SOUND - LIGHT
z	-2,457 ^a
Asymp. Sig. (2-tailed)	,014

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

Рис. 3.9. Пакет SPSS. Результаты критерия знаковых рангов Уилкоксона для парных наблюдений

Результаты. На рис. 3.9 приведены таблицы расчетов для критерия знаковых рангов Уилкоксона. В таблице **Ranks** для отрицательных **Negative Ranks** и положительных **Positive Ranks** рангов указано их число, сумма рангов и средние ранги (частное от деления суммы отрицательных (положительных) рангов на число соответствующих рангов). В таблице **Test Statistics** указано значение z -аппроксимации для распределения статистики критерия и асимптотический уровень значимости против двусторонних альтернатив.

Полученный с помощью этого критерия минимальный уровень значимости для проверки нулевой гипотезы о совпадении распределений достаточно мал, что позволяет скорее отвергнуть гипотезу, чем принять ее.

Комментарий. При использовании этой процедуры для малых выборок ее результаты (минимальные уровни значимости) должны рассматриваться как приближительные.

Дополнительная литература

1. Гаек Я., Шидак З. Теория ранговых критериев. — М.: Наука, 1971. — 376 с.
2. Рунион Р. Справочник по непараметрической статистике. Современный подход. — М.: Финансы и статистика, 1982. — 198 с.
3. Холлендер М., Вулф Д. Непараметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.

Начала теории оценивания

4.1. Введение

Что такое оценивание. Статистика имеет дело с данными, подверженными случайной изменчивости. Их поведение может характеризоваться законом распределения вероятностей, если данные являются выборкой, или более сложными моделями (факторными, регрессионными и т.п.), если данные неоднородны. Эти законы распределения вероятностей и модели, как правило, содержат неизвестные величины (параметры) — среднее значение, дисперсию, вклады факторов, коэффициенты функциональных зависимостей и т.п. Исследователя обычно интересуют либо сами эти параметры, либо некоторые заранее известные функции от них. К сожалению, в силу случайной изменчивости наблюдаемых данных нельзя, основываясь только на них, указать совершенно точное значение параметров. Приходится довольствоваться лишь приближенными значениями. Термин «оценить» в статистике означает «указать приближенное значение».

Определение. *Оцениванием в статистике называется указание приближенного значения интересующего нас параметра (или функции от некоторых параметров) на основе наблюдаемых данных. Оценка — это правило вычисления приближенного значения параметра (или функции от некоторых параметров) по наблюдаемым данным.*

Примеры оценок. Мы уже сталкивались с наиболее простыми и распространенными оценками — выборочным средним, выборочной дисперсией, выборочной медианой и др., — в п. 1.8 (хотя само слово «оценка» мы там не произносили). Так, выборочное среднее является оценкой среднего распределения случайной величины, породившей выборку, выборочная дисперсия является оценкой дисперсии этого распределения и т.д.

Требования к оценкам. Методов для определения приближенного значения параметра (т.е. оценок этого параметра) можно придумать великое множество. Поэтому при построении оценок и выборе их для практического применения к оценкам предъявляются определенные

требования, например требования точности (близости к истинному значению параметра), несмещенности (чтобы математическое ожидание оценки было равно истинному значению параметра), состоятельности (чтобы при увеличении числа наблюдений оценка сходилась по вероятности к истинному значению параметра) и т.д. Обсуждению свойств оценок посвящен п. 4.5.

Замечание. К сожалению, наилучших во всех отношениях оценок не бывает. Например, оценка, замечательно ведущая себя при некоторых предположениях об исходных данных, при отклонениях от этих предположений может приводить к сильно искаженным результатам. Например, выборочное среднее — широко распространенная оценка среднего распределения по выборке — обладает многими свойствами оптимальности для нормально распределенных выборок, но очень плохо реагирует на наличие в выборке выбросов, т.е. резко выделяющихся значений (обычно они порождены грубыми ошибками в измерениях и иными причинами). Поэтому в последнее время интенсивно развиваются методы устойчивого (робастного) оценивания. Главная задача этих методов — получение надежных и эффективных оценок, пригодных для ситуаций, когда данные отклоняются от моделей выборок, содержат засорения или грубые наблюдения. Эти вопросы подробно рассмотрены в [108] и [116]. А изложение классических результатов теории оценивания можно найти в [16], [64] и др.

О содержании этой главы и следующих глав. В этой главе мы расскажем об оценках и их свойствах в самой простой ситуации — когда имеются независимые наблюдения некоторой случайной величины и мы хотим по ним оценить параметры распределения этой случайной величины. Будут рассмотрены некоторые важнейшие фундаментальные основы теории оценивания (закон больших чисел, центральная предельная теорема), разобраны начала некоторых подходов к оцениванию параметров вероятностных распределений по выборке (метод наибольшего правдоподобия, метод моментов, метод квантилей) и кратко рассказано об основных свойствах оценок и доверительном оценивании.

В гл. 5 будет более подробно рассмотрено оценивание параметров для нормально распределенных выборок. А в гл. 6—9 разбираются более сложные случаи, когда оценке подлежат параметры регрессионных и факторных моделей, а также меры связи (зависимости) переменных.

4.2. Закон больших чисел

Рассмотрим сначала самую простую задачу оценивания — оценку вероятности некоторого события. Хотя в основе любого статистического вывода лежит понятие вероятности, мы лишь в немногих случаях можем определить вероятность события непосредственно. Как обсуждалось в гл. 1, иногда эту вероятность можно установить из соображений

симметрии, равной возможности (карты, кости, домино и прочие азартные игры) и т.п. Но универсального метода, который позволял бы для произвольного события указать его вероятность, не существует. Теорема Бернулли дает возможность приближенной оценки вероятности, если для интересующего нас события A можно проводить независимые повторные испытания.

Теорема Бернулли. Пусть в каждом из n испытаний вероятность $p = P(A)$ события A остается неизменной и результат каждого испытания независим от остальных. Обозначим через S случайное число тех испытаний (из общего числа n), в которых произошло событие A . Обычно кратко говорят, что S — число «успехов» в n испытаниях Бернулли. Теорема Бернулли утверждает, что при большом n относительная частота S/n события A приближенно равна вероятности события A , т.е. $S/n \simeq p$, где $p = P(A)$.

Замечание. Исторически эту теорему можно считать первой теоремой теории вероятностей. Она содержалась в сочинении Якоба Бернулли (1654 – 1705) «Искусство предположений» («Ars. Conjectandi»), изданном в 1713 г. уже после смерти автора (русский перевод последней, четвертой части этого сочинения, см. в [15]). В истории теории вероятностей это сочинение сыграло важнейшую роль. Оно завершается обсуждением упомянутой теоремы и ее доказательством, которое было довольно сложным.

В наше время теорема Бернулли представляется частным вариантом более общей закономерности — закона больших чисел. Благодаря развитию науки для установления этого важного факта теперь не требуется больших усилий.

Вероятностный предел. Рассмотрим теперь, что означает использованное в формулировке теоремы Бернулли выражение «приближенно равно при больших n ». Читатель, знакомый с математическим анализом, мог уже переформулировать это утверждение в привычную форму: если $n \rightarrow \infty$, то $S/n \rightarrow p$, где S — число появлений события A в n независимых испытаниях. В теории вероятностей и статистике такие обозначения также используются весьма широко. Однако понятие предела толкуется здесь, как правило, в своем, особом смысле, *отличном* от того, который вкладывается в него в математическом анализе.

Действительно, вспомним принятое в математическом анализе определение предела последовательности. Мы говорим, что $a_n \rightarrow a$ при $n \rightarrow \infty$, если для любого $\varepsilon > 0$ найдется такое N , что при $n > N$ будет выполняться неравенство $|A_n - a| < \varepsilon$. Для теоремы Бернулли это значило бы, что для достаточно больших n действует соотношение

$$\left| \frac{S}{n} - p \right| < \varepsilon.$$

К сожалению, это утверждение неверно. Хотя и с малой вероятностью, но значения p и S/n могут отличаться значительно. Например, с положительной вероятностью S может быть равно 0. Поэтому нельзя рассчитывать на неперемнное выполнение соотношения $|\frac{S}{n} - p| < \varepsilon$. Поэтому для случайных последовательностей используется другое понятие предела:

$$P\left(\left|\frac{S}{n} - p\right| < \varepsilon\right) \rightarrow 1$$

(для любого $\varepsilon > 0$) при $n \rightarrow \infty$. Когда требуется отличать это понятие предела от того, которое используется в математическом анализе, говорят: «последовательность случайных величин сходится по вероятности».

Итак, событие $|\frac{S}{n} - p| < \varepsilon$ не является достоверным, но теорема Бернулли утверждает, что оно практически достоверно при достаточно больших n .

Закон больших чисел. При рассмотрении биномиального распределения в п. 2.1 мы вводили случайные величины X_i , $i = 1, \dots, n$, связанные с отдельными испытаниями: $X_i = 1$ в случае «успеха» в испытании i и $X_i = 0$ — в противном случае. Ясно, что мы можем представить S в виде суммы $X_1 + \dots + X_n$, где случайные величины X_1, \dots, X_n независимы и одинаково распределены, причем для любого i : $MX_i = p$. Тогда мы можем переформулировать утверждение теоремы Бернулли в виде:

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - MX\right| < \varepsilon\right) \rightarrow 1 \quad \text{при } n \rightarrow \infty.$$

Итак, здесь *среднее арифметическое от большого числа независимых одинаково распределенных случайных слагаемых оказалось близким к их математическому ожиданию*. На самом деле это утверждение верно не только для величин X_i , полученных из испытаний Бернулли, а является гораздо более общим. Ниже мы докажем его для любых величин X_i , имеющих дисперсию. А с помощью небольших математических усилий условие наличия дисперсии можно заменить и более слабым.

Как мы говорили в п. 1.8, среднее арифметическое является выборочным аналогом математического ожидания. Иначе говоря, если в формуле, определяющей математическое ожидание, заменить истинную функцию распределения F случайных величин X_i на выборочную (эмпирическую) функцию распределения F_n , то получится формула среднего арифметического. На самом деле стремление при больших n значения

выборочной характеристики распределения к значению соответствующей теоретической характеристики (часто говорят — к ее истинному значению) справедливо не только для среднего арифметического. При весьма слабых предположениях на свойства F и интересующей нас характеристики распределения *при больших n значение выборочной характеристики распределения стремится к значению соответствующей теоретической характеристики*. Это утверждение очень важно для теории вероятностей и статистики, оно называется *законом больших чисел*.

Пример. Мореплаватели только сравнительно недавно получили возможность определять координаты своего корабля вдали от берегов. Если широту корабля несложно установить с помощью астрономических наблюдений, то для определения долготы, т.е. угла поворота земного шара, при котором совмещаются местный меридиан и гринвичский, надо точно знать гринвичское время. Следовательно, до появления радио было необходимо иметь на корабле часы, точно идущие по гринвичскому времени.

Однако до XIX века существовавшие часы не обеспечивали необходимой для измерения долготы точности. Лишь в XIX веке были сконструированы особо точные часы — хронометр. И когда в 1831 г. в кругосветное плавание для составления карт отправлялся корабль «Бигль» (эта экспедиция сейчас широко известна благодаря участию в ней молодого тогда Ч. Дарвина), капитан корабля Фиц Рой, человек просвещенный и ученый, взял с собой 24(!) хронометра. Гринвичское время капитан определял усреднением показателей всех хронометров. И он был прав, поскольку по закону больших чисел среднее арифметическое от большого числа случайных слагаемых близко к среднему арифметическому от их математических ожиданий (как правило, ближе, чем для каждого слагаемого в отдельности). Подробнее мы обсудим это ниже.

Вернемся теперь к закону больших чисел и сформулируем простейший его вариант — теорему Чебышева.

Теорема Чебышева. Пусть X_1, \dots, X_n — независимые одинаково распределенные случайные величины, имеющие математическое ожидание и дисперсию. Общее значение математического ожидания этих величин обозначим через a . Тогда для любого $\varepsilon > 0$ при $n \rightarrow \infty$

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - a\right| < \varepsilon\right) \rightarrow 1.$$

В статистике среднее арифметическое величин X_1, \dots, X_n обозначают \bar{X} . Так что кратко теорему Чебышева можно записать так: $\bar{X} \rightarrow a$.

Доказательство. Для доказательства теоремы нам потребуется неравенство Чебышева: пусть ξ — неотрицательная случайная величина, имеющая математическое ожидание. В таком случае для любого $\varepsilon > 0$

$$P(\xi \geq \varepsilon) \leq \frac{M\xi}{\varepsilon}.$$

Проведем доказательство этого неравенства для случая, когда непрерывная случайная величина имеет плотность распределения вероятностей $f(x)$:

$$P(\xi \geq \varepsilon) = \int_{\varepsilon}^{\infty} f(x) dx \leq \int_{\varepsilon}^{\infty} \frac{x}{\varepsilon} f(x) dx \leq \frac{1}{\varepsilon} \int_{\varepsilon}^{\infty} x f(x) dx = \frac{M\xi}{\varepsilon},$$

что и требовалось доказать.

Применив это неравенство к неотрицательной случайной величине $\xi = (X - MX)^2$, получим, что

$$P(|X - MX| \geq \varepsilon) = P((X - MX)^2 \geq \varepsilon^2) \leq \frac{1}{\varepsilon^2} M(X - MX)^2 = \frac{1}{\varepsilon^2} DX,$$

т.е. для любой случайной величины, имеющей математическое ожидание и дисперсию, $P(|X - MX| > \varepsilon) < \frac{DX}{\varepsilon^2}$.

Применим это утверждение к \bar{X} . Легко видеть, что $M\bar{X} = a$, $D\bar{X} = \sigma^2/n$, где $\sigma^2 = DX_i$. По неравенству Чебышева

$$P(|\bar{X} - a| \geq \varepsilon) \leq \frac{D\bar{X}}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \rightarrow 0.$$

Поэтому вероятность противоположного события $\{|\bar{X} - a| < \varepsilon\}$ стремится к 1, что и требовалось доказать.

Продолжение примера. Вернемся к измерению времени на «Бигле». Показание каждого прибора x_i , $i = 1, \dots, n$ — это измерение, независимое от других хронометров. Подразумевается, что конструкция хронометра такова, что в работе этого прибора отсутствует систематическая ошибка. Это значит, что одни экземпляры хронометров могут «уходить», другие «отставать», но эти ошибки случайные, связанные с изготовлением данного образца. Математически это условие формулируется так: $MX_i = a$, $i = 1, \dots, n$. Качество конструкции и технологии изготовления хронометров характеризуется тем, насколько однородна по точности хода вся продукция в целом. Математически это выражается разбросом показаний отдельных приборов, т.е. дисперсией случайных величин X_i . При доказательстве закона больших чисел мы выяснили, что $D\bar{X}$ в n раз меньше DX_i . Поэтому «среднее время» \bar{X} ближе к истинному, чем можно ожидать того от отдельных значений x_i .

Доказательство теоремы Бернулли. Из теоремы Чебышева, как уже говорилось, легко вывести теорему Бернулли. Пусть S — число успехов в n испытаниях Бернулли, p — вероятность успеха в отдельном испытании. Введем случайные величины X_i , $i = 1, \dots, n$, связанные с отдельными испытаниями. Пусть $X_i = 1$, если испытание i закончилось «успехом», и $X_i = 0$ — в противном случае. Ясно, что $S = X_1 + \dots + X_n$, а случайные величины X_1, \dots, X_n независимы и одинаково распределены. Легко видеть, что $MX_i = p$, $DX_i = p(1 - p)$, а $\frac{S}{n} = \bar{X}$. По теореме Чебышева $P(|S/n - p| < \varepsilon) \rightarrow 1$ при $n \rightarrow \infty$, что и требовалось доказать.

Центральная предельная теорема. Пусть θ — некоторая теоретическая характеристика распределения, θ_n — ее выборочный аналог, полученный по выборке объема n . Закон больших чисел говорит, что при $n \rightarrow \infty$ $\theta_n \rightarrow \theta$ с вероятностью 1. Однако для практических задач одного утверждения о сходимости недостаточно — хотелось бы знать,

насколько далеко θ_n может отклоняться от θ при конкретных значениях n . Например, мы можем захотеть построить интервал, в который $\theta_n - \theta$ попадает с вероятностью 99%, либо найти среднее квадратическое отклонение величины $\theta_n - \theta$, чтобы затем, скажем, указывать возможные границы для неизвестного нам θ по известному значению θ_n .

Лучше всего, когда мы можем точно вычислить распределение случайной величины $\theta_n - \theta$. Иногда это удается сделать (например, в п. 5.2 мы найдем распределения выборочного среднего и выборочной дисперсии для нормального распределения), однако это бывает очень редко. Обычно функцию распределения $\theta_n - \theta$ можно получить только моделированием на ЭВМ. Однако асимптотическое распределение $\theta_n - \theta$ (точнее, $\sqrt{n}(\theta_n - \theta)$) известно достаточно хорошо. Оказывается, при весьма слабых предположениях на функцию распределения F и характеристику θ случайная величина

$$\sqrt{n}(\theta_n - \theta)$$

имеет асимптотически (при $n \rightarrow \infty$) нормальное распределение с некоторыми параметрами (a, σ^2) . Это утверждение носит гордое имя *центральной предельной теоремы*. Действительно, это одно из ключевых положений теории вероятностей и статистики, оно весьма важно как в теоретических исследованиях, так и в прикладных задачах. В этой книге мы еще неоднократно будем встречаться с различными следствиями центральной предельной теоремы.

Замечание. Множитель \sqrt{n} в приведенной выше формуле показывает, как меняется распределение $\theta_n - \theta$, а значит, точность статистических выводов, основанных на θ_n , при увеличении объема выборки n . Мы видим, что увеличение точности (например, уменьшение длины доверительного интервала, см. ниже) происходит пропорционально $1/\sqrt{n}$, а не $1/n$, т.е. происходит гораздо медленнее, чем рост числа наблюдений. Отсюда следует, что если мы хотим увеличить точность выводов в 10 раз чисто статистическими средствами, мы, как правило, должны увеличить объем выборки в 100 раз. Подробнее об этом будет сказано ниже.

4.3. Статистические параметры

4.3.1. Параметры распределения

В математической статистике и теории вероятностей слово *параметр* (параметры) имеет два близких по смыслу, но все же различных значения. *Параметрами распределения вероятностей* называют набор чисел, значения которых полностью определяют это распреде-

ление как конкретный элемент некоторого семейства вероятностных распределений.

Например, параметрами нормального распределения вероятностей на числовой прямой обычно выступает его математическое ожидание (скажем, a) и дисперсия (скажем, b). В этом случае нормальная плотность как функция аргумента x , изменяющегося от $-\infty$ до $+\infty$, зависит от x и параметров (a, b) :

$$\varphi(x) = \frac{1}{\sqrt{2\pi b}} \exp\left(-\frac{(x-a)^2}{2b}\right).$$

В дальнейшем параметр (или всю их совокупность) будем обозначать одной буквой, скажем, θ . Если параметр один, то θ — число. Если параметров несколько, допустим, r , то θ обозначает их совокупность, скажем, $\theta = (\theta_1, \dots, \theta_r)$. Обычно параметризацию семейства распределений вводят так, чтобы между значениями параметров и элементами семейства устанавливалось взаимно-однозначное соответствие, т.е. чтобы разным наборам $\theta = (\theta_1, \dots, \theta_r)$ и $\theta' = (\theta'_1, \dots, \theta'_r)$ соответствовали разные распределения. В остальном выбор параметров (способов параметризации) диктуется конкретными обстоятельствами. Например, для нормального распределения на прямой возможна и параметризация с помощью параметров a и $\sigma = \sqrt{b}$.

Оценки параметров. Любые характеристики распределения вероятностей могут быть выражены через его параметры. Поэтому одна из основных задач математической статистики — по наблюдениям над независимыми реализациями случайной величины (т.е. по выборке) сделать выводы о параметрах ее распределения, например, указать их приближенные значения.

Вместо словосочетания «приближенное значение» в статистике используется термин «оценка». Так что «указать приближенные значения параметров» означает оценить их, указать оценки. Основой для этого должны служить только зарегистрированные во время эксперимента значения, которые приняли наблюдаемые случайные величины. Если x_1, \dots, x_n — совокупность независимых одинаково распределенных случайных величин (выборка), закон распределения вероятностей которых зависит от неизвестного параметра θ , то в качестве оценки могут выступать функции от аргументов x_1, \dots, x_n , скажем, $t(x_1, \dots, x_n)$. При этом надо, чтобы

$$t(x_1, \dots, x_n) \simeq \theta. \quad (4.1)$$

4.3.2. Параметры модели

Выборка представляет собой простейшую, но далеко не единственную модель случайных данных. Например, нам уже известна задача

сравнения двух выборок. В этой задаче мы можем использовать предположения (математическую модель), согласно которым законы распределения этих выборок отличаются только сдвигом одного распределения относительно другого. Если мы захотим проверить гипотезу о том, что этот сдвиг равен нулю, либо оценить величину сдвига, то эта величина (неизвестная экспериментатору) будет выступать в данном случае *параметром модели*. Задача оценивания параметров модели является очень важной на практике. В этой книге (гл. 6—8) мы будем рассматривать наиболее распространенные модели — регрессионные и факторные. В каждой из них имеются несколько параметров модели, которые нужно оценить.

Надо отметить, что даже точное знание значений параметров модели не всегда позволяет идентифицировать закон случайности, т.е. то распределение вероятностей, которому подчиняются случайные наблюдения. Например, знание величины смещения одной выборки относительно другой не дает нам сведений о распределениях этих выборок. В этом отличие параметров модели от параметров распределения.

4.4. Оценивание параметров распределения по выборке

Вопросы оценки параметров статистических моделей будут рассмотрены в следующих главах. Здесь же мы обсудим подробнее методы оценивания параметров распределения по имеющейся выборке.

В математической статистике есть много подходов, которые придают высказанному выше требованию (4.1) точную математическую форму. Ни один из них не может считаться универсальным или наилучшим. В зависимости от целей эти методы можно разделить на две группы. Первую группу составляют методы оценивания параметров по конечной выборке, вторую — методы оценивания по неограниченно растущей выборке. С практической точки зрения вторая группа подходов важнее, так как интуитивно понятно, что для получения сколь-либо надежных выводов о параметрах и характеристиках распределения надо иметь достаточно информации, т.е. проделать большое количество экспериментов. Кроме того, с теоретической точки зрения вторая группа подходов проще, так как при больших n исчезают многие проблемы, относящиеся к конечным выборкам. Основой для выводов в этом случае служит закон больших чисел — при больших n значения выборочных характеристик распределения приближаются к неизвестным нам теоретическим значениям этих характеристик.

Если посмотреть с этих позиций на теорему Чебышева, мы увидим, что она дает способ оценки по выборке теоретического значения математического ожидания, — его оценкой является среднее значение наблюдений: $\bar{x} \simeq a$. Выведем аналогичный результат для дисперсии распределения.

Оценка дисперсии распределения. Пусть x_1, \dots, x_n — совокупность независимых реализаций случайной величины ξ . Согласно закону больших чисел, для получения приближенного значения дисперсии $D\xi = M(\xi - M\xi)^2$ надо в определении дисперсии заменить теоретическую функцию распределения F на ее выборочный аналог F_n . Иначе говоря, требуется заменить операцию математического ожидания M усреднением по выборке. Сначала сделаем это по отношению к M , стоящему внутри скобок. Вместо $(\xi - M\xi)^2$ получим совокупность

$$(x_1 - \bar{x})^2, \quad (x_2 - \bar{x})^2, \dots, \quad (x_n - \bar{x})^2.$$

Остается применить усреднение вместо внешнего символа M . Получаем приближенное выражение для дисперсии: $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.

Докажем закон больших чисел для дисперсии. Нам надо показать, что при $n \rightarrow \infty$

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \rightarrow D\xi. \quad (4.2)$$

Для этого прежде преобразуем $\sum_{i=1}^n (x_i - \bar{x})^2$ следующим образом:

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n [(x_i - a)^2 - (\bar{x} - a)]^2 = \\ &= \sum_{i=1}^n (x_i - a)^2 - 2(\bar{x} - a) \sum_{i=1}^n (x_i - a) + n(\bar{x} - a)^2 = \sum_{i=1}^n (x_i - a)^2 - n(\bar{x} - a)^2. \end{aligned}$$

Поэтому левая часть соотношения (4.2) равна

$$\frac{1}{n} \sum_{i=1}^n (x_i - a)^2 - (\bar{x} - a)^2. \quad (4.3)$$

Так как $\bar{x} \rightarrow a$, второй член выражения (4.3) стремится при $n \rightarrow \infty$ к нулю. Первый же член выражения (4.3) при $n \rightarrow \infty$ сходится к $M(\xi - a)^2$, т.е. к $D\xi$, что и доказывает утверждение (4.2).

Выражение $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ можно назвать выборочной дисперсией (иногда говорят — *дисперсия выборки*). Однако чаще вместо него используют

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

Понятно, что уменьшение n на 1 в знаменателе левой части (4.2) сказывается на предельном поведении этого выражения и $s^2 \rightarrow D\xi$ при $n \rightarrow \infty$. В то же самое время s^2 обладает тем свойством, что

$$Ms^2 = D\xi \quad \text{при любом } n, \quad (4.4)$$

что считается достоинством. Говорят, что s^2 является *несмещенной оценкой* $D\xi$.

Для доказательства (4.4) надо обратиться к (4.3) и учесть, что $M(\bar{x} - a)^2 = D\bar{x}$, так как $M\bar{x} = a$. Как отмечалось ранее, $D\bar{x} = \frac{1}{n}D\xi$, поэтому $M \sum_{i=1}^n (x_i - \bar{x})^2 = nD\xi - D\xi = (n-1)D\xi$. Отсюда следует (4.4).

Оценки параметров распределения. Пусть мы имеем выборку из распределения, принадлежащего некоторому параметрическому семейству $F(\theta)$, и хотим по выборке оценить неизвестные нам параметры θ этого распределения. Для этого часто используется следующий прием. Выбирают какую-либо характеристику распределения T (среднее, медиану, квантиль и т.д.), выражаемую через функцию распределения. Но поскольку функция распределения F зависит от θ , то и значение характеристики T есть функция от неизвестного нам значения θ . Выборочный аналог этой характеристики T_n на основании закона больших чисел будет близок к ее теоретическому значению, если объем наблюдений достаточно велик. В связи с этим рассмотрим уравнение, правой частью которого является теоретическое значение характеристики, а левой — ее выборочное значение: $T(\theta) = T_n$. Если параметр θ одномерный, то, разрешая подобное уравнение, получим оценку θ . Если параметр θ многомерный (т.е. параметров распределения несколько), то для их нахождения выбираются несколько характеристик распределения и составляется система из соответствующего количества уравнений.

В качестве характеристик распределения часто используют моменты (метод моментов), реже — квантили (метод квантилей). Проследим за действием этих методов на примере оценивания по выборке параметров нормального распределения (оба параметра неизвестны).

Метод моментов. Пусть X_1, \dots, X_n — независимые случайные величины, распределенные по нормальному закону с параметрами a и σ^2 (кратко — по закону $N(a, \sigma^2)$). В качестве характеристик распределения будем использовать первый и второй моменты ($M\xi$ и $M\xi^2$). Теоретические значения этих характеристик равны a и $\sigma^2 + a^2$. Приравнявая выборочные моменты к их теоретическим аналогам, получим систему уравнений относительно a и σ^2 :

$$\begin{cases} a = \frac{1}{n} \sum_{i=1}^n x_i, \\ a^2 + \sigma^2 = \frac{1}{n} \sum_{i=1}^n x_i^2. \end{cases}$$

Решение системы, т.е. моментные оценки a , σ^2 , обозначим через a^* , σ^{2*} . Легко видеть, что

$$a^* = \bar{x}, \quad \sigma^{2*} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Заметим, что мы получили бы для a и σ^2 иные выражения, если бы в качестве характеристик распределения взяли другие моменты (а не первый и второй, как в приведенном случае).

Метод квантилей. Чтобы использовать метод квантилей, надо прежде решить, какими квантилями мы будем пользоваться. Для нормальной выборки (и вообще для выборок, в которых параметрами служат сдвиг и масштаб) обычно используют медиану и квартили — верхнюю и нижнюю.

Случайную величину ξ , распределенную по закону $N(a, \sigma^2)$, можно представить в виде $\xi = a + \sigma\eta$, где η подчиняется $N(0, 1)$. Для стандартного распределения $N(0, 1)$ медиана равна 0, а нижняя и верхняя квартили равны $\pm\Phi^{-1}(0.75)$ соответственно. Поэтому для $N(a, \sigma^2)$ медиана равна a , квартили (верхняя, нижняя) равны $a \pm \sigma\Phi^{-1}(0.75)$.

Видно, что σ равна половине разности верхней и нижней квартилей распределения, деленной на $\Phi^{-1}(0.75)$.

Обозначим через $Q_n(0.5)$ медиану выборки x_1, \dots, x_n , через $Q_n(0.25)$ и $Q_n(0.75)$ ее нижнюю и верхнюю квартили. Приравняв к указанным выше теоретическим характеристикам их выборочные аналоги, получим оценки для a и σ :

$$a^* = Q_n(0.5) = \text{med}(x_1, \dots, x_n), \quad \sigma^* = \frac{1}{2\Phi^{-1}(0.75)} [Q_n(0.75) - Q_n(0.25)].$$

4.5. Свойства оценок. Доверительное оценивание

Поскольку, как мы видели, для одних и тех же параметров распределения возможны и употребительны разные оценки, хотелось бы как-то сравнивать их между собой и выбирать из них те, которые лучше или которые обладают желательными свойствами. Ниже мы укажем те свойства, которые обычно имеют часто используемые оценки. Пусть θ_n — оценка характеристики распределения θ , полученная по выборке объема n . Тогда:

- оценка θ_n называется *состоятельной*, если $\theta_n \rightarrow \theta$ по вероятности, когда $n \rightarrow \infty$;
- оценка θ_n называется *несмещенной*, если $M\theta_n = \theta$.

Следует заметить, что если состоятельность — практически обязательное свойство всех используемых на практике оценок (несостоятельные оценки употребляются крайне редко), то свойство несмещенности является лишь желательным. Многие часто применяемые оценки свойством несмещенности не обладают.

Эффективность оценок. Прежде чем ставить вопрос о выборе наилучшей оценки, надо научиться сравнивать оценки между собой. Единого способа сравнения оценок не существует; приходится использовать различные подходы. Чаще всего в качестве критерия качества оценки θ_n параметра θ выбирают малость величины $M(\theta_n - \theta)^2$, а наилучшей оценкой считают такую оценку, для которой эта величина минимальна. Более общий подход состоит в том, что вместо величины $(\theta_n - \theta)^2$ выбирают другую неотрицательную функцию «штрафа» $W(\theta_n, \theta)$ за отклонение θ_n от θ (иногда говорят, функцию потерь), и наилучшей оценкой считают такую, для которой математическое ожидание величины штрафа $M W(\theta_n, \theta)$ минимально.

Оценки, для которых минимальна некоторой функции потерь, часто называют *оптимальными* или *эффективными*. Не следует приписывать этим определениям какие-либо магические свойства, считая, что такие оценки заведомо лучше всех других. На самом деле оптимальные свойства оценок получены при определенных предположениях, которые на практике могут и не выполняться или выполняться лишь приближенно. При этом свойства подобных оценок могут оказаться не столь хорошими.

Например, среднее арифметическое элементов выборки является «эффективной» оценкой параметра a для выборки из нормального распределения $N(a, \sigma^2)$: эта оценка несмещенная и обладает минимальной дисперсией. Но при отклонении распределения от нормального (например, при наличии «выбросов», т.е. резко выделяющихся значений) свойства этой оценки становятся неудовлетворительными, так как ее значения очень сильно зависят от «выбросов».

Доверительное оценивание. Во многих случаях представляет интерес не получение точечной оценки $\hat{\theta}$ неизвестного параметра θ , а указание области (например, интервала на числовой прямой), в которой этот параметр находится с вероятностью, не меньшей заданной (скажем, 95 или 99%). Построить такую область можно следующим образом. Выберем число α , $0 < \alpha < 1$ — вероятность, с которой параметр θ должен попасть в построенную нами область. Пусть мы имеем оценку $\hat{\theta}$ неизвестного параметра θ и для каждого значения θ можем указать область $A(\theta, \alpha)$, в которую оценка $\hat{\theta}$ попадает с вероятностью не меньше α :

$$P_{\theta} \left\{ \hat{\theta} \in A(\theta, \alpha) \right\} \geq \alpha \quad \text{для любого } \theta.$$

Тогда *доверительной областью* (в одномерном случае — *доверительным интервалом*) с уровнем доверия α для неизвестного нам истинного значения θ , построенной по наблюдаемому в опыте значению оценки $\hat{\theta}$, является множество

$$\{\theta \mid \hat{\theta} \in A(\theta, \alpha)\}.$$

Можно сказать, что процесс доверительного оценивания является как бы обращением процесса проверки статистических гипотез: там мы по известному значению параметра θ строили множество $A(\theta)$, в которое с заданной вероятностью попадает некоторая статистика $\hat{\theta}$, а здесь мы по таким множествам строим область, которая накрывает с заданной вероятностью само значение θ .

Примеры построения доверительных интервалов мы приведем в гл. 5 (см. п. 5.2).

4.6. Метод наибольшего правдоподобия

Как мы видели в п. 4.4, разные методы оценивания одних и тех же параметров распределения могут давать разные результаты. Когда есть несколько путей к одной цели, естественно, хочется выбрать наилучший. При определенных ограничениях таким методом является метод наибольшего правдоподобия, основанный на оптимальном использовании имеющейся в выборке информации о параметрах распределения.

Пусть X_1, \dots, X_n — выборка из распределения, плотность которого в точке x зависит от неизвестного параметра θ . Обозначим плотность отдельного наблюдения X_i ($i = 1, \dots, n$) через $p(x, \theta)$. Поскольку случайные величины X_1, \dots, X_n независимы, плотность вероятностей вектора (X_1, \dots, X_n) равна

$$p(x_1, \theta) p(x_2, \theta) \dots p(x_n, \theta), \quad (4.5)$$

где θ — неизвестное нам истинное значение параметра.

Метод наибольшего правдоподобия состоит в следующем. Подставим в (4.5) вместо переменных (x_1, \dots, x_n) элементы выборки, т.е. реализации случайных величин X_1, \dots, X_n , а параметр θ в (4.5) будем рассматривать как переменную величину, изменяющуюся в заданной области значений. В таком случае плотность (4.5) превращается в величину, которую мы будем называть *правдоподобием*:

$$p(X_1, \theta) p(X_2, \theta) \dots p(X_n, \theta). \quad (4.6)$$

Оно, естественно, является функцией переменного θ . Метод наибольшего правдоподобия рекомендует выбирать в качестве оценки $\hat{\theta}$

неизвестного истинного значения параметра θ из (4.5) такое значение, при котором правдоподобие достигает максимума:

$$p(X_1, \theta) p(X_2, \theta) \dots p(X_n, \theta) \rightarrow \max_{\theta}.$$

Ясно, что такой выбор $\hat{\theta}$ происходит в зависимости от значений X_1, \dots, X_n , поэтому $\hat{\theta}$ является функцией от X_1, \dots, X_n , т.е. случайной величиной.

Пример: применение к нормальной модели. Прежде чем обсуждать теоретические свойства метода наибольшего правдоподобия, рассмотрим его действие на примере нормальной выборки $N(a, b)$. В этом случае функция правдоподобия равна

$$\left(\frac{1}{\sqrt{2\pi b}}\right)^n \exp\left\{-\frac{1}{2b} \sum_{i=1}^n (x_i - a)^2\right\}. \quad (4.7)$$

Надо выбрать параметры a, b так, чтобы выражение (4.7) было максимальным (при заданных значениях x_1, \dots, x_n). Заметим, что при произвольном фиксированном b выражение (4.7) будет иметь наибольшее из возможных для него значений, если $\sum_{i=1}^n (x_i - a)^2$ примет наименьшее значение (относительно a). Это произойдет при $a = \bar{x}$. Следовательно, оценка наибольшего правдоподобия \hat{a} для a равна \bar{x} .

Для того чтобы найти оценку наибольшего правдоподобия параметра b , вычислим

$$\max_b \left[(2\pi b)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2b} \sum_{i=1}^n (x_i - \bar{x})^2\right\} \right].$$

Эта задача без труда решается при помощи средств математического анализа. (Надо взять производную по b , приравнять ее нулю и решить полученное уравнение относительно b .) После всех вычислений получим $\hat{b} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.

Итак, оценка наибольшего правдоподобия для (a, b) равна

$$\left(\bar{x}, \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right).$$

В данном случае оценки, полученные методом наибольшего правдоподобия и методом моментов, совпали. Так бывает далеко не всегда.

Пояснения к методу. Попытаемся выяснить теоретически причину действия метода наибольшего правдоподобия: почему при больших n полученные этим методом оценки параметра θ в (4.5) близки к его истинному значению и как в этом участвует закон больших чисел.

Отметим, что функции $p(X_1, \theta) \dots p(X_n, \theta)$ и

$$\frac{1}{n} \ln[p(X_1, \theta) \dots p(X_n, \theta)] \quad (4.8)$$

достигают максимума при одном и том же значении θ , так как логарифм является монотонно возрастающей функцией. Представив логарифм

произведения в виде суммы логарифмов, получаем, что для нахождения оценки максимального правдоподобия можно искать такое значение θ , при котором выражение

$$\frac{1}{n} \sum_{i=1}^n \ln p(X_i, \theta) \quad (4.9)$$

достигает максимума.

По закону больших чисел выражение (4.9) как среднее арифметическое независимых одинаково распределенных величин сходится к их математическому ожиданию, т.е. при больших n оказывается близким к $M \ln p(X_i, \theta)$. Поэтому оценка максимального правдоподобия близка к такому значению параметра θ , при котором величина $M \ln p(X_i, \theta)$ достигает максимального значения как функция θ . Остается только указать это значение параметра.

Обозначим через θ^0 то неизвестное истинное значение параметра, которое мы пытаемся оценить. По определению математического ожидания,

$$M \ln p(X, \theta) = \int p(x, \theta^0) \ln p(x, \theta) dx.$$

Остается исследовать, при каком θ стоящий справа интеграл достигает максимального значения. Оказывается, что максимум достигается при $\theta = \theta^0$: $\int p(x, \theta^0) \ln p(x, \theta^0) dx \geq \int p(x, \theta^0) \ln p(x, \theta) dx$ для любого θ и, более того, для любой функции плотности $q(x)$:

$$\int p(x, \theta^0) \ln p(x, \theta^0) dx \geq \int p(x, \theta^0) \ln q(x) dx. \quad (4.10)$$

Здесь могут возникнуть некоторые сложности из-за того, что функция $q(x)$ при некоторых x может обращаться в нуль, а при таких значениях x не существует логарифма. Однако это затруднение успешно преодолевается.

Неравенства вида (4.10) были впервые обнаружены в пятидесятые годы при создании теории информации, поэтому они называются «неравенствами теории информации». Вы можете попытаться самостоятельно доказать аналог неравенства (4.10) для дискретного случая: пусть p_1, \dots, p_n и q_1, \dots, q_n — два набора положительных величин, причем $\sum_{i=1}^n p_i = 1$, $\sum_{i=1}^n q_i = 1$. Тогда

$$\sum_{i=1}^n p_i \ln q_i \leq \sum_{i=1}^n p_i \ln p_i.$$

Для доказательства можно воспользоваться методом математической индукции.

Итак, из неравенства теории информации (4.10) вытекает, что выражение (4.9) достигает максимума, когда для любого x : $p(x, \theta)$ равно

$p(x, \theta^0)$, т.е. когда $\theta = \theta^0$. Поэтому оценка наибольшего правдоподобия при больших объемах выборки оказывается близкой к истинному значению параметра.

Замечание. Р. Фишер доказал, что в определенном смысле оценки наибольшего правдоподобия наилучшим образом используют информацию о параметрах, содержащуюся в наблюдениях (см., например, [57]). Его работы сделали метод наибольшего правдоподобия очень популярным. Было открыто, что для многих задач самой различной статистической природы метод наибольшего правдоподобия дает хорошие результаты. Задачи эти подчас столь разнородны, что не покрываются единой теорией, которая описывала бы свойства метода и указывала границы его применимости.

Однако далеко не во всех практических задачах метод наибольшего правдоподобия (равно как и другие «наиболее эффективные» для данного семейства распределений методы) дает удовлетворительные результаты. Дело в том, что предположение о принадлежности неизвестной плотности распределения определенному параметрическому семейству (нормальному, показательному или какому-то другому) на практике выполняется лишь приближенно. Метод, который принимает это предположение безоговорочно, может привести к результатам, не имеющим даже приблизительно правильного характера. Так может происходить при определенных, хоть и небольших, отклонениях от начальных предположений. Особенно чувствительны к такого рода нарушениям должны быть оптимальные методы — волюно выражаясь, они используют всю информацию, ничего не оставляя в качестве запаса прочности.

4.7. Оценивание параметров вероятностных распределений в пакете SPSS

При построении оценок параметров распределений к ним можно предъявлять различные требования, такие как несмещенность, эффективность, устойчивость к отклонениям от модели и т.п. Статистическая наука постоянно предлагает новые концепции и подходы к оцениванию, а также конкретные алгоритмы их реализации. Свой вклад в разнообразие оценок вносят и различные способы параметризации распределений. Все это порождает множество различных оценок одних и тех же параметров. Поэтому трудно ожидать, что в том или ином статистическом пакете обязательно найдется процедура, в точности реализующая требуемый алгоритм.

Однако многие пакеты выводят значения наиболее распространенных оценок параметров стандартных вероятностных распределений (см. п. 2.7.). В примере 4.1 мы рассмотрим, как эти возможности реализованы в пакете SPSS.

Во многих случаях требуемые оценки параметров распределения можно получить по соответствующим формулам самостоятельно, воспользовавшись тем, что практически все пакеты дают стандартные оценки младших моментов и процентилей распределения (см. примеры 1.1к и 1.2к). При нахождении значений оценок могут оказаться очень полезными различные вспомогательные процедуры преобразования данных, средства решения систем линейных и нелинейных уравнений и т.п. Использование некоторых из этих возможностей показано в примере 4.2к.

Задача оценивания параметров нормального распределения в статистических пакетах рассматривается отдельно в гл. 5.

В п. 2.7 отмечалось, что в SPSS нет отдельного меню для блока процедур, работающих с распределениями вероятностей. Доступ к процедурам вычисления значений функций распределения и квантилей, а также генерация случайных выборок осуществляется через процедуру **Compute** из меню **Transform** редактора данных (см. п. 1.9.2 и 2.7.2). Однако для большинства вероятностных распределений, с которыми в целом поддерживает работу SPSS, прямое вычисление оценок параметров в пакете не предусмотрено. Исключение составляют четыре распределения: нормальное, Пуассона, равномерное и экспоненциальное, с которыми работает процедура **1-Sample K-S** (критерий согласия Колмогорова—Смирнова для одной выборки). Работа этой процедуры будет разобрана ниже.

Пример 4.1к. Сгенерируем выборку размера $n = 100$ из экспоненциального распределения со средним значением $b = 3$ и оценим по ней значение этого параметра.

Подготовка данных. Генерация выборок из заданного распределения обсуждалась в примере 2.2к. В SPSS используется следующая параметризация для экспоненциального распределения:

$$p(x, \theta) = \theta e^{-\theta x} \quad (x \geq 0),$$

где θ часто называют параметром масштаба (scale) или «отношением риска». Связь между θ и b указана в п. 2.3: $\theta = 1/b$. Следовательно, в данном примере $\theta = 1/3$. Для генерации требуемой выборки сначала в редакторе данных необходимо сформировать произвольную переменную длины $n = 100$, а затем в процедуре **Compute** присвоить переменной **exp0** значения выражения **RV.EXP(1/3)**. В результате получим переменную **exp0** со значениями выборки из экспоненциального закона в редакторе пакета.

Выбор процедуры. Так как требуемая оценка является средним значением выборки, то получить ее в пакете можно множеством различных способов, например процедурами **Frequencies**, **Descriptives**, **Explore** из блока **Descriptive Statistics** меню **Analyze**. (Часть этих процедур разобрана в

п. 1.9.) Расскажем, как решить эту задачу с помощью более общей процедуры — 1-Sample K-S. Она оценивает параметры заданного распределения и проверяет согласие распределения выборки с гипотетическим в том случае, когда это распределение нам известно точно. (Эта тема подробно обсуждается в гл. 10.) Вызов процедуры осуществляется из блока Nonparametric Test меню Analyze (см. рис. 3.2).

Заполнение полей ввода данных. Окно ввода данных и параметров процедуры представлено на рис. 4.1.

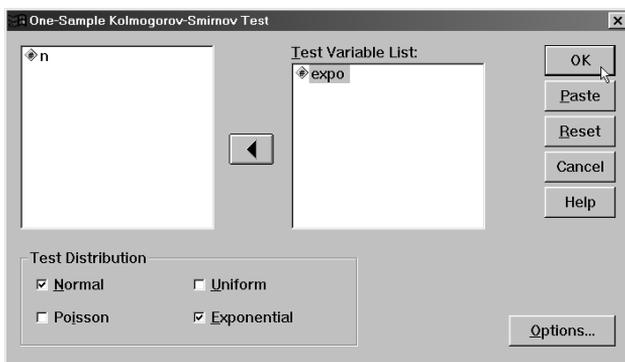


Рис. 4.1. Пакет SPSS. Окно ввода данных и параметров процедуры «1-Sample K-S»

В этом окне необходимо стандартным образом указать переменную для анализа в поле **Test Variable List** и в блоке **Test Distribution** (выбор распределения) указать **Exponential**.

Результаты. Таблица выдачи результатов процедуры представлена на рис. 4.2. Она включает оценку среднего значения распределения, а также значение статистики Колмогорова—Смирнова **Kolmogorov-Smirnov Z** и ее асимптотический минимальный уровень значимости **Asymp. Sig. (2-tailed)** против двусторонних альтернатив.

Комментарии. 1. Основное назначение данной процедуры — проверка согласия выборочных данных с точно указанным распределением. Эта часть ее работы подробно обсуждается в гл. 10. Здесь же заметим, что пользоваться результатами проверки согласия в этой процедуре следует крайне осторожно, учитывая особенности ее реализации. (Процедура не позволяет уточнить пользователю, с каким же именно распределением будет проводиться проверка согласия. Так, в рассматриваемом примере уместно проводить проверку согласия с экспоненциальным распределением, среднее значение которого равно 3. А процедура проводит проверку согласия с экспоненциальным распределением, среднее значение которого подобрано (оценено) по выборке и равно 2.4175 (см. рис. 4.8). К сожалению, это типичная ошибка, свойственная реализациям критерия Колмогорова—Смирнова во многих известных статистических пакетах [100].)

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
EXPO	100	2,4175	2,3245	,11	12,16

One-Sample Kolmogorov-Smirnov Test

		EXPO
N		100
Exponential ^{a,b}	Mean	2,4175
Most Extreme Differences	Absolute	,059
	Positive	,041
	Negative	-,059
Kolmogorov-Smirnov Z		,593
Asymp. Sig. (2-tailed)		,874

a. Test Distribution is Exponential.

b. Calculated from data.

Рис. 4.2. Пакет SPSS. Результаты расчетов процедуры «1-Sample K-S»

2. Укажем параметризацию других распределений, фигурирующих в процедуре 1-Sample K-S.

Распределение Пуассона имеет стандартную параметризацию, при которой

$$P(\xi = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \dots, \quad \lambda > 0.$$

В качестве оценки параметра λ по выборке x_1, x_2, \dots, x_n в пакете используется несмещенная эффективная оценка максимального правдоподобия:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Для непрерывного равномерного распределения (17) Uniform в пакете используется следующая параметризация:

$$f(x, a, b) = \frac{1}{b - a}, \quad (a \leq x \leq b),$$

где параметры a и b задают левую и правую границы распределения. В качестве оценок параметров a и b в пакете используются оценки максимального правдоподобия:

$$\hat{a} = x_{(1)}, \quad \hat{b} = x_{(n)},$$

где $x_{(1)}$ и $x_{(n)}$ — минимальный и максимальный элементы выборки. Указанные оценки являются смещенными (их математические ожидания не равны a и b). Несмещенными оценками с минимальной дисперсией для этих параметров являются величины:

$$a^* = \frac{1}{n-1}(nx_{(1)} - x_{(n)}), \quad b^* = \frac{1}{n-1}(nx_{(n)} - x_{(1)}).$$

Для нормального распределения в пакете используется стандартная параметризация плотности распределения $f(x, a, \sigma)$:

$$f(x, a, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad -\infty < x < +\infty,$$

где a — математическое ожидание, а σ — стандартное отклонение. В качестве оценки параметра a используется оценка, равная среднему значению выборки

$$a^* = \bar{x}.$$

В качестве оценки параметра σ используется величина:

$$\sigma^* = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Некоторые свойства этих оценок указаны в п. 5.3.

Приведенные примеры показывают то разнообразие подходов, которое используется в пакете при построении оценок параметров распределения.

Покажем некоторые возможности прямых арифметических и функциональных преобразований пакета для непосредственных вычислений на примере построения оценки одного из параметров логнормального распределения.

Пример 4.2к. По выборке размера $n = 18$ из логнормального распределения с плотностью вероятности

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

построим оценку максимального правдоподобия параметра μ .

Подготовка данных. Пусть выборка размером $n = 18$ из логнормального распределения находится в переменной `lognor` редактора данных пакета. Выше указывалось, что требуемая оценка вычисляется по формуле $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \ln(x_i)$, где x_i — элементы выборки.

Выбор процедуры. Проще всего получить требуемую оценку, используя процедуру `Compute`.

Заполнение полей ввода данных. В окне ввода данных и параметров процедуры `Compute` (см. рис. 2.10) задать в поле `Target Variable` имя переменной, в которой будут находиться логарифмы исходных данных, например, `ln`. В поле `Numeric Expression` задать функцию `LN(lognor)`. После выполнения процедуры в редакторе данных появится переменная с именем `ln`, в которой будут находиться логарифмы исходной переменной `lognor`.

Для вычисления оценки $\hat{\mu}$ применить к переменной `ln` процедуру `Frequencies` из блока `Descriptive Statistics` меню `Analyze` редактора данных. Работа этой процедуры подробно обсуждалась в п. 1.9.

Комментарии. 1. Процедура `Compute` включает широкий круг стандартных математических функций, работающих со скалярными и векторными переменными. Результатом работы этих функций всегда является переменная, размер которой совпадает с максимальным размером переменных, загруженных в редактор данных. Так, например, вычисляя значение функции стандартного нормального распределения вероятностей в точке 1.5 с помощью выражения `CDFNORM(1.5)`, в результате будет выдана векторная переменная, каждое значение которой равно 0.93, а размер этой переменной будет определяться переменными, ранее загруженными в редактор данных. Такой порядок имеет свои плюсы (когда надо получить значения функции распределения сразу в нескольких точках, указывая в качестве аргумента функции вектор, содержащий координаты точек) и очевидные минусы.

2. Наряду со стандартными математическими функциями эта процедура позволяет работать с широким спектром специальных статистических функций, включая функции распределения вероятностей и обратные к ним (функции квантилей), датчики случайных чисел и др.

3. Процедура `Compute` эффективно используется для выбора требуемого массива наблюдений из общего массива данных загруженного в редактор пакета. Для этого в ней можно использовать арифметические и логические операции для формирования фильтров отбора.

Дополнительная литература

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика. Основы эконометрики. Т. 1. Теория вероятностей и прикладная статистика. — М.: Юнити-Дана, 2001. — 656 с.

2. Бикел П., Доксум К. Математическая статистика. — М.: Финансы и статистика, 1983. Вып. 1. — 280 с.; Вып. 2. — 254 с.

3. Бородин А.Н. Элементарный курс теории вероятностей и математической статистики. — СПб.: Лань, 2005. — 256 с.

4. Королев В.Ю. Теория вероятностей и математическая статистика: учебник. — М.: ТК Велби, Проспект, 2006 — 165 с.

5. Справочник по прикладной статистике: в 2 т.; под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. — М.: Финансы и статистика, 1989, 1990.

Анализ одной и двух нормальных выборок

Нормальное распределение играет особую роль в теории вероятностей и математической статистике. Как показывает практика, самые разнообразные статистические данные с хорошей степенью точности можно считать выборками из нормального распределения. Примерами могут служить помехи в электроаппаратуре, ошибки измерений, разброс попадания снарядов при стрельбе по заданной цели, рост наудачу взятого человека, скорость реакции на раздражитель и т.д. В гл. 2 отмечалось, что можно предполагать нормальное распределение у случайной величины, если на ее отклонение от некоторого заданного значения влияет множество различных факторов, причем влияние каждого из них вносит малый вклад в это отклонение, а их действия независимы или почти независимы.

Кроме того, в силу центральной предельной теоремы и ее разновидностей (см. [24], [30], [118]) распределение целого ряда широко распространенных в статистике функций от случайных величин (статистик, оценок) хорошо аппроксимируется нормальным распределением.

Прежде чем перейти к подробному разбору конкретных методов анализа нормальных выборок, кратко охарактеризуем основные его цели и возможные результаты.

5.1. Об исследовании нормальных выборок

О проверке нормальности распределения. Для исследования нормальных (т.е. подчиняющихся нормальному распределению) данных математической статистикой выработаны эффективные методы. Строго говоря, эти методы непригодны для данных другой природы (т.е. они могут давать для них неправильные результаты). Поэтому, когда мы готовимся применить ориентированные на нормальное распределение методы к имеющимся наблюдениям, полезно выяснить, похоже ли распределение этих наблюдений на нормальное. С полной уверенностью сказать это все равно будет невозможно, но по крайней мере от грубых ошибок такие проверки могут нас уберечь.

Методы установления закона распределения (или типа закона распределения) выборки получили название *критериев согласия*. К ним относятся критерии типа Колмогорова—Смирнова, хи-квадрат и омега-квадрат, критерии асимметрии и эксцесса и др. Они подробно разбираются нами в гл. 10. Одной из главных особенностей этих методов является требование достаточно больших объемов (сотни или даже тысячи) анализируемых данных для получения эффективных выводов. Другими словами, для небольших объемов данных эти методы способны отвергнуть предположение о нормальности только при довольно резких отклонениях от нормального распределения. Если же истинный закон распределения данных не очень сильно отличается от нормального, то эти критерии не отвергнут предположение о нормальности. В этой главе мы ограничимся только рассказом об одном, самом наглядном и распространенном на практике методе проверки на нормальность — глазомерном (см. п. 5.2).

Рассматриваемые задачи. Анализ одной нормальной совокупности сводится к двум взаимосвязанным типам задач: получению оценок параметров нормального распределения и доверительных интервалов для них и проверки гипотез о том что эти параметры равны заданным значениям. Мы рассмотрим эти задачи в п. 5.3 и 5.4. Кроме того, в п. 5.4 мы рассмотрим и задачу проверки того, равны ли средние и дисперсии у двух нормальных выборок.

Стоит сказать, что методы, используемые для решения этих задач (критерии Стьюдента, Фишера и т.д.), очень широко используются и в более сложных задачах — в регрессионном, факторном и других видах анализа данных. Материал данной главы позволит вам хорошо разобраться в их сути.

Замечание. Стоит заметить, что для нормально распределенных выборок самыми эффективными оценками параметров нормального распределения являются хорошо известные нам простые оценки — выборочное среднее и выборочная дисперсия. Однако эти оценки имеют весьма существенный недостаток — они неустойчивы к грубым (ошибочным) наблюдениям или выбросам. Поэтому при их использовании следует соблюдать определенную осторожность и внимательно изучать другие сопутствующие описательные характеристики выборки (см. п. 1.8).

Напомним еще раз те свойства нормального распределения, которые непосредственно используются для анализа нормальных выборок.

Нормальный закон распределения. Напомним, что случайная величина ξ имеет нормальный (гауссовский) закон распределения, если ее функция распределения $F(x)$ задается формулой: $F(x) = \Phi((x - a)/\sigma)$, где $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-t^2/2} dt$ — функция Лапласа, a и σ^2 — параметры

закона распределения. Как отмечалось выше, параметры a и σ^2 имеют непосредственный вероятностный смысл: это соответственно математическое ожидание и дисперсия случайной величины ξ .

Свойства нормального распределения уже обсуждались в гл. 2. Приведем из них те, которые нам понадобятся в этой главе.

1. Если $\eta \sim N(0, 1)$, а $\xi = a + \sigma\eta$, то $\xi \sim N(a, \sigma^2)$. (Другими словами, линейное преобразование $\xi = a + \sigma\eta$ случайной величины η , имеющей стандартное нормальное распределение, приводит к случайной величине ξ , имеющей нормальное распределение с параметрами a и σ^2).
2. Если ξ_1 и ξ_2 — независимые нормально распределенные случайные величины с параметрами a_1, σ_1^2 и a_2, σ_2^2 соответственно, то их сумма $\xi_1 + \xi_2$ тоже распределена по нормальному закону, притом с параметрами $a_1 + a_2$ и $\sigma_1^2 + \sigma_2^2$.

5.2. Глазомерный метод проверки нормальности

Для того чтобы убедиться, что выборка действительно имеет нормальный характер распределения (т.е. о ней можно говорить как о выборке из гауссовского распределения с некоторыми значениями a и σ^2), можно использовать простой графический прием представления данных. В его основе лежат следующие рассуждения.

Рассмотрим зависимость $y = \Phi\left(\frac{x-a}{\sigma}\right)$. Значения функции Лапласа $\Phi(u)$ и обратной к ней Φ^{-1} нетрудно найти по таблицам (см. гл. 2). Применим к рассматриваемой зависимости функцию Φ^{-1} и введем переменную $z = \Phi^{-1}(y)$. Тогда зависимость превращается в линейную:

$$z = \frac{x - a}{\sigma}.$$

Для проверки гипотезы о нормальном характере закона распределения выборки x_1, \dots, x_n воспользуемся тем, что выборочная функция распределения $F_n(x)$ при больших объемах выборки n равномерно близка к теоретической функции распределения. Для удобства дальнейших рассуждений перейдем от выборки к вариационному ряду $x_{(1)}, \dots, x_{(n)}$. Как мы отмечали в гл. 1, $F_n(x)$ — кусочно-постоянная функция, которая в каждой из точек x_i совершает скачок, равный $1/n$, причем при $x < x_{(1)}$ $F_n(x) = 0$, а при $x > x_{(n)}$ $F_n(x) = 1$. Для проверки нормальности выборки мы можем применить функцию Φ^{-1} к серединам этих скачков (значения функции надо взять из таблицы квантилей функции Лапласа). В результате мы получим точки $(x_{(i)}, \Phi^{-1}\left(\frac{2i-1}{2n}\right))$ в плоскости (x, z) .

В зависимости от того, насколько хорошо эти точки «ложатся» на прямую линию, мы можем судить о нормальности распределения выборки.

Даже небольшой опыт работы с реальными выборками позволяет человеку достаточно уверенно выделять среди них отклоняющиеся от нормальных. В сомнительных случаях проверку на нормальность можно продолжить, прибегнув и к другим статистическим критериям (см. также гл. 10). В заключение заметим, что в основе обсуждаемого графического метода лежит удивительное свойство человеческого глаза обнаруживать сходство геометрического образа с прямой линией.

Замечание. Применение функции Φ^{-1} к серединам скачков функции F_n в определенной степени вызвано тем, что мы не могли применить Φ^{-1} ни к самой функции $F_n(x)$, ни к верхним или нижним «концам» ее скачков. Дело в том, что $\Phi^{-1}(0) = -\infty$, а $\Phi^{-1}(1) = \infty$.

Пример. Проверим с помощью изложенного метода гипотезу о том, что время реакции на свет распределено по нормальному закону. Данные этой задачи приведены в табл. 3.1 (см. п. 3.3). Имеем выборку ($x_1 = 181$, $x_2 = 194$, $x_3 = 173$, $x_4 = 153$, $x_5 = 168$, $x_6 = 176$, $x_7 = 163$, $x_8 = 152$, $x_9 = 155$, $x_{10} = 156$, $x_{11} = 178$, $x_{12} = 160$, $x_{13} = 164$, $x_{14} = 169$, $x_{15} = 155$, $x_{16} = 122$, $x_{17} = 144$). Перейдем к вариационному ряду $x_{(i)}$ и нанесем наблюдения на ось x . Далее с помощью таблицы квантилей функции Лапласа вычислим $\Phi^{-1}(1/34)$, $\Phi^{-1}(3/34)$, ..., $\Phi^{-1}(33/34)$. Заметим, что $\Phi^{-1}((2k-1)/2n) = -\Phi^{-1}((2n-2k+1)/2n)$. Отсюда имеем:

$$\begin{aligned} \Phi^{-1}(17/34) &= \Phi^{-1}(1/2) = 0, \\ \Phi^{-1}(19/34) &= -\Phi^{-1}(15/34) = 0.1479, & \Phi^{-1}(21/34) &= -\Phi^{-1}(13/34) = 0.2993, \\ \Phi^{-1}(23/34) &= -\Phi^{-1}(11/34) = 0.4578, & \Phi^{-1}(25/34) &= -\Phi^{-1}(9/34) = 0.6289, \\ \Phi^{-1}(27/34) &= -\Phi^{-1}(7/34) = 0.8208, & \Phi^{-1}(29/34) &= -\Phi^{-1}(5/34) = 1.0494, \\ \Phi^{-1}(31/34) &= -\Phi^{-1}(3/34) = 1.3517, & \Phi^{-1}(33/34) &= -\Phi^{-1}(1/34) = 1.8895. \end{aligned}$$

На рис. 5.1 приведены значения $F_n(x)$ на плоскости (x, z) . Глазомерный метод позволяет нам судить, насколько правдоподобна гипотеза о нормальности распределения выборки. Однако четкого критерия отклонения гипотезы он не дает.

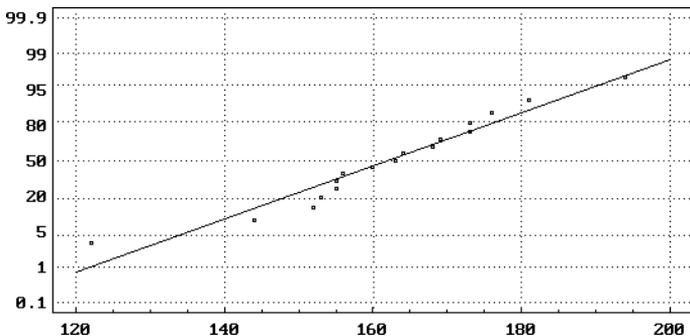


Рис. 5.1. Значения скачков эмпирической функции распределения $F_n(x)$ на плоскости (x, z) (вдоль оси ординат приведены значения $\Phi(z)$ в процентах)

В целом отметим, что детальная проверка гипотезы о нормальности выборки требует довольно значительных объемов выборки (как минимум, порядка сотни наблюдений), и исследователю при обработке данных прежде всего необходимо руководствоваться априорными соображениями о законе распределения.

5.3. Оценки параметров нормального распределения и их свойства

В практических задачах часто возникает необходимость проверки гипотез, связанных со значениями параметров одной или нескольких нормальных выборок. Решение этих задач основано на свойствах оценок параметров нормального распределения a и σ^2 . Поэтому прежде чем формулировать постановки задач, связанных с проверкой гипотез, изучим свойства оценок параметров нормального распределения.

Пусть x_1, \dots, x_n — выборка из нормального распределения с параметрами a и σ^2 . Как отмечалось выше, если случайная величина $\xi \sim N(a, \sigma^2)$, то $M\xi = a$ и $D\xi = \sigma^2$. Поэтому в качестве оценок параметров a и σ^2 , т.е. их приближенных значений, вычисленных по выборочным данным, можно использовать, например, выборочное среднее и дисперсию. Иногда в качестве оценок указанных параметров рассматривают и некоторые другие функции от выборки x_1, \dots, x_n . Например, в качестве оценки параметра a часто используют медиану выборки x_1, \dots, x_n или среднее значение выборки без максимального и минимального элементов, т.е. $\frac{1}{n-2} \sum_{i=2}^{n-1} x_{(i)}$. В качестве оценки σ^2 вместо обычно используемой оценки $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ можно рассматривать величину $[\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|]^2$ и т.д.

О том, чем можно руководствоваться при выборе той или иной оценки неизвестного параметра и какие оценки лучше, упоминалось в гл. 4. Сейчас мы изучим свойства оценок \bar{x} и s^2 , начав с \bar{x} .

Свойства выборочного среднего. Мы уже знаем, что по закону больших чисел (см. гл. 4) выборочное среднее \bar{x} стремится к a с увеличением объема выборки n , т.е. \bar{x} приблизительно равно a при больших объемах выборки. Нас будет интересовать, насколько точным является это приближенное равенство. Близость \bar{x} к a подразумевает существование некоторого малого числа ε , такого, что

$$|\bar{x} - a| < \varepsilon. \quad (5.1)$$

Так как \bar{x} является случайной величиной, $|\bar{x} - a|$ хоть и с малой вероятностью, но все же может оказаться больше ε (мы уже обсуждали

это в гл. 4). Поэтому соотношение (5.1) может быть лишь практически достоверным, т.е. выполняется с вероятностью, близкой к единице — для достаточно больших n . Для выяснения вероятности выполнения неравенства (5.1) надо найти распределение оценки \bar{x} .

Из свойств нормального распределения, приведенных в п. 5.1, легко следует, что \bar{x} также имеет нормальное распределение. При этом

$$M\bar{x} = a, \quad D\bar{x} = D\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n^2} D \sum_{i=1}^n x_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Чтобы найти вероятность выполнения неравенства (5.1), рассмотрим величину $\eta = \sqrt{n}(\bar{x} - a)/\sigma$. По отмеченным свойствам нормального закона эта случайная величина имеет распределение $N(0, 1)$. Предположим сначала, что нам известна величина σ . (На практике это довольно редкий случай. Мы начнем с него, чтобы яснее изложить статистическую идею.)

Для любого малого α , $\alpha > 0$ можно указать с помощью таблиц нормального распределения такое число z , что $P(|\eta| < z) = 1 - 2\alpha$. Чтобы связь z и α была более явной, обозначим это число как $z_{1-\alpha}$. Нетрудно видеть, что $z_{1-\alpha}$ — это квантиль уровня $1 - \alpha$ стандартного нормального распределения. На рис. 5.2 изображена функция распределения $y = \Phi(x)$ стандартного нормального распределения $N(0, 1)$ и отмечена точка $z_{1-\alpha}$. При этом в силу симметрии распределения $z_\alpha = -z_{1-\alpha}$. Каждый отмеченный отрезок на оси ординат имеет длину, равную α .

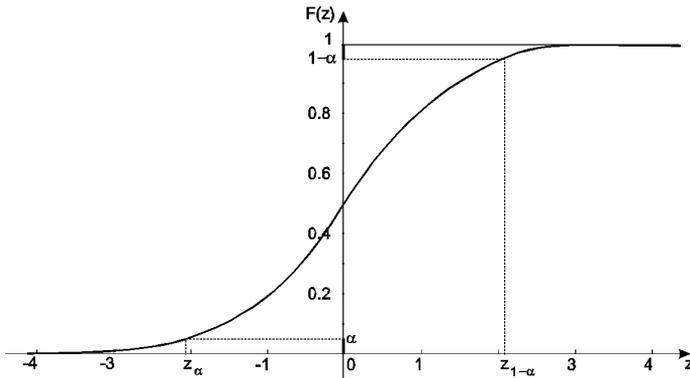


Рис. 5.2. Квантили стандартного нормального распределения

Заменяя η выражением $\sqrt{n} \frac{(\bar{x} - a)}{\sigma}$, находим, что

$$P\left(\left|\sqrt{n} \frac{(\bar{x} - a)}{\sigma}\right| < z_{1-\alpha}\right) = 1 - 2\alpha$$

или

$$P\left(|\bar{x} - a| < \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right) = 1 - 2\alpha.$$

Это означает, что с вероятностью $1 - 2\alpha$ точность приближения \bar{x} к a не ниже, чем $\sigma z_{1-\alpha}/\sqrt{n}$. При этом значение вероятности $1 - 2\alpha$ может быть выбрано сколь угодно близким к единице.

Заметим, что по отношению к неизвестному a решение неравенства

$$|\bar{x} - a| < \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$$

представляет собой интервал $\left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right)$ с центром \bar{x} и длиной $\frac{2\sigma}{\sqrt{n}} z_{1-\alpha}$. Этот интервал называют *доверительным интервалом* для неизвестного a с коэффициентом доверия $1 - 2\alpha$.

Точность оценивания. Выясним, как влияет на точность оценивания параметра a объем выборки n , разброс σ , а также коэффициент доверия $1 - 2\alpha$:

- а) при увеличении n (числа повторных измерений, объема выборки) точность тоже увеличивается. К сожалению, увеличение точности (т.е. уменьшение длины доверительного интервала) пропорционально $1/\sqrt{n}$, а не $1/n$, т.е. происходит гораздо медленнее, чем рост числа наблюдений. Например, если мы хотим увеличить точность выводов в 10 раз чисто статистическими средствами, мы должны увеличить объем выборки в 100 раз;
- б) чем больше σ , тем ниже точность. Зависимость точности от этого параметра носит линейный характер;
- в) чем выше коэффициент доверия $1 - 2\alpha$, тем больше квантиль $z_{1-\alpha}$, т.е. тем ниже точность. При этом между $1 - \alpha$ и $z_{1-\alpha}$ существует нелинейная связь (см. рис. 5.2). С уменьшением α значение $z_{1-\alpha}$ резко увеличивается ($z_{1-\alpha} \rightarrow \infty$ при $\alpha \rightarrow 0$). Поэтому с большой уверенностью (с высокой доверительной вероятностью) мы можем гарантировать лишь относительно невысокую точность. (Доверительный интервал окажется широким.) И наоборот: когда мы указываем для неизвестного a относительно узкие пределы, мы рискуем совершить ошибку — с относительно большой вероятностью.

Для доверительной вероятности (для коэффициента доверия) нет какого-либо наилучшего значения, которого мы могли бы придерживаться. Поэтому обычно указывают несколько вариантов точности приближения для различных коэффициентов доверия. Обычно в качестве значений $1 - 2\alpha$ используют величины 0.9, 0.95, 0.99 и т.д.

Оценка среднего при неизвестной дисперсии. Теперь обратимся к широко распространенной на практике оценке параметра a , когда значение σ^2 неизвестно. Заметим, что в описанном выше случае, где σ считалось известным, все рассуждения основывались на том, что случайная величина $\eta = \sqrt{n}(\bar{x} - a)/\sigma$ имеет известное нам распределение (не зависящее от неизвестных величин a и σ^2). При этом значение σ^2

вошло в конечные выводы о точности оценки параметра a . Естественно попытаться заменить теперь значение σ^2 его оценкой s^2 и сконструировать соответствующий доверительный интервал для параметра a .

Рассмотрим аналог случайной величины η , когда значение σ^2 неизвестно, а именно случайную величину

$$t = \sqrt{n} \frac{(\bar{x} - a)}{s}.$$

Ее часто называют *студентовской дробью*, или *студентовским отношением*. Замечательно то, что распределение t также не зависит от неизвестных параметров a , σ^2 , хотя уже и не является гауссовским. Отсутствие зависимости между законом распределения случайной величины t (несмотря на то, что a входит в выражение t) и параметрами a и σ^2 легко проверить. Как отмечалось выше, случайная величина x_i , имеющая распределение $N(a, \sigma^2)$, может быть записана в виде

$$x_i = a + \sigma \xi_i,$$

где ξ_i имеет стандартное нормальное распределение $N(0, 1)$. Отсюда следует, что $\bar{x} = a + \sigma \bar{\xi}$, а

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sigma^2 \sum_{i=1}^n (\xi_i - \bar{\xi})^2. \quad (5.2)$$

Поэтому

$$t = \sqrt{n} \frac{(\bar{x} - a)}{s} = \frac{\sqrt{n} \sigma \bar{\xi}}{\frac{\sigma}{\sqrt{n-1}} \sqrt{\sum_{i=1}^n (\xi_i - \bar{\xi})^2}} = \sqrt{n} \frac{\bar{\xi}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (\xi_i - \bar{\xi})^2}}.$$

Видно, что t является функцией от стандартно распределенных величин ξ_1, \dots, ξ_n и поэтому не связано с параметрами a, σ^2 . Единственный параметр, от которого зависит распределение t , — это объем выборки n .

Для каждого значения n распределение случайной величины t может быть вычислено. Его называют распределением Стьюдента с числом степеней свободы $n - 1$. По таблицам этого распределения при заданном коэффициенте доверия $1 - 2\alpha$ можно найти квантиль $t_{1-\alpha}$, такую, что

$$P(|t| < t_{1-\alpha}) = 1 - 2\alpha.$$

Отсюда получаем, что

$$P\left(\sqrt{n} \left| \frac{\bar{x} - a}{s} \right| < t_{1-\alpha}\right) = 1 - 2\alpha,$$

или

$$P\left(|\bar{x} - a| < \frac{s}{\sqrt{n}} t_{1-\alpha}\right) = 1 - 2\alpha.$$

Как и в случае с известной дисперсией σ^2 , последние соотношения характеризуют точность приближения \bar{x} к a при заданном коэффициенте доверия $1 - 2\alpha$. А именно, неизвестное нам значение параметра a с коэффициентом доверия $1 - 2\alpha$ принадлежит доверительному интервалу $\left(\bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha}, \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha}\right)$ с центром \bar{x} и длиной $\frac{2s}{\sqrt{n}} t_{1-\alpha}$. О влиянии величин n , σ и a на точность оценивания можно сказать то же самое, что и в случае с известной дисперсией σ^2 .

Рассмотрим теперь свойства оценки дисперсии s^2 и построим доверительный интервал для величины σ^2 .

Выше было показано, что, представляя x_i в виде $x_i = a + \sigma\xi_i$, величину s^2 можно записать как

$$s^2 = \frac{\sigma^2}{n-1} \sum_{i=1}^n (\xi_i - \bar{\xi})^2.$$

Заметим, что каждая случайная величина $\xi_i - \bar{\xi}$ имеет нормальное распределение, так как она является линейной комбинацией независимых нормально распределенных случайных величин.

Как отмечалось в гл. 2, посвященной функциям распределения случайных величин, сумма квадратов n независимых случайных величин η_i , $i = 1, \dots, n$, с распределением $N(0, 1)$ каждая, имеет распределение хи-квадрат (χ^2) с n степенями свободы. Однако мы не можем прямо воспользоваться этим фактом при построении доверительного интервала для σ^2 , так как величины $\xi_1 - \bar{\xi}$, $\xi_2 - \bar{\xi}$, \dots , $\xi_n - \bar{\xi}$ не являются независимыми. Действительно, в каждое выражение

$$\xi_j - \bar{\xi} = \xi_j - \frac{1}{n}(\xi_1 + \dots + \xi_i + \dots + \xi_n)$$

входят остальные случайные величины.

Но все же оказывается, что сумму $\sum_{i=1}^n (\xi_i - \bar{\xi})^2$ можно представить в виде суммы независимых квадратов $\sum_{i=1}^{n-1} \eta_i^2$, где η_i ($i = 1, \dots, n-1$) — независимые случайные величины с распределением $N(0, 1)$. Таким образом, получается, что величина $\sum_{i=1}^n (\xi_i - \bar{\xi})^2$ имеет распределение χ^2 с $n - 1$ степенями свободы.

Для случайной величины с распределением χ^2 и с помощью таблиц распределения можно найти квантили χ_α^2 и $\chi_{1-\alpha}^2$ так, что

$$P(\chi_\alpha^2 < \chi^2 < \chi_{1-\alpha}^2) = 1 - 2\alpha.$$

(Здесь для обозначения случайной величины мы использовали тот же символ, что и для функции распределения. Это соглашение удобно и часто применяется в статистике.)

Перепишем выражение (5.2) с использованием s^2 :

$$\frac{s^2(n-1)}{\sigma^2} = \sum_{i=1}^n (\xi_i - \bar{\xi})^2.$$

Из сказанного выше заключаем, что случайная величина $\frac{s^2(n-1)}{\sigma^2}$ имеет распределение χ^2 с $n-1$ степенями свободы. Поэтому при заданном коэффициенте доверия $1-2\alpha$

$$P \left\{ \chi_\alpha^2 < \frac{s^2(n-1)}{\sigma^2} < \chi_{1-\alpha}^2 \right\} = 1 - 2\alpha,$$

или

$$P \left\{ \frac{1}{n-1} \chi_\alpha^2 < \frac{s^2}{\sigma^2} < \frac{1}{n-1} \chi_{1-\alpha}^2 \right\} = 1 - 2\alpha.$$

Этому утверждению часто придают другую форму, тождественно преобразовав неравенство в скобках:

$$P \left\{ s^2 \frac{(n-1)}{\chi_{1-\alpha}^2} < \sigma^2 < s^2 \frac{(n-1)}{\chi_\alpha^2} \right\} = 1 - 2\alpha.$$

Таким образом, доверительный интервал для дисперсии имеет вид:

$$\left(s^2 \frac{(n-1)}{\chi_{1-\alpha}^2}, s^2 \frac{(n-1)}{\chi_\alpha^2} \right). \quad (5.3)$$

5.4. Проверка гипотез, связанных с параметрами нормального распределения

5.4.1. Одна выборка

Вернемся к задаче проверки статистических гипотез, связанных с нормальным распределением. Так как конкретное нормальное распределение полностью задается значением параметров a и σ^2 , рассмотрим сначала задачу проверки гипотезы о значениях параметров нормального распределения. Эта задача тесно связана с построением доверительных интервалов для параметров нормального распределения.

Критерий Стьюдента. Проверим гипотезу о равенстве среднего значения выборки из нормального распределения заданной величине. Здесь, как и в случае построения доверительного интервала для a , возможны два случая:

- 1) когда σ^2 известно;
- 2) когда σ^2 неизвестно.

Если дисперсия известна. Статистическая формулировка задачи в первом случае следующая. Пусть x_1, \dots, x_n — выборка из нормального распределения $N(a, \sigma^2)$ с некоторыми параметрами a и σ^2 .

Гипотеза H заключается в том, что среднее значение a равно заданному числу a_0 ($H : a = a_0$). Рассмотрим двустороннюю альтернативу: $a \neq a_0$. Выберем уровень значимости α и рассмотрим следующую статистику:

$$\eta = \sqrt{n} \frac{(\bar{x} - a_0)}{\sigma}.$$

(Напоминаем, что σ нам сейчас известно.) Легко видеть, что η имеет стандартное нормальное распределение. Пусть $z_{1-\alpha/2}$ — квантиль уровня $1 - \alpha/2$ этого распределения. Теперь критерий, основанный на статистике η , для проверки гипотезы H формулируется так:

- на уровне значимости α , $\alpha > 0$ гипотеза H принимается, если

$$\sqrt{n} \left| \frac{(\bar{x} - a_0)}{\sigma} \right| < z_{1-\alpha/2};$$

- в противном случае гипотеза отклоняется.

Другими словами, если гипотетическое значение a_0 попадает в доверительный интервал для a с коэффициентом доверия $1 - \alpha$, то гипотеза принимается при уровне значимости α , в противном случае — отвергается.

Если дисперсия неизвестна (т.е. во втором случае), вместо статистики η рассмотрим статистику t

$$t = \sqrt{n} \frac{\bar{x} - a_0}{s}.$$

Статистика t имеет распределение Стьюдента с $n - 1$ степенью свободы. Для заданного уровня значимости α находим процентную точку $t_{1-\alpha/2}$ распределения Стьюдента с $n - 1$ степенью свободы. Критерий для проверки H , основанный на статистике t , будет таков.

Гипотеза H принимается, если

$$\sqrt{n} \left| \frac{(\bar{x} - a_0)}{s} \right| < t_{1-\alpha/2},$$

в противном случае — отвергается. (Напомним, что из этого же соотношения $|t| < t_{1-\alpha/2}$ строился и доверительный интервал для среднего значения при неизвестной дисперсии.)

Сопоставляя доверительные интервалы и теорию проверки статистических гипотез, можно сказать, что доверительный интервал для неизвестного параметра (с доверительной вероятностью $1 - \alpha$) составляют

те значения параметра, которые совместимы с нашими наблюдениями при проверке соответствующих гипотез на уровне значимости α , $\alpha > 0$.

Аналогичным образом обстоит дело с проверкой гипотезы о значении дисперсии нормальной выборки.

5.4.2. Две выборки

Критерий Стьюдента. Рассмотрим теперь задачу сравнения средних значений двух нормальных выборок.

Пусть $x_1, \dots, x_n; y_1, \dots, y_m$ — нормальные независимые выборки из законов распределения с параметрами (a_1, σ_1^2) и (a_2, σ_2^2) соответственно. Рассмотрим проверку гипотезы $H : a_1 = a_2$ против альтернативы $a_1 \neq a_2$. Заметим, что более общий случай $H : a_1 = a_2 + \Delta$, где Δ — заданное число, сводится к предыдущему путем преобразования выборки y_1, \dots, y_m в выборку $y_1 + \Delta, \dots, y_m + \Delta$.

Относительно параметров σ_1^2 и σ_2^2 выделим следующие четыре варианта предположений:

- а) обе дисперсии известны и равны между собой;
- б) обе дисперсии известны, но не равны между собой;
- в) обе дисперсии неизвестны, но предполагается, что они равны между собой;
- г) обе дисперсии неизвестны, их равенство не предполагается.

Для построения критерия проверки гипотезы H проведем следующие рассуждения. От выборок x_1, \dots, x_n и y_1, \dots, y_m перейдем к выборочным средним \bar{x} и \bar{y} . Согласно свойствам нормального распределения и выдвинутой гипотезе, величины \bar{x} и \bar{y} имеют нормальные распределения с одним и тем же средним и дисперсиями σ_1^2/n и σ_2^2/m .

Далее перейдем к статистике, основанной на выборочных средних \bar{x} , \bar{y} и дисперсиях σ_1^2 , σ_2^2 (если они известны) или их оценках s_1^2 , s_2^2 (если дисперсии неизвестны). Статистику мы выберем так, чтобы ее распределение при гипотезе не зависело от неизвестных нам значений параметра. Это позволит нам указать распределение статистики и вычислить его квантили. Наиболее естественными статистиками для перечисленных выше случаев будут следующие:

а) $\frac{\bar{x} - \bar{y}}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}$. Статистика имеет стандартное нормальное распределение, так как является линейной комбинацией независимых нормальных величин. Гипотеза H принимается на уровне значимости α , если

$$\left| \frac{\bar{x} - \bar{y}}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \right| < z_{1-\alpha/2};$$

в противном случае гипотеза отвергается в пользу альтернативы $a_1 \neq a_2$;

б) $\frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}$. Статистика имеет также стандартное нормальное распределение. Правило принятия гипотезы аналогично правилу пункта а);

в) в случае, когда обе дисперсии неизвестны, но предполагаются равными между собой, мы имеем две оценки s_1^2 и s_2^2 одной и той же величины дисперсии $\sigma_1^2 = \sigma_2^2$ (назовем ее, скажем, σ^2). В связи с этим разумно перейти к объединенной оценке σ^2 :

$$s^2 = \frac{s_1^2(n-1) + s_2^2(m-1)}{(n-1) + (m-1)}.$$

Случайная величина $(n+m-2)s^2/\sigma^2$ имеет распределение χ^2 с $n+m-2$ степенями свободы. Критерий для проверки гипотезы $H : a_1 = a_2$ опирается на статистику

$$\frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

которая имеет распределение Стьюдента с $n+m-2$ степенями свободы;

г) в случае неизвестных дисперсий, равенство которых не предполагается, используется аналог статистики пункта б) с заменой неизвестных дисперсий их оценками:

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}.$$

В этой ситуации указать точное распределение введенной статистики затруднительно. Известно, однако, что это распределение близко к распределению Стьюдента с числом степеней свободы, равным

$$\frac{(s_1^2/n + s_2^2/m)^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_2^2/m)^2}{m-1}}.$$

Критерий проверки гипотезы устроен так же, как и в пункте в).

Замечание. Обратим внимание на то, что указанное число степеней свободы является случайной величиной и ее значение, вообще говоря, дробное. Распределения Стьюдента с дробным положительным числом степеней свободы может быть определено, например, с помощью функции плотности распределения, в которой вместо целого числа степеней свободы n фигурирует произвольное положительное число ν (см. п. 2.6.2). Таблицы для дробного числа

степеней свободы составлять не принято. Для нахождения квантилей указанного выше распределения приходится пользоваться приближенными методами. Их описание дано, например, в [19].

Критерий Фишера. Кратко остановимся на вопросе проверки гипотезы о равенстве дисперсий двух нормальных выборок.

Рассмотрим отношение оценок дисперсий первой и второй выборок s_1^2 и s_2^2 :

$$F = \frac{s_1^2}{s_2^2},$$

называемое дисперсионным отношением Фишера, или просто статистикой Фишера. В случае справедливости нулевой гипотезы о равенстве дисперсий нормальных выборок величина F имеет F -распределение с числом степеней свободы $(n - 1, m - 1)$, где n и m — объемы первой и второй выборок соответственно. При нарушении нулевой гипотезы величина F имеет тенденцию к увеличению (уменьшению) в зависимости от того, больше или меньше единицы значение величины σ_2^2/σ_1^2 .

Критерий проверки нулевой гипотезы при заданном уровне значимости α против двусторонних альтернатив $\sigma_1^2 \neq \sigma_2^2$ сводится к следующему: принять гипотезу, если

$$F_{\alpha/2, n-1, m-1} \leq F \leq F_{1-\alpha/2, n-1, m-1};$$

в противном случае отвергнуть гипотезу. Здесь $F_{\alpha/2, n-1, m-1}$ — это квантиль F -распределения уровня $\alpha/2$ с $(n - 1, m - 1)$ числом степеней свободы.

Другое правило проверки гипотезы основывается на использовании доверительного интервала для σ_1^2/σ_2^2 . Если единица (значение отношения σ_1^2/σ_2^2 при гипотезе) принадлежит доверительному интервалу для σ_1^2/σ_2^2 , то гипотеза принимается. В противном случае она отвергается.

5.4.3. Парные данные

В п. 3.6 мы подробно описали, что такое парные данные и каково их обычное экспериментальное происхождение. Там же мы рассмотрели два непараметрических статистических критерия для проверки гипотезы об отсутствии закономерного различия между наблюдениями в паре (иначе говоря — гипотезы об отсутствии эффекта обработки). Типичный пример того, как могут возникать парные данные, дают опыты, в которых наблюдения над объектами (т.е. измерения определенной характеристики) производят дважды: до и после воздействия.

Пусть x_i и y_i — результаты этих измерений для объекта номер i , $i = 1, \dots, n$, где n — численность экспериментальной группы (число объектов). Как обычно, все наблюдения мы считаем случайными величинами (реализациями случайных величин) и предполагаем, что методика эксперимента обеспечивает их независимость для разных объектов. Но наблюдения, входящие в одну пару, мы не можем считать независимыми, поскольку они относятся к одному и тому же объекту. Эти два наблюдения отражают свойства общего для них индивидуального объекта и потому могут быть зависимы друг от друга. Напомним обозначения, введенные для парных данных в п. 3.6.

Данные — совокупность пар случайных величин $(x_1, y_1), \dots, (x_n, y_n)$, где n — объем совокупности (число пар). Обозначим $z_i = y_i - x_i$, $i = 1, \dots, n$.

Допущения (частично повторяют допущения п. 3.6, а частично их усиливают).

1. Все z_i , $i = 1, \dots, n$ — взаимно независимы.
2. Предположим, что z_i можно представить в виде:

$$z_i = \theta + e_i,$$

где e_1, \dots, e_n — независимые случайные величины, θ — неизвестная постоянная (неслучайная) величина (означающая результат воздействия, эффект обработки). Иначе говоря, мы принимаем аддитивную модель для отражения результатов воздействия.

3. Случайные величины e_1, \dots, e_n распределены по нормальному закону $N(0, \sigma^2)$, где дисперсия σ^2 обычно неизвестна. Это предположение дополняет и усиливает перечень свойств случайных величин e_1, \dots, e_n , принятый в п. 3.6.2.

Приняв эти допущения, мы свели задачу о парных данных к задаче об одной нормальной выборке, уже рассмотренной в п. 5.4.1.

В отношении неизвестного θ возможны два вопроса: проверка гипотезы о θ и оценивание θ . Анализ обычно начинают с проверки гипотезы $H: \theta = 0$ (или $\theta = \theta_0$, где θ_0 задано). Если гипотеза оказывается отвергнутой (несовместимой с наблюдениями), обращаются к оцениванию неизвестного θ . Обе эти задачи мы уже обсуждали в п. 5.4.1 об одной нормальной выборке, так что нет нужды повторяться.

Пример (продолжение примера из п. 3.6). Проиллюстрируем описанный метод на примере сравнения времени реакции на звук и на свет, рассмотренном в п. 3.3.2, сопоставим полученные результаты с теми, которые мы имели при применении к этим данным критерия знаков и критерия знаковых рангов Уилкоксона.

В табл. 3.6 приведены значения выборки: $z_1 = -42$, $z_2 = 90$, $z_3 = -36$, $z_4 = -30$, $z_5 = -12$, $z_6 = 8$, $z_7 = -52$, $z_8 = -20$, $z_9 = -45$, $z_{10} = -35$, $z_{11} = -19$, $z_{12} = -23$, $z_{13} = -10$, $z_{14} = -7$, $z_{15} = -0$, $z_{16} = 7$, $z_{17} = -19$. Прежде всего проверим, насколько эти данные согласуются с нормальным законом распределения, используя для этого описанный выше глазомерный критерий. Как видно из рис. 5.3, точки скачков эмпирической функции распределения $F_n(\cdot)$, изображенные в соответствующих координатах, в общем, группируются около прямой линии. Исключение составляет наблюдение $z_2 = 90$. Оно далеко отстоит от основного массива и воспринимается как «выброс». Возможно, что произошла какая-то ошибка в самом эксперименте либо при регистрации — передаче его результата. Это число следовало бы проверить.

Если такой возможности нет и нам приходится действовать чисто статистическими средствами, то выявленный выброс $z_2 = 90$ из дальнейшего анализа следует исключить. Этот путь мы подробно рассмотрим ниже. А сейчас обработаем исходную выборку как гауссовскую. (Сделав вид, что мы либо не проводили проверки на нормальность, либо ничего не заметили.)

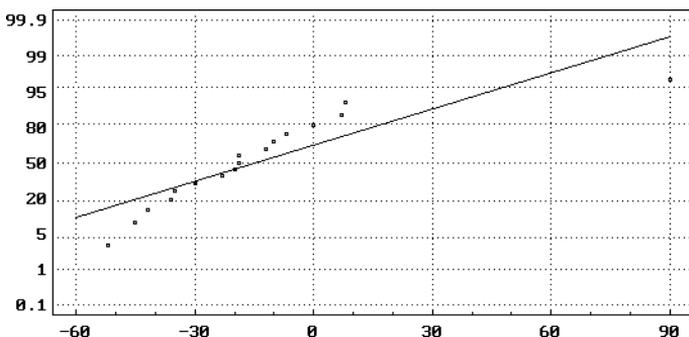


Рис. 5.3. Эмпирическая функция распределения на нормальной вероятностной бумаге

Проверим гипотезу $H : \theta = 0$ против альтернативы $\theta < 0$. Заметим, что в данной ситуации целесообразно рассмотреть именно одностороннюю альтернативу, так как данные определенно говорят о том, что значение θ может быть отрицательным. Вычисления дают: $\bar{z} = -14.4$, $s^2 = 1033.7.6$, $s = 32.15$.

Для проверки гипотезы $H : \theta = 0$ составляем отношение Стьюдента:

$$t = \frac{\bar{z} - \theta}{s} \sqrt{n} = -\frac{14.4}{32.15} \sqrt{17} = -1.846.$$

При справедливости гипотезы статистика t подчиняется распределению Стьюдента с 16 степенями свободы.

Вычислим минимальный уровень значимости, при котором может быть отвергнута гипотеза H . По определению, он равен $P(t < t_{\text{набл.}})$, так как мы рассматриваем одностороннюю альтернативу $\theta < 0$. Воспользовавшись таблицами распределения Стьюдента с 16 степенями свободы, находим, что наименьший уровень значимости, на котором может быть отвергнута гипотеза $H : \theta = 0$ против альтернативы $\theta < 0$, приблизительно равен 0.04.

Обсуждение. Вспомним, что в той же задаче наименьший уровень значимости для критерия знаков оказался равен приблизительно 0.01, т.е. был существенно меньше. На первый взгляд это вызывает удивление. Ведь применяя критерий, опирающийся на известное распределение выборки (в данном случае нормальное), мы должны были бы получить более сильный результат, чем с помощью непараметрического критерия, использующего меньше сведений о выборке. Однако, как оказалось, предположение о нормальном законе распределения только уменьшило нашу уверенность в вопросе принятия или отвержения гипотезы.

Дело здесь в следующем. Обратим внимание на то, что основной вклад в величину выборочной дисперсии s^2 вносит всего одно наблюдение $z_2 = 90$. Это значение и раньше вызывало у нас подозрения. Возможно, что оно порождено ошибкой при регистрации данных. Возможно, что этот испытуемый по своим данным резко отличается от всех остальных. Если мы исключим это наблюдение из нашей выборки и заново проведем расчеты, величины \bar{z} и s изменятся следующим образом:

$$\bar{z} \simeq -22, \quad s^2 \simeq 331.6, \quad s \simeq 18.2, \quad t \simeq -\frac{22}{18.2}\sqrt{16} \simeq -4.835.$$

Воспользовавшись таблицами распределения Стьюдента с 15 степенями свободы, получаем, что наименьший уровень значимости в этом случае менее 0.0005, т.е. данный метод позволяет с гораздо большей уверенностью сделать вывод об имеющемся различии исследуемых характеристик, чем критерий знаков.

Процедура с исключением подозрительных наблюдений называется *отбраковкой грубых наблюдений*. Критерии, используемые для подобной процедуры, можно найти в [19], [77], [86]. Необходимость отбраковки вызвана тем, что традиционные оценки параметров нормального распределения чувствительны к грубым ошибкам, даже если таких ошибок немного в выборке. При использовании непараметрических методов в отбраковке грубых наблюдений, как правило, нет необходимости.

Из приведенного сравнения двух критериев можно сделать вывод о том, что у каждого из них есть свои достоинства и недостатки, поэтому применение их должно основываться на анализе конкретной ситуации. В дальнейшем мы не раз будем обращать внимание на сравнение разных статистических методов и правил, направленных к общей цели. Но уже сейчас можно сделать общий вывод: чем меньше предположений, тем надежнее статистический вывод (тем надежнее он защищен от ошибок исследователя).

5.5. Анализ нормальных выборок в пакете SPSS

Процедуры работы с нормальными выборками входят практически во все статистические пакеты. Кроме разделов пакетов, непосредственно относящихся к этому вопросу, они могут составлять часть разделов описательных методов статистики, дисперсионного и регрессионного анализа, критериев согласия и др. Ниже на примерах будут рассмотрены некоторые из основных процедур анализа нормальных выборок. Процедуры проверки нормальности распределения выборки (критерии согласия) обсуждаются в гл. 10.

Основные процедуры для анализа нормальных выборок в пакете SPSS сосредоточены в двух блоках **Descriptive Statistics** и **Compare Means** меню **Analyze**. Обратим особое внимание на процедуру **Explore** из блока **Descriptive Statistics**. Она позволяет строить доверительные интервалы для среднего для одной и нескольких выборок, выводит на график гистограмму и строит график на нормальной вероятностной бумаге, а также вычисляет различные устойчивые оценки для среднего значения. В блоке **Compare Means** (см. рис. 5.7) практически все процедуры предназначены для работы с одной или несколькими нормальными выборками и парными данными. Ниже на примерах будут разобраны процедуры **One-Sample T-Test** (критерий Стьюдента для одной выборки) и **Independent Samples T-Test** (критерий Стьюдента для нескольких независимых выборок).

Пример 5.1к. Построим 95% доверительные интервалы для среднего значения и дисперсии по выборке диаметров головок заклепок. Проверим гипотезу о равенстве среднего значения выборки заданной величине 13.4.

Подготовка данных. См. пример 1.1к.

Выбор процедуры. В блоке **Compare means** (см. рис. 5.7) меню **Analyze** редактора пакета выбрать процедуру **One-Sample T-Test**.

Заполнение полей ввода данных. Окно ввода данных и параметров этой процедуры представлено на рис. 5.4.

В этом окне перенести переменную **d** из левого поля в поле **Test Variable**. Для получения доверительного интервала для среднего значения в поле **Test Value** указать значение 0. Если необходимо проверить гипотезу о равенстве среднего заданному числу 13.4, в поле **Test Value** указать это число. Кнопка **(Option)** позволяет скорректировать уровень доверия в поле **Confidence Interval**. По умолчанию он равен 95%.

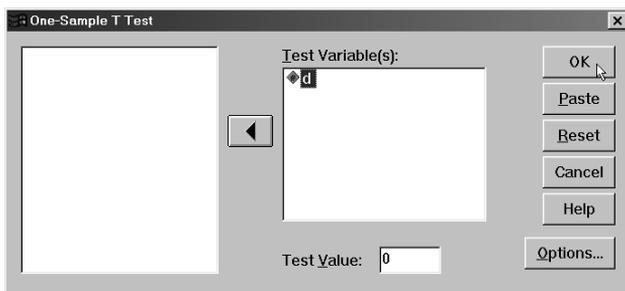


Рис. 5.4. Пакет SPSS. Окно ввода данных и параметров процедуры «One-Sample T-Test»

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
D	200	13,4215	,1344	1,E-02

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
D	1411,8	199	,000	13,4215	13,4028	13,4402

Рис. 5.5. Пакет SPSS. Результаты расчетов процедуры «One-Sample T-Test» для случая Test Value равно нулю

Результаты. Процедура выдает две таблицы в окно навигатора вывода результатов (см. рис. 5.5).

Таблица **One-Sample Statistics** включает объем выборки **N**, ее среднее значение **Mean**, стандартное отклонение **Std. Deviation** и стандартную ошибку среднего значения **Std. Error Mean**. Последняя величина равна стандартному отклонению, деленному на квадратный корень из объема выборки.

Таблица **One-Sample Test** содержит значения t-статистики Стьюдента **t**, ее число степеней свободы **df**, минимальный уровень значимости t-статистики против двусторонних альтернатив **Sig. (2-tailed)**, разность между выборочным средним и значением, указанным в поле **Test Value** экрана ввода параметров процедуры, а также нижнюю **Lower** и верхнюю **Upper** границы 95% доверительного интервала для разности **95% Confidence Interval of the Difference**. Если в поле **Test Value** указан 0, получаем просто доверительный интервал для среднего.

One-Sample Test

Test Value = 13.4						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
D	2,262	199	,025	2,Е-02	3,Е-03	4,Е-02

Рис. 5.6. Пакет SPSS. Результаты расчетов процедуры «One-Sample T-Test» для случая Test Value равно 13.4

Результаты проверки гипотезы о равенстве среднего значения 13.4 (Test Value = 13.4) приведены на рис. 5.6.

Минимальный уровень значимости t-статистики в этом случае равен 0.025. Таким образом, гипотеза о равенстве среднего значения выборки 13.4 отвергается при 95% уровне доверия, хотя при 99% уровне доверия у нас нет оснований отвергнуть гипотезу.

Пакет SPSS не предоставляет простой возможности для построения доверительного интервала для дисперсии выборки. Для расчета границ доверительного интервала (5.3) процедура **Compute** пакета может выдать значения $\chi^2_{1-\alpha}$ и χ^2_{α} с заданным числом степеней свободы. Остальные вычисления в (5.3) надо провести вручную, используя полученное значение стандартного отклонения в качестве оценки величины s .

Пример 5.2к. Проведем анализ однородности двух нормальных выборок для данных о росте девушек и юношей. Проверим гипотезу о равенстве их средних значений и дисперсий.

Данные для этого примера были получены авторами во время чтения курса статистических методов студентам факультета психологии МГУ им. М.В. Ломоносова. Нами были собраны сведения о росте девушек и юношей одного из курсов. Выборка, относящаяся к девушкам, более многочисленна, ее размер оказался равным $n = 53$. Объем выборки ростов юношей оказался $m = 20$. Полученные данные представлены в двух таблицах в порядке их регистрации.

Для выборки ростов девушек с помощью глазомерного метода проверки нормальности можно убедиться в соответствии этих данных нормальному закону распределения. Выборка ростов юношей недостаточно велика, чтобы можно было с уверенностью судить о ее законе распределения. Аналогия с первой выборкой дает разумное основание предполагать и ее нормальной.

Подготовка данных. В редакторе данных пакета ввести данные табл. 5.1 и 5.2 в одну переменную **height**, как это показано на рис. 5.7. Создать переменную **sex** и в ней указать признак пола респондента,

Таблица 5.1

Рост девушек, см													
165	164	158	168	162	166	167	154	165	164	172	167	164	157
164	164	166	173	164	160	164	157	152	175	165	174	163	155
163	162	178	166	165	163	168	161	164	173	161	161	160	164
166	170	167	159	158	164	161	163	163	165	170			

Таблица 5.2

Рост юношей, см													
182	183	168	174	165	174	163	168	179	185	171	174	180	175
179	181	169	184	172	174								

например, 0 – для девушек и 1 – для юношей. Это стандартная форма ввода данных в SPSS для всех процедур анализа выборок, за исключением анализа парных данных.

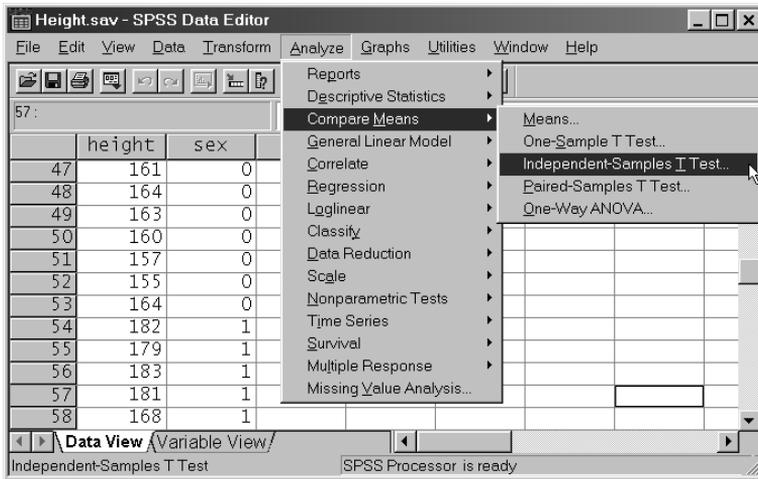


Рис. 5.7. Пакет SPSS. Форма ввода данных для процедуры «Independent-Samples T-Test». Меню блока «Compare Means»

Выбор процедуры. Выбрать процедуру Independent-Samples T-Test в блоке Compare Means, как это показано на рис. 5.7.

Заполнение полей ввода данных. Окно ввода данных и параметров этой процедуры показано на рис. 5.8. В нем перенесите переменную height в поле Test Variable, а переменную sex — в поле Grouping Variable. Кнопкой **Define Groups** вызовите окно, в котором определите значение 0 для первой группы Group 1 и значение 1 — для второй группы Group 2. Кнопка **Option** позволяет скорректировать уровень доверия для доверительного интервала.

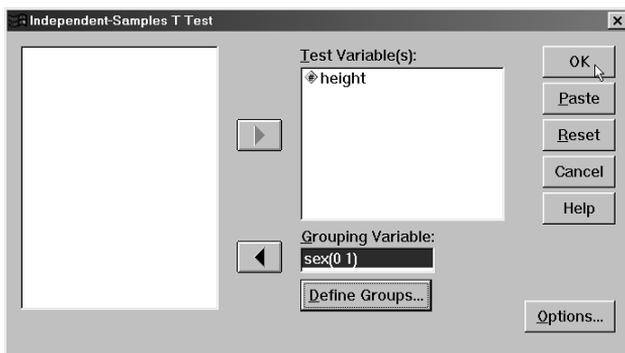


Рис. 5.8. Пакет SPSS. Окно ввода данных и параметров процедуры «Independent-Samples T-Test»

Group Statistics

	SEX	N	Mean	Std. Deviation	Std. Error Mean
HEIGHT	0	53	164,23	5,18	,71
	1	20	175,00	6,46	1,45

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
HEIGHT	Equal variances assumed	2,936	,091	-7,397	71	,000	-10,77	1,46	-13,68	-7,87
	Equal variances not assumed			-6,688	28,693	,000	-10,77	1,61	-14,07	-7,48

Рис. 5.9. Пакет SPSS. Результаты расчетов процедуры «Independent-Samples T-Test»

Результаты. Выдача результатов включает две таблицы, представленные на рис. 5.9.

Таблица **Group Statistics** (статистики для групп) полностью аналогична таблице **One-Sample Statistics**, разобранный в примере 5.1к, но только включает показатели для двух выборок.

Таблица **Independent Samples Test** содержит результаты для двух разных способов расчета. Первый — предполагает равенство дисперсий выборок **Equal variances assumed**, а второй — отсутствие этого условия **Equal variances not assumed**. Для подсказки, какими результатами воспользоваться, в таблице приводится значение F-статистики и ее минимальный уровень значимости для проверки гипотезы о равенстве дисперсий выборок (**Levene's Test for Equality of Variances**).

Расчеты процедуры включают t-статистику Стьюдента **t**, число степеней свободы **df**, минимальный уровень значимости статистики против двусторонних альтернатив **Sig. (2-tailed)**, разность средних значений выбо-

рок **Mean Difference**, стандартную ошибку разности **Std. Error Difference** и 95% доверительный интервал для разности.

Обратим внимание, что оба 95% доверительных интервала для разности в таблице **Independent Samples Test** не включают значение 0, т.е. гипотеза о равенстве средних значений при этом уровне значимости может быть отвергнута. О том же свидетельствуют уровни значимости *t*-критерия Стьюдента для каждого из способов расчетов.

Минимальный уровень значимости критерия Левена показывает, что у нас нет оснований отвергнуть гипотезу о равенстве дисперсий выборок.

Комментарий. Существуют различные критерии для проверки гипотезы о равенстве дисперсий нормальных выборок. Описанный в п. 5.4.2 критерий Фишера весьма чувствителен к нарушениям нормального закона распределения выборок и, в частности, к наличию в выборке нехарактерных или аномальных значений. Критерий Левена более устойчив к нарушениям начальных предположений о выборках. В его основе лежит вычисление модулей отклонений значений выборки от выборочных средних и дальнейший дисперсионный анализ этих данных.

Дополнительная литература

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика. Основы эконометрики. Т. 1. Теория вероятностей и прикладная статистика. — М.: Юнити-Дана, 2001. — 656 с.

2. Бикел П., Доксум К. Математическая статистика. — М.: Финансы и статистика, 1983. Вып. 1. — 280 с.; Вып. 2. — 254 с.

3. Бородин А.Н. Элементарный курс теории вероятностей и математической статистики. — СПб.: Лань, 2005. — 256 с.

4. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. — М.: ФИЗМАТЛИТ, 2006. — 816 с.

5. Справочник по прикладной статистике: в 2 т.; под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. — М.: Финансы и статистика, 1989, 1990.

Однофакторный анализ

6.1. Постановка задачи

Задача 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 (называемые задачами *однофакторного анализа*) весьма часто встречаются на практике. Типичный пример — сравнение по достигаемым результатам нескольких различных способов действия, направленных на достижение одной цели, скажем, нескольких школьных учебников или нескольких лекарств.

Терминология. Для описания задач однофакторного анализа установилась следующая терминология:

- то, что, как мы считаем, должно оказывать влияние на конечный результат, называют *фактором* или *факторами*, если их несколько (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*.

Заметим, что термин «способ обработки» часто имеет прямое толкование: например, если фактором является агротехнический прием, то он может быть способом обработки почвы (химическими удобрениями, мелиоративной обработки и т.п.). В дальнейшем для единообразия будем говорить о сравнении нескольких способов обработки.

Данные. Для сравнения влияния факторов на результат необходим определенный статистический материал. Обычно его получают следующим образом: каждый из k способов обработки применяют несколько раз (не обязательно одно и то же число раз) к исследуемому объекту и регистрируют результаты. Итогом подобных испытаний являются k выборок, вообще говоря, разных объемов (численностей).

Наиболее распространенным и удобным способом представления подобных данных является таблица (см. табл. 6.1). В зависимости от количества влияющих факторов (в данном случае фактор один), говорят, что данные сведены в таблицу с одним, двумя и т.д. входами.

Таблица 6.1

Обработки (соответствуют уровням фактора)	1	2	...	k
Результаты измерений	x_{11}	x_{12}	...	x_{1k}
	x_{21}	x_{22}	...	x_{2k}
	\vdots	\vdots		\vdots
	$x_{n_1 1}$	$x_{n_2 2}$...	$x_{n_k k}$

Здесь n_1, \dots, n_k — объемы выборок, $N = n_1 + n_2 + \dots + n_k$ — общее число наблюдений.

Статистические предположения. Наше отношение к полученным значениям x_{ij} может быть различно по нескольким причинам. Во-первых, оно зависит от того, в какой шкале проведены эти измерения. (Этот вопрос подробно разбирается в гл. 9.) Во-вторых, можно делать различные предположения о характере случайной изменчивости наблюдений x_{ij} — об их законе распределения и его зависимости от различных способов обработки.

Как уже отмечалось при анализе двухвыборочных задач в п. 3.5, опыт показывает, что при изменении способа обработки наибольшей изменчивости в первую очередь, как правило, подвержено положение случайной величины, которое можно характеризовать медианой или средним значением. Следуя этому эмпирическому правилу, в однофакторных задачах также обычно предполагают, что все наблюдения принадлежат некоторому *сдвиговому семейству распределений*. Часто в качестве такого семейства рассматривается семейство нормальных распределений и для обработки данных применяются методы *дисперсионного анализа* (см. п. 6.5). В других случаях предположение о нормальности распределений не является правомерным, и тогда используют различные непараметрические методы анализа, из которых наиболее разработаны ранговые методы (см. п. 6.2—6.4).

Указанные выше моменты приводят к различным постановкам задач однофакторного анализа, однако общая стратегия анализа во всех случаях примерно одинакова.

Стратегия анализа и возможные результаты. Одной из главных конечных целей в задачах однофакторного анализа является оценка величины влияния конкретного способа обработки на изучаемый отклик. Эта задача также может быть сформулирована в форме сравнения

влияния двух или нескольких способов обработки между собой, т.е. оценки различия (в статистике говорят — *контраста*) между действием различных уровней фактора. Так, сравнивая влияние нескольких агротехнических приемов обработки почвы на урожайность, нас может интересовать не сама величина урожайности (которая зависит еще и от погодных условий), а только насколько она больше или меньше для разных способов обработки почвы.

Но прежде чем судить о количественном влиянии фактора на изменяемый признак, полезно спросить себя, есть ли такое влияние вообще. Нельзя ли объяснить расхождения наблюдаемых в опыте значений для разных уровней фактора действием чистой случайности? Ведь внутренне присущая явлению изменчивость уже привела к тому, что результаты оказываются различными даже при неизменном значении фактора (т.е. в каждом столбце табл. 6.1). Может быть, той же причиной можно объяснить и различие между ее столбцами? На статистическом языке это предположение означает, что все данные табл. 6.1 принадлежат одному и тому же распределению. Это предположение обычно именуют *нулевой гипотезой* и обозначают H_0 . Для проверки нулевой гипотезы могут быть использованы различные критерии: как традиционные, опирающиеся на предположение о нормальности распределения данных (F -отношение), так и непараметрические, не требующие подобных допущений (ранговые критерии Краскела—Уоллиса, Джонкхиера и др.).

Если нулевая гипотеза об отсутствии эффектов обработки отвергается, то проводится оценка действия этих эффектов или контрастов между ними и строятся доверительные интервалы для этих характеристик. На этом этапе наибольший интерес представляет вопрос точности и достоверности полученных оценок. Здесь также можно строить оценки, основанные на предположении о нормальности распределения исходных данных и свободные от этого допущения. На практике целесообразно вычислить и те и другие оценки, а при заметном отличии этих оценок между собой предпочтение следует отдавать непараметрическим оценкам, как более надежным.

Если же критерии не позволяют отвергнуть нулевую гипотезу об отсутствии эффектов обработки, то обычно на этом анализ может быть завершен. Но иногда вывод об отсутствии эффектов обработки нас не может устроить, так как он противоречит теоретическим предпосылкам или результатам предыдущих исследований. Тогда следует выяснить, нет ли каких-либо еще факторов, влияющих на имеющиеся наблюдения. Может быть, влияние эффекта обработки не удалось обнаружить лишь потому, что его влияние незаметно на фоне различий, вызванных действием не учтенного нами фактора. Например, при изучении влияния

способов обработки почвы на урожайность таким фактором может быть тип почвы. В гл. 7 мы расскажем о методах двухфакторного анализа, используемых для решения задач, в которых на конечный результат влияют не один, а два фактора.

Кроме того, может быть полезно последовательно проводить сравнение между собой только двух способов обработки с помощью методов, описанных в гл. 3 и 5. Этот процесс может показать, что, наряду со способами обработки, различия между влияниями которых статистически незначимы, могут быть выявлены и значимо отличающиеся уровни факторов. Это может помочь по-новому сформулировать задачу, объединив несколько способов обработки между собой.

Углубленный анализ. После выполнения однофакторного анализа может быть полезно провести углубленное исследование его результатов. При этом могут ставиться две цели.

1. Проверка корректности применения использованного метода анализа. Например, может проверяться предположение об одинаковом разбросе (дисперсии) наблюдений при разных способах обработки. Об используемых для этого критериях мы упоминаем в п. 6.7.2. А при применении методов, основанных на предположении о нормальности распределения данных, может быть проведено исследование нормальности остатков (т.е. данных, из которых вычтен эффект обработки). Если предположение о нормальности остатков вызовет сильное сомнение, следует использовать ранговые или знаковые процедуры анализа данных.

2. Выделение однородных по воздействию методов обработки — с его помощью можно разбить все способы обработки на однородные (гомогенные) группы. Мы расскажем о методах решения этой задачи в п. 6.7.1.

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 отношения «больше—меньше» между наблюдениями, так как они не зависят от распределения наблюдений. При этом вся информация, которую мы используем из табл. 6.1, содержится в тех *рангах*, что получают числа x_{ij} при упорядочении всей их 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* (см. гл. 9), 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 (и потому нет трудностей

в назначении рангов). При наличии совпадений (и использовании средних рангов) теоретическая схема действует как приближенная, а надежность ее выводов снижается тем больше, чем больше совпадений. Ниже мы укажем, какие поправки делаются при наличии совпадений.

Упорядочим величины x_{ij} (все равно как — от большего к меньшему либо от меньшего к большему). Обозначим через r_{ij} ранг числа x_{ij} во всей совокупности. Тогда табл. 6.1 преобразуется в табл. 6.2. Важно отметить, что при выполнении гипотезы H_0 любые возможные расположения рангов по местам в табл. 6.2 равновероятны.

Таблица 6.2

Обработки	1	2	...	k
Ранги результатов измерений	r_{11}	r_{12}	...	r_{1k}
	r_{21}	r_{22}	...	r_{2k}
	⋮	⋮		⋮
	$r_{n_1 1}$	$r_{n_2 2}$...	$r_{n_k k}$

Согласно сформулированной стратегии анализа возникает вопрос: нельзя ли объяснить наблюдаемое в опыте расположение рангов в табл. 6.2 действием чистой случайности? Этот вопрос можно переформулировать в виде статистической гипотезы о том, что все k представленных выборок (столбцы табл. 6.1) однородны, т.е. являются выборками из одного и того же закона распределения. Наша задача — указать статистический критерий, с помощью которого можно было бы судить о справедливости выдвинутой гипотезы.

Общая методика проверки статистических гипотез (см. п. 3.2) рекомендует нам сконструировать некоторую статистику, т.е. в данном случае функцию от рангов r_{ij} , которая бы легла в основу критерия проверки гипотезы. Основное требование к этой статистике следующее: ее распределение при гипотезе H_0 должно заметно отличаться от ее распределения при альтернативах. Последние слова подчеркивают, что статистический критерий для проверки H_0 должен быть направлен против определенной совокупности альтернатив.

Как уже отмечалось, все реализации табл. 6.2 равновероятны при H_0 . Это дает возможность рассчитать закон распределения при H_0 любой ранговой статистики (насколько это позволяют компьютерные средства).

Ниже будут разобраны два ранговых критерия проверки однородности, направленные против различных совокупностей альтернатив (п. 6.2.1 и 6.2.2). Построение непараметрических оценок эффектов обработки изложено в п. 6.4. Параметрические методы (дисперсионный однофакторный анализ) описаны в п. 6.5 и 6.6.

6.2. Непараметрические критерии проверки однородности

6.2.1. Критерий Краскела–Уоллиса (произвольные альтернативы)

Если мы не можем сказать что-либо определенное об альтернативах к H_0 , можно воспользоваться для ее проверки свободным от распределения критерием Краскела–Уоллиса. Для этого заменим наблюдения x_{ij} их рангами r_{ij} , упорядочивая всю совокупность $\|x_{ij}\|$ в порядке возрастания (для определенности). Затем для каждой обработки j (т.е. для каждого столбца исходной таблицы) надо вычислить

$$R_j = \sum_{i=1}^{n_j} r_{ij} \quad \text{и} \quad R_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} r_{ij},$$

где $R_{.j}$ — это средний ранг, рассчитанный по столбцу. Если между столбцами нет систематических различий, средние ранги $R_{.j}$, $j = 1, \dots, k$ не должны значительно отличаться от среднего ранга, рассчитанного по всей совокупности $\|r_{ij}\|$. Ясно, что последний равен $(N + 1)/2$. Поэтому величины

$$\left(R_{.1} - \frac{N + 1}{2}\right)^2, \dots, \left(R_{.k} - \frac{N + 1}{2}\right)^2$$

при H_0 в совокупности должны быть небольшими. Составляя общую характеристику, разумно учесть различия в числе наблюдений для разных обработок и взять в качестве меры отступления от чистой случайности величину

$$H = \frac{12}{N(N + 1)} \sum_{j=1}^k n_j \left(R_{.j} - \frac{N + 1}{2}\right)^2. \quad (6.1)$$

Эта величина называется *статистикой Краскела–Уоллиса*. Множитель $12/[N(N + 1)]$ нужен для стабилизации ее распределения при большом числе наблюдений (см. ниже). Другая форма для вычисления H :

$$H = \frac{12}{N(N + 1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N + 1). \quad (6.2)$$

Таблицы и асимптотика. Небольшие таблицы распределения статистики H при гипотезе H_0 можно найти в сборниках статистических таблиц. При больших объемах n_1, \dots, n_k , которые находятся за пределами таблиц, случайная величина H (при гипотезе H_0) приближенно

распределена как хи-квадрат с $(k - 1)$ степенями свободы (сведения о более точной аппроксимации можно найти в [65]). Так что при использовании этого приближения мы отвергаем H_0 (на уровне значимости α), если $H_{\text{набл.}} > \chi_{1-\alpha}^2$, где $\chi_{1-\alpha}^2$ — квантиль уровня $(1 - \alpha)$ распределения хи-квадрат с $(k - 1)$ степенями свободы.

Совпадающие значения. Если в табл. 6.1 есть совпадающие значения, надо при ранжировании и переходе к табл. 6.2 использовать средние ранги. Если совпадений много, рекомендуют использовать модифицированную форму статистики H' :

$$H' = \frac{H}{1 - \left(\sum_{j=1}^g T_j / [N^3 - N] \right)}, \quad (6.3)$$

где g — число групп совпадающих наблюдений, $T_j = (t_j^3 - t_j)$, t_j — число совпадающих наблюдений в группе с номером j . Более подробные сведения по этому поводу можно найти, например, в [115].

Замечание. При $k = 2$ статистика Краскела–Уоллиса H по своему действию эквивалентна статистике Уилкоксона W .

6.2.2. Критерий Джонкхиера (альтернативы с упорядочением)

Нередко исследователю заранее известно, что имеющиеся группы результатов упорядочены по возрастанию влияния фактора. Пусть, для определенности, первый столбец табл. 6.1 отвечает наименьшему уровню фактора, последний — наибольшему, а промежуточные столбцы получили номера, соответствующие их положению. В таких случаях можно использовать критерий Джонкхиера, более чувствительный (более мощный) против альтернатив об упорядоченном влиянии фактора. Разумеется, против других альтернатив свойства этого критерия могут оказаться хуже свойств критерия Краскела–Уоллиса.

Статистика Джонкхиера. Разберем сначала, как устроена статистика этого критерия в случае, когда сравниваются только два способа обработки. Таблица 6.1 в этом случае имеет два столбца. Фактически здесь речь идет о проверке однородности двух выборок. Напомним, что в гл. 3 для решения этой задачи была предложена статистика Манна–Уитни. А именно: пусть x_1, \dots, x_m и y_1, \dots, y_n — две выборки. Положим:

$$\varphi(x_i, y_j) = \begin{cases} 1, & \text{если } x_i < y_j; \\ 1/2, & \text{если } x_i = y_j; \\ 0, & \text{если } x_i > y_j. \end{cases}$$

Статистикой Манна–Уитни называют величину

$$U = \sum_{\substack{i=1, \dots, m \\ j=1, \dots, n}} \varphi(x, y).$$

Обратившись теперь к общему случаю, когда сравниваются k способов обработки, поступим следующим образом. Для каждой пары натуральных чисел u и v , где $1 \leq u < v \leq k$, составляем по выборкам с номерами u, v статистику Манна–Уитни.

$$U_{u,v} = \sum_{\substack{i=1, \dots, m_u \\ j=1, \dots, n_v}} \varphi(x_{iu}, y_{jv}).$$

Определим статистику Джонкхиера J как

$$J = \sum_{1 \leq u < v \leq k} U_{u,v}.$$

Свидетельством в пользу альтернативы упорядоченности эффектов (против гипотезы однородности) служат большие значения статистики J , полученные в эксперименте.

Таблицы и аппроксимация. При небольших объемах выборок и небольшом k распределение статистики J табулировано (см., например, [115]). Для больших выборок в отношении J действует нормальная аппроксимация: $J \stackrel{\text{ac}}{\approx} N(MJ, DJ)$, где MJ и DJ равны:

$$MJ = \frac{1}{4} \left(N^2 - \sum_{j=1}^k n_j^2 \right), \quad DJ = \frac{1}{72} \left[N^2(2N + 3) - \sum_{j=1}^k n_j^2(2n_j + 3) \right].$$

Свидетельством против гипотезы однородности служат большие (сравнительно с процентными точками стандартного нормального распределения) значения статистики $(J - MJ)/\sqrt{DJ}$, полученные в эксперименте (сведения о более точной аппроксимации можно найти в [65]).

6.3. Практический пример

Проиллюстрируем применение описанных выше критериев на следующем примере. Для выяснения влияния денежного стимулирования на производительность труда шести однородным группам из пяти человек каждая были предложены задачи одинаковой трудности. Задачи предлагались каждому испытуемому независимо от всех остальных. Группы отличаются между собой величиной денежного вознаграждения за решаемую задачу. В следующей таблице приведено число решенных задач членами каждой группы. Данные приведены из [33].

Таблица 6.3

Величина вознаграждения (от меньшей к большей)

группа 1	группа 2	группа 3	группа 4	группа 5	группа 6
10	8	12	12	24	19
11	10	17	15	16	18
9	16	14	16	22	27
13	13	9	16	18	25
7	12	16	19	20	24

Проверим гипотезу об отсутствии влияния денежного вознаграждения на число решенных задач. Отметим, что величины, приведенные в таблице, имеют смысл и сами по себе, а не только в сравнении с другими величинами. Это широко распространенная ситуация, в которой также часто целесообразно применять ранговые критерии Краскела–Уоллеса или Джонкхиера, хотя при переходе от величины x_{ij} к их рангам уже происходит определенная потеря информации. Однако часто подобная потеря информации, во-первых, не столь значительна, а во-вторых, компенсируется тем, что от обычно неизвестного закона распределения величин x_{ij} мы переходим к величинам r_{ij} , распределение которых при гипотезе H_0 известно. Если же мы можем полагать, что величины x_{ij} имеют нормальный (гауссовский) закон распределения, для их исследования можно применить методы дисперсионного анализа, рассматриваемые ниже в п. 6.5 и 6.6.

Применение критерия Краскела–Уоллеса. В связи с наличием в табл. 6.3 совпадений мы будем вынуждены воспользоваться средними рангами. Так, значение $x_{ij} = 10$ встречается в табл. 6.3 дважды, и при упорядочении x_{ij} оно «делит пятое и шестое места». Поэтому средний ранг $x_{ij} = 10$ равен 5.5. В результате ранжирования получим табл. 6.4. В двух нижних строках приведены суммы рангов R_j и средние ранги $R_{.j} = R_j/n_j$ по столбцам.

Таблица 6.4

Таблица рангов наблюдений

группа 1	группа 2	группа 3	группа 4	группа 5	группа 6
5,5	2	9	9	27,5	23,5
7	5,5	20	14	17	21,5
3,5	17	13	17	26	30
11,5	11,5	3,5	17	21,5	29
1	9	17	23,5	25	27,5
$R_1 = 28,5$	$R_2 = 45$	$R_3 = 62,5$	$R_4 = 80,5$	$R_5 = 117$	$R_6 = 131,5$
$R_{.1} = 5,7$	$R_{.2} = 9$	$R_{.3} = 12,5$	$R_{.4} = 16,1$	$R_{.5} = 23,4$	$R_{.6} = 26,3$

Для вычисления статистики Краскела–Уоллиса H удобнее воспользоваться формулой (6.2). В нашем случае общее число наблюдений $N = 30$, число наблюдений при заданном значении фактора $n_j = 5$, $j = 1, \dots, 6$. Подставляя эти значения, получаем: $H = 17682/155 - 93 = 21.077$.

Как было указано, величина H асимптотически имеет распределение χ^2 с числом степеней свободы, равным в данном случае 5. По таблице распределения χ^2 находим, что минимальный уровень значимости α чуть больше 0.001. Заметим, что этот вывод является приближенным в связи с тем, что в табл. 6.3 было определенное число совпадений наблюдений x_{ij} . Для учета влияния связей можно воспользоваться статистикой H' (6.3). В нашем случае имеем следующие восемь групп совпадающих наблюдений:

9, 9; 10, 10; 12, 12; 13, 13; 16, 16, 16, 16, 16; 18, 18; 19, 19; 24, 24.

Соответственно: $T_1 = (2^3 - 2) = 6$, $T_2 = (2^3 - 2) = 6$, $T_3 = (3^3 - 3) = 24$, $T_4 = 6$, $T_5 = (5^3 - 5) = 120$, $T_6 = 6$, $T_7 = 6$, $T_8 = 6$. Знаменатель дроби в выражении для H' равен: $1 - \sum_{j=1}^8 T_j / (30^3 - 30) = 1 - 6/899$, а само значение H' приблизительно равно 21.2186.

Так как скорректированное значение H' статистики Краскела–Уоллиса несущественно отличается от значения H , мы можем отвергнуть гипотезу на минимальном уровне значимости около 0.001.

Применение критерия Джонкхиера. Заметим, что в данном примере можно предположить монотонное влияние материального стимулирования на результаты, а поэтому оправданно применение критерия Джонкхиера. Итак, выберем в качестве альтернативы к нулевой гипотезе предположение, что чем выше уровень стимулирования, тем выше производительность. Для вычисления статистики Джонкхиера J найдем значения статистики Манна–Уитни U для всех комбинаций индексов u и v , где u и v меняются от 1 до 6, причем $u < v$. Простой расчет дает:

$$\begin{aligned} U_{12} = 17 & \quad U_{23} = 17 & \quad U_{34} = 16.5 & \quad U_{45} = 22 \\ U_{13} = 18.5 & \quad U_{24} = 20.5 & \quad U_{35} = 23,5 & \quad U_{46} = 23,5 \\ U_{14} = 24 & \quad U_{25} = 24.5 & \quad U_{36} = 25 & \quad U_{56} = 18 \\ U_{15} = 25 & \quad U_{26} = 25 \\ U_{16} = 25 \end{aligned}$$

Отсюда

$$J = \sum_{\substack{u=1, \dots, 6 \\ v=1, \dots, 6 \\ u < v}} U_{u,v} = 325.$$

Для нахождения минимального уровня значимости критерия воспользуемся нормальной аппроксимацией. Величина $J^* = (J - MJ)/\sqrt{DJ}$ асимптотически имеет стандартное нормальное распределение, где выражения для MJ и DJ были указаны выше. В результате расчетов получаем $MJ = 187.5$, $DJ = 27.5$. Следовательно, $J^* \simeq (325 - 187.5)/27.5 \simeq 5$. С помощью таблиц стандартного нормального распределения находим, что вычисленное значение соответствует минимальному уровню значимости $\alpha \simeq 3 \cdot 10^{-7}$. Заметим, что мы получили более сильный результат по сравнению с применением критерия Краскела–Уоллиса. Если в первом случае мы отвергали гипотезу об однородности на уровне значимости не менее $1 \cdot 10^{-3}$, то во втором случае минимальный уровень значимости понизился почти на 4 порядка.

Замечание. Оба критерия достаточно определенно отвергают гипотезу об однородности выборок. Однако для исследователя гораздо больший интерес представляет не сам факт существования влияния, а вопрос о количественном влиянии способа обработки на результаты. Ниже будет разобрана довольно распространенная модель аддитивного влияния фактора на отклик и построены оценки эффектов обработки.

6.4. Оценивание эффектов обработки (непараметрический подход)

Для описания данных табл. 6.1 в большинстве случаев оказывается приемлемой *аддитивная модель*. Она предполагает, что значение отклика x_{ij} можно представить в виде суммы вклада (воздействия) фактора и независимой от вкладов факторов случайной величины. Иначе говоря, каждое наблюдение x_{ij} является суммой вида:

$$x_{ij} = a_j + e_{ij}, \quad j = 1, \dots, k, \quad i = 1, \dots, n, \quad (6.4)$$

где a_1, a_2, \dots, a_k — неслучайные неизвестные величины, являющиеся результатом действия соответствующих обработок, e_{ij} — независимые одинаково распределенные случайные величины, отражающие внутренне присущую наблюдениям изменчивость. Случайные величины e_{ij} непосредственно не наблюдаемы, нам известны лишь значения x_{ij} .

Теоретически ясная картина получается в том случае, когда общий для всех e_{ij} закон распределения оказывается непрерывным (еще более точные выводы можно сделать, когда указанный закон распределения нормален — эту возможность мы рассмотрим отдельно в п. 6.5). На практике эти предпосылки не всегда соблюдаются. В таком случае и выводы становятся приближенными.

Для дальнейших рассуждений удобнее вместо a_j — влияния обработки j на результаты — рассматривать влияние обработки на от-

клонения x_{ij} от среднего уровня. Введем величину среднего уровня μ следующим образом:

$$\mu = \frac{1}{k} \sum_{i=1}^k a_i.$$

Будем называть величину $\tau_j = a_j - \mu$ отклонением от среднего уровня при j -й обработке. Ясно, что $\tau_1 + \tau_2 + \dots + \tau_k = 0$. Тогда $x_{ij} = a_j + e_{ij}$ можно записать в виде:

$$x_{ij} = \mu + \tau_j + e_{ij}, \quad j = 1, \dots, k, \quad i = 1, \dots, n.$$

Хотя в полученной модели имеется $k + 1$ параметров, общее количество независимых параметров не изменилось, так как $\sum_{i=1}^k \tau_i = 0$.

Теперь вопрос о различии обработок сводится к выяснению различий между τ_1, \dots, τ_k . Гипотеза об однородности данных означает равенства $a_1 = a_2 = \dots = a_k$, т.е. $\tau_1 = \tau_2 = \dots = \tau_k = 0$. Альтернатива об упорядоченности эффектов обработки превращается в $\tau_1 \leq \tau_2 \leq \dots \leq \tau_k$, а различие между эффектами i -й и j -й обработок, естественно, характеризуется величиной $a_i - a_j = \tau_i - \tau_j$.

Оценки сдвига. Рассмотрим сначала на примере построение простейших оценок различия между эффектами обработки двух выборок. Заметим, речь в этом случае идет о сдвиге одной выборки относительно другой. В качестве оценки этого сдвига можно взять *медиану Ходжеса–Лемана*, т.е. величину z_{ij} :

$$z_{ij} = \text{med}(x_{ui} - x_{vj}, u = 1, \dots, n_i, v = 1, \dots, n_j).$$

Отметим, что $z_{ij} = -z_{ij}$. Статистика z_{ij} может служить оценкой величины $\tau_i - \tau_j$, однако у нее есть существенный недостаток. Проиллюстрируем его на описанном выше примере о влиянии материального стимулирования на производительность. Вычислим величины z_{14} , z_{46} , z_{16} . Так, z_{14} является медианой 25 разностей значений 1-го и 4-го столбцов табл. 6.3. После простых подсчетов получим $z_{14} = -6$, $z_{46} = -8$ и $z_{16} = -13$. Заметим, что сдвиг первой выборки относительно шестой можно представить в виде суммы сдвигов первой выборки относительно четвертой и четвертой относительно шестой. Действительно, $\tau_1 - \tau_6 = (\tau_1 - \tau_4) + (\tau_4 - \tau_6)$. Поэтому естественно было бы ожидать, что аналогичное равенство будет выполняться и для оценок сдвига. Однако оценки z_{ij} этому разумному требованию не удовлетворяют. Так, $z_{14} + z_{46} \neq z_{16}$. Поэтому оценки z_{ij} часто используют в скорректированном варианте.

Скорректированные оценки сдвига. Введем величину

$$\overline{\Delta}_i = \frac{\sum_{u=1}^k n_u z_{iu}}{N}, \quad i = 1, \dots, k,$$

где $z_{ii} = 0$, $i = 1, \dots, k$. $\overline{\Delta}_i$ отражает сдвиг выборки i относительно всех остальных выборок, усредненный с весами n_1, \dots, n_k .

Будем называть взвешенной скорректированной оценкой величины $\tau_i - \tau_j$ величину $W_{ij} = \overline{\Delta}_i - \overline{\Delta}_j$. Ее также называют оценкой Спетволля. Исходную оценку z_{ij} при этом называют нескорректированной оценкой $\tau_i - \tau_j$. Отметим, что оценки W_{ij} удовлетворяют соотношению

$$W_{ij} + W_{jh} = W_{ih}$$

для всех i, j, h от 1 до k . Однако у оценок Спетволля есть свой недостаток: оценка сдвига одной выборки относительно другой зависит от всех остальных выборок.

Вычислим, например, оценку W_{14} величины $\tau_1 - \tau_4$ в рассмотренной выше задаче. Для этого нам необходимо прежде всего знать значения оценок z_{1u} и z_{4v} при всех u и v , изменяющихся от 1 до k . Для нашего примера имеем:

$$\begin{aligned} z_{11} = 0, \quad z_{12} = -2, \quad z_{13} = -4, \quad z_{14} = -6, \quad z_{15} = -10, \quad z_{16} = -13, \\ z_{41} = 6, \quad z_{42} = 4, \quad z_{43} = 2, \quad z_{44} = 0, \quad z_{45} = -4, \quad z_{46} = -8. \end{aligned}$$

Таким образом,

$$\begin{aligned} \overline{\Delta}_1 &= \frac{5}{30} - (z_{11} + z_{12} + z_{13} + z_{14} + z_{15} + z_{16}) = -5\frac{5}{6}, \\ \overline{\Delta}_4 &= \frac{5}{30} - (z_{41} + z_{42} + z_{43} + z_{44} + z_{45} + z_{46}) = 0, \quad W_{14} = \overline{\Delta}_1 - \overline{\Delta}_4 = -5\frac{5}{6}. \end{aligned}$$

Контрасты. Довольно часто в задачах однофакторного анализа представляют интерес не сами оценки величин τ_i , а некоторые их линейные комбинации. Для их определения вводится понятие *контраста*. Контрастом параметров τ в модели аддитивного влияния фактора на отклик называется величина θ :

$$\theta = \sum_{j=1}^k c_j \tau_j,$$

где $\sum_{j=1}^k c_j = 0$ и c_1, \dots, c_k — заданные константы. Ясно, что разность $\tau_i - \tau_j$ является простейшим примером контраста, когда $c_i = 1$, $c_j = -1$, $c_u = 0$ при всех u , не равных i и j .

Чаще бывает удобно задавать θ в другой, эквивалентной форме, а именно

$$\theta = \sum_{i=1}^k \sum_{j=1}^k d_{ij} (\tau_i - \tau_j),$$

где $d_{ij} = c_i/k$ при $j = 1, \dots, k$, $i = 1, \dots, k$. Учитывая построенные выше взвешенные скорректированные оценки W_{ij} для разностей $\tau_i - \tau_j$, естественно определить оценку контраста θ как

$$\theta^* = \sum_{i=1}^k \sum_{j=1}^k d_{ij} W_{ij}.$$

Сведения о свойствах оценок θ^* и W_{ij} можно найти в [115].

6.5. Дисперсионный анализ

До сих пор, рассматривая аддитивную модель однофакторного анализа (6.4): $x_{ij} = a_j + e_{ij}$, мы предполагали только непрерывность закона распределения величин e_{ij} , при том, что e_{ij} — независимы и одинаково распределены. Часто о распределении e_{ij} можно сказать больше, а именно, величины $e_{ij} \sim N(0, \sigma^2)$, т.е. имеют нормальное распределение с нулевым средним и общей для всех дисперсией σ^2 , которая нам неизвестна. Дополнительная информация о законе распределения случайных величин e_{ij} позволяет использовать более сильные методы в модели однофакторного анализа как для проверки гипотез, так и для оценки параметров. Совокупность этих методов носит название *однофакторного дисперсионного анализа*.

Это название связано с тем, что анализ модели (6.4) основан на сопоставлении двух оценок дисперсии σ^2 . Одна из них действует вне зависимости от того, верна или нет гипотеза $H_0 : a_1 = \dots = a_k$. Другая оценка существенно использует это предположение. Она дает близкий к σ^2 результат только в том случае, если гипотеза верна. Сопоставляя друг с другом эти две оценки, мы можем заключить, что H_0 следует отвергнуть, если они оказываются заметно (значимо) различны. Реализация и уточнение этой идеи и будут осуществлены далее.

Построение оценок дисперсии. Вспомнив известное нам о статистической обработке одной нормальной выборки, мы можем сказать, что каждая однородная группа табл. 6.1 (каждый ее столбец) дает оценку σ^2 . Для этого надо по каждому столбцу найти выборочную сумму квадратов отклонений от среднего арифметического. Положим,

$$x_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}, \quad j = 1, \dots, k,$$

и далее вычислим $\sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2$. Анализируя одну нормальную выборку, мы нашли, что такую сумму квадратов можно представить в виде произведения $\sigma^2 \chi^2$, где случайная величина χ^2 имеет распределение χ^2

с $n_j - 1$ степенями свободы. Поскольку данные в разных столбцах получены независимо, объединенная сумма квадратов $\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2$ имеет распределение $\sigma^2 \chi^2$ с $N - k$ степенями свободы. Отсюда получаем первую (основную) оценку σ^2 :

$$\sigma^{2*} = \frac{1}{N - k} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2. \quad (6.5)$$

При выводе не было упоминания о гипотезе H_0 , следовательно, $\sigma^{2*} \approx \sigma^2$ независимо от того, верна гипотеза H_0 или нет.

Чтобы получить другую оценку σ^2 , обратимся вновь к столбцам табл. 6.1, точнее — к их средним значениям $x_{.j}$. Согласно свойствам нормального распределения,

$$x_{.j} \sim N(a_j, \sigma^2/n_j). \quad (6.6)$$

Кроме того, $x_{.j}$ и $\sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2$ статистически независимы. Найдем центр совокупности (6.6) с учетом «весов» средних значений n_j , т.е. найдем, при каком z достигается минимум выражения

$$\sum_{j=1}^k (x_{.j} - z)^2 n_j \rightarrow \min_z. \quad (6.7)$$

С помощью стандартных средств математического анализа легко видеть, что минимум (6.7) достигается при $z = \bar{x}$, где

$$\bar{x} = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij}. \quad (6.8)$$

Заметим, что при выполнении гипотезы H_0 значение выражения (6.7) при $z = \bar{x}$ имеет распределение $\sigma^2 \chi^2(k - 1)$, где $\chi^2(k - 1)$ — распределение хи-квадрат с $(k - 1)$ степенями свободы. Отсюда находим вторую оценку для σ^2 :

$$\sigma^{2**} = \frac{1}{k - 1} \sum_{j=1}^k n_j (x_{.j} - \bar{x})^2. \quad (6.9)$$

Поскольку, как было отмечено, случайные величины $x_{.j}$ независимы от (6.5), то же верно и для их комбинаций. Поэтому оценка (6.9) является независимой от (6.5).

При нарушении H_0 оценка σ^{2**} имеет тенденцию к возрастанию, тем большему, чем больше отклонение от H_0 . Можно показать, что распределение оценки (6.9) — это так называемое нецентральное распределение хи-квадрат с $k - 1$ степенями свободы и параметром нецентральности $\frac{1}{k-1} \sum_{j=1}^k n_j (a_j - \bar{a})^2$.

Замечание. Нецентральное распределение хи-квадрат с k степенями свободы имеет сумма квадратов k независимых нормальных величин с единичной дисперсией и не обязательно нулевым средним. Параметр нецентральности в этом случае — сумма квадратов средних этих нормальных величин.

F -отношение. Поскольку мы имеем для оценки σ^2 две независимые оценки, имеющие при гипотезе H_0 распределение хи-квадрат, их частное $F = \sigma^{2**}/\sigma^{2*}$, или, подробнее,

$$F = \frac{\frac{1}{k-1} \sum_{j=1}^k n_j (x_{.j} - \bar{x})^2}{\frac{1}{N-k} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2}, \quad (6.10)$$

должно иметь F -распределение с $(k-1, N-k)$ степенями свободы. Заметим, что статистика (6.10) уже не зависит от σ^2 . Как следует из обсуждения свойств σ^{2**} , дробь (6.10) получает тем большую тенденцию к возрастанию, чем сильнее нарушается гипотеза H_0 . Поэтому против H_0 говорят большие (неправдоподобно большие) значения F , рассчитанные по наблюдениям, далее — $F_{\text{набл.}}$. Следовательно, для проверки H_0 надо было бы вычислить $P(F \geq F_{\text{набл.}})$, т.е. вероятность получить за счет действия случайности значение статистики F большее или равное $F_{\text{набл.}}$. Гипотезу H_0 следует отвергнуть, если вероятность $P(F \geq F_{\text{набл.}})$ — мала. К сожалению, мы не располагаем столь подробными таблицами F -распределения, в них приводятся только процентные точки. Поэтому вместо вычисления $P(F \geq F_{\text{набл.}})$ приходится сравнивать $F_{\text{набл.}}$ с соответствующими α процентными точками.

6.6. Оценивание эффектов обработки в нормальной модели

6.6.1. Доверительные интервалы

Если гипотеза H_0 оказалась несовместимой с наблюдениями, есть основания для обсуждения значений параметров a_1, \dots, a_k . Ранее мы уже видели, что их оценками могут служить внутригрупповые средние $x_{.j}$, которые имеют распределения $N(a_j, \sigma^2/n_j)$ и статистически независимы от оценки дисперсии σ^{2*} (6.5). Поэтому отношение

$$t = \frac{x_{.j} - a_j}{\sigma^*} \sqrt{n_j} \quad (6.11)$$

подчиняется распределению Стьюдента с $N-k$ степенями свободы. С помощью (6.11) можно указать доверительный интервал для a_j с

произвольным коэффициентом доверия $1 - 2\alpha$:

$$P \left\{ \left| \sqrt{n_j} \frac{x_{.j} - a_j}{\sigma^*} \right| < t_{1-\alpha} \right\} = 1 - 2\alpha.$$

Здесь $t_{1-\alpha}$ — квантиль уровня $(1 - \alpha)$, соответствующего распределению Стьюдента. Отсюда получаем доверительный вывод об a_j (с коэффициентом доверия $1 - 2\alpha$):

$$|x_{.j} - a_j| < \frac{\sigma^*}{\sqrt{n_j}} t_{1-\alpha}. \quad (6.12)$$

Доверительные интервалы для контрастов. Можно указать доверительный интервал также и для любой линейной комбинации $\theta = \sum_{j=1}^k c_j a_j$, где c_1, \dots, c_k — произвольные коэффициенты. В частности, нередко приходится обращаться к сравнениям групп попарно, т.е. к разностям $a_j - a_l$, ($j, l = 1, \dots, k$). В любом случае стьюдентово отношение (с $N - k$ степенями свободы) имеет вид:

$$t = \frac{\theta^* - \theta}{\sigma^* \sqrt{\sum_{j=1}^k c_j^2 / n_j}}, \quad (6.13)$$

где $\theta^* = \sum_{j=1}^k c_j x_{.j}$. С помощью (6.13) доверительные суждения о различных θ получаем аналогично сказанному ранее:

$$|\theta^* - \theta| < \sigma^* \sqrt{\sum_{j=1}^k c_j^2 / n_j} t_{1-\alpha}. \quad (6.14)$$

6.6.2. Метод Шеффе множественных сравнений

Метод п. 6.6.1 не позволяет указать вероятность, с которой одновременно выполняются несколько неравенств типа (6.14). А задачи, в которых требуется нахождение такой вероятности, возникают достаточно часто. Например, это необходимо, когда требуется сравнить попарно все выборки, чтобы выделить все заведомо различные. Ниже мы расскажем об одном из методов (методе Шеффе), позволяющем получать совместные доверительные интервалы для контрастов.

Из отмеченных свойств групповых средних $x_{.j}$ следует, что случайная величина $\sum_{j=1}^k n_j (x_{.j} - a_j)^2$ имеет вид $\sigma^2 \chi^2(k)$, при этом она не зависит от σ^{2*} . Поэтому величина

$$F = \frac{\frac{1}{k} \sum_{j=1}^k n_j (x_{.j} - a_j)^2}{\sigma^{2*}}$$

имеет F -распределение (с k и $N - k$ степенями свободы).

Выбирая коэффициент доверия $1 - \alpha$ и соответствующую ему квантиль F -распределений $F_{1-\alpha}$, получим

$$P \left\{ \sum_{j=1}^k n_j (a_j - x_j)^2 < k\sigma^{2*} F_{1-\alpha} \right\} = 1 - \alpha. \quad (6.15)$$

Множество точек $a = (a_1, \dots, a_k)$ k -мерного пространства, удовлетворяющих (6.15), образует эллипсоид с центром $a^* = (x_1, \dots, x_k)$. Проведем к нему необходимое нам число пар параллельных касательных плоскостей. Уравнение каждой пары таких плоскостей имеет вид:

$$\sum_{j=1}^k c_j (a_j - x_j) = \pm d. \quad (6.16)$$

Эти пары плоскостей, пересекаясь, выделяют в пространстве многогранное множество R , описанное вокруг эллипсоида. Как эллипсоид, так и R — случайные множества. Их размеры и центры зависят от статистик (x_1, \dots, x_k) и σ^{2*} . Истинное значение $a = (a_1, \dots, a_k)$, согласно определению, попадает в эллипсоид с вероятностью $1 - \alpha$. Ясно, что вероятность накрытия a многогранником R не ниже $1 - \alpha$.

Точка a находится внутри R в том и только в том случае, если для ее координат выполняются все соотношения

$$\sum_{j=1}^k c_j x_j - d < \sum_{j=1}^k c_j a_j < \sum_{j=1}^k c_j x_j + d$$

из выделенного выбора плоскостей (6.16). Если мы рассмотрим вообще все плоскости, многогранник превратится в эллипсоид. Остается для каждого $c = (c_1, \dots, c_k)$ определить соответствующее d (6.16). Такое $d > 0$ есть максимальное значение выражения $\sum_{j=1}^k c_j (a_j - x_j)$ при условии, что точка $a = (a_1, \dots, a_k)$ лежит на поверхности эллипсоида, т.е. удовлетворяет соотношению $\sum_j^{n_j} (a_j - x_j)^2 = k\sigma^{2*} F_{1-\alpha}$. Расчет дает

$$d(c_1, \dots, c_k) = k\sigma^{2*} F_{1-\alpha} \sum_{j=1}^k c_j^2 / n_j.$$

Вывод. Для любой совокупности векторов (c_1, \dots, c_k) вероятность одновременного выполнения всех неравенств

$$\left| \sum_{j=1}^k c_j (a_j - x_j) \right| < \sqrt{k\sigma^{2*} F_{1-\alpha}} \sqrt{\sum_{j=1}^k c_j^2 / n_j} \quad (6.17)$$

не меньше, чем $1 - \alpha$.

Правило (6.17) позволяет сделать вывод о всех интересующих нас контрастах одновременно. В частности, мы можем выделить среди разностей $a_j - a_l$ те, которые значимо отличаются от нуля (на выбранном уровне значимости). Тем самым мы получаем возможность не только быть уверенными в существовании различия между группами (что бывает, если мы отвергли H_0), но и указать значимо различающиеся выборки (методы обработки).

6.7. Однофакторный анализ в пакете SPSS

В пакете довольно широко отражены различные методы и модели многофакторного анализа. Доступ к ним осуществляется из разных пунктов меню **Analyze**. Процедуры, относящиеся к однофакторному анализу, есть в блоках **Compare Means**, **Nonparametric Tests**, **Descriptive Statistics** (процедура **Explore**).

Разберем работу наиболее простых и употребительных однофакторных процедур на рассмотренном выше примере.

Пример 6.1к. Проверим гипотезу с помощью критерия Краскела–Уоллиса об отсутствии эффектов обработки для данных о влиянии стимулирования на производительность труда (табл. 6.3).

Подготовка данных. Данные для однофакторного анализа в SPSS вводятся следующим образом. В редакторе базы данных создаются две переменные. В первой (**product**) — находятся наблюдаемые значения для всех респондентов, а во второй (**group**) — указан номер группы, в которую они входили, т.е. уровень фактора (см. рис. 6.1).

Выбор процедуры. В блоке **Nonparametric Tests** меню **Analyze** выбрать процедуру **K Independent Samples**, как это показано на рис. 6.1.

Заполнение полей ввода данных. На рис. 6.2 приведен экран ввода данных и параметров этой процедуры. Следует перенести переменную **product** в поле **Test Variable List**, переменную **group** — в поле **Grouping Variable** и нажать кнопку **(Define Range)**. Открывшееся при этом окно настройки позволяет уточнить число уровней фактора (число групп испытуемых), которые будут включены в анализ. Рассматривая все группы, укажем в качестве минимального номера группы — 1, а в качестве максимального — 6. При этом переменная **group** в поле **Grouping Variable** примет вид, указанный на рис. 6.2.

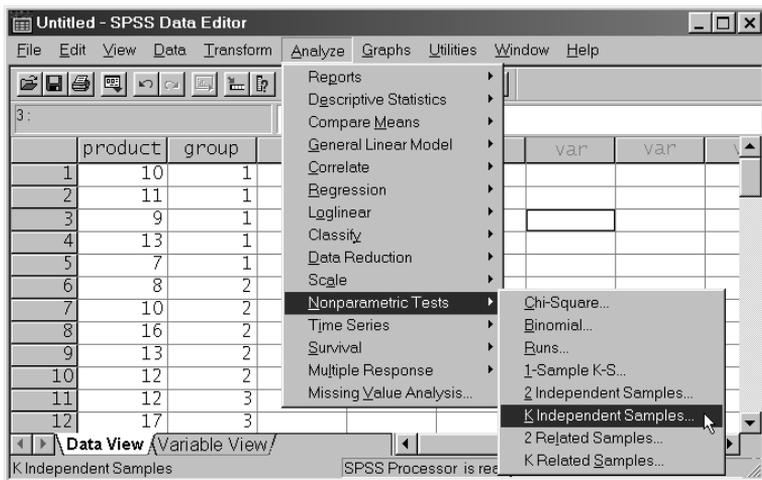


Рис. 6.1. Пакет SPSS. Форма ввода данных для однофакторного анализа

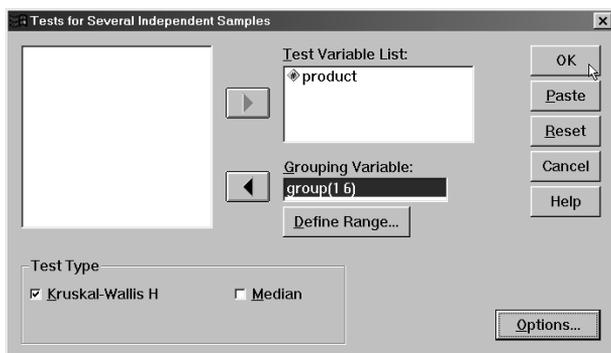


Рис. 6.2. Пакет SPSS. Окно ввода данных и параметров процедуры «K Independent Samples»

В блоке **Test Type** (тип теста) указать критерий Краскела—Уоллиса (*Kruskal-Wallis H*). Кнопка **Option** позволяет включить в выдачу результатов процедуры таблицу описательных статистик.

Результаты. После заполнения полей ввода и нажатия кнопки **OK** в окне навигатора вывода результатов появятся результаты обработки. Они включают две таблицы. Первая (рис. 6.3) содержит информацию о числе наблюдений **N** и средний ранг наблюдений **Mean Rank** в каждой группе. Эта величина является отношением суммы рангов наблюдений группы к числу наблюдений в группе.

Ranks

	GROUP	N	Mean Rank
PRODUCT	1	5	5,70
	2	5	9,00
	3	5	12,50
	4	5	16,10
	5	5	23,40
	6	5	26,30
	Total	30	

Рис. 6.3. Пакет SPSS. Результаты процедуры «K Independent Samples»

Test Statistics^{a, b}

	PRODUCT
Chi-Square	21,219
df	5
Asymp. Sig.	,001

a. Kruskal Wallis Test
b. Grouping Variable: GROUP

Рис. 6.4. Пакет SPSS. Результаты процедуры «K Independent Samples»

Вторая таблица (рис. 6.4) показывает значение статистики критерия **Chi-Square**, число степеней свободы **df** и асимптотический уровень значимости критерия (**Asymp. Sig.**).

Критерий показывает, что гипотезу об отсутствии влияния стимулирования на производительность следует отвергнуть.

Комментарии. 1. Ввод данных в процедуру в описанном выше виде является более гибким по сравнению с вводом в виде таблицы (матрицы). Его преимущества особенно ощутимы в тех случаях, когда производится изменение порядка группировки данных по результатам предварительного анализа. Примером изменения порядка группировки может являться объединение данных, соответствующих нескольким способам обработки, в один блок по причине отсутствия значимых различий между этими способами обработки.

2. В пакете отсутствует процедура, реализующая оценивание эффектов обработки непараметрическими методами.

Пример 6.2к. Проведем однофакторный дисперсионный анализ для данных примера 6.1к: проверим нулевую гипотезу об отсутствии эффектов обработки и построим 95% доверительные интервалы для эффектов обработки.

Подготовка данных. См. пример 6.1к.

Выбор процедуры. В блоке **Compare Means** меню **Analyze** выбрать процедуру **One-Way ANOVA**. (Сокращение **ANOVA** происходит от выражения «Analysis of variance». В отечественной литературе наряду с этим термином часто используется термин «дисперсионный анализ».)

Заполнение полей ввода данных. Окно ввода данных и параметров этой процедуры приведено на рис. 6.5. В нем следует перенести переменную `product` в поле `Dependent List`, а переменную `group` — в поле `Factor`. Для получения доверительных интервалов для эффектов обработки нажать кнопку `(Option)` и в открывшемся окне указать выдачу описательных статистик. (О назначении других кнопок в нижней части окна (рис. 6.5) сказано ниже в углубленном анализе.)

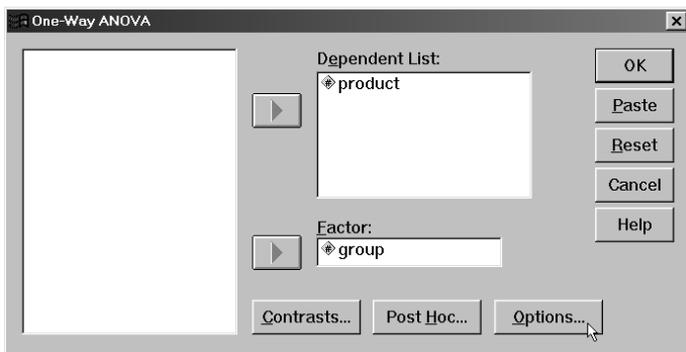


Рис. 6.5. Пакет SPSS. Окно ввода данных и параметров процедуры «One-Way ANOVA»

Результаты. Процедура, в зависимости от ее настройки, создает в окне навигатора вывода результатов несколько таблиц. К ним относится таблица дисперсионного анализа ANOVA (рис. 6.6).

ANOVA

PRODUCT					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	590,800	5	118,160	12,637	,000
Within Groups	224,400	24	9,350		
Total	815,200	29			

Рис. 6.6. Пакет SPSS. Таблица дисперсионного анализа (ANOVA-таблица) процедуры «One-Way ANOVA»

Таблица дисперсионного анализа. Дадим определения величин, приведенных в таблице дисперсионного анализа. Сначала рассмотрим второй столбец `Sum of Squares` (Сумма квадратов). В последней строке `Total` (общая) указана общая сумма квадратов разностей наблюдений и их среднего значения:

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2.$$

В строке **Between groups (между группами)** приведен вклад в общую сумму квадратов, обусловленный различиями в уровнях фактора a_j . Часто эту величину называют *суммой квадратов между группами*:

$$\sum_{j=1}^k n_j (x_{.j} - \bar{x})^2, \quad (6.18)$$

где $x_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$, а \bar{x} определяется выражением (6.8).

В строке **Within groups (внутри групп)** указан вклад в общую сумму квадратов, вызванный случайной изменчивостью данных внутри групп:

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2. \quad (6.19)$$

Легко видеть, что сумма величин первой и второй строк столбца **Sum of Squares** таблицы дисперсионного анализа (рис. 6.6) дает величину в третьей строке этого столбца. Таким образом, смысл анализа вариации данных сводится к выяснению разложения общей суммы квадратов отклонений на две части. Первая из них интерпретируется как вариация, обусловленная введенной моделью, а вторая — как случайная изменчивость данных внутри самой модели.

В случае справедливости нулевой гипотезы каждая из величин в первом столбце таблицы имеет распределение $\sigma^2 \chi^2$ со своим числом степеней свободы (оно указано в третьем столбце **df (степени свободы)** таблицы). Наконец, в четвертом столбце таблицы **Mean Square (средние квадраты)** находятся частные от деления величин второго столбца на соответствующие величины третьего столбца. Согласно формулам (6.5) и (6.9), нормированные средние квадраты между группами являются оценкой σ^{2**} , а средние квадраты внутри групп являются оценкой σ^{2*} . Отношение двух этих оценок носит название *F-отношения* (6.10), и его значение, приведенное в пятом столбце таблицы дисперсионного анализа, как раз и используется для проверки нулевой гипотезы. В последнем столбце таблицы **Sig. (уровень значимости)** указывается минимальный уровень значимости указанной *F*-статистики. В нашем случае он практически равен нулю. Как обычно, если значимость *F*-статистики близка к нулю, есть основание отвергнуть нулевую гипотезу.

Для данных нашего примера из приведенной таблицы дисперсионного анализа (рис. 6.6) можно сделать вывод, что нулевая гипотеза об отсутствии эффектов обработки должна быть отвергнута, так как вероятность получения указанного или большего значения *F*-отношения (уровень значимости *F*-статистики) при нулевой гипотезе практически равна нулю. Таким образом, представляет интерес получение оце-

нок эффектов обработки и построение для них доверительных интервалов. Эту информацию предоставляет таблица описательных статистик (рис. 6.7). В ней указано число наблюдений в каждой группе (для каждого уровня фактора), среднее значение по группе (Mean), его стандартное отклонение (Std. Deviation) и стандартная ошибка среднего (Std. Error). В последующих столбцах таблицы приведены нижние (Lower Bound) и верхние (Upper Bound) границы 95% доверительного интервала для внутригрупповых средних.

Descriptives

PRODUCT		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
1	5					10,00	2,24		
2	5	11,80	3,03	1,36	8,03	15,57	8	16	
3	5	13,60	3,21	1,44	9,62	17,58	9	17	
4	5	15,60	2,51	1,12	12,48	18,72	12	19	
5	5	20,00	3,16	1,41	16,07	23,93	16	24	
6	5	22,60	3,91	1,75	17,74	27,46	18	27	
Total	30	15,60	5,30	,97	13,62	17,58	7	27	

Рис. 6.7. Пакет SPSS. Таблица описательных статистик процедуры «One-Way ANOVA»

Как отмечалось в п. 6.6.1, оценками эффектов обработки могут служить внутригрупповые средние $x_{.j}$, находящиеся в третьем столбце таблицы рис. 6.7.

Углубленный анализ. Процедура включает тест однородности дисперсий для каждого уровня фактора (Test of Homogeneity Variances) в виде критерия Левена **Levene Statistics**. Этот относительно более современный критерий более устойчив к отклонениям от нормальности распределения данных, чем более известные критерии Кокрена, Бартлетта и Хартли (см. [19], [87], [132]). Включение этого критерия в процедуру осуществляется после нажатия кнопки (Option).

В процедуре реализовано много различных методов множественных сравнений эффектов обработки. Выбор этих методов осуществляется после нажатия кнопки **Post Hoc** в окне ввода параметров процедуры (рис. 6.5).

Процедура также дает возможность оценить различные контрасты между уровнями факторов. Задать контрасты (линейные функции специального вида от эффектов обработки) можно, нажав кнопку **Contrasts** (см. рис. 6.5).

Дополнительная литература

1. Болдин М.В., Симонова Г.И., Тюрин Ю.Н. Знаковый статистический анализ линейных моделей. — М.: Наука; Физматлит, 1997. — 288 с.
2. Холлендер М., Вулф Д. Непараметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.

Двухфакторный анализ

7.1. Связь задач двухфакторного и однофакторного анализа

Продолжая тему исследования зависимостей, начатую в гл. 6, рассмотрим задачу о действии на измеряемую величину (отклик) двух факторов. В этой задаче мы предполагаем, что на отклик могут влиять два фактора, каждый из которых принимает конечное число значений (уровней), и интересуемся тем, как влияют эти факторы на изучаемый отклик и влияют ли вообще. Такие задачи характерны как для промышленных и технологических экспериментов, так и для гуманитарных исследований. Остановимся более подробно на одном из распространенных случаев возникновения задач двухфакторного анализа.

Бывает, что в рамках однофакторной модели (см. гл. 6) влияние интересующего нас фактора не проявляется, хотя содержательные соображения указывают, что такое влияние должно быть. Иногда это влияние проявляется, но точность выводов о количественной стороне этого влияния недостаточна. Причиной такого явления может быть большой внутригрупповой разброс, на фоне которого действие фактора остается незаметным или почти незаметным. Очень часто этот разброс вызывается не только случайными причинами, но также действием еще одного фактора. Если мы в состоянии указать такой фактор, можно попытаться включить его в модель, чтобы уменьшить статистическую неоднородность наблюдений и благодаря этому выявить действие на отклик закономерных причин. Конечно, не всегда удастся поправить дело введением одного «мешающего» фактора и переходом к двухфакторным схемам, как выше. Иногда приходится рассматривать и трех-, и многофакторные модели. Замысел во всех этих случаях остается прежним.

К задачам двухфакторного или многофакторного анализа часто приводят также исследования по оптимизации технологических процессов. При этом чаще всего заранее известно, что оба фактора оказывают значимое влияние на отклик, а исследователя интересует численная оценка этого влияния с целью выбора оптимального уровня факторов. Особенности подобных задач подробно изложены в [43].

Иногда факторы разделяют на важные и мешающие, но это совсем не обязательно. В ряде задач факторы содержательно равноправны для экспериментатора. Эти нюансы мало влияют на статистические модели, они могут сказаться только на постановках статистических вопросов.

Замечания. 1. В практических ситуациях вполне возможен не только переход от однофакторной постановки задачи к двухфакторной, но и наоборот. Если при решении двухфакторной задачи оказывается, что влияние одного из факторов несущественно, то задача сводится к однофакторной.

2. Один из методов борьбы с нежелательными воздействиями мешающих факторов основан на специальном планировании процедуры сбора экспериментальных данных. Его цель — свести к нулю влияние мешающих факторов на отклик за счет усреднения положительных и отрицательных вкладов указанных факторов. А именно, при фиксированном уровне фактора проводят испытания на такой группе объектов наблюдения, внутри которой влияния мешающих факторов, будучи различными, в среднем уравнивают друг друга. При этом в таблицу однофакторного анализа заносится среднее значение измеряемой величины по взятой группе объектов. Однако чаще всего информация о характере влияния мешающих факторов на исследуемый отклик отсутствует, а поэтому такой подбор оказывается невозможным. Другой способ — случайное формирование соответствующих групп объектов наблюдения, когда из большого количества потенциально пригодных объектов случайным образом выбираются те, которые образуют требуемую группу. Этот метод позволяет по-прежнему использовать однофакторный анализ, однако возникающие в нем осложнения, описанные выше, часто делают предпочтительным другой способ устранения влияния мешающих факторов, который основан на прямом количественном учете влияния наиболее существенных из указанных факторов. Если в задаче удается выделить один главный мешающий фактор, то она сводится к задаче двухфакторного анализа. Влияние остальных факторов желательно удалить с помощью процедуры случайного выбора объектов наблюдения (см., например, [39], [114]).

7.2. Таблица двухфакторного анализа

Рассмотрим, как изменяется таблица однофакторного анализа, приведенная в п. 6.1, при включении в модель действия мешающего фактора.

Назовем главный фактор фактором A , а мешающий фактор — фактором B . Пусть фактор A принимает k , а фактор B — n различных значений. Фактор B разбивает все объекты наблюдения на n блоков, каждый блок образуют наблюдения, проведенные при одном уровне фактора B . В блоке отклики могут значимо различаться только за счет применения к ним различных *обработок*, т.е. за счет различных уровней фактора A . Уровни фактора A (обработки) отображаются в таблице по столбцам, а уровни фактора B (блоки) — по строкам. Традиционная терминология «блок-обработка» в применении к факторам B и A сло-

Таблица 7.1

Блоки	Обработки			
	1	2	...	k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
⋮	⋮	⋮		⋮
n	x_{n1}	x_{n2}	...	x_{nk}

жилась как результат различного отношения к этим факторам, один из которых является мешающим, а другой определяющим.

Таблица 7.1, содержащая $n \times k$ наблюдений (по одному наблюдению в клетке) является основной таблицей двухфакторного анализа. Ее отличие от таблицы однофакторного анализа заключается в том, что наблюдения в любом столбце не являются однородными, т.е. могут не образовывать выборки (если влияние мешающего фактора значимо). Для описания такой двухфакторной таблицы требуются более сложные вероятностные модели, чем для однофакторного анализа.

Замечание. Таблица 7.1 на самом деле является *простейшей* таблицей двухфакторного анализа. На практике часто рассматриваются, скажем, таблицы с повторными изменениями (там в каждой клетке табл. 7.1 могут содержаться несколько наблюдений). Более подробно об этом можно прочесть в [39], [114].

7.3. Аддитивная модель данных двухфакторного эксперимента при независимом действии факторов

Для описания данных табл. 7.1 двухфакторного эксперимента в большинстве случаев оказывается приемлемой аддитивная модель. Она предполагает, что значение отклика x_{ij} является суммой самостоятельных вкладов соответствующих уровней каждого из факторов и независимых от этих факторов случайных величин. Последние отражают внутреннюю изменчивость отклика при фиксированных уровнях факторов, которая может порождаться различными причинами.

Таким образом, каждое наблюдение x_{ij} представляется в виде:

$$x_{ij} = b_i + t_j + e_{ij}; \quad i = 1, \dots, n; \quad j = 1, \dots, k. \quad (7.1)$$

При этом числа b_1, \dots, b_n являются результатом влияния на отклик мешающего фактора B , действие которого разбивает все данные на блоки. Поэтому величины b_1, \dots, b_n называют *эффектами блоков*. Числа t_1, \dots, t_k отражают действие на отклик интересующего нас фактора A и именуются *эффектами обработки*. Относительно случайных величин

e_{ij} предполагается, что они одинаково распределены и независимы в совокупности. Различные методы двухфакторного анализа требуют от их распределения либо только непрерывности, либо принадлежности к нормальному семейству распределений $N(0, \sigma^2)$ со средним 0 и некоторой неизвестной дисперсией σ^2 . Оба эти случая будут рассмотрены ниже.

Замечание. Требования одинаковой распределенности величин e_{ij} можно ослабить, предполагая, что в каждом блоке отклики x_{ij} принадлежат к своему непрерывному семейству распределений F_i , а параметр сдвига для конкретного наблюдения в блоке определяется числами t_1, \dots, t_k , т.е. эффектами обработки. Некоторые ослабления можно сделать и в условии независимости e_{ij} (см., например, [113]). Для простоты изложения мы будем использовать в дальнейшем первоначальные предположения о величинах e_{ij} .

Заметим, что даже в случае справедливости представления (7.1) величины вкладов факторов b_i и t_j не могут быть восстановлены однозначно. Действительно, увеличение всех b_i на одну и ту же константу и одновременно уменьшение всех t_j на эту константу оставляет выражение (7.1) неизменным. Для однозначной определенности вкладов факторов удобно перейти к представлению наблюдений в виде:

$$x_{ij} = \mu + \beta_i + \tau_j + e_{ij}; \quad i = 1, \dots, n; \quad j = 1, \dots, k, \quad (7.2)$$

считая, что $\sum_{i=1}^n \beta_i = 0$, $\sum_{j=1}^k \tau_j = 0$. При этом параметр μ интерпретируется как среднее значение, присущее всем величинам x_{ij} , а β_i и τ_j — как отклонения от μ в результате действия факторов B и A .

Гипотеза. Как и в случае однофакторного анализа, целесообразно прежде всего проверить гипотезу о значимости эффектов обработки. Сформулируем нулевую гипотезу в виде: $H_0 : \tau_1 = \tau_2 = \dots = \tau_k = 0$. Другими словами, предположим, что влияние фактора A отсутствует. Ниже будут рассмотрены критерии проверки этой гипотезы как в непараметрическом случае, так и в случае, когда величины x_{ij} принадлежат нормальному семейству распределений.

7.4. Непараметрические критерии проверки гипотезы об отсутствии эффектов обработки

7.4.1. Критерий Фридмана (произвольные альтернативы)

Непараметрический критерий Фридмана для проверки гипотезы H_0 против альтернативы о наличии влияния фактора A используется в

случае, если о распределении случайных величин e_{ij} , $i = 1, \dots, n$, $j = 1, \dots, k$ в модели (7.2) известно только то, что оно непрерывно, а сами величины e_{ij} независимы в совокупности. (То, что e_{ij} одинаково распределены, было оговорено раньше.) Критерий основан на идее перехода от значений величин x_{ij} в таблице двухфакторного анализа к их рангам. В отличие от однофакторного анализа, ранжирование происходит не по всей совокупности величин x_{ij} , а поблочно, т.е. рассматривается каждая отдельная строка табл. 7.1 и при фиксированном индексе i осуществляется ранжирование величин x_{ij} при $j = 1, \dots, k$. Тем самым устраняется влияние «мешающего» фактора B , значение которого для каждой строки таблицы постоянно.

Обозначим полученные ранги величин x_{ij} через r_{ij} . Ясно, что значения r_{ij} изменяются от 1 до k , а соответствующая строка рангов представляет собой некоторую перестановку чисел $1, 2, \dots, k$. Для простоты изложения будем предполагать, что среди элементов x_{ij} , стоящих в одной строке таблицы (7.1), нет совпадающих (в противном случае следует использовать средние ранги). При гипотезе $H_0 : \tau_1 = \tau_2 = \dots = \tau_k = 0$ каждая строка рангов $r_{i1}, r_{i2}, \dots, r_{ik}$ будет представлять случайную перестановку чисел от 1 до k , причем все $k!$ перестановок равновероятны. Введем величину: $r_{.j} = \frac{1}{n} (\sum_{i=1}^n r_{ij})$, являющуюся средним значением рангов по столбцу j . При гипотезе H_0 в силу равновероятности всех перестановок рангов в каждой строке значение $r_{.j}$ для каждого j не должно сильно отличаться от величины $r_{..} = (k+1)/2$, которая представляет собой общий средний ранг всех элементов таблицы рангов. (Действительно, сумма рангов по всей таблице есть $nk(k+1)/2$. Средний ранг получается делением на число nk элементов таблицы.)

Статистика Фридмана S для проверки гипотезы H_0 имеет следующий вид:

$$S = \frac{12n}{k(k+1)} \sum_{j=1}^k (r_{.j} - r_{..})^2. \quad (7.3)$$

Здесь множитель, стоящий перед знаком суммы, добавлен для того, чтобы S имело простое асимптотическое распределение. В вычислительном плане более удобна другая форма записи величины S , а именно:

$$S = \left[\frac{12}{nk(k+1)} \sum_{j=1}^k \left(\sum_{i=1}^n r_{ij} \right)^2 \right] - 3n(k+1). \quad (7.4)$$

Как отмечалось выше, при справедливости гипотезы H_0 величины $(r_{.j} - r_{..})^2$ в выражении (7.3) с большой вероятностью сравнительно малы для всех j , и, следовательно, значение S сравнительно невелико.

А при нарушении H_0 суммы рангов в одних столбцах будут тяготеть к превышению значения среднего ранга $r_{..}$, а в других — к уменьшению этого значения, в зависимости от знака величины $\tau_j \neq 0$. Это приводит к возрастанию статистики Фридмана S . Из этих соображений вытекает вид критерия Фридмана для проверки гипотезы $H_0 : \tau_1 = \tau_2 = \dots = \tau_k = 0$ против альтернативы наличия эффектов обработки.

Правило проверки гипотезы. Гипотеза H_0 принимается на уровне значимости α , если $S < S(\alpha, k, n)$, и отвергается в пользу альтернативы при $S \geq S(\alpha, k, n)$. Критическое значение $S(\alpha, k, n)$ находят как решение уравнения $P\{S \geq S(\alpha, k, n)\} = \alpha$, где вероятность P вычисляется при справедливости гипотезы H_0 .

Таблицы и аппроксимация. Для небольших значений n, k величина $S(\alpha, k, n)$ может быть найдена из таблиц [32] и [115]. При больших n для выбора критических значений приходится пользоваться аппроксимацией. Она основана на том факте, что при справедливости гипотезы H_0 и $n \rightarrow \infty$ статистика Фридмана S асимптотически распределена как хи-квадрат с $(k - 1)$ степенями свободы (сведения о более точной аппроксимации можно найти в [65]). В этом случае критерий для проверки гипотезы H_0 сводится к следующему: принять H_0 на уровне значимости α , если $S < \chi^2_{(1-\alpha)}(k - 1)$, и отклонить H_0 в противном случае. Здесь $\chi^2_{(1-\alpha)}(k - 1)$ — квантиль уровня $1 - \alpha$, или $(1 - \alpha)$ -квантиль случайной величины χ^2 с $(k - 1)$ степенями свободы.

Совпадающие значения. Если в строках таблицы двухфакторного анализа имеются совпадающие значения, при переходе к таблице рангов используются средние ранги, а вместо статистики S используется ее модификация, выражение для которой можно найти в [115].

7.4.2. Критерий Пейджа (альтернативы с упорядочением)

Назначение. Часто целью исследования является установление преимущества одного метода обработки над другим. Если таких обработок несколько, возможно предположение, что их эффективность возрастает в определенном направлении, например, по мере увеличения интенсивности воздействия. Для того чтобы подтвердить или опровергнуть такое предположение, снова обратимся к проверке H_0 . Но на этот раз постараемся выбрать критерий, чувствительный именно к альтернативам о возрастании (вариант: убывании) эффекта. Против такой специальной и более узкой группы альтернатив можно предложить ориентированный именно на эту ситуацию критерий Пейджа.

Критерий Пейджа предназначен для проверки гипотезы H_0 об отсутствии эффектов обработки ($H_0 : \tau_1 = \tau_2 = \dots = \tau_k$) против альтернатив с упорядочением: $\tau_1 \leq \tau_2 \leq \dots \leq \tau_k$, где хотя бы одно из неравенств строгое.

Статистика Пейджа. Введем величину r_j как $r_j = \sum_{i=1}^n r_{ij}$. Статистика Пейджа L по определению есть:

$$L = \sum_{j=1}^k jr_j = r_1 + 2r_2 + \dots + kr_k. \quad (7.5)$$

Вид критерия. Критерий проверки гипотезы H_0 против альтернатив с упорядочением на уровне значимости α имеет вид:

- принять H_0 , если $L < l(\alpha, k, n)$;
- отклонить H_0 в пользу альтернативы, если $L \geq l(\alpha, k, n)$,

где функция $l(\alpha, k, n)$ удовлетворяет уравнению $P\{L \geq l(\alpha, k, n)\} = \alpha$.

Таблицы и асимптотика. Для значений $k = 3$, $n = 2(1)20$ и $k = 4(1)8$, $n = 2(1)12$ таблица приближенных значений $l(\alpha, k, n)$ дана в [115]. В случае больших значений k и n для нахождения процентных точек следует использовать асимптотическое распределение статистики L . Рассмотрим величину L^* :

$$L^* = \frac{L - nk(k+1)^2/4}{[n(k^3 - k)^2/144(k-1)]^{1/2}}. \quad (7.6)$$

При справедливости H_0 статистика L^* имеет при $n \rightarrow \infty$ асимптотическое распределение $N(0, 1)$ (сведения о более точной аппроксимации можно найти в [65]). Следовательно, приближенный критерий для проверки H_0 против альтернатив с упорядочением на уровне значимости α имеет вид: принять H_0 , если $L^* < z_\alpha$, в противном случае — отклонить H_0 в пользу альтернативы. Здесь z_α — α -процентная точка стандартного нормального распределения.

Если в пределах строки исходной двухфакторной таблицы встречаются совпадающие значения, надо использовать средние ранги. Чем больше таких совпадений, тем более приближенными становятся выводы.

7.5. Практический пример

Покажем, как используются описанные выше критерии на практике. В табл. 7.2 приведены данные из [115]. Они являются результатом исследова-

Таблица 7.2

Частота тремора руки (Гц) как функция веса браслета

Вес браслета (фунт)	0	1.25	2.5	5	7.5
Испытуемый\Обработка	1	2	3	4	5
1	3.01	2.85	2.62	2.63	2.58
2	3.47	3.43	3.15	2.83	2.70
3	3.35	3.14	3.02	2.71	2.78
4	3.10	2.86	2.58	2.49	2.36
5	3.41	3.32	3.08	2.96	2.67
6	3.07	3.06	2.85	2.50	2.43

дования зависимости частоты самопроизвольного дрожания мышц рук (тремора) от тяжести специального браслета, надеваемого на запястье.

Каждое табличное значение — среднее из 5 экспериментальных измерений частоты тремора у испытуемого. Каждая обработка соответствует весу браслета, измеренного в фунтах. Перейдем от табл. 7.2 к соответствующей таблице рангов 7.3.

Таблица 7.3

Испытуемый\Обработка	1	2	3	4	5
1	5	4	2	3	1
2	5	4	3	2	1
3	5	4	3	1	2
4	5	4	3	2	1
5	5	4	3	2	1
6	5	4	3	2	1
r_j	30	24	17	12	7
$r_{.j}$	5	4	2.8333	2	1.1667

В двух последних строках табл. 7.3 приведены соответственно суммы рангов по каждому столбцу и средние суммы рангов по столбцам. Подставляя эти значения в выражение (7.4), вычислим статистику Фридмана S (здесь $n = 6$, $k = 5$):

$$S = \left[\frac{12}{nk(k-1)} \sum_{j=1}^k r_j^2 \right] - 3n(k+1) = 22.5333.$$

Для проверки с помощью статистики S гипотезы H_0 против произвольных альтернатив воспользуемся ее асимптотическим распределением χ^2 с $(k-1)$ степенями свободы. При $\alpha = 0.05$ соответствующая процентная точка распределения $\chi^2(4)$ есть $\chi^2(4, 0.05) = 9.488$, при $\alpha = 0.01$ — $\chi^2(4, 0.01) = 13.292$, при $\alpha = 0.001$ — $\chi^2(4, 0.001) = 18.51$. Учитывая, что $S > \chi^2(4, 0.001)$, мы отвергаем гипотезу в пользу альтернативы на уровне значимости $\alpha = 0.001$. Согласно таблицам распреде-

ления $\chi^2(4)$, минимальный уровень значимости, при котором гипотеза отвергается в пользу альтернативы, равен $\alpha = 0.00016$.

Теперь применим к данным табл. 7.2 критерий Пейджа, поскольку есть априорные основания считать, что частота тремора уменьшается при увеличении веса браслета. Чтобы непосредственно применить формулу (7.5), построенную для возрастающего влияния уровня фактора, мы должны произвести перенумерацию столбцов табл. 7.3 в обратном порядке. То есть номер $j = 1$ будет соответствовать пятому столбцу табл. 7.3, номер $j = 2$ — четвертому столбцу и т.д. Соответственно статистика Пейджа L равна: $L = \sum_{j=1}^k jr_j = 7 + 2 \cdot 12 + 3 \cdot 17 + 4 \cdot 24 + 5 \cdot 30 = 328$.

Из таблицы критических значений статистики Пейджа в [115] находим, что для $\alpha = 0.01$ $l(0.01, 5, 6) = 299$, а при $\alpha = 0.001$ $l(0.001, 5, 6) = 307$. Так как $L \geq l(0.001, 5, 6)$, то, следовательно, гипотеза H_0 должна быть отвергнута в пользу альтернативы $\tau_1 \leq \tau_2 \leq \dots \leq \tau_k$ на уровне значимости $\alpha = 0.001$.

Для нахождения приближенного значения минимального уровня значимости критерия Пейджа воспользуемся нормальной аппроксимацией распределения статистики L^* . В нашем примере значения n и k в выражении (7.6) равны соответственно 6 и 5. Следовательно:

$$L^* = \frac{328 - 6 \cdot 5 \cdot (5 + 1)^2 / 4}{[6 \cdot (125 - 5)^2 / (144 \cdot 4)]^{1/2}} \simeq 4.75.$$

Согласно таблицам стандартного нормального распределения, минимальный уровень значимости, на котором может быть отвергнута гипотеза с помощью критерия Пейджа, равен $\alpha = 0.000001$, что на два порядка меньше, чем для критерия Фридмана. Это иллюстрирует положение, что в случае упорядоченных альтернатив критерий Пейджа обладает большей мощностью, чем критерий Фридмана.

7.6. Двухфакторный дисперсионный анализ

Если есть основания предполагать, что случайные величины e_{ij} в модели двухфакторного анализа (7.1) имеют нормальное распределение с нулевым средним и неизвестной одинаковой при всех i и j дисперсией σ^2 , можно предложить более мощный критерий для проверки гипотезы $H_0 : \tau_1 = \tau_2 = \dots = \tau_k$ и построить более эффективные оценки параметров μ , τ_j и β_i . Используемые для этого методы аналогичны тем, которые были рассмотрены при решении задач однофакторного диспе-

рсионного анализа в п. 6.5 гл. 6. В связи с этим здесь мы дадим только их краткое описание, достаточное для решения прикладных задач.

Получение оценок дисперсии. Так же как и в задаче однофакторного дисперсионного анализа, проверка гипотезы H_0 основывается на сравнении двух независимых оценок σ^2 . При этом одна из оценок σ^{2*} действует вне зависимости от того, верна ли гипотеза H_0 , а другая — σ^{2**} — только в случае справедливости гипотезы.

Оптимальная в классе несмещенных оценок оценка σ^{2*} может быть получена с помощью метода наименьших квадратов. Для этого сначала оценим неизвестные значения параметров μ , β_i и τ_j в модели (7.2). А именно, найдем значения $\hat{\mu}$, $\hat{\beta}_i$ и $\hat{\tau}_j$ такие, что при них достигается минимума выражение:

$$\sum_{i,j} (x_{ij} - \mu - \beta_i - \tau_j)^2 \quad (7.7)$$

при условии, что $\sum_{i=1}^n \beta_i = \sum_{j=1}^k \tau_j = 0$. Минимальная величина (7.7), равная $\sum_{i,j} (x_{ij} - \hat{\mu} - \hat{\beta}_i - \hat{\tau}_j)^2$, выражает разброс наблюдений относительно подобранных ожидаемых значений.

Решение задачи (7.7) осуществляется стандартными методами математического анализа и приводит к следующим оценкам $\hat{\mu}$, $\hat{\beta}_i$ и $\hat{\tau}_j$:

$$\begin{aligned} \hat{\mu} = x_{..} &= \frac{1}{nk} \sum_{i=1}^n \sum_{j=1}^k x_{ij} & \hat{\beta}_i = x_{i.} - x_{..} &= \frac{1}{k} \sum_{j=1}^k x_{ij} - x_{..} \\ \hat{\tau}_j = x_{.j} - x_{..} &= \frac{1}{n} \sum_{i=1}^n x_{ij} - x_{..} \end{aligned} \quad (7.8)$$

Полученные оценки параметров модели имеют следующие распределения:

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{nk}\right); \quad \hat{\beta}_i \sim N\left(\beta_i, \frac{\sigma^2(n-1)}{nk}\right); \quad \hat{\tau}_j \sim N\left(\tau_j, \frac{\sigma^2(k-1)}{nk}\right).$$

Для получения оценки σ^{2*} можно использовать величину:

$$\sum_{i=1}^n \sum_{j=1}^k (x_{ij} - \hat{\mu} - \hat{\beta}_i - \hat{\tau}_j)^2 = \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - x_{i.} - x_{.j} + x_{..})^2,$$

которая имеет распределение $\sigma^2 \chi^2$ с числом степеней свободы $nk - (n-1) - (k-1) - 1 = (n-1)(k-1)$. Сама оценка σ^{2*} равна:

$$\sigma^{2*} = \frac{1}{(n-1)(k-1)} \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - x_{i.} - x_{.j} + x_{..})^2. \quad (7.9)$$

Выражение (7.9) дает несмещенную оценку σ^{2*} , которая справедлива как при выполнении гипотезы H_0 , так и при ее нарушении.

Для получения второй оценки величины σ^2 , независимой от оценки σ^{2*} , воспользуемся тем, что случайные величины $x_{.1}, \dots, x_{.k}$, являющиеся средними значениями по соответствующим столбцам таблицы двухфакторного анализа, при нулевой гипотезе независимы и одинаково распределены по нормальному закону $N(\mu, \sigma^2/n)$. На их основе мы стандартным образом (см. гл. 5 и 6) можем сконструировать статистику для оценки σ^2 : $n \sum_{j=1}^k (x_{.j} - x_{..})^2$, имеющую распределение $\sigma^2 \chi^2$ с $(k-1)$ степенями свободы. При этом сама оценка σ^{2**} есть:

$$\sigma^{2**} = \frac{n}{k-1} \sum_{j=1}^k (x_{.j} - x_{..})^2. \quad (7.10)$$

При H_0 выражение (7.10) тоже дает несмещенную оценку σ^2 . При нарушении же H_0 статистика (7.10) приобретает тенденцию к увеличению — тем большую, чем больше различие между эффектами обработки $\tau_1, \tau_2, \dots, \tau_k$.

Критерий для проверки гипотезы H_0 : $\tau_1 = \tau_2 = \dots = \tau_k$. Составляя, так же как в гл. 6, F -отношение двух оценок дисперсий, получаем:

$$F = \frac{\frac{n}{k-1} \sum_{j=1}^k (x_{.j} - x_{..})^2}{\frac{1}{(n-1)(k-1)} \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - x_{i.} - x_{.j} + x_{..})^2}.$$

При гипотезе величина F имеет F -распределение с числом степеней свободы $(k-1)$ и $(n-1)(k-1)$. Критерий для проверки гипотезы H_0 имеет при этом следующий вид:

- отвергнуть гипотезу H_0 на уровне значимости α , если $F \geq F_{1-\alpha}$;
- не отвергать гипотезу H_0 на уровне значимости α , если $F < F_{1-\alpha}$.

Здесь $F_{1-\alpha}$ обозначает квантиль уровня $1 - \alpha$ F -распределения с числом степеней свободы $((k-1)$ и $(n-1)(k-1))$.

Замечание. Обратим внимание на то, что полная сумма квадратов отклонений величин x_{ij} от их общего среднего $x_{..}$ представима в виде:

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - x_{..})^2 &= \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - x_{i.} - x_{.j} + x_{..})^2 + \\ &\sum_{i=1}^n \sum_{j=1}^k (x_{i.} - x_{..})^2 + \sum_{i=1}^n \sum_{j=1}^k (x_{.j} - x_{..})^2 = \end{aligned}$$

$$\sum_{i=1}^n \sum_{j=1}^k (x_{ij} - x_{i.} - x_{.j} + x_{..})^2 + k \sum_{i=1}^n (x_{i.} - x_{..})^2 + n \sum_{j=1}^k (x_{.j} - x_{..})^2.$$

Отсюда и идет название «дисперсионный анализ», т.е. анализ разложения дисперсии (вариации, изменчивости) на части, обусловленные влиянием факторов, и часть, обусловленную случайной изменчивостью самих данных.

Выше в (7.8) были получены оценки параметров нормальной (гауссовской) модели линейного дисперсионного анализа и указано их распределение. Последнее позволяет легко построить индивидуальные доверительные интервалы. При этом в качестве оценки дисперсии следует использовать величину σ^{2*} .

Следует отметить, что выводы дисперсионного анализа о равенстве или неравенстве эффектов τ_1, \dots, τ_n довольно устойчивы даже при нарушении основных предположений о нормальном распределении и о равенстве дисперсий.

7.7. Двухфакторный анализ в пакете SPSS

Пример 7.1к. С помощью критерия Фридмана проверим нулевую гипотезу об отсутствии эффектов обработки для данных о зависимости частоты самопроизвольного дрожания мышц рук (тремора) от тяжести специального браслета, надеваемого на запястье (табл. 7.2).

Подготовка данных. Данные для критерия Фридмана должны быть введены так, как показано на рис. 7.1. То есть для каждого из k способов обработки должна быть заведена отдельная переменная, например var1 — var5.

Обратим внимание на то, что процедуру критерия Фридмана можно применять только к данным, состоящим из равного числа наблюдений для каждого из k способов обработки в каждом из n блоков. Подобные планы эксперимента часто называют *сбалансированными*.

Выбор процедуры. В блоке Nonparametric Tests выбрать процедуру K Related Samples (связанные выборки).

Заполнение полей ввода данных. Окно ввода данных процедуры приведено на рис. 7.2.

В этом окне перенесите все переменные (var1 – var5) в поле Test Variable, как это показано на рис. 7.2. В блоке Test Type (тип теста) следует отметить критерий Фридмана (Friedman). Кнопка Statistics в этом окне позволяет дополнительно задать вывод таблицы описательных статистик.

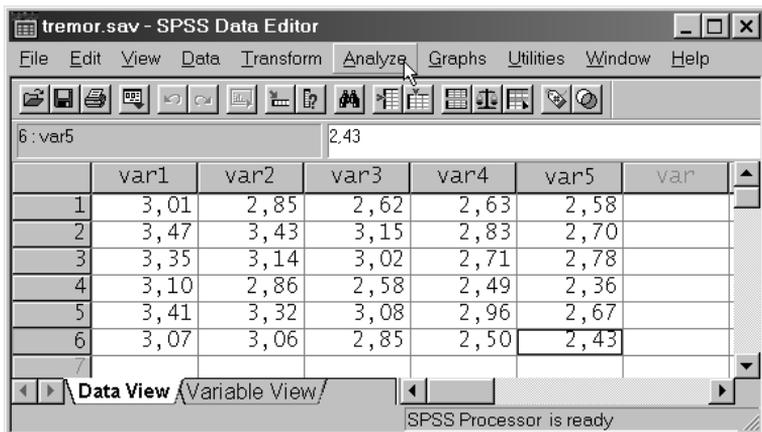


Рис. 7.1. Пакет SPSS. Форма ввода данных для критерия Фридмана

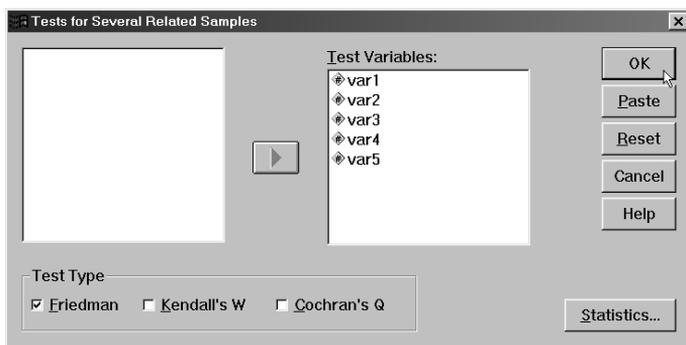


Рис. 7.2. Пакет SPSS. Окно ввода данных и параметров процедуры «K Related Samples»

Test Statistics^a

N	6
Chi-Square	22,533
df	4
Asymp. Sig.	,000

a. Friedman Test

Рис. 7.3. Пакет SPSS. Результаты критерия Фридмана (процедура «K Related Samples»)

Результаты. После заполнения полей ввода и нажатия кнопки **OK** в окно навигатора вывода результатов будут выведены две таблицы. В первой из них указаны средние ранги для каждой анализируемой переменной (способа обработки). Во второй — статистика критерия Фридмана **Chi-Square**, ее число степеней свободы и асимптотический уровень значимости (**Asymp. Sig**), как это показано на рис. 7.3.

Полученный уровень значимости говорит, что нулевую гипотезу об отсутствии эффектов обработки следует отвергнуть.

В пакете SPSS нет отдельной процедуры для двухфакторного дисперсионного анализа. Выполнить этот анализ можно с помощью гораздо более общей процедуры **Univariate** из блока **General Linear Model**. Мы не будем здесь разбирать ее работу.

Дополнительная литература

1. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика. Основы эконометрики. Т. 1. Теория вероятностей и прикладная статистика. — М.: Юнити-Дана, 2001. — 656 с.

2. *Гусев А.Н.* Дисперсионный анализ в экспериментальной психологии: учеб. пособие. — М.: Учебно-методический коллектор «Психология», 2000. — 136 с.

3. *Холлендер М., Вулф Д.* Непараметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.

Линейный регрессионный анализ

8.1. Модель линейного регрессионного анализа

Линейный регрессионный анализ объединяет широкий круг задач, связанных с построением функциональных зависимостей между двумя группами числовых переменных: x_1, \dots, x_p и y_1, \dots, y_q . Для краткости мы объединим x_1, \dots, x_p в многомерную переменную \mathbf{x} , а y_1, \dots, y_q — в переменную \mathbf{y} , и будем говорить об исследовании зависимости между \mathbf{x} и \mathbf{y} . При этом мы будем считать \mathbf{x} независимой переменной, влияющей на значения \mathbf{y} . В связи с этим мы будем называть \mathbf{y} *откликом*, а $\mathbf{x} = (x_1, \dots, x_p)$ — *факторами*, влияющими на отклик.

Исходные данные. Статистический подход к задаче построения (точнее, восстановления) функциональной зависимости \mathbf{y} от \mathbf{x} основывается на предположении, что нам известны некоторые исходные (экспериментальные) данные $(\mathbf{x}_i, \mathbf{y}_i)$, где \mathbf{y}_i — значение отклика при заданном значении фактора \mathbf{x}_i , i изменяется от 1 до n . Пару значений $(\mathbf{x}_i, \mathbf{y}_i)$ часто называют результатом одного измерения, а n — числом измерений.

Регрессионная модель. Мы будем предполагать, что наблюдаемое в опыте значение отклика \mathbf{y} можно мысленно разделить на две части: одна из них закономерно зависит от \mathbf{x} , т.е. является функцией \mathbf{x} ; другая часть — случайна по отношению к \mathbf{x} . Обозначим первую через $f(\mathbf{x})$, вторую через ε и представим отклик в виде

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon, \quad (8.1)$$

где ε — некоторая случайная величина. Случайное слагаемое ε выражает либо внутренне присущую отклику изменчивость, либо влияние на него факторов, не учтенных в соотношении (8.1), либо то и другое вместе. Иногда ε называют ошибкой эксперимента, связывая ее присутствие с несовершенством метода измерения \mathbf{y} .

Применяя соотношение (8.1) к имеющимся у нас исходным данным, получаем:

$$\mathbf{y}_i = f(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \dots, n. \quad (8.2)$$

Предположения об ошибках. Разделение y_i на закономерную и случайную составляющие можно сделать только мысленно. Реально ни $f(\mathbf{x}_i)$, ни ε_i в отдельности нам не известны, в опыте мы узнаем только их сумму. В связи с этим нам необходимо сделать определенные уточнения относительно величин ε_i . В классической модели регрессионного анализа предполагается, что:

- а) все опыты были проведены независимо друг от друга в том смысле, что случайности, вызвавшие отклонение отклика от закономерности в одном опыте, не оказывали влияния на подобные отклонения в других опытах;
- б) статистическая природа этих случайных составляющих оставалась неизменной во всех опытах.

Из этих предположений очевидно вытекает, что случайные величины $\varepsilon_1, \dots, \varepsilon_n$ статистически независимы и одинаково распределены.

В последние десятилетия активно развиваются методы, позволяющие находить решение задачи при изменении и ослаблении этих предположений (см., например, [27]).

Предположения о регрессионной функции. Для того чтобы задача о подборе функции отклика f была осмысленной, мы должны определить набор допустимых функций $f(\mathbf{x})$. Как правило, предполагают, что множество допустимых функций является параметрическим семейством $f(\mathbf{x}, \theta)$, где $\theta \in \Theta$ — параметр семейства. Тогда соотношение (8.2) можно переписать в виде:

$$y_i = f(\mathbf{x}_i, \theta) + \varepsilon_i, \quad i = 1, \dots, n, \quad (8.3)$$

и восстановление зависимости между \mathbf{x} и \mathbf{y} оказывается эквивалентным указанию значения θ (точнее, ее оценки $\hat{\theta}$) по исходным данным $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, \dots, n$. Знание $\hat{\theta}$ позволит нам по заданному значению фактора \mathbf{x} предсказывать отклик \mathbf{y} , точнее, его закономерную часть.

Например, в наиболее простой задаче одномерной линейной регрессии (она подробно рассматривается в п. 8.2) мы предполагаем зависимость между x и y вида $y = ax + b + \varepsilon$, где a и b — неизвестные параметры. Тогда θ — это двумерный параметр (a, b) .

В этой книге мы рассмотрим широко распространенную в практических задачах ситуацию, когда функция $f(\mathbf{x}, \theta)$ линейно зависит от параметров θ , т.е. $f(\mathbf{x}, \theta) = A(\mathbf{x})\theta$, где $A(\mathbf{x})$ — некоторая известная матрица, элементы которой зависят от \mathbf{x} , θ — вектор, составленный из неизвестных параметров. Эта задача носит название *линейного регрессионного анализа*. С кратким обзором методов построения регрес-

сионных зависимостей в случае, когда $f(\mathbf{x}, \theta)$ нелинейна по θ , можно познакомиться в [41] или более подробно в [36].

Активный и пассивный эксперименты. Ситуация, в которой экспериментатор может выбирать значения факторов \mathbf{x}_i по своему желанию и таким образом планировать будущие эксперименты, называется *активным экспериментом*. В этом случае значения факторов \mathbf{x}_i обычно рассматриваются как неслучайные. Более того, сообразуясь с целями эксперимента, экспериментатор может выбрать его план (т.е. значения x_1, \dots, x_n) наилучшим образом.

В отличие от этой ситуации в *пассивном эксперименте* значения фактора складываются вне воли экспериментатора, под действием других обстоятельств. Поэтому значения \mathbf{x}_i иногда приходится толковать как случайные величины, что накладывает особые черты на интерпретацию результатов. Сама же математическая обработка совокупности $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, \dots, n$ от этого не меняется.

8.2. О стратегиях, методах и проблемах регрессионного анализа

Предваряя подробный разбор методов регрессионного анализа, расскажем, не вдаваясь в подробности, об общем порядке решения регрессионных задач. При первом чтении данный параграф можно пропустить.

Простая регрессия. Самый простой случай регрессионных задач — это исследование связи между одной независимой (одномерной) переменной x и одной зависимой переменной (откликом) y . Эта задача носит название *простой регрессии*. Исходными данными этой задачи являются два набора наблюдений x_1, x_2, \dots, x_n — значения x и y_1, y_2, \dots, y_n — соответствующие значения y . Мы сначала расскажем о последовательности действий при решении задач простой регрессии.

Подбор модели. Первым шагом решения задачи является предположение о возможном виде функциональной связи между x и y . Примерами таких предположений могут являться зависимости: $y = a + bx$, $y = a + bx + cx^2$, $y = e^{a+bx}$, $y = 1/(a + bx)$ и т.д., где a, b, c и т.д. — неизвестные параметры, которые надо определить по исходным данным. Компьютерные программы регрессионного анализа, как правило, содержат достаточно обширные списки подобных функций или позволяют задавать вид зависимости формулой.

Для подбора вида зависимости между x и y полезно построить и изучить график, на котором изображены точки с координатами

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Иногда примерный вид зависимости бывает известен из теоретических соображений или предыдущих исследований аналогичных данных.

Оценка параметров модели. После выбора конкретного вида функциональной зависимости $f(x, \theta)$ можно по исходным данным x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_n провести расчет (оценку) θ , т.е. входящих в f неизвестных коэффициентов (параметров). Тем самым мы полностью определили подобранную регрессионную функцию:

$$y = f(x, \hat{\theta}), \quad \text{где } \hat{\theta} \text{ — оценка } \theta.$$

Анализ адекватности модели. После подбора регрессионной модели желательно выяснить, насколько хорошо выбранная модель описывает имеющиеся данные. К сожалению, единого общего правила для этого нет. На практике первое впечатление о правильности подобранной модели могут дать изучение некоторых численных характеристик (коэффициента детерминации, F -отношения, доверительных интервалов для оценок). Однако эти показатели скорее позволяют отвергнуть совсем неудачную модель, чем подтвердить правильность выбора функциональной зависимости. Более обоснованное решение можно принять, сравнив имеющиеся значения y_i со значениями \hat{y}_i , полученными с помощью подобранной регрессионной функции: $\hat{y}_i = f(x_i, \hat{\theta})$. Разности между наблюдаемыми и предсказанными значениями y :

$$r_i = y_i - \hat{y}_i = y_i - f(x_i, \hat{\theta}), \quad i = 1, \dots, n$$

называют *остатками*. Например, для линейной зависимости $y = a + bx$ значения остатков вычисляются в виде: $r_i = y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}x_i)$, где \hat{a} и \hat{b} — оценки коэффициентов a и b .

Анализ остатков. Анализ остатков позволяет получить представление, насколько хорошо подобрана сама модель и насколько правильно выбран метод оценки коэффициентов. Согласно общим предположениям регрессионного анализа, остатки должны вести себя как независимые (в действительности почти независимые) одинаково распределенные случайные величины. В классических методах регрессионного анализа предполагается также нормальный закон распределения остатков.

Исследование остатков полезно начинать с изучения их графика. Он может показать наличие какой-то зависимости, не учтенной в модели. Скажем, при подборе простой линейной зависимости между x и y график остатков может показать необходимость перехода к нелинейной модели (квадратичной, полиномиальной, экспоненциальной) или включения в модель периодических компонент.

Для проверки нормальности распределения остатков чаще всего используется график на нормальной вероятностной бумаге (п. 5.2, 5.5), а также критерии типа Колмогорова—Смирнова, хи-квадрат и др., подробно разобранные в гл. 10.

Для проверки независимости остатков обычно используются критерий серий и критерий Дарбина—Уотсона. Их описание можно найти в [41]. В случае выявления сильной корреляции остатков следует перейти от регрессионной модели к моделям типа авторегрессии-скользящего среднего и возможно использовать разностные и сезонные операторы удаления тренда. Эти методики подробно описаны в гл. 12 и 14.

Выбросы. График остатков хорошо показывает и резко отклоняющиеся от модели наблюдения — *выбросы*. Подобным наблюдениям надо уделять особо пристальное внимание, так как их присутствие может грубо исказить значения оценок (особенно если для их получения используется метод наименьших квадратов). Устранение эффектов выбросов может проводиться либо с помощью удаления этих точек из анализируемых данных (эта процедура называется *цензурированием*), либо с помощью применения методов оценивания параметров, устойчивых к подобным грубым отклонениям. Иллюстрацией эффекта выброса является пример 8.2к, разобранный в п. 8.7.

Множественная регрессия. В более общем случае задача регрессионного анализа предполагает установление линейной зависимости между группой независимых переменных x_1, x_2, \dots, x_k (здесь индекс k означает номер переменной, а не номер наблюдения этой переменной) и одномерным откликом y . Эта обширная тема, носящая название *множественной регрессии*, не нашла отражения в данной книге. С ней можно познакомиться в [36], [41]. Заметим, что для решения этой задачи существуют мощные компьютерные процедуры, они имеются и в разбираемых нами пакетах.

Стратегия анализа адекватности подобранной модели в задаче множественной регрессии в целом аналогична задаче простой регрессии и сводится к детальному анализу остатков.

Замечания. 1. Имеются процедуры решения задач множественной регрессии, реализующие автоматический выбор тех переменных, которые оказывают существенное влияние на отклик, и отсеивание несущественных переменных. Эти методы носят название *шаговой регрессии*, они весьма эффективны на практике.

2. Наибольшие трудности в задачах поиска зависимости от нескольких переменных возникают, когда сами эти переменные сильно взаимозависимы. Это весьма характерная ситуация для многих экономических задач. Показателем подобной зависимости служит матрица корреляций переменных x_1, x_2, \dots, x_k . Самой простой рекомендацией при сильно зависимых переменных является

удаление части из них и проведение повторных расчетов. Затем проводится сравнение полученных результатов. Другой особенностью подобных задач может являться эффект, когда каждая из переменных x_1, x_2, \dots, x_k действует на отклик не только независимо от других, но и порождает совместное воздействие. Для учета этого в модель, кроме переменных x_1, x_2, \dots, x_k можно включать их совместные произведения, например, переменные $x_1 \cdot x_2, x_1 \cdot x_3, x_2 \cdot x_3$ и т.д. Однако в задачах множественной регрессии лучше стремиться сократить общее число переменных, от которых будет искаться зависимость, так как это существенно упрощает последующий анализ модели.

Нелинейная регрессия. Скажем еще несколько слов о задаче *нелинейной регрессии*. В этом случае параметры модели θ входят в подбираемую регрессионную функцию $f(\mathbf{x}, \theta)$ нелинейным образом. Поэтому нахождение оценок параметров модели $\hat{\theta}$ в аналитическом виде обычно невозможно, так что эти оценки вычисляются на компьютере методом итеративного приближения. Используемые здесь вычислительные алгоритмы довольно сложны и не всегда работают успешно. Кроме того, огромный произвол в выборе вида самой нелинейной зависимости весьма затрудняет осмысленный подбор этой зависимости. На наш взгляд, использование методов нелинейной регрессии оправданно в основном, когда вид регрессионной зависимости заранее известен из теоретических соображений.

8.3. Простая линейная регрессия

Проиллюстрируем основные идеи обработки регрессионного эксперимента (8.3) на примере простой линейной регрессии. Так называют задачу регрессии, в которой \mathbf{x} и \mathbf{y} — одномерные величины (поэтому мы будем обозначать их x и y), а функция $f(x, \theta)$ имеет вид $A + bx$, где $\theta = (A, b)$. В этом случае соотношение (8.3) принимает вид:

$$y_i = A + bx_i + \varepsilon_i \quad i = 1, \dots, n. \quad (8.4)$$

Здесь x_1, \dots, x_n — заданные числа (значения фактора); y_1, \dots, y_n — наблюдаемые значения отклика; $\varepsilon_1, \dots, \varepsilon_n$ — независимые (ненаблюдаемые) одинаково распределенные случайные величины.

Гауссовская модель. При решении задачи (8.4) (как и во многих других случаях) используются два основных подхода: непараметрический и гауссовский, они различаются характером предположений относительно закона распределения случайных величин ε . Сначала мы рассмотрим гауссовскую модель простой линейной регрессии. В ней дополнительно к вышесказанному предполагается, что величины ε_i распределены по нормальному закону $N(0, \sigma^2)$ с некоторой неизвестной дисперсией σ^2 .

Метод наименьших квадратов. При выборе методов определения параметров регрессионной модели можно руководствоваться различными подходами. Один из наиболее естественных и распространенных состоит в том, что при «хорошем» выборе оценки $\hat{\theta}$ параметра модели θ величины $y_i - f(\mathbf{x}_i, \theta)$ (в случае простой линейной регрессии — величины $y_i - A - bx_i$) должны в совокупности быть близки к нулю. Мэру близости совокупности этих величин (они обычно называются *остатками*) к нулю можно выбирать по-разному (например, максимум модулей, сумму модулей и т.д.), но наиболее простые формулы расчета получаются, если в качестве этой меры выбрать сумму квадратов:

$$\sum_{i=1}^n [y_i - A - bx_i]^2 \rightarrow \min_{A, b}.$$

Определение. *Методом наименьших квадратов называется способ подбора параметров регрессионной модели исходя из минимизации суммы квадратов остатков.*

Сам по себе метод наименьших квадратов не связан с какими-либо предположениями о распределении случайных ошибок $\varepsilon_1, \dots, \varepsilon_n$, он может применяться и тогда, когда мы не считаем эти ошибки случайными (например, в задачах сглаживания экспериментальных данных). Однако мы будем рассматривать метод наименьших квадратов в связи с гауссовской моделью. Причины этого следующие:

- именно в гауссовской модели метод наименьших квадратов обладает определенными свойствами оптимальности (мы их обсуждать не будем);
- в гауссовской модели получаемые с помощью этого метода оценки неизвестных параметров обладают ясными статистическими свойствами (которые мы обсудим).

Оценки метода наименьших квадратов. Чтобы упростить дальнейшие формулы, перепишем соотношение (8.4) в виде

$$y_i = a + b(x_i - \bar{x}) + \varepsilon_i \quad i = 1, \dots, n, \quad (8.5)$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $a = A + b\bar{x}$. Этот переход означает перенос начала отсчета на оси абсцисс в точку \bar{x} , которая служит центром совокупности (выборки) x_1, \dots, x_n .

Для нахождения оценок по методу наименьших квадратов нам надо выяснить, при каких (a, b) достигается минимум выражения

$$\sum_{i=1}^n [y_i - a - b(x_i - \bar{x})]^2. \quad (8.6)$$

Приравнивая нулю частные производные по a и b выражения (8.6), получим систему уравнений относительно неизвестных a и b :

$$\begin{cases} \sum_{i=1}^n [y_i - a - b(x_i - \bar{x})] = 0, \\ \sum_{i=1}^n (x_i - \bar{x}) [y_i - a - b(x_i - \bar{x})] = 0. \end{cases}$$

Ее решение (\hat{a}, \hat{b}) легко найти:

$$\hat{a} = \bar{y} \quad (\text{где } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i), \quad (8.7)$$

$$\hat{b} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (8.8)$$

Величины \hat{a} , \hat{b} и будут полученными по методу наименьших квадратов оценками неизвестных нам величин a и b .

Свойства оценок. Естественно, возникает вопрос: как соотносятся полученные значения \hat{a} и \hat{b} с истинными значениями a и b или, другими словами, каково качество оценок метода наименьших квадратов \hat{a} и \hat{b} . Для ответа на этот вопрос укажем некоторые свойства этих оценок.

- 1) $M\hat{a} = a$ и $M\hat{b} = b$;
- 2) $D\hat{a} = \sigma^2/n$, и $D\hat{b} = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2$;
- 3) $\text{cov}(\hat{a}, \hat{b}) = 0$;
- 4) случайные величины \hat{a} и \hat{b} обе распределены по нормальному закону;
- 5) \hat{a} и \hat{b} независимы как случайные величины.

Доказательства утверждений 1–3 могут быть получены прямым вычислением, используя выражения (8.7) и (8.8). Покажем, например, что $M\hat{b} = b$.

$$M\hat{b} = M \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) M(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

поскольку величины x_1, \dots, x_n и \bar{x} не случайны и содержащие только их выражения можно вынести из-под знака математического ожидания. Далее, поскольку $M\varepsilon_i = 0$ и $M\bar{\varepsilon} = 0$, то

$$M(y_i - \bar{y}) = My_i - M\bar{y} = a + b(x_i - \bar{x}) - a = b(x_i - \bar{x}).$$

Подставляя это выражение в предыдущую формулу, находим, что $M\hat{b} = b$.

Заметим, что свойства 1–3 не используют предположения о нормальном характере ошибок в модели (8.4) или (8.5). Зато свойство 4 верно только в гауссовском случае. Доказательство свойства 4 следует из вида формул (8.7), (8.8), которые по отношению к y_1, \dots, y_n имеют вид линейных функций, а линейные комбинации независимых нормальных случайных величин, как мы отмечали ранее, сами распределены нормально.

Свойство 5 есть следствие нормальности ошибок и свойства 3. Независимость оценок \hat{a} , \hat{b} заметно упрощает дальнейший анализ. В первую очередь ради этого модель (8.4) была заменена на (8.5).

В совокупности свойства 1–4 дают важные результаты, характеризующие качество оценок \hat{a} и \hat{b} :

$$\hat{a} \sim N\left(a, \frac{\sigma^2}{n}\right), \quad \hat{b} \sim N\left(b, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right). \quad (8.9)$$

Оценка дисперсии. В модели (8.5), кроме a и b , есть еще один неизвестный параметр — дисперсия σ^2 ошибок наблюдения. Этот параметр явно входит в соотношения (8.9) и тем самым влияет на точность оценок. Поэтому σ^2 , в свою очередь, требует оценивания. Ключ к этому дает остаточная сумма квадратов

$$\sum_{i=1}^n [y_i - \hat{a} - \hat{b}(x_i - \bar{x})]^2. \quad (8.10)$$

Можно доказать, что в гауссовской модели выражение (8.10) является независимой от \hat{a} и \hat{b} случайной величиной, имеющей распределение $\sigma^2 \chi^2(n-2)$, где $\chi^2(n-2)$ — распределение хи-квадрат с $n-2$ степенями свободы. Благодаря этому свойству мы можем построить для σ^2 несмещенную оценку s^2 :

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n [y_i - \hat{a} - \hat{b}(x_i - \bar{x})]^2. \quad (8.11)$$

Поскольку s^2 не зависит от \hat{a} и \hat{b} , отношения

$$\sqrt{n} \frac{\hat{a} - a}{s} \quad \text{и} \quad \frac{\hat{b} - b}{s} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (8.12)$$

имеют распределение Стьюдента с $(n-2)$ степенями свободы. Это позволяет легко построить для параметров a и b доверительные интервалы и указать тем самым, каковы статистические свойства погрешности при их оценивании посредством (8.7), (8.8).

Проверка гипотез о коэффициенте наклона. Наиболее часто в задаче простой линейной регрессии возникает вопрос о равенстве нулю коэффициента наклона. Со статистической точки зрения это означает проверку гипотезы $H : b = 0$. Важность этой гипотезы объясняется тем, что в этом случае переменная y изменяется чисто случайно, не завися от значения x .

Против двусторонних альтернатив $b \neq 0$ гипотезу H следует отвергнуть на уровне значимости α , если число 0 не входит в доверительный интервал для b , который мы стандартным образом строим с помощью указанного выше стьюдентова отношения (8.12). Другая редакция этой идеи, с использованием F -отношения, дана, например, в [41].

Замечание. Стоит обратить внимание на сходство результата (8.11) с тем, что мы уже встречали, имея дело с нормальной выборкой. Пусть сейчас y_1, \dots, y_n — выборка из $N(a, \sigma^2)$. Оценками a, σ^2 служат соответственно $\hat{a} = \bar{y}$ и $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$. При этом a и s^2 независимы как случайные величины, и $\sum_{i=1}^n (y_i - \bar{y})^2$ распределена как $\sigma^2 \chi^2(n-1)$. Для большего сходства с (8.11) s^2 можно записать в виде $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{a})^2$. Отмеченная параллель с нормальной выборкой простирается и на более сложные линейные гауссовские модели.

8.4. О проверке предпосылок в задаче регрессионного анализа

Уверенность в том, что соотношение (8.4) или (8.5) и другие предпосылки правильно отражают условия опыта, никогда не бывает полной. Поэтому нужны средства для проверки хотя бы некоторых из основных постулатов. Всех их из-за ограниченности информации, доставляемой единственным экспериментом, который мы обсуждаем, проверить нельзя. Эти постулаты сложились на основе коллективного предыдущего опыта.

Независимость наблюдений. Наиболее фундаментальным является предположение о том, что результаты отдельных измерений представляют собой независимые случайные величины. Проверить эту предпосылку статистическими средствами достаточно трудно, а при неизвестном виде зависимости между наблюдениями — практически невозможно. Ее выполнение должно быть обеспечено всей методикой опыта.

Одинаковая распределенность ошибок. Второе по важности значение имеет предположение о том, что ошибки эксперимента как случайные величины распределены одинаково. Иначе говоря, это означает, что измерения отклика имеют равную точность при всех значениях фактора — если случайную составляющую отклика мыслить как ошибку при его измерении. Если же эти случайные составляющие мы трактуем как выражение изменчивости, внутренне присущей переменной y , то обсуждаемое предположение означает, что эта изменчивость не испытывает влияния со стороны факторов. Это требование тоже трудно поддается статистическому контролю и должно поддерживаться методикой эксперимента. В тех случаях, когда невыполнимость этого условия ясна, классическая регрессионная схема использована быть не может. Исключение составляет скорее теоретически мыслимый, чем практически возможный случай, когда известна зависимость от x распределения ε . В других случаях статистическая неоднородность может помешать применению регрессионного анализа.

Вид функциональной зависимости. Следующим по важности является предположение о виде функциональной зависимости (8.3). Решающее значение имеет правильный выбор выражения для $f(\cdot, \cdot)$, особенно когда речь идет о прогнозе отклика вне области, в которой проводились измерения. Важно выбрать функцию $f(x, \theta)$ так, чтобы она не просто хорошо описывала закономерную часть отклика, но и имела «физический» смысл, т.е. открывала какую-то объективную закономерность. Впрочем, полезны бывают и чисто эмпирические, «подгоночные» формулы, поскольку они позволяют в сжатой форме приближенно выразить зависимость y от x . Поэтому выбор типа регрессионной зависимости (8.3) является самой острой проблемой в любом исследовании. О том, как можно проверить его корректность, мы будем говорить ниже, на примере простой линейной регрессии (8.4).

Нормальность распределений ошибок. Остается сказать о последней предпосылке, которая и выделяет гауссовский регрессионный анализ. Речь идет о том предположении, что случайные величины $\varepsilon_1, \dots, \varepsilon_n$ распределены по нормальному закону. На буквальном выполнении этого условия настаивать нет необходимости. Но без его хотя бы приближенного осуществления нельзя использовать те статистические выводы, которые мы сумели сделать в п. 8.3. В случае одномерной регрессии для проверки этого условия можно воспользоваться тем, что при справедливости предположений модели остатки $y_i - \hat{y}_i$, где $\hat{y}_i = \hat{a} - \hat{b}(x_i - \bar{x})$, должны вести себя практически так же, как независимые одинаково распределенные случайные величины.

Проверка адекватности линейной регрессии. Обратимся к проверке адекватности модели регрессии на примере простой линейной регрессии (8.4). Основой для этого служат видимые отклонения от установленной закономерности, т.е. величины $y_i - \hat{y}_i$, $i = 1, \dots, n$, где

$$\hat{y}_i = \hat{a} + \hat{b}(x_i - \bar{x}). \quad (8.13)$$

Поскольку фактор x — одномерная переменная, точки $(x_i, y_i - \hat{y}_i)$ можно изобразить на чертеже. Такое наглядное представление наблюдений позволяет иногда обнаружить в поведении остатков какую-либо зависимость от x . Однако глазомерный анализ остатков возможен не всегда и не является правилом с контролируемыми свойствами. Нужны более точные методы. Мы расскажем об одном из таких методов, который можно применять, если при составлении плана эксперимента предусматриваются многократные измерения отклика при некоторых значениях факторов.

Проверка адекватности регрессионной модели при наличии повторных наблюдений. При наличии повторных наблюдений при некоторых (а еще лучше при всех) значениях факторов у нас появляется возможность получить еще одну оценку величины изменчивости случайной составляющей ε и сравнить ее с полученной ранее оценкой дисперсии σ^2 .

Предположим, что в модели (8.5) при каждом значении $x = x_i$, $i = 1, \dots, n$ проводится m независимых измерений отклика. Их результаты при данном i удобно обозначить через y_{i1}, \dots, y_{im} . При этом y_{ij} как случайные величины независимы при всех $j = 1, \dots, m$, $i = 1, \dots, n$. (Можно изучить и такой случай, когда число измерений при данном x_i находится в зависимости от i . Это несколько усложнило бы следующие ниже формулы, не меняя их принципиально.) От выборки y_{i1}, \dots, y_{im} перейдем к

$$y_i = \frac{1}{m} \sum_{j=1}^m y_{ij}, \quad s_i^2 = \frac{1}{m-1} \sum_{j=1}^m (y_{ij} - y_i)^2. \quad (8.14)$$

Мы уже вспоминали, что величины $(m-1)s_i^2$, $i = 1, \dots, n$ распределены как $\sigma^2 \chi^2(m-1)$ и стохастически независимы от y_i . Объединяя, мы получим, что

$$(m-1) \sum_{i=1}^n s_i^2 = \sigma^2 \chi^2[n(m-1)]. \quad (8.15)$$

Как мы видели в п. 8.3, другую оценку дисперсии ошибок дает остаточная сумма квадратов:

$$\sum_{i=1}^n [y_i - \hat{a} - \hat{b}(x_i - \bar{x})]^2 = \frac{\sigma^2}{m} \chi^2(n-2). \quad (8.16)$$

Мы использовали формулу (8.10). Роль y_i в ней играет теперь y_i , причем $Dy_i = \sigma^2/m$. Подчеркнем, что соотношение (8.16) действует, только если регрессионная модель (8.4) или (8.5) выбрана правильно. В противном случае в остаточную сумму квадратов, кроме случайных ошибок, входят и систематические, а потому она получает тенденцию к возрастанию.

Выражения (8.15) и (8.16) позволяют составить F -отношение (как мы поступали неоднократно, обсуждая дисперсионный анализ):

$$F = \frac{\frac{m}{n-2} \sum_{i=1}^n [y_i - \hat{a} - \hat{b}(x_i - \bar{x})]^2}{\frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - y_i)^2} \quad (8.17)$$

с числом степеней свободы $(n-2, n(m-1))$.

Гипотеза о линейности должна быть отвергнута, если наблюдаемое в опыте значение F (8.17) оказывается неправдоподобно большим с точки зрения F -распределения с $n-2, n(m-1)$ степенями свободы.

Более подробную информацию о критериях проверки адекватности модели, основанных на анализе остатков $y_i - \hat{y}_i$, можно найти в [41].

8.5. Непараметрическая линейная регрессия

Мы уже говорили выше, что, когда есть сомнения в приложимости гауссовской модели, вместо метода наименьших квадратов следует использовать другие. Здесь будет рассказано об одном из таких методов, основанном на рангах наблюдений.

Модель. Рассмотрим схему простой линейной регрессии

$$y_i = A + bx_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (8.18)$$

где $\varepsilon_1, \dots, \varepsilon_n$ — независимые одинаково распределенные (далее — н.о.р.) случайные величины. Будем считать, что они распределены непрерывно (не уточняя далее, по какому именно закону). Выводы о зависимости между y и x будем основывать на рангах y . Ясно, что в таком случае ничего определенного о величине A сказать не удастся, так как изменение всех y_i на одну и ту же постоянную величину не изменяет рангов y_1, \dots, y_n . Предметом интереса остается только коэффициент наклона b . Постараемся найти его оценку в схеме (8.18).

Оценка коэффициента наклона. Для дальнейшего удобно так занумеровать наблюдения, чтобы

$$x_1 < x_2 < \dots < x_n.$$

При такой нумерации легче следить за поведением остатков.

Если из наблюдаемых величин y_i вычесть истинные значения bx_i , то остатки $y_i - bx_i = A + \varepsilon_i$, $i = 1, \dots, n$ образуют последовательность н.о.р. случайных величин. Не зная b , мы будем вычитать из y_i переменную величину βx_i , где β изменяется по нашему произволу. Остатки $y_i - \beta x_i$, $i = 1, \dots, n$ будут похожи на совокупность н.о.р. случайных величин, когда β близко к b — и тем более похожи, чем ближе β к b . Если нет, то остатки будут проявлять тенденцию к возрастанию или убыванию вместе с номером i (это зависит от знака разности $b - \beta$). В этом легко убедиться, переписав $y_i - \beta x_i$ в следующем виде:

$$y_i - \beta x_i = y_i - bx_i + x_i(b - \beta) = A + \varepsilon_i + x_i(b - \beta).$$

Так, при положительном значении разности $(b - \beta)$ остатки $y_i - \beta x_i$ будут тем больше, чем больше номер i , учитывая, что x_i упорядочены в порядке возрастания.

Тенденцию изменения значений $y_i - \beta x_i$ с изменением номера i или отсутствие таковой можно обнаружить с помощью коэффициентов корреляции. Если закон распределения не известен, надо использовать коэффициенты ранговой корреляции, и ниже эта возможность будет использована. (Подробнее о коэффициентах ранговой корреляции смотри параграф 9.3.) Но прежде посмотрим, к чему приводит этот подход при использовании обычного коэффициента корреляции Пирсона (см. п. 1.8.1). Выборочный коэффициент корреляции Пирсона по совокупности $(x, y_i - \beta x_i)$ имеет вид:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) [(y_i - \bar{y}) - \beta(x_i - \bar{x})]}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n [(y_i - \bar{y}) - \beta(x_i - \bar{x})]^2}}.$$

Наименьшей зависимости остатков $y_i - \beta x_i$ от x_i ($i = 1, \dots, n$) соответствует значение $r = 0$. По отношению к β это дает уравнение

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \beta \sum_{i=1}^n (x_i - \bar{x})^2$$

Его решение — уже известное нам выражение (8.8). Итак, использование коэффициента корреляции К. Пирсона приводит для b к оценке наименьших квадратов. Поэтому можно предположить, что использование коэффициента ранговой корреляции тоже будет успешным.

Итак, для двух рядов чисел

$$\begin{aligned} y_1 - \beta x_1, y_2 - \beta x_2, \dots, y_n - \beta x_n \\ x_1, x_2, \dots, x_n \end{aligned} \quad (8.19)$$

составим коэффициенты ранговой корреляции: ρ Спирмена и τ Кендэла. Коэффициент ранговой корреляции ρ Спирмена получается заменой величин $y_i - \beta x_i$ и x_i в коэффициенте выборочной корреляции Пирсона на их ранги. В данном случае, учитывая, что x_i упорядочены в порядке возрастания, ранг x_i равен i (при условии отсутствия совпадений между x_i) Таким образом,

$$\rho = \frac{\sum_{i=1}^n (i - \frac{n+1}{2}) (R_i - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^n (i - \frac{n+1}{2})^2} \sqrt{\sum_{i=1}^n (R_i - \frac{n+1}{2})^2}}, \quad (8.20)$$

где R_i — ранг величины $y_i - \beta x_i$. Поскольку R_i принимает значения от 1 до n , найдем: $\sum_{i=1}^n (R_i - \frac{n+1}{2})^2 = \sum_{i=1}^n (i - \frac{n+1}{2})^2 = \frac{n(n^2-1)}{12}$. Преобразовав числитель выражения (8.20), окончательно запишем ρ в виде:

$$\rho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^n (i - R_i)^2. \quad (8.21)$$

Коэффициент корреляции τ Кендэла определяется как

$$\tau = \frac{2(P-Q)}{n(n-1)} = \frac{2K}{n(n-1)}, \quad (8.22)$$

где P и Q — соответственно число согласованных и несогласованных пар $(y_i - \beta x_i, x_i)$ и $(y_j - \beta x_j, x_j)$ для всех i, j таких, что $i < j$. Здесь пары $(y_i - \beta x_i, x_i)$ и $(y_j - \beta x_j, x_j)$ называются согласованными, если $x_i > x_j$ и $y_i - \beta x_i > y_j - \beta x_j$, либо $x_i < x_j$ и $y_i - \beta x_i < y_j - \beta x_j$. В противном случае пары называются несогласованными.

Величина $K = P - Q$ называется *статистикой Кендэла*. Ее можно записать в следующем виде, учитывая, что $x_1 < \dots < x_n$:

$$K = \sum_{1 \leq i < j \leq n} \text{sign}(y_j - \beta x_j - y_i + \beta x_i) = \sum_{1 \leq i < j \leq n} \text{sign}(R_j - R_i).$$

Чтобы подчеркнуть зависимость коэффициентов τ и ρ от β , будем далее писать $\tau(\beta)$ и $\rho(\beta)$. Измеренная с помощью этих коэффициентов ранговой корреляции зависимость между рядами (8.19) будет наименьшей, если выбрать β так, чтобы

$$\tau(\beta) = 0, \quad (8.23)$$

или

$$\rho(\beta) = 0. \quad (8.24)$$

Чтобы решить уравнение (8.23) или (8.24), надо представить себе зависимость $\tau(\beta)$, $\rho(\beta)$ от β . Выясним, как выглядят эти функции.

При β отрицательных и очень больших по абсолютной величине порядок следования разностей $y_i - \beta x_i$, $i = 1, \dots, n$ определяется исключительно числами x_1, \dots, x_n и совпадает с порядком их следования. Следовательно, при таких β ($\beta \rightarrow -\infty$) оба коэффициента ранговой корреляции $\tau(\beta)$ и $\rho(\beta)$ равны единице.

Пусть теперь β начинает возрастать (уходит из области очень больших отрицательных чисел, приближаясь к положительной полуоси). Первое изменение порядка следования остатков $y_1 - \beta x_1, \dots, y_n - \beta x_n$ произойдет при первом совпадении двух из них:

$$y_i - \beta x_i = y_j - \beta x_j \quad (8.25)$$

для каких-то i, j . Оба коэффициента ранговой корреляции при этом уменьшатся.

При дальнейшем увеличении β такие изменения $\tau(\beta)$, $\rho(\beta)$ будут происходить всякий раз, как будет достигаться равенство (8.25). Следовательно, значения β , при которых (скачком) изменяются $\tau(\beta)$ и $\rho(\beta)$, суть

$$\beta_{ij} = \frac{y_j - y_i}{x_j - x_i} \quad \text{где } 1 \leq i < j \leq n, \quad (8.26)$$

если все числа x_1, \dots, x_n различны между собой. (Если среди них есть совпадающие, в выражении (8.26) участвуют лишь такие i, j для которых $x_i - x_j \neq 0$. Точек изменения функций $\tau(\beta)$, $\rho(\beta)$ оказывается в этом случае меньше, чем число сочетаний C_n^2 , но величины скачков могут быть больше).

Функции $\tau(\beta)$, $\rho(\beta)$ таковы, что их симметрично расположенные скачки равны по величине. Поэтому их графики проходят через ноль при таком $\hat{\beta}$, что левее $\hat{\beta}$ и правее него остаются по одинаковому количеству точек разрыва (8.22). Иначе говоря:

$$\hat{\beta} = \text{med} \left\{ \frac{y_j - y_i}{x_j - x_i}, \quad \text{все } 1 \leq i < j \leq n \mid x_i \neq x_j \right\}. \quad (8.27)$$

Выражение (8.27) дает оценку коэффициента наклона (новую по сравнению с (8.8)). Можно показать, что в условиях гауссовской модели она менее точна, чем (8.8), но зато (8.27) применима в гораздо более широких условиях.

Доверительные интервалы для b . Основываясь на функциях $\tau(\beta)$, $\rho(\beta)$, можно построить доверительные интервалы для неизвестного b . Выберем коэффициент доверия $1 - 2\alpha$. Пусть для данного n (объем наблюдений) τ_α (соответственно, ρ_α) обозначает верхнее критическое значение для коэффициента ранговой корреляции τ (соответственно ρ). Тем самым,

$$P\{|\tau| \leq \tau_\alpha\} = 1 - 2\alpha \quad \text{и} \quad P\{|\rho| \leq \rho_\alpha\} = 1 - 2\alpha. \quad (8.28)$$

(Дискретный характер распределения вероятностей между возможными значениями τ , ρ приводит к тому, что соотношения (8.28) выполняются не для всех α . Надо либо выбрать такое α , чтобы (8.28) имело место, либо же в качестве τ_α (или ρ_α) взять минимальное значение, при котором $P\{|\tau| \leq \tau_\alpha\} \geq 1 - 2\alpha$ (для ρ_α — аналогично).

Доверительные интервалы для b с коэффициентом доверия не меньше $1 - 2\alpha$ имеют вид:

$$\{\beta : |\tau(\beta)| \leq \tau_\alpha\} \quad \text{или} \quad \{\beta : |\rho(\beta)| \leq \rho_\alpha\}, \quad (8.29)$$

в зависимости от выбора коэффициента ранговой корреляции.

На рис. 8.1 изображен график $\tau(\beta)$ при $n = 5$. Точки скачков функции $\tau(\beta)$ выделяют доверительный интервал.

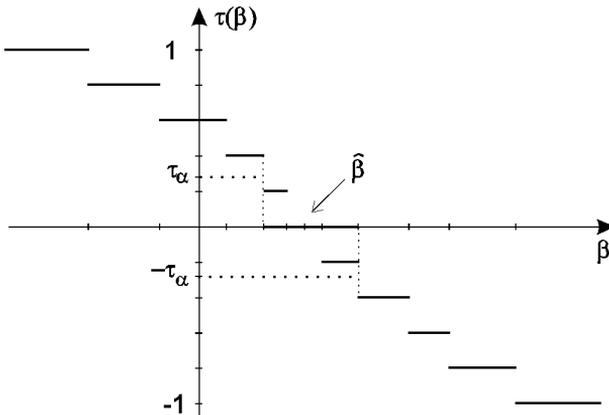


Рис. 8.1. Доверительный интервал для коэффициента корреляции $\tau(\beta)$ при $n = 5$

График функции $\rho(\beta)$ сложнее, так как величины его скачков не постоянны. В дальнейшем для построения доверительного интервала

будем использовать коэффициент ранговой корреляции τ , так как по указанной причине с ним действовать проще. Обсуждение доверительного интервала для ρ приведено, например, в [113].

Учитывая, что таблицы распределения чаще составлены не для величины τ , а для статистики Кендэла K , введем функцию

$$K(\beta) = \frac{n(n-1)}{2} \tau(\beta),$$

для которой справедливо все сказанное ранее о $\tau(\beta)$. То есть доверительный интервал для b с коэффициентом доверия $1 - 2\alpha$ имеет вид:

$$\{\beta : |K(\beta)| \leq K_\alpha\},$$

где K_α есть решение уравнения $P\{|K| \leq K_\alpha\} = 1 - 2\alpha$. При этом вероятность P рассматривается в случае справедливости выдвинутой гипотезы о независимости двух рядов чисел (8.19). В [115] приведена таблица вероятностей хвостов распределения статистики K для $n = 4(1)40$. Чтобы воспользоваться этими таблицами, заметим, что $K_\alpha + 2$ удовлетворяет соотношению $P(K \geq K_\alpha + 2) = \alpha/2$.

Затем совокупность чисел $\frac{y_j - y_i}{x_j - x_i}$, $1 \leq i < j \leq n$, надо расположить в порядке возрастания. Мы предположим сейчас, что среди чисел x_1, \dots, x_n нет совпадающих. Обозначим элементы этой упорядоченной совокупности через $S^{(1)} \leq S^{(2)} \leq \dots \leq S^{(N)}$, $N = \frac{n(n-1)}{2}$. Положим $M_1 = (N - K_\alpha)/2$, $M_2 = (N + K_\alpha)/2$. В этих обозначениях доверительный интервал для b (8.29) имеет явный вид:

$$\{S^{(M_1)} < \beta < S^{(M_2+1)}\}.$$

При этом $P\{S^{(M_1)} < \beta < S^{(M_2+1)}\} = 1 - \alpha$. В случае больших n для K приходится использовать приближенное выражение, основанное на нормальной аппроксимации распределения коэффициента ранговой корреляции τ при гипотезе независимости. Получаем, что

$$K_\alpha \sim \sqrt{\frac{n(n-1)(2n+5)}{18}} u_{1-\alpha/2},$$

где $u_{1-\alpha/2}$ — квантиль уровня $1 - \alpha/2$ стандартного нормального распределения, т.е. решение уравнения $\Phi(u_{1-\alpha/2}) = 1 - \alpha/2$, где Φ — функция Лапласа $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-t^2/2} dt$.

Поправки при совпадениях. Укажем поправки, которые надо сделать при построении доверительного интервала для коэффициента b в том случае, когда среди x_i имеются совпадения. Наличие совпадений среди x_i соответствует повторным наблюдениям в этих точках. Обозначим через g число групп совпадающих значений x_i (т.е. число связей среди иксов), а через t_l — число совпадающих элементов в группе с

номером $l : l = 1, \dots, g$. Тогда значение K_α , получаемое при использовании нормальной аппроксимации для распределения коэффициента ранговой корреляции τ при гипотезе независимости, имеет вид:

$$K_\alpha \sim \frac{\sqrt{n(n-1)(2n+5) - \sum_{l=1}^g t_l(t_l-1)(2t_l+5)}}{18} u_{1-\alpha/2}. \quad (8.30)$$

Этот результат был получен П. Сеном [138]. Соответствующие значения M_1 и M_2 равны:

$$M_1 = \left[\frac{N - K_\alpha}{2} \right], \quad M_2 = \left[\frac{N + K_\alpha}{2} \right] + 1. \quad (8.31)$$

Ниже будет проиллюстрировано применение изложенных методов в практической задаче.

8.6. Практический пример

В качестве примера рассмотрим использование линейного регрессионного анализа в задаче восстановления зависимости между входом и выходом измерительно-регистрирующей системы. Подобные задачи широко распространены в экспериментальных исследованиях, во многих предметных областях они называются по-своему: градуировка, калибровка, тарировка и т.д. Необходимость применения статистических методов для решения подобных задач в последнее время возросла как в связи с усложнением средств измерений, так и в связи с повышением требований к их точности и надежности. А использование ЭВМ значительно упростило и расширило возможности обработки результатов подобных экспериментов.

Рассмотрим измерительно-регистрирующий тракт тензовесов, используемых для измерения сил и моментов сил, действующих на тело при продувке его в аэродинамической трубе. Для этих измерений в тензовесах используются тензодатчики, определенным образом расположенные на конструкции весов. В основу работы тензодатчика положен эффект изменения сопротивления чувствительного элемента при его сжатии или расширении. Через все тензодатчики пропускают электрический ток, а сигналы тензодатчиков (показывающие напряжения на тензоэлементах) через усилитель и аналого-цифровой преобразователь регистрируют с помощью компьютера.

Хотя характеристики каждого звена тензовесов можно измерить, рассчитать на основе этих измерений свойства связи между входом и выходом измерительной системы (т.е. между силами и моментами сил, действующих на продуваемое тело, и напряжениями на тензодатчиках) весьма трудно, а оценить точность этих расчетов еще труднее. Гораздо

проще эта задача решается с помощью градуировочного эксперимента: на тензовесы оказывается воздействие эталонной силой (моментом сил) и фиксируется значение отклика на выходе системы. Варьируя значения эталонной силы в пределах рабочего диапазона тензовесов, мы получаем данные, по которым следует восстановить вид зависимости между входом и выходом измерительной системы.

Таблица 8.1

Данные калибровочного эксперимента одной компоненты тензовесов

Эталонная сила x_i $i = 1, \dots, 6$	0.0	0.2	0.4	0.6	0.8	1.0
Значение отклика y_{ij}						
$j = 1$	31.0	110.0	186.5	266.7	345.5	425.6
$j = 2$	29.8	111.0	191.0	269.7	349.3	425.9
$j = 3$	29.1	109.6	187.1	270.1	349.7	426.5
$j = 4$	29.0	111.0	190.3	270.2	349.9	426.5
$j = 5$	29.15	109.6	186.7	266.55	347.05	427.0
$j = 6$	28.2	110.35	190.95	270.25	349.8	427.0
Средние значения y_i .	29.38	110.26	188.76	268.92	348.54	426.42
Значения s_i^2	0.894	0.408	4.858	3.191	3.364	0.326

В табл. 8.1 приведены данные градуировочного эксперимента одной компоненты тензовесов, предназначенной для измерения силы лобового сопротивления. В ходе эксперимента значения эталонной силы x изменялись от 0 до 1 кг с шагом 0.2 кг, и для каждого значения силы регистрировалось значение отклика y в десятках мВ. Измерения повторялись 6 раз. В таблице приведены также средние отклики y_i и стандартные отклонения s_i^2 . Графическое изображение этих данных дано на рис. 8.2.

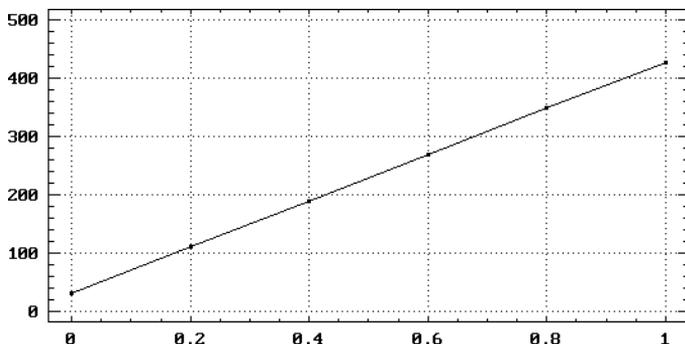


Рис. 8.2. Графическая зависимость y_i от x_i .

Поскольку при правильном расположении чувствительных элементов на балках усилия на тензодатчики должны линейно зависеть от

действующих на тело сил и моментов сил, а тензодатчики осуществляют линейное преобразование силы в напряжения электрического тока, естественно искать связь между силой x и результирующим напряжением y в виде:

$$y = A + bx + \varepsilon, \quad (8.32)$$

т.е. решать задачу простой линейной регрессии. Учитывая структуру экспериментальных данных, перепишем (8.32) следующим образом: $y_{ij} = A + bx_i + \varepsilon_{ij}$, $i = 1, \dots, 6$, $j = 1, \dots, 6$, и приведем его к виду, аналогичному (8.5):

$$y_{ij} = a + b(x_i - \bar{x}) + \varepsilon_{ij} \quad i = 1, \dots, 6, \quad j = 1, \dots, 6.$$

Отметим, что требование независимости величин ε_{ij} должно обеспечиваться методикой проведения калибровочного эксперимента, когда съём каждого из значений y_{ij} осуществляется независимо от остальных. Величины ε_{ij} отражают как суммарное влияние внешних факторов, так и погрешности, возникающие в измерительно-регистрирующем тракте. Учитывая характер формирования случайных отклонений, величины ε_{ij} в рабочем диапазоне имеют обычно один и тот же закон распределения, который принято считать нормальным. Следовательно, у нас есть все основания для применения классического метода линейной регрессии.

Запишем выражение (8.6) для случая, когда в каждой точке x_i ($i = 1, \dots, n$) сделано одинаковое число измерений y_{ij} ($j = 1, \dots, m$). Имеем:

$$\sum_{i=1}^n \sum_{j=1}^m [y_{ij} - a - b(x_i - \bar{x})]. \quad (8.33)$$

Приравнивая к нулю производные по переменным a и b в выражении (8.33), получаем:

$$\sum_{i=1}^n \sum_{j=1}^m (y_{ij} - a - b(x_i - \bar{x})) = 0, \quad (8.34)$$

$$\sum_{i=1}^n \sum_{j=1}^m (x_i - \bar{x})(y_{ij} - a - b(x_i - \bar{x})) = 0.$$

Проводя суммирование в уравнениях (8.34) по индексу j и деление каждого из уравнений на компоненту m , имеем:

$$\sum_{i=1}^n (y_{i.} - a - b(x_i - \bar{x})) = 0, \quad (8.35)$$

$$\sum_{i=1}^n (x_i - \bar{x})(y_{i.} - a - b(x_i - \bar{x})) = 0,$$

где $y_{i.}$ определено в (8.14).

Полученная система уравнений отличается от системы, рассмотренной в п. 8.3, заменой y_i на $y_{i.}$. Таким образом, задача простой линейной регрессии с m наблюдениями в каждой точке x_i сводится к задаче с одним наблюдением в точке x_i , если в качестве этого наблюдения рассматривать величину $y_{i.} = \frac{1}{m} \sum_{j=1}^m y_{ij}$. Оценки параметров a и b , являющиеся решением системы (8.35), согласно (8.7), (8.8) суть

$$\hat{a} = \bar{y}, \quad \text{где } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_{i.} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m y_{ij}, \quad (8.36)$$

$$\hat{b} = \frac{\sum_{i=1}^n (y_{i.} - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (8.37)$$

Подставляя в (8.36) и (8.37) соответствующие значения из табл. 8.1, получаем $\hat{a} = 228.711$, $\hat{b} = 397.174$.

Статистические свойства оценок \hat{a} и \hat{b} указаны в п. 8.3, а именно:

$$\hat{a} \sim N\left(a, \frac{\sigma_1^2}{n}\right), \quad \hat{b} \sim N\left(b, \frac{\sigma_1^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right),$$

где σ_1^2 — дисперсия $\frac{1}{m} \sum_{j=1}^m \varepsilon_{ij}$. То есть $\sigma_1^2 = \frac{\sigma^2}{m}$, где σ^2 — дисперсия ε_{ij} .

Для построения доверительных интервалов для истинных значений коэффициентов a и b и проверки качества выбранной модели мы должны построить оценки дисперсии σ^2 или σ_1^2 . Согласно (8.11), несмещенной оценкой σ_1^2 является:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_{i.} - \hat{a} - \hat{b}(x_i - \bar{x}))^2.$$

Производя необходимые вычисления, получаем $s = 0.64526$.

Таким образом, используя выражение (8.12) и положения п. 5.3, получаем границы доверительных интервалов для a и b , а именно:

$$\hat{a} - \frac{s}{\sqrt{n}} t_{1-\alpha/2} < a < \hat{a} + \frac{s}{\sqrt{n}} t_{1-\alpha/2},$$

$$\hat{b} - \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} t_{1-\alpha/2} < b < \hat{b} + \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} t_{1-\alpha/2},$$

где $t_{1-\alpha/2}$ есть квантиль распределения Стьюдента с 4 степенями свободы при коэффициенте доверия $1 - \alpha$. Выбирая, например, $\alpha = 0.05$, по таблице (см. [19]) находим $t_{1-\alpha/2} \simeq 2.79$. Отсюда 95% доверительные интервалы для a и b равны:

$$227.8 < a < 229.6, \quad 394.5 < b < 399.9. \quad (8.38)$$

Как указывалось в п. 8.4, для оценки адекватности выбранной модели необходимо получить еще одну независимую от s^2 оценку дисперсии σ^2 . Это можно сделать, подставляя в выражение (8.15) значения s_i^2 из таблицы (8.1). То есть:

$$(m-1) \sum_{i=1}^n s_i^2 \simeq \sigma^2 \chi^2(n(m-1)).$$

Для проверки качества подобранной линейной модели составим F -отношение согласно выражению (8.17):

$$F = \frac{\frac{m}{n-2} \sum_{i=1}^n [y_i - \hat{a} - \hat{b}(x_i - \bar{x})]^2}{\frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - y_i.)^2} = \frac{ms^2}{\frac{1}{n} \sum_{i=1}^n s_i^2}.$$

Подставляя имеющиеся значения, получаем

$$F = \frac{60.64256}{(1/6)13.041} = 1.781256.$$

Учитывая, что величина F имеет F -распределение с (4, 3) степенями свободы, сравним полученное значение с процентными точками указанного распределения. По таблице [19] находим, что 2.5% точка F распределения равна 3.2499, 5% точка равна 2.6896 и 10% точка равна 2.1422. Мы видим, что полученное нами значение $F = 1.781256$ меньше приведенных процентных точек, что свидетельствует о хорошем качестве приближения данных линейной зависимостью.

Заметим, что в рассмотренной задаче основной интерес представляет коэффициент наклона (усиления) b , так как значение a зависит от регулировки аппаратуры и его можно менять по соображениям удобства экспериментатора.

Обсуждение. Представляет интерес сравнение полученной оценки коэффициента b и оценки, полученной с помощью непараметрического метода, изложенного в п. 8.4. Непараметрический метод оценки коэффициента b предполагает представление массива x_i в виде $x_1, x_1, x_1, x_1, x_1, x_1, x_2, \dots, x_2, x_3, \dots, x_3, \dots, x_6, \dots, x_6$, где каждое значение x_i повторяется 6 раз. Таким образом, объем массива x равен $N = 36$. Рассмотрим массив:

$$\beta_{ij} = \left\{ \frac{y_j - y_i}{x_j - x_i}, \text{ все } 1 \leq i < j \leq N, \text{ для которых } x_i \neq x_j \right\}.$$

Объем массива β_{ij} в нашем случае, с учетом повторений значений в массиве x , равен: $C_{36}^2 - 6C_6^2 = 540$. Согласно (8.27), новой оценкой коэффициента b будет являться величина $\tilde{\beta}$, равная:

$$\tilde{\beta} = \underset{\substack{i, j=1, \dots, n \\ x_i \neq x_j}}{\text{med}} \left\{ \frac{y_j - y_i}{x_j - x_i} \right\} = \frac{\beta^{(270)} + \beta^{(271)}}{2}.$$

Здесь $\beta^{(k)}$ обозначает k -й член упорядоченного в порядке возрастания массива β_{ij} . Расчет показывает, что $\tilde{\beta} = 397.5$. Сравнивая полученное значение $\tilde{\beta}$ с полученным ранее $\hat{\beta} = 397.174$ и доверительным интервалом для b , полученным в гауссовской модели, видим довольно хорошее согласие результатов. Интересно сравнить доверительный интервал для b в непараметрическом случае с полученным ранее в (8.38).

Для построения нового доверительного интервала воспользуемся выражениями (8.30), (8.31). В нашем случае $g = 6$, $t_l = 6$, при $l = 1, \dots, g$. Выбирая значение $\alpha = 0.05$ по таблице [19], получаем $u_{1-\alpha/2} = 1.96$. Следовательно, согласно (8.30):

$$K_\alpha \sim \frac{\sqrt{n(n-1)(2n+5) - \sum_{l=1}^g t_l(t_l-1)(2t_l+5)}}{18} u_{1-\alpha/2} \sim 141.$$

Отсюда доверительный интервал для b , согласно (8.31), имеет вид: $\beta^{(198)} < b < \beta^{(341)}$, или

$$395.8 < b < 399.4. \quad (8.39)$$

Сравнение выражений (8.38) и (8.39) показывает, что доверительный интервал, построенный непараметрическим методом, оказывается более узким. Причиной этого может быть либо действие случая, либо неполное согласие обрабатываемых данных с гауссовской моделью линейной регрессии. Чтобы в этом разобраться, следовало бы подвергнуть анализу совокупность видимых отклонений от линии регрессии (см. п. 8.4). Но мы не станем этого делать, а просто прервем исследование, удовлетворившись уже полученными результатами.

8.7. Регрессионный анализ в пакете SPSS

Выше на примере задачи простой линейной регрессии были разобраны основные понятия и методы решения регрессионных задач. Как отмечалось, эти задачи весьма разнородны по своим постановкам и по возможным алгоритмам построения оценок и проверки адекватности моделей. Краткий обзор основных подходов к исследованию регрессионных задач можно найти в [41]. Там же приведена краткая справка о регрессионных программах в таких широко распространенных пакетах, как BMDP-79, SPSS, SAS, Minitab. В последние годы появилась литература, описывающая работу некоторых отечественных пакетов, содержащих в основном методы регрессионного анализа [36], [84], [121]. В целом отметим, что комплектация статистических пакетов регрессионными программами сильно варьируется. Регрессионные процедуры SPSS довольно многообразны и снабжены эффективными дополнительными инструментами исследования. Однако от этого их использование усложняется и требует высокой статистической квалификации.

В пакете SPSS полностью отсутствуют непараметрические методы регрессионного анализа.

Часть процедур регрессионного анализа сосредоточена в базовом модуле пакета SPSS Base, а часть — в дополнительном модуле SPSS Trends. Кратко перечислим процедуры регрессионного анализа в базовом модуле пакета. В основном они сосредоточены в блоке **Regression** меню **Analyze** редактора данных пакета. Здесь представлены процедуры **Linear**, **Curve Estimation**, **Binary Logistic**, **Multinomial Logistic**, **Ordinal**, **Probit**, **Nonlinear**, **Weight Estimation**, **2-stage Least Squares**. Еще ряд процедур, использующих регрессионный анализ в более общих ситуациях, когда, скажем, часть предикторов является количественными, а часть — качественными или когда остатки регрессионной модели коррелированы, представлены в блоках **General Linear Model** (общая линейная модель), **Loglinear** (логлинейный анализ) и **Time Series** (анализ временных рядов) меню **Analyze** редактора пакета.

Укажем назначение и основные особенности наиболее употребительных из этих процедур.

Linear. Эта процедура решает задачи множественной линейной регрессии. Она позволяет задавать различные стратегии подбора предикторов в множественной регрессии (шаговая регрессия). Процедура снабжена обширным инструментарием для выяснения устойчивости подобранной модели и диагностики мультиколлинеарности предикторов. Вопросы, связанные с этой процедурой, описаны в [41], [68], [1], [42], [40].

Curve Estimation. Эта процедура позволяет установить различные типы функциональной связи между откликом (зависимой переменной) и одним предиктором (независимой переменной). Как частный случай, она включает простую линейную регрессию. Работа этой процедуры разбирается в примере 8.1к. Процедура включает как линейные по параметрам модели, так и нелинейные.

Процедуры **Binary Logistic**, **Multinomial Logistic**, **Ordinal**, **Probit** предназначены для работы с данными, измеренными в номинальных и порядковых шкалах (см. п. 9.1). Чаще всего эти процедуры используются в социологических, маркетинговых и медицинских исследованиях. Примером типичной задачи, для решения которой привлекаются эти процедуры, является установление связи между предпочтениями потребителей того или иного товара и, скажем, их полом, социальным статусом, образованием, уровнем доходов и т.п.

Nonlinear (нелинейная регрессия). Эта процедура вычисляет оценки наименьших квадратов для параметров в заданной пользователем нелинейной регрессионной модели (см. [12], [36], [41]). Задание этой модели в пакете осуществляется в окне ввода данных и параметров процедуры, очень похожем на аналогичное окно разобранный нами процедуры **Compute** (см. пример 2.2к).

Для эффективного использования большинства перечисленных процедур требуется определенная квалификация. Например, процедура **Linear** включает довольно много понятий, далеко выходящих за рамки этой книги и начальных курсов математической статистики. К таким понятиям относятся **Deleted Residual** (удаленные остатки), балансирующие статистики: **Cook's distances** (расстояние Кука), **Leverage values** (точки балансировки); меры влияния: **Standardized Df Beta** и **Df Fit**; частные корреляции и т.п. Все эти понятия весьма полезны для выявления влияния отдельных значений отклика и предикторов на оценки регрессионной модели, но требуют отдельного обстоятельного разговора. Поэтому мы не будем здесь разбирать процедуры регрессионного анализа SPSS, кроме упомянутой выше процедуры **Curve Estimation**.

Пример 8.1к. Для данных урожайности зерновых культур в СССР подобрать модель простой линейной регрессии и построить на базе подобранной модели прогноз на несколько лет вперед.

Подготовка данных. Пусть данные табл. 1.2 находятся в текстовом (ASCII) файле **zerno.txt** в виде двух столбцов, первый из которых содержит значение года, а второй — значение урожайности. Для загрузки этих данных в пакет SPSS выберем в меню пакета пункт **FILE**, а в нем подпункт **Read ASCII Data...** На экран будет выведен запрос открытия файла, его вид — такой же, как в большинстве Windows-программ. Только в нижней части запроса имеется переключатель **File Format** (Формат файла), позволяющий выбирать между фиксированным (**Fixed**) и свободным (**Freefield**) форматами файла. Установим значение этого переключателя **Fixed**, выбрав фиксированный формат файла. Этот формат предполагает, что значения каждой переменной в строках файла записаны в тех же столбцах, что и в первой строке файла. Затем щелкнем мышью кнопку запроса **Define** (Определить) и перейдем к определению формата записи переменных в файле. На экране откроется диалоговое окно **Define Fixed Variables** (Определить переменные фиксированного формата) (см. рис. 8.3).

В этом окне необходимо задать имена переменных. В нашем случае мы создали переменные **date** и **zerno**, а также указали начальную и конечную колонки столбца, в котором лежит каждая переменная, и формат этой переменной. Завершив описание переменных, щелкнем мышью кнопку окна **OK**. Произойдет загрузка данных в SPSS, и они отразятся в электронной таблице пакета (рис. 8.4).

Комментарий. При загрузке ASCII-файлов могут возникнуть две проблемы. Первая — несовпадение разделителя целой и дробной части числа в исходном файле и в установке в Windows. (Обычно для этих целей используется либо точка, либо запятая.) При этом не происходит корректной загрузки данных в SPSS. Для исправления ситуации обратитесь в пункт **Стандарты** (Windows Inter-

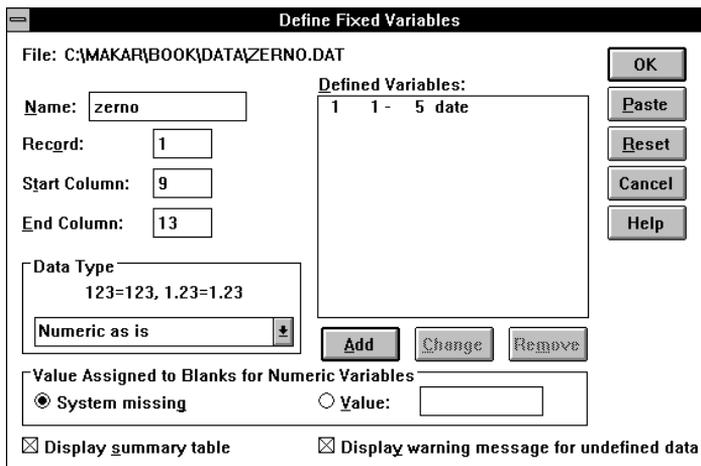


Рис. 8.3. Окно определения переменных фиксированного формата процедуры загрузки текстовых файлов в SPSS

1.zerno		5.6	
	date	zerno	var
1	1945	5.6	
2	1946	4.6	
3	1947	7.3	
4	1948	6.7	
5	1949	6.9	
6	1950	7.9	

Рис. 8.4. Таблица SPSS с данными об урожайности зерновых

national Settings) **Панели Управления (Control Panel)** и поменяйте тип десятичного разделителя в пункте **Формат чисел**. Вторая проблема — частичное (округленное) отображение чисел в электронной таблице. Для исправления этой ситуации обратитесь в пункт меню SPSS **Data Define Variable Type** и увеличьте в выведенном окне значение поля **Decimal Places**. Это значение задает число позиций, отведенных для десятичной части числа в электронной таблице.

Построение графика. Анализ данных и подбор возможной модели (функциональной связи) начнем с построения графика. Возможности SPSS по построению и оформлению графиков очень широки, их описание занимает в документации более 250 страниц. Мы не будем вдаваться в детали оформления графиков, а будем излагать лишь общий порядок действий и показывать полученные результаты.

Для построения графика исходных данных в пункте меню **Graphs (Диаграммы)** можно выбрать один из двух возможных типов процедур:

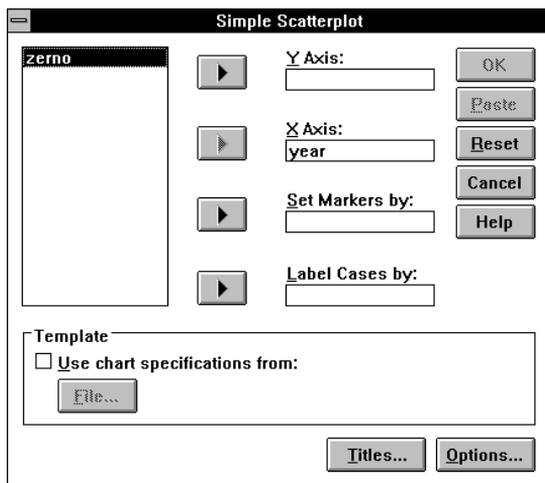


Рис. 8.5. Диалоговое окно процедуры Simple Scatterplot в SPSS

Simple Scatterplot (Простой график рассеивания) или Sequence (График последовательности). В этой задаче будет рассмотрена работа процедуры Simple Scatterplot. Ее диалоговое окно приведено на рис. 8.5.

Выделяя щелчком мыши требуемые переменные и нажимая соответствующие кнопки запроса , присвоим значения переменных *date* и *zerno* осям X и Y соответственно. На рис. 8.6 изображен полученный результат (после небольшого дополнительного оформления).



Рис. 8.6. График урожайности зерновых в SPSS

Выбор процедуры. На графике рис. 8.6 видно, что анализируемые данные со временем возрастают. Характер этого роста похож на линейный. Поэтому мы попытаемся решить задачу простой линейной регрессии, в которой предиктором будет служить год, а откликом —

урожайность зерновых культур. Для этого следует выбрать в меню пакета пункт **Analyze** (Анализ) и далее в открывшемся подменю — пункт **Regression** (Регрессия). Здесь в еще одном подменю можно выбрать один из двух методов: **Linear Regression** (линейная регрессия) или **Curve Estimation** (оценка кривой).

Процедура **Linear Regression** предоставляет широкие возможности при анализе адекватности классической модели простой и множественной линейной регрессии, включая выделение возможных «выбросов», проверку нормальности и некоррелированности остатков. А процедура **Curve Estimation** больше нацелена на выделение различных функциональных взаимосвязей. Поэтому мы продолжим свой анализ с помощью процедуры **Curve Estimation**, диалоговое окно которой приведено на рис. 8.7.

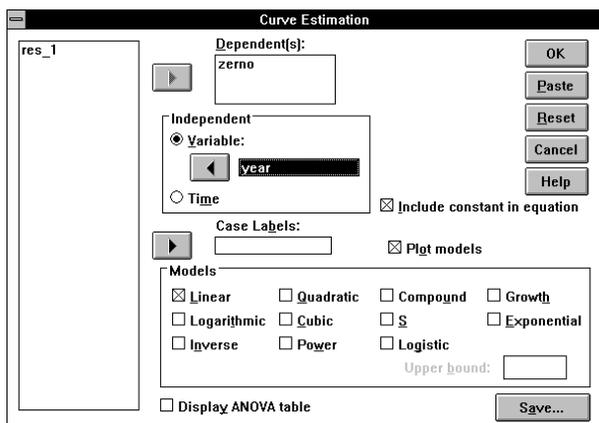


Рис. 8.7. Диалоговое окно процедуры **Curve Estimation** в SPSS

Задание параметров процедуры. В окне рис. 8.7 следует перенести переменную *zerno* в поле **Dependent(s)** (зависимая переменная). Для этого надо выделить ее щелчком мыши и нажать кнопку . Аналогичным образом в поле **Independent** (независимая переменная) надо поместить переменную *date*. В прямоугольнике **Models** (модели) выберем линейную модель **Linear**, а также укажем включение константы в эту модель, установив флажок **Include constant in equation**. Затем нажмем кнопку окна **OK**.

Замечание. Поле независимой переменной можно оставить незаполненным, поставив переключатель типа независимой переменной в положение **Time** — в этом случае зависимая переменная будет трактоваться как временной ряд.

Модели функциональной взаимосвязи. Дадим формулы взаимосвязей, приведенных в прямоугольнике **Models** на рис. 8.7. Пусть x — независимая переменная или время, b_i и u — константы (параметры моделей). Тогда формулы моделей можно записать так:

- Linear (линейная): $y = b_0 + b_1x$;
 Logarithmic (логарифмическая): $y = b_0 + b_1 \ln(x)$;
 Inverse (обратная): $y = b_0 + (b_1/x)$;
 Quadratic (квадратичная): $y = b_0 + b_1x + b_2x^2$;
 Cubic (кубическая): $y = b_0 + b_1x + b_2x^2 + b_3x^3$;
 Power (степенная): $y = b_0 \cdot (x^{b_1})$ или $\ln(y) = \ln(b_0) + b_1 \ln(x)$;
 Compound (показательная): $y = b_0(b_1)^x$ или $\ln(y) = \ln(b_0) \cdot [\ln(b_1)] \cdot x$;
 S (S-образная): $y = e^{(b_0+b_1/x)}$ или $\ln(y) = b_0 + b_1/x$;
 Logistic (логистическая): $y = 1/(1/u + b_0 \cdot (b_1^x))$ или $\ln(1/y - 1/u) = \ln(b_0) + [\ln(b_1)] \cdot x$;
 Growth (роста): $y = e^{(b_0+b_1x)}$ или $\ln(y) = b_0 + b_1x$;
 Exponential (экспоненциальная): $y = b_0 \cdot (e^{b_1x})$ или $\ln(y) = \ln(b_0) + b_1 \cdot x$.

Замечание. Кнопка диалогового окна (рис. 8.7) позволяет сохранить в виде отдельных переменных значения подобранной модели Predicted values, остатки Residuals и доверительные интервалы Prediction intervals, которые будут помещены в электронную таблицу пакета.

Результаты. После выполнения процедуры Curve Estimation в окне Output вывода результатов появится ряд вычисленных статистических характеристик, включая коэффициент корреляции R , коэффициент детерминации R^2 , таблицу анализа вариации, значения оценок коэффициентов модели и их статистические характеристики (рис. 8.8). Одновременно график ряда с подобранной кривой тренда будет помещен в окно Chart Carousel (рис. 8.9).

```

Dependent variable.. ZERNO                Method.. LINEAR

Listwise Deletion of Missing Data

Multiple R                .91521
R Square                 .83760
Adjusted R Square        .83383
Standard Error           1.60940

Analysis of Variance:

                DF    Sum of Squares    Mean Square
Regression      1      574.46135      574.46135
Residuals      43      111.37777      2.59018

F =      221.78428      Signif F = .0000

----- Variables in the Equation -----
Variable                B          SE B          Beta          T      Sig T
YEAR                   .275112     .018473     .915207     14.892  .0000
(Constant)            -528.949728  36.337744     -14.556  .0000

```

Рис. 8.8. Результаты работы процедуры Curve Estimation в окне выдачи результатов в SPSS

Результаты расчетов (см. рис. 8.8) показывают, что линейная модель тренда объясняет примерно 83% общей вариации данных, а получен-

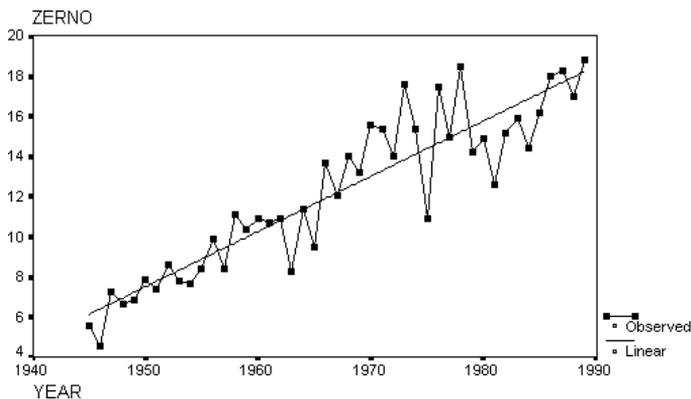


Рис. 8.9. Окно Chart Carousel с результатами работы процедуры Curve Estimation в SPSS

ные оценки коэффициентов модели значимо отличаются от нуля. В частности, значение коэффициента B при переменной $year$ (т.е. средний прирост урожайности за год) равен примерно 0.275 (ц/га).

Анализ остатков. Дальнейший анализ модели связан с исследованием остатков. Мы остановимся здесь лишь на проверке нормальности остатков, не касаясь подробно вопросов их возможной коррелированности. (Последнее требует дополнительных теоретических понятий, которые мы не рассматриваем в этой книге.)

Замечание. Учитывая небольшую длину исследуемого ряда, вряд ли можно ожидать здесь высокой точности и достоверности результатов. Однако подобный анализ позволит понять, как далеко мы могли отклониться от условий применения метода наименьших квадратов для удаления тренда и, тем самым, насколько можно верить полученным результатам.

Проверка нормальности распределения остатков. Для проверки соответствия распределения остатков нормальному распределению построим график остатков на нормальной вероятностной бумаге. Для этого надо в пункте меню **Graphs** выбрать процедуру **Normal P-P** (график на нормальной вероятностной бумаге). В диалоговом окне этой процедуры (рис. 8.10) необходимо указать в поле **Variables** обрабатываемую переменную.

Результаты работы процедуры приведены на рис. 8.11. Они показывают, что при хорошем согласии распределения данных с нормальным распределением «на хвостах» (зоны в районе нуля и единицы на графике), а также в центре есть определенные отклонения в промежуточных зонах. Подобные отклонения не сильно влияют на точность получен-

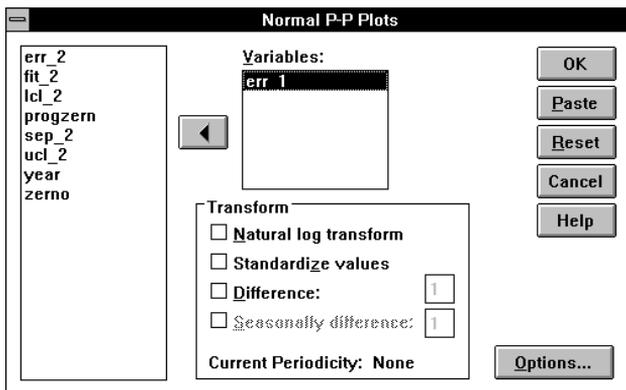


Рис. 8.10. Диалоговое окно процедуры Normal P-P в SPSS

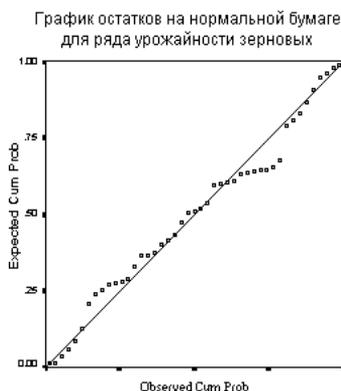


Рис. 8.11. Результаты процедуры Normal P-P в SPSS

ных результатов. Поэтому результаты, полученные с помощью метода наименьших квадратов, можно считать приемлемыми.

Замечания. 1. Процедура Normal P-P (см. рис. 8.11) предоставляет возможность проводить различные преобразования указанной переменной: переходить к логарифмической шкале, стандартизировать значения переменной, использовать простые и сезонные разностные операторы.

2. Малый объем наблюдений, естественно, не позволяет сделать обоснованных выводов о распределении остатков. Строя график остатков на нормальной вероятностной бумаге, мы прежде всего пытаемся выяснить, насколько сильно нарушается предположение о нормальности. Если эти нарушения невелики, то полученные выводы можно считать достаточно надежными. В противном случае возникает вопрос о целесообразности применения выбранного метода обработки и замене его на методы, не чувствительные к распределению данных и устойчивые к различным отклонениям.

Более подробное обсуждение проверки нормальности распределения с помощью критерия Колмогорова—Смирнова проводится в гл. 10.

Дополнительная литература

1. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика. Основы эконометрики. Т. 1. Теория вероятностей и прикладная статистика. — М.: Юнити-Дана, 2001. — 656 с.
2. *Болдин М.В., Симонова Г.И., Тюрин Ю.Н.* Знаковый статистический анализ линейных моделей. — М.: Наука. Физматлит, 1997. — 288 с.
3. *Вучков И., Бояджијева Л., Солаков Е.* Прикладной линейный регрессионный анализ. — М.: Финансы и статистика, 1987. — 239 с.
4. *Демиденко Е.З.* Линейная и нелинейная регрессия. — М.: Финансы и статистика, 1981. — 302 с.
5. *Дугерти К.* Введение в эконометрику: пер. с англ. — М.: ИНФРА-М, 1999. — XIV, 402 с.
6. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ: пер. с англ. 3-е изд. — М.: Издательский дом Вильямс, 2007. — 912 с.
7. *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика. Начальный курс: учебник. 4-е изд. — М.: Дело, 2000. — 400 с.

Независимость признаков

Во многих практических задачах мы исследуем объекты, обладающие несколькими (двумя или более) признаками, и хотим выяснить, насколько эти признаки связаны между собой. Например, у каждого человека есть возраст и место рождения, уровень образования и годовой доход, пол и социальная принадлежность и т.п. Вопрос состоит в том, можно ли по степени выраженности одного признака судить о выраженности другого, либо же эти признаки следует считать проявляющимися независимо (в вероятностном смысле). Ответы на такие вопросы могут иметь значительную практическую ценность. Например, если мы установим, что признаки «профессия» и «политические убеждения» зависимы, то окажется, что социологические опросы по предсказанию результатов парламентских выборов следует проводить с учетом профессиональной принадлежности опрашиваемых — это позволит уменьшить размер представительной (репрезентативной) выборки.

9.1. О шкалах измерений

Измерения. Прежде чем говорить о зависимости или независимости признаков, надо эти признаки измерить. Это может быть нетривиальной задачей: действительно, как измерить «профессию», «политические убеждения» или «степень доверия»? Поэтому сначала мы обсудим вопрос о *шкалах измерений*, в которых измеряются различные признаки.

«Измерить все, что измеримо, и сделать измеримым все, что таковым еще не является» — такую программу точному естествознанию наметил Г. Галилей еще в XVII веке. Галилей ясно понимал, что измерения составляют основу наших знаний о природе. Но чем дальше, тем большее место измерения занимают и в науках о человеке и обществе, поставляя твердую основу для дальнейших исследований. Разумеется, в гуманитарных науках измерения более сложны, чем в естественных. Дело не только в том, что трудным может быть процесс измерения. Сложности касаются в основном истолкования результатов измерений. Например, в психологии многое приходится измерять с помощью психологических тестов, а по своему содержанию тестовый балл очевидно отличается от результатов измерения с помощью секундомера или линейки. Впрочем,

и между двумя последними тоже есть серьезная формальная разница, о чем будет сказано позже. Осознание подобных различий привело к понятию *шкалы измерений*.

Непрерывные и дискретные шкалы. Начнем с того, что имеется общего у всех видов измерений — их результатом всегда является число, будь то школьная оценка, тестовый балл, календарная дата, температура тела, расстояние на местности и т.д. Что же касается их различий, то первым бросается в глаза различие в «запасе» возможных значений при разных измерениях. Так, школьные оценки (у нас) могут принимать только 4 значения (2, 3, 4 и 5). Тестовым баллом может быть любое целое число (из того промежутка, который определяется количеством вопросов и тем, как оцениваются ответы). Показателем температуры может быть любое действительное число (если отвлечься от пределов, которые задают физические соображения) и т.д. Итак, шкалы измерений могут иметь различные множества значений. С этой позиции различают шкалы конечные и бесконечные, дискретные и непрерывные.

Запас допустимых операций в шкале. Но главные различия шкал не в этом. Важнее то, что по отношению к результатам измерений в разных шкалах осмысленными являются разные арифметические действия. Рассмотрим, например, измерение времени. Каждому моменту времени соответствует календарная дата, скажем, число t . (В разных календарных системах данному моменту времени могут соответствовать разные числа, но сейчас это не имеет значения, поскольку далее мы будем говорить о каком-нибудь одном календаре, хотя бы о привычном григорианском.) Пусть t и s — даты двух событий, два числа. Нам понятно, что означает их разность $(t - s)$ — это временной интервал между событиями. Следовательно, операция вычитания допустима в шкале измерения времени, потому что приводит к осмысленному результату. Можно также сравнить числа t и s по величине (по принципу больше-меньше) — таким путем мы узнаем, какое из событий произошло раньше, какое позже. Следовательно, в этой шкале операция сравнения чисел является допустимой. Но в комбинациях типа $t + s$, $2t$, ts и т.д. мы никакого смысла не находим. Поэтому эти операции в данной шкале допустимыми не считаются.

Сказанное об измерении времени полностью приложимо, например, к измерению температуры. Но в случае измерения длины (и других размеров) положение оказывается иным. Пусть x и y — длины двух предметов, скажем, труб или рельсов. Нам понятно, что означает не только $x - y$, но и $2x$, $x + y$ и многое другое. Например, $x + y$ есть длина

трубы, которую можно получить, соединив трубы длины x и длины y , и т.д. В этой шкале запас допустимых операций особенно богат.

Порядковые шкалы. Для изучения психических и физических характеристик человека, например его способностей к умственной или физической деятельности, нередко прибегают к специально организованным пробам или испытаниям, называемыми *тестами*. Результатом такого теста является число, называемое *тестовым баллом*. При замене выбранного теста другим, предназначенным для измерения той же характеристики, тестовый балл данного испытуемого, скорее всего, изменится. Но что-то при таком изменении должно сохраниться, ведь объект измерения тот же, что и прежде. В частности, должно сохраниться соотношение между тестовыми баллами, которые получают в этих условиях два испытуемых. Если два теста измеряют одну и ту же характеристику (мы признаем, что это ситуация скорее воображаемая, чем реальная), тот из испытуемых, кто обладает этой характеристикой в большей мере, получит и большие тестовые баллы. Для тестовых баллов, как и для школьных оценок, осмысленными (допустимыми) оказываются только их сравнения. Операции вроде сложения и вычитания для этих шкал не имеют смысла. Например, нельзя сказать, что школьник, получивший четверку, знает предмет на единицу лучше, чем тот, кто получил тройку, ибо для знаний нет единицы измерения. Мы можем лишь сказать, что первый ученик знает предмет лучше, чем второй.

Описанные шкалы, в которых существен лишь взаимный порядок, в котором следуют результаты измерений, а не их количественные значения, часто называют *порядковыми*, или *ординальными* шкалами.

Номинальные шкалы. Еще одним важным видом шкал являются *номинальные* шкалы. В них числа служат только для различения отдельных возможностей, заменяя названия и имена. Никаких содержательных соотношений, кроме $x = y$ или $x \neq y$, между значениями в этих шкалах нет. Конечно, выбор чисел вместо названий или других способов идентификации не обязателен. Но бывает, что к нему приходится прибегать поневоле. Например, в полиграфии и текстильном деле используют сотни цветов и оттенков. Они должны быть стандартизованы и иметь отличительные обозначения. Существуют альбомы, содержащие такие цветовые образцы. Указывать и называть какой-либо цвет можно только с помощью его номера в таком альбоме, поскольку существующие в языке названия цветов слишком малочисленны и неопределенны.

Виды шкал. Мы уже ввели два вида шкал: порядковые и номинальные. Кроме того, мы будем рассматривать еще и *количественные* шкалы, такие как описанные выше шкалы времени, температуры, длины

и т.д. С помощью принципа, положенного в основу классификации шкал (т.е. объема допустимых операций над числами), мы могли бы проводить тонкие различия между шкалами. Однако с позиции статистики это пока неоправданно, так как статистические методы еще не имеют столь тонкой приспособленности. Они разработаны для больших групп шкал: количественных, порядковых и номинальных, которые мы и будем рассматривать далее.

Замечание. Классификацию шкал измерений можно обсудить и с другой точки зрения (разумеется, родственной первой) — в зависимости от числа и характера тех соглашений, которые приходится делать при создании каждой шкалы. Для календаря, например, надо выбрать начальный момент, от которого будет отсчитываться время (вперед, в будущее, и назад, в прошлое). Реальное содержание измерения от этого не должно зависеть. В частности, разность двух дат не меняется при перемене начала отсчета (в отличие от их суммы, например). Именно поэтому вычитание в этой шкале является допустимой операцией. Подробнее мы развивать данную тему не будем и ограничимся этими беглыми замечаниями.

В дальнейшем мы рассмотрим, как решаются вопросы о статистической независимости признаков в трех шкалах: номинальной, порядковой и количественной.

9.2. Инструменты и стратегия исследования связи признаков

Классификация типа данных. Методы определения связи признаков заметно отличаются в зависимости от вида шкалы измерений этих признаков:

- для изучения связи признаков, измеренных в номинальной шкале, например, признаков вида «да или нет», применяются таблицы сопряженности, статистика Фишера—Пирсона X^2 , различные меры связи признаков (коэффициенты Юла, Крамера, Чупрова и др.) и логарифмически линейные модели (см. п. 9.3);
- для признаков, измеренных в порядковой шкале — данных типа «лучше – хуже», тестовых баллов и т.д., — применяются ранжирование и коэффициенты корреляции Спирмена и Кендэла (см. п. 9.4);
- для данных, измеренных в количественных шкалах, применяются коэффициент корреляции Пирсона и модель простой линейной регрессии.

Таким образом, первым шагом анализа является классификация типа данных, т.е. отнесение их к той или иной шкале измерений —

номинальной, порядковой или количественной (см. п. 9.1). Однако и на этом первом шаге на практике часто делаются ошибки. Типичной из них является вычисление и сравнение средних значений тестовых баллов, например школьных оценок. Эти данные относятся к порядковой шкале, в которой операция усреднения не имеет ясного смысла.

Проверка гипотезы об отсутствии связи признаков. Следующим шагом исследования является проверка гипотезы об отсутствии связи (независимости) между признаками. Методы подобной проверки довольно хорошо проработаны как с теоретической, так и с практической точки зрения. Гипотеза об отсутствии связи отвергается в случае, когда статистика Фишера—Пирсона X^2 принимает неоправданно большие значения или соответствующие коэффициенты корреляции заметно отклоняются от нуля. Эти вопросы подробно разбираются в п. 9.3—9.5.

Замечание. Следует помнить, что коэффициенты корреляции не всегда позволяют отличить зависимость от независимости. В первую очередь это относится к сложным типам зависимости.

Оценка силы связи. Если гипотеза о независимости признаков отвергается, то обычно имеет смысл выяснить степень силы связи признаков. Для этого используются различные *меры связи* — обычный коэффициент корреляции для признаков, измеренных в количественных шкалах, ранговые коэффициенты корреляции Кендэла и Спирмена для признаков, измеренных в порядковых шкалах, и различные показатели типа ϕ -коэффициента, коэффициента λ Гудмена—Краскела и др. Если модуль меры связи лежит в интервале от 0.8 до единицы, то это свидетельствует о сильной связи признаков, если он находится в интервале [0.3, 0.7] — о неярко выраженной связи, а меры связи, близкие к нулю, означают отсутствие зависимости или очень слабую зависимость признаков.

9.3. Связь номинальных признаков (таблицы сопряженности)

Наиболее типичной ситуацией, в которой встречаются номинальные признаки, является обработка социологических анкет. В ходе социологического обследования появляются тысячи анкет, содержащие различные комбинации таких признаков, как профессия, образование, пол, предпочтительный вид отдыха, использование свободного времени и т.п. Эти комбинации появляются с разной частотой. Возникает необходимость осмыслить этот хаос, связать один признак с другим.

Иногда такие признаки связаны жестко: если профессия — шахтер или сталевар, то пол, несомненно, мужской. Тем самым по некоторым значениям признака «профессия» можно узнать значение признака «пол». Другая крайность — отсутствие связи, т.е. зависимости одного признака от другого. (Если глаза серые, то каков пол?)

Исследователя в подобных задачах обычно интересует, насколько точно можно предсказать значение одного признака по значению другого. Если точное предсказание невозможно, надо указать распределение вероятностей между возможными значениями второго признака при данном значении первого. Этой проблеме должна предшествовать более простая: надо сначала проверить, существует ли вообще какая-либо связь между этими признаками, или же они ведут себя независимо друг от друга? Статистический способ ответа на этот вопрос основан на изучении выборки (см. п. 1.8), т.е. конечной совокупности объектов, наудачу извлеченных из генеральной совокупности.

Пример. Рассмотрим пример, подробно описанный в [91], в котором каждый испытуемый мог выбрать инструкцию, регламентирующую его дальнейшую работу. Предварительно у каждого испытуемого был определен тип нервной системы. Результаты этого опыта приведены в следующей ниже таблице, которая заодно дает пример таблицы сопряженности признаков.

Таблица 9.1

Предпочтение различных видов инструкций в группах высокорезактивных (+P) и низкорезактивных (-P) индивидов (по Чижковской, 1974)

Вид инструкции	Группы испытуемых		В сумме
	+P	-P	
Детальная, подробно регламентирующая последовательные действия	63	42	105
Итоговая, обобщенная, краткая	34	56	90
В сумме	97	98	195

Здесь каждый признак (свойства нервной системы, свойства инструкции) имеет два уровня, вместе они образуют таблицу размера 2×2 (как говорят, два на два). В каждой из ее четырех клеток показано, сколько раз встречалась данная комбинация признаков. На полях таблицы указаны суммарные значения (т.е. сколько раз встретился тот или иной уровень признака). Общее количество испытуемых (в данном случае 195) помещено в правом нижнем углу таблицы. Оно получается как сумма чисел, стоящих на полях. Аналогично устроены и более сложные таблицы сопряженности, с большим числом факторов и уровней.

Для данного примера естественен вопрос: есть ли связь между свойствами нервной системы и предпочтением того или иного вида инструкций? Если бы связь существовала и была совершенно твердой, в таблице 2×2 ненулевые клетки располагались бы только на диагонали (одной или другой). При связи не столь сильной некоторое число наблюдений попадает и во внедиагональ-

ные клетки. Чем слабее связь, тем менее четко проявляется эта тенденция. Присутствует ли эта тенденция в приведенной таблице?

Статистическая независимость признаков. Начнем с того, что в противовес представлению о взаимосвязи признаков введем гипотезу, отрицающую эту связь. Это гипотеза о независимости признаков (в дальнейшем — «нулевая» гипотеза H_0). Уточним задачу, ограничиваясь (для простоты) двумя признаками. Пусть признак A имеет r градаций (или уровней), которые мы назовем A_1, A_2, \dots, A_r , признак B подразделяется на s градаций B_1, B_2, \dots, B_s . В предыдущем примере каждый из двух признаков (вид инструкции, тип нервной системы) имел по два уровня.

Определение. Признаки A и B называют независимыми, если (при случайном выборе объекта) оказываются независимыми события «признак A принимает значение A_i » и «признак B принимает значение B_j », притом для всех пар i, j .

Если сказать короче, то признаки A и B называются независимыми, если (при случайном выборе объекта):

$$P(A_i B_j) = P(A_i) P(B_j) \quad (9.1)$$

для всех A_i и B_j . Иначе говоря, независимость признаков означает, что значение, принятое признаком A , не влияет на вероятности возможных значений признака B , т.е.:

$$P(B_j/A_i) = P(B_j) \quad (9.2)$$

для всех пар A_i, B_j .

Непосредственно проверить соотношения между вероятностями (9.1) или (9.2) мы не можем, поскольку этих вероятностей не знаем.

Таблица сопряженности. Предположим, однако, что в нашем распоряжении имеется выборка из интересующей нас генеральной совокупности. По этой выборке мы можем определить частоты событий A_i и B_j по отдельности и в любых комбинациях.

Обозначим через n_{ij} частоту события $A_i B_j$, т.е. количество объектов выборки, обладающих комбинацией уровней A_i и B_j признаков A и B . Ясно, что число появлений признака A_i (частота события A_i) равно:

$$\sum_{j=1}^s n_{ij} = n_{i1} + n_{i2} + \dots + n_{is}. \quad (9.3)$$

Обозначим эту сумму через $n_{i.}$. Аналогично, частота появления B_j равна

$$n_{.j} = n_{1j} + n_{2j} + \dots + n_{rj}. \quad (9.4)$$

Сделаем общее соглашение: пусть замена индекса точкой означает результат суммирования по этому индексу. Тогда:

$$n_{..} = \sum_{i=1}^r n_{i.} = \sum_{j=1}^s n_{.j} = \sum_{i=1}^r \sum_{j=1}^s n_{ij}$$

обозначает общее число наблюдений, т.е. объем выборки. Часто вместо $n_{..}$ мы будем писать просто n .

Выборочные частоты обычно представляют в виде таблицы, приведенной ниже.

Определение. Таблицу 9.2 называют таблицей сопряженности признаков A и B .

Таблица 9.2

Таблица сопряженности признаков A и B

$A \setminus B$	B_1	B_2	B_j	B_s	
A_1	n_{11}	n_{12}	n_{1j}	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	n_{2j}	n_{2s}	$n_{2.}$
A_i	n_{i1}	n_{i2}	n_{ij}	n_{is}	$n_{i.}$
A_r	n_{r1}	n_{r2}	n_{rj}	n_{rs}	$n_{r.}$
	$n_{.1}$	$n_{.2}$	$n_{.j}$	$n_{.s}$	$n_{..}$

Введем аналогичные обозначения и для вероятностей. Положим,

$$p_{ij} = P(A_i B_j). \quad (9.5)$$

Теперь

$$P(A_i) = \sum_{j=1}^s p_{ij} = p_{i.}, \quad P(B_j) = \sum_{i=1}^r p_{ij} = p_{.j}. \quad (9.6)$$

Гипотеза о независимости признаков в принятых обозначениях записывается так:

$$p_{ij} = p_{i.} p_{.j} \quad (9.7)$$

для всех пар (i, j) , $i = 1, \dots, r$, $j = 1, \dots, s$.

Ожидаемые частоты. Мы хотим знать, выполняются ли соотношения (9.1) или (9.7) для наших признаков. Судить об этом можно, основываясь на выборочных частотах, представленных в таблице сопряженности. При большом объеме выборки эти частоты близки к вероятностям. Поэтому для частот из табл. 9.2 соотношения (9.1) и (9.7) превращаются в приближенные равенства (если, конечно, гипотеза о независимости верна). Остается найти способ, чтобы судить о том, выполняются эти приближенные равенства или нет.

Итак, по теореме Бернулли, при $n \rightarrow \infty$:

$$\frac{n_{ij}}{n} \rightarrow p_{ij}; \quad \frac{n_{i.}}{n} \rightarrow p_{i.}; \quad \frac{n_{.j}}{n} \rightarrow p_{.j}, \quad (9.8)$$

а поэтому для независимых признаков: $n_{ij} \simeq n_{i.}n_{.j}/n$.

Определение. Величины $n_{i.}n_{.j}/n$ называются ожидаемыми частотами (имеется в виду, ожидаемыми при выполнении гипотезы).

При выполнении гипотезы ожидаемые частоты не должны сильно отличаться от наблюдаемых частот n_{ij} . Наша задача сейчас состоит в том, чтобы решить, выполняются ли в действительности (для наблюдаемой таблицы) эти приближенные соотношения.

Ожидаемые частоты полезно ввести в исходную таблицу, чтобы иметь возможность сравнить их с наблюдаемыми. Скажем, приведенная выше табл. 9.1 принимает вид:

Таблица 9.3

Предпочтение различных видов инструкций в группах высокорективных (+P) и низкорективных (-P) индивидов (с ожидаемыми частотами)

Вид инструкции	Тип испытуемого		
	+P	-P	
Детальная	63 / 52.2	42 / 52.7	105
Краткая	34 / 44.8	56 / 45.2	90
Всего	97	98	195

Если видимые различия между наблюдаемыми частотами и частотами, рассчитанными на основании гипотезы о независимости признаков, можно объяснить случайными колебаниями (т.е. действием случайной изменчивости), то отвергать гипотезу независимости нет оснований. (В просторечии даже говорят, что гипотеза H_0 принимается.) Итак, осталось условиться, как сопоставлять два ряда частот, как измерить различие между ними.

Сопоставление ожидаемых и наблюдаемых частот. Вопрос о сравнении наблюденных в опыте частот с теми, которые предписывает теория (ради проверки этой теории), возникает не только при анализе таблиц сопряженности, но и во многих других задачах. Со времени К. Пирсона (начало века) и Р. Фишера (двадцатые годы) стал общепринятым следующий способ сопоставления наблюдаемых частот с частотами, рассчитанными по модели (их также иногда называют теоретическими).

Чтобы сформулировать критерий Пирсона—Фишера в общем и легко запоминающемся виде, обозначим наблюдаемые частоты через H ; ожидаемые, или теоретические, частоты обозначим буквой T . Если модель правильно описывает действительность, числа H и T должны

быть близки друг к другу. Следовательно, сумма квадратов отклонений $(H - T)^2$ не должна быть большой. Разумно в общую сумму отдельные слагаемые вносить с различными весами, поскольку чем больше T , тем больше H может от него отклоняться за счет действия случая, без отступления от модели. Поэтому в качестве меры близости наблюдаемых и ожидаемых частот разумно рассмотреть величину:

$$X^2 = \sum \frac{(H - T)^2}{T}, \quad (9.9)$$

где сумма берется по всем ячейкам таблицы сопряженности. В данном случае X^2 есть мера согласия опытных данных с теоретической моделью.

Если в конкретном опыте величина X^2 оказывается чрезмерно большой, приходится признать, что ожидаемые частоты слишком сильно отличаются от наблюдаемых. Тем самым гипотеза, на основании которой были рассчитаны ожидаемые частоты, оказывается в противоречии с опытом. Поэтому ее следует признать неправильной и отвергнуть.

Остается лишь разобраться в том, какие значения для X^2 надо считать чрезмерно большими (неправдоподобно большими), а какие нет. Для этого надо знать распределение случайной величины X^2 как в случае, когда гипотеза верна, так и в случае ее нарушения. Ответ в первом случае дает приводимая ниже теорема. После ее обсуждения мы рассмотрим и второй вопрос.

Теорема (К. Пирсон, Р. Фишер). *Если верна модель, по которой рассчитаны теоретические частоты T , то при неограниченном росте числа наблюдений распределение случайной величины X^2 стремится к распределению хи-квадрат. Число степеней свободы этого распределения определяется как разность между числом событий и числом связей, налагаемых моделью.*

Число степеней свободы распределения хи-квадрат. В нашем примере *число событий* — это число ячеек в таблице сопряженности, т.е. число событий вида $A_i B_j$. Оно равно rs . Подсчитаем *число связей*. Во-первых, $\sum_{i,j} n_{ij} = n$ (одна связь). Во-вторых, определяя n_i (и n_j), мы воспользовались соотношениями

$$\sum_{j=1}^s n_{ij} = n_i \quad \text{и} \quad \sum_{i=1}^r n_{ij} = n_j.$$

Число таких независимых соотношений равно $r - 1$ для первой группы соотношений и $s - 1$ для второй. Действительно, хотя число соотношений в первой группе равно r , любое одно из них (благодаря существованию соотношения $\sum_{i,j} n_{ij} = n$) является следствием остальных.

Итак, число степеней свободы распределения хи-квадрат при проверке независимости равно:

$$rs - (r - 1) - (s - 1) - 1 = (r - 1)(s - 1).$$

Другая форма статистики X^2 . Для статистики X^2 существует другая форма, порой более удобная для расчетов:

$$Y^2 = 2 \sum H \ln \frac{H}{T}. \quad (9.10)$$

Сумма снова берется по всем ячейкам таблицы сопряженности. При гипотезе статистика Y^2 распределена в пределе так же, как и X^2 , т.е. по закону хи-квадрат. Правило для подсчета числа степеней свободы X^2 действует и для Y^2 . Вообще величины X^2 и Y^2 при расчетах мало отличаются друг от друга, если гипотеза верна, т.е. если наблюдаемые частоты близки к ожидаемым.

Пределы использования аппроксимации распределения для статистик X^2 и Y^2 . Как было сказано, распределение хи-квадрат является предельным для случайных величин X^2 и Y^2 . Поэтому использовать его как приближение для реальных распределений X^2 , Y^2 можно только при большом числе наблюдений n . Считается достаточным, чтобы по всем ячейкам теоретические частоты были бы не меньше 5. Есть данные, что это ограничение в задаче независимости признаков можно снизить до 3, так что должно выполняться соотношение: $n_{i \cdot n \cdot j} / n \geq 3$. Требования к ожидаемым частотам определенно смягчаются при увеличении числа степеней свободы.

Независимые признаки. Посмотрим, как выглядят общие результаты Пирсона–Фишера применительно к задаче о независимости признаков. Составим статистики:

$$X^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - \frac{n_{i \cdot n \cdot j}}{n})^2}{\frac{n_{i \cdot n \cdot j}}{n}}, \quad (9.11)$$

$$Y^2 = 2 \sum_{i=1}^r \sum_{j=1}^s n_{ij} \ln \left(\frac{n_{ij}}{\frac{n_{i \cdot n \cdot j}}{n}} \right). \quad (9.12)$$

После упрощений они выглядят так:

$$X^2 = n \left[\sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}^2}{n_{i \cdot n \cdot j}} - 1 \right],$$

$$Y^2 = 2 \left[\sum_{i,j} n_{ij} \ln n_{ij} - \sum_i n_{i\cdot} \ln n_{i\cdot} - \sum_j n_{\cdot j} \ln n_{\cdot j} + n \ln n \right].$$

Теорема Пирсона–Фишера утверждает, что если признаки A и B (имеющие r, s уровней соответственно) независимы, то статистики X^2 , Y^2 имеют (приблизительно, при большом числе n) распределение хи-квадрат с $(r-1)(s-1)$ степенями свободы.

Зависимые признаки. Чтобы понять, как ведут себя статистики X^2 (или Y^2) при больших n , когда гипотеза независимости неверна, надо преобразовать выражение (9.11) и затем воспользоваться свойствами (9.8). Получим, что:

$$\frac{X^2}{n} = \sum_{i=1}^r \sum_{j=1}^s \frac{\left(\frac{n_{ij}}{n} - \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} \right)^2}{\frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}} \approx \sum_{i=1}^r \sum_{j=1}^s \frac{(p_{ij} - p_{i\cdot} p_{\cdot j})^2}{p_{i\cdot} p_{\cdot j}}. \quad (9.13)$$

Если гипотеза H_0 неверна (и только тогда), правая часть (9.13) отлична от нуля. В этом случае X^2 стремится к бесконечности (при $n \rightarrow \infty$). Следовательно, при большом конечном n для зависимых признаков мы будем получать в опытах большое значение величины X^2 . Аналогичное рассуждение верно и для Y^2 . Таким образом, при больших n :

- для независимых признаков статистика X^2 распределена (практически) по закону хи-квадрат;
- для зависимых признаков X^2 неограниченно возрастает при увеличении n .

Поэтому большие (неправдоподобно большие для хи-квадрат) значения X^2 указывают на взаимную зависимость признаков.

Правило проверки гипотезы о независимости. Какие же значения X^2 (или Y^2) надо считать настолько большими, что они несовместимы с гипотезой H_0 ? Очевидно, те, появление которых при гипотезе маловероятно, т.е. те, которые превосходят критические значения распределения хи-квадрат, соответствующие выбранному уровню значимости. Итак, для проверки гипотезы о независимости признаков надо вычислить одну из статистик X^2 или Y^2 и сравнить ее значение с соответствующими критическими значениями распределения хи-квадрат, взятыми из таблиц.

Продолжение примера. В примере, приведенном выше, расчет дает $X^2 = 9.58$. Число степеней свободы для таблицы 2×2 равно 1. Верхние процентные точки распределения хи-квадрат (χ^2) с одной степенью свободы таковы:

Процент	10%	5%	2.5%	1%	0.5%	0.1%
Пр.точка	2.71	3.84	5.02	6.63	7.88	10.83

Мы видим, что $P\{\chi^2 \geq X^2\} < 0.005$. Это значит, что вероятность получить чисто случайно для независимых признаков такое же, как в опыте, или даже большее значение не превышает 0.005. Можно считать поэтому, что в нашем примере признаки не являются независимыми, т.е. связь между ними проявляется. (Иногда говорят, что данная таблица *значима*.)

Таблицы 2×2 . В частном случае таблиц сопряженности, когда признаки A и B принимают только по 2 значения A_1, A_2 и B_1, B_2 (обычно первое из них — наличие признака, а второе — его отсутствие), статистика X^2 упрощается:

$$X^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

В этой ситуации статистики X^2, Y^2 имеют распределение χ^2 с одной степенью свободы (если признаки независимы).

Видимо, лучшее согласие с предельным распределением имеет модифицированная статистика:

$$X^{*2} = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - \frac{n}{2})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

(Это X^2 с поправкой на непрерывность; иногда говорят — с поправкой на группировку.)

Меры связи признаков. Как всегда в статистике, принятие какой-либо гипотезы не означает ее доказательства. Оно означает лишь, что имеющиеся данные и принятые методики проверки не позволяют отвергнуть гипотезу. Вполне возможно, и так часто и бывает, что при увеличении числа наблюдений гипотезу (в данном случае независимости) придется отклонить. Для статистики X^2 (по закону больших чисел) это будет означать, что

$$\lim_{n \rightarrow \infty} \frac{1}{n} X^2 = \sum_{i,j} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

настолько отличается от нуля, что этого не может скрыть свойственная X^2 случайная изменчивость. Участвующая в этом выражении сумма квадратов, естественно, должна рассматриваться как одна из характеристик различия между таблицами $\|p_{ij}\|$ и $\|p_{i.}p_{.j}\|$.

В реальных задачах исследователя интересует взаимодействие признаков. Если признаки оказались взаимосвязаны (гипотеза об их неза-

висимости проверена и отвергнута), исследователя интересует сила их связи. Для описания такой связи было предложено много различных коэффициентов, называемых *мерами связи*. К сожалению, ни один из них не может передать всей сложной картины взаимодействия, особенно для таблиц с большим числом признаков и уровней признаков. В связи с этим и, главное, с появлением более точных методов анализа таблиц сопряженности (например, логарифмически линейных моделей) интерес к этим мерам связи заметно снизился.

Мы немного расскажем об этих мерах на примере таблиц 2×2 , для которых они полезнее, чем для более сложных. Самый старый из них — коэффициент связи Юла (1900, 1912):

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}.$$

С ростом n ($n \rightarrow \infty$) $Q \rightarrow (p_{11}p_{22} - p_{12}p_{21}) / (p_{11}p_{22} + p_{12}p_{21})$.

Используется также мера связи $\varphi = \sqrt{\frac{1}{n}X^2}$, вероятностный смысл которой был отмечен ранее.

Кроме этих, были предложены коэффициенты Крамера, Чупрова, λ -меры и τ -меры Гудмена и Краскела и др. Подробную информацию по изложенным выше вопросам можно найти в [7], [53], [83], [102], [106].

9.4. Связь признаков, измеренных в шкале порядков

Ранги. Обсуждая измерения в порядковых (ординальных) шкалах, мы убедились, что реальным содержанием этих измерений является тот порядок, в котором выстраиваются объекты (по степени выраженности измеряемого признака). Предположим, к примеру, что для изучения двигательных возможностей группы детей мы предложили каждому ребенку сложить что-то определенное из кубиков и палочек. Ясно, что время, затраченное на выполнение задания, тем больше, чем менее развиты способности к тонким движениям рук и пальцев. Поэтому упорядочение испытуемых по затраченному времени совпадает с их упорядочением по развитию этих способностей. При другом подобном задании затраченное время будет другим, но порядок сохранится (за вычетом влияния на результат случайных обстоятельств).

Сказанное означает, что для нас имеют значение не столько результаты (числа) X_1, \dots, X_n измерения определенного признака A для объектов $O(1), \dots, O(n)$, сколько ранги r_1, \dots, r_n чисел X_1, \dots, X_n . (Здесь r_i — ранг X_i среди чисел X_1, \dots, X_n .)

Независимость признаков. Представим себе, что теперь мы имеем дело с двумя разными признаками A и B , измерения которых проведены в порядковой шкале. Нас интересует, как влияет величина одного признака на степень выраженности другого. Если такого влияния нет, признаки естественно назвать независимыми. Как проверить гипотезу о независимости порядковых признаков (гипотезу H_0)? Первым решение этой задачи предложил психолог Ч. Спирмен в 1900 г.

Пусть, как уже говорилось выше, X_1, \dots, X_n суть значения признака A для объектов $O(1), \dots, O(n)$, а Y_1, \dots, Y_n — значения признака B для тех же объектов. Каждый объект $O(i)$, $i = 1, \dots, n$ теперь характеризуется парой чисел (X_i, Y_i) — своими значениями признаков A и B . От чисел Y_1, \dots, Y_n (так же, как ранее для признака A) переходим к их рангам s_1, \dots, s_n . (Здесь s_i — ранг Y_i среди Y_1, \dots, Y_n .) Будем считать, что среди чисел X_1, \dots, X_n (и среди чисел Y_1, \dots, Y_n) нет повторяющихся, так что переход к рангам вопросов не вызывает. Для измерений в непрерывных шкалах эта ситуация типична.

Замечание. Ранговые последовательности могут возникать и иначе, непосредственно. Ч. Спирмен, например, обсуждал связь между способностями к музыке и математике. Группу детей мы можем упорядочить дважды — сначала по успехам в музыке, затем — в математике. (В школьном классе мы можем попросить учителей составить два таких списка.) Места, которые займет ученик N в обоих списках, и будут его рангами r, s .

Распределение набора рангов для независимых признаков. Теперь каждому объекту $O(i)$ приписана пара натуральных чисел (r_i, s_i) . Если признаки A и B взаимосвязаны, то порядок, в котором следуют числа x_1, \dots, x_n , в определенной степени влияет на порядок, в котором следуют числа y_1, \dots, y_n . Иными словами, последовательность рангов r_1, \dots, r_n в какой-то мере влияет на ранговую последовательность s_1, \dots, s_n . Чем более тесно связаны эти признаки, тем в большей степени последовательность r_1, \dots, r_n предопределяет последовательность s_1, \dots, s_n .

Если же признаки такой связи не проявляют, то порядок среди игроков случаен по отношению к порядку среди иксов. В этом случае все $n!$ перестановок чисел $1, 2, \dots, n$, которые могут выступать как ранги s_1, \dots, s_n , оказываются равновероятными, т.е. равновероятными при любом порядке чисел r_1, \dots, r_n . Это центральный момент обсуждения: при гипотезе H_0 и любом наборе r_1, \dots, r_n все возможные последовательности s_1, \dots, s_n равновозможны (т.е. вероятность распределена между ними равномерно).

Вторым важным моментом является выбор меры сходства для двух наборов рангов. Здесь много математических возможностей. Наиболее

популярны две меры сходства, которые приводят к коэффициентам ранговой корреляции Спирмена и Кендэла соответственно. С этими ранговыми коэффициентами мы уже встречались в п. 8.4. Начнем с той меры, которую предложил Ч. Спирмен.

Коэффициент Спирмена. Близость двух рядов чисел r_1, \dots, r_n и s_1, \dots, s_n отражает величина

$$S = \sum_{i=1}^n (r_i - s_i)^2.$$

Она принимает наименьшее возможное значение $S = 0$ тогда и только тогда, когда последовательности полностью совпадают. Наибольшее возможное значение $S = \frac{1}{3}(n^3 - n)$ величина S принимает, когда эти последовательности полностью противоположны. (Это значит, что для $r_i = 1$ значение $s_i = n$; для $r_i = 2$ соответствующие $s_i = n - 1$ и т.д.) Кроме степени сходства последовательностей (r_1, \dots, r_n) и (s_1, \dots, s_n) , на S оказывает влияние также и численность группы n . Чтобы ослабить влияние переменной n , переходят к *коэффициенту ранговой корреляции Спирмена*:

$$\rho = 1 - \frac{6S}{n^3 - n}.$$

Коэффициент ρ по абсолютной величине ограничен единицей: $|\rho| \leq 1$. Свои крайние значения $\rho = \pm 1$ он принимает в указанных выше случаях полной предсказуемости одной ранговой последовательности по другой.

Заметим, что значение S не зависит от первоначальной нумерации объектов. В качестве таковой часто удобно выбрать упорядочение по одному из признаков. Тогда последовательность рангов по этому признаку превратится в последовательность $1, 2, \dots, n$. Вторую последовательность обозначим, скажем, z_1, \dots, z_n . При этом

$$S = \sum_{i=1}^n (r_i - s_i)^2 = \sum_{k=1}^n (k - z_k)^2.$$

Коэффициент Кендэла. Другой коэффициент ранговой корреляции получил популярность после работ М. Кендэла, в особенности после выхода его книги [53]. Этот коэффициент в качестве меры сходства между двумя ранжировками использует минимальное число перестановок соседних объектов, которые надо сделать, чтобы одно упорядочение объектов превратить в другое.

Для определения коэффициента ранговой корреляции по Кендэлу сначала введем статистику Кендэла K . Выберем в качестве первоначальной

чальной нумерации упорядочение объектов по признаку A и подсчитаем K , сопоставляя $(1, 2, \dots, n)$ и (z_1, z_2, \dots, z_n) . Оказывается, что K равно числу *инверсий* в ряду (z_1, \dots, z_n) . Пусть, например, $n = 4$ и $(z_1, \dots, z_4) = (4, 3, 1, 2)$. Инверсии (нарушения порядка) суть: (4 прежде 3) — одна инверсия, (4 прежде 1) — еще одна и (4 прежде 2). Итого, первый элемент последовательности дает три инверсии. Далее подсчитаем число инверсий, которые образует второй элемент последовательности: (3 прежде 1), (3 прежде 2) — итого две инверсии. Единица, как полагается, стоит прежде 2, и потому пара (1, 2) инверсии не образует. Всего инверсий в данном случае $3 + 2 = 5$. Таким образом, $K = 5$. Наименьшее возможное значение $K = 0$, наибольшее $K = n(n - 1)/2$. Как и для S , эти значения получаются при полном совпадении и полной противоположности ранговых последовательностей. Чтобы ослабить влияние n на величину K , от K переходят к коэффициенту ранговой корреляции τ (по Кендэлу):

$$\tau = 1 - \frac{4K}{n(n - 1)}.$$

Как и ρ , τ может изменяться от -1 до $+1$; свои крайние значения τ принимает в указанных выше случаях.

Распределение коэффициентов корреляции ρ и τ . Мы уже отмечали, что в случае независимых признаков вероятность между всеми $n!$ возможными значениями (z_1, \dots, z_n) распределяется равномерно. Это дает возможность (по крайней мере принципиальную) рассчитать закон распределения вероятностей между возможными значениями ρ или τ в условиях H_0 . Для малых значений n это несложная задача, но с ростом n число комбинаций $n!$, которые надо учесть, быстро увеличивается. (Например, $10! = 3628800$). Тем не менее составлены достаточные для практических нужд таблицы распределений случайных величин ρ и τ в случае H_0 . Для небольших n эти таблицы точные, для других значений — приближенные (о чем ниже). Правильнее сказать, что в сборниках статистических таблиц приводят обычно распределения не самих ρ и τ , а определяющих их статистик S и K (либо их вариантов).

Проверка независимости признаков. Теперь обсудим, как с помощью коэффициентов ранговой корреляции можно проверить гипотезу H_0 о независимости признаков. Для этого надо знать характер распределения вероятностей для этих коэффициентов ρ и τ при H_0 и при отступлении от H_0 .

Вероятность распределяется на отрезке $[-1, 1]$. При H_0 распределение этих величин симметрично и концентрируется около нуля (тем сильнее, чем больше n). Если признаки зависимы, распределение веро-

ятностей может быть иным. Поведение коэффициентов ранговой корреляции в этом случае легко проследить лишь для наиболее простого вида связи — монотонной (положительной или отрицательной). Для монотонной положительной связи значение одного признака тем больше, чем больше значение другого (при отрицательной — наоборот). Такая альтернатива независимости легко обнаруживается с помощью коэффициентов ранговой корреляции, абсолютное значение которого в этом случае должно быть близко к единице. Если же зависимость между признаками более сложная, ее влияние на ранжировки может быть не столь простым. Поэтому с помощью коэффициентов ранговой корреляции далеко не всякую зависимость можно отличить от независимости. Все же мы можем сказать, что появление в эксперименте больших (по модулю) наблюдаемых значений коэффициентов ранговой корреляции свидетельствует против гипотезы независимости в пользу связи между признаками (положительной либо отрицательной, смотря по знаку коэффициента).

Для проверки H_0 надо вычислить выборочное значение коэффициента ранговой корреляции и сравнить его с критическим значением для данного уровня значимости, которое следует извлечь из таблиц. Гипотезу H_0 надо отвергнуть (на выбранном уровне значимости), если полученное в опыте значение коэффициента ранговой корреляции превосходит критическое (по модулю).

При больших n критические значения не табулированы, их приходится вычислять по приближенным формулам. Как правило, в таблицах критических значений такие формулы приводятся. Они основаны на том, что при H_0 и больших n случайные величины $\sqrt{n-1}\rho$ и $\sqrt{\frac{9n(n+1)}{2(2n+5)}}\tau$ распределены (приближенно) по стандартному нормальному закону $N(0, 1)$.

Дополнительную информацию по изложенным в этом пункте вопросам можно найти в [32], [106], [113], [115].

9.5. Связь признаков в количественных шкалах

9.5.1. Коэффициент корреляции

Количественные шкалы. Количественными шкалами мы будем называть шкалы отношений и интервальные:

- *интервальной шкалой* называют такую шкалу с непрерывным множеством значений, в которой о двух сопоставляемых объек-

тах можно сказать не только, одинаковы они или различны (как в номинальных шкалах), не только в каком из них признак более выражен (как в порядковых шкалах), но и *насколько* более этот признак выражен;

- *шкалой отношений* называют такую шкалу с непрерывным множеством значений, в которой о двух сопоставляемых объектах можно сказать не только, одинаковы они или различны, не только в каком из них признак более выражен, но и *во сколько раз* более этот признак выражен.

Примером интервальной шкалы является измерение времени или температуры. Сопоставляя календарные даты двух событий, можно сказать, сколько лет, дней, часов и т.д. прошло между ними, т.е. насколько одно событие произошло позже (раньше) другого. Чтобы задать интервальную шкалу, надо выбрать начальную точку отсчета и единицу измерения. В температурной шкале Цельсия начало отсчета — нуль градусов — температура замерзания воды; за сто единиц принят интервал температур от замерзания до кипения воды (при нормальном давлении). Однако отношения измерений не всегда имеют смысл, так, мы не можем сказать, что температура в десять градусов Цельсия «в два раза больше» температуры в пять градусов.

Если же нулевая точка шкалы выбрана не условно, а имеет естественный «физический» смысл, то по результатам измерения можно сказать, во сколько раз один объект превосходит другой по степени выраженности измеряемого признака. Таковы большинство шкал, применяемых в физике и технике: измерение массы, длины и т.п. Эти шкалы называются шкалами отношений.

Независимость признаков. Обсудим, как выразить числом степень взаимной зависимости или установить взаимную независимость двух признаков, измеренных в количественных шкалах. Предположим, что есть некая генеральная совокупность, каждый элемент которой обладает двумя количественными признаками, скажем A и B . Станем наудачу извлекать объекты из этой совокупности. Обозначим через α и β значения, которые при этом принимают признаки A и B . Ясно, что α и β — это случайные величины.

Определение. *Признаки, измеренные в количественной шкале, называются независимыми, если независимы (статистически) случайные величины α и β .*

Как говорилось в гл. 1, случайные величины α и β статистически независимы (для краткости — просто независимы), если независимы любые события U и V , которые выражаются с помощью α и β соответственно. Для независимости α и β достаточно (и необходимо), чтобы были независимы все события вида $U = (a_1 < \alpha < a_2)$, $V = (b_1 < \beta < b_2)$, где $a_1 < a_2$, $b_1 < b_2$ — произвольные числа. Напомним, что неза-

висимыми считаются такие события U, V , что $P(UV) = P(U)P(V)$. Следовательно, условие независимости α и β выглядит так:

$$P(a_1 < \alpha < a_2, b_1 < \beta < b_2) = P(a_1 < \alpha < a_2)P(b_1 < \beta < b_2). \quad (9.14)$$

В основу статистических проверок независимости признаков можно положить проверку того или другого следствия из соотношения (9.14).

Коэффициент корреляции. Из гл. 1 мы знаем, что для независимых случайных величин α, β их ковариация

$$\text{cov}(\alpha, \beta) = M\alpha\beta - M\alpha M\beta$$

равна нулю, а для зависимых случайных величин она может (хотя и не обязательно) отличаться от нуля. Поэтому ненулевое значение ковариации означает зависимость случайных величин. Однако обращение в нуль ковариации не гарантирует независимости: бывают зависимые случайные величины, ковариация которых равна 0 (упражнение: придумайте пример). Кроме того, ковариация вообще может не существовать (так же, как и математические ожидания). Так что обращение в нуль ковариации признаков не является достаточным для их независимости, а только необходимым (и то лишь если ковариация существует).

Однако использование ковариации в качестве меры связи признаков не совсем удобно, так как при переходе к другим единицам измерения (например, от метров к сантиметрам) ковариация тоже изменяется. Поэтому в качестве меры связи признаков обычно используют не $\text{cov}(\alpha, \beta)$, а безразмерную величину — коэффициент корреляции $\rho(\alpha, \beta)$:

$$\rho = \frac{\text{cov}(\alpha, \beta)}{\sqrt{D\alpha}\sqrt{D\beta}}. \quad (9.15)$$

Свойства коэффициента корреляции мы уже описывали в гл. 1. Напомним, что коэффициент корреляции может принимать значения от -1 до 1 , при этом он может быть равен -1 или 1 , лишь если случайные величины α и β линейно связаны, т.е. существуют такие числа t, k , что $P(\beta = t\alpha + k) = 1$. Для независимых случайных величин коэффициент корреляции (если он существует) равен нулю.

Выборочный коэффициент корреляции. Чтобы вычислить ρ по формуле (9.15), надо знать ковариацию и дисперсию признаков. На практике они обычно неизвестны. Информация о признаках α, β обычно представлена выборкой $(\alpha_1, \beta_1), \dots, (\alpha_n, \beta_n)$, которую получают, научаду отбирая n объектов и измеряя значения их признаков.

По выборке можно найти выборочный аналог теоретического коэффициента корреляции — коэффициент корреляции выборки, или *выборочный коэффициент корреляции*. Как мы говорили в гл. 1, его

вычисляют, заменяя усреднения по генеральной совокупности (математические ожидания) усреднениями по выборке. Выборочные аналоги для дисперсий, согласно п. 1.8, суть:

$$s_{\alpha}^2 = \frac{1}{n} \sum_{i=1}^n (\alpha_i - \bar{\alpha})^2, \quad s_{\beta}^2 = \frac{1}{n} \sum_{i=1}^n (\beta_i - \bar{\beta})^2.$$

Как обычно, черта сверху означает усреднение по выборке. Выборочным аналогом для $M\alpha\beta$ служит $\overline{\alpha\beta} = \frac{1}{n} \sum_{i=1}^n \alpha_i \beta_i$. Это позволяет записать выборочный коэффициент корреляции в виде:

$$r = \frac{\overline{\alpha\beta} - \bar{\alpha}\bar{\beta}}{s_{\alpha}s_{\beta}}. \quad (9.16)$$

Можно r выразить и по-другому, например:

$$r = \frac{\sum_{i=1}^n (\alpha_i - \bar{\alpha})(\beta_i - \bar{\beta})}{\sqrt{\sum_{i=1}^n (\alpha_i - \bar{\alpha})^2} \sqrt{\sum_{i=1}^n (\beta_i - \bar{\beta})^2}}. \quad (9.17)$$

В силу закона больших чисел $r \rightarrow \rho$ при неограниченном росте объема выборки, т.е. при $n \rightarrow \infty$. Более того, центральная предельная теорема позволяет заключить, что случайная величина $\sqrt{n}(r - \rho)$ распределена приблизительно нормально, причем асимптотическое среднее этого нормального закона равно 0. Можно указать и асимптотическую дисперсию, но выражение ее довольно сложное. Практически им не пользуются. К вопросу о предельном распределении r мы еще вернемся.

9.5.2. Нормальная корреляция

Коэффициент корреляции не всегда выполняет свою роль измерителя связи между признаками, так как случай $\rho = 0$ еще не означает статистической независимости α и β . Но если совместное распределение пары случайных величин (α, β) оказывается нормальным, то равенство $\rho = 0$ влечет за собой статистическую независимость α и β .

Общее условие независимости признаков. Укажем, как выражается независимость случайных величин в терминах их совместной и частных плотностей. Пусть совместная плотность пары случайных величин (α, β) есть $p(x, y)$. Тогда плотность распределения случайной величины α (частная плотность) есть

$$p_1(x) = \int_{-\infty}^{\infty} p(x, y) dy.$$

Аналогично плотность распределения β равна

$$p_2(y) = \int_{-\infty}^{\infty} p(x, y) dx.$$

Напомним определение независимости случайных величин α и β : для любых чисел $a < b$, $c < d$

$$P(a < \alpha < b, c < \beta < d) = P(a < \alpha < b) P(c < \beta < d).$$

Если записать вероятности этих событий через соответствующие плотности, мы получим следующее условие независимости α и β :

$$\int_a^b \int_c^d p(x, y) dx dy = \int_a^b p_1(x) dx \int_c^d p_2(y) dy.$$

Отсюда можно заключить, если привлечь более глубокие сведения из интегрального исчисления, что необходимым и достаточным условием независимости служит условие равенства совместной плотности произведению частных плотностей:

$$p(x, y) = p_1(x) p_2(y).$$

Условие независимости нормальных признаков. Обратимся к виду общей плотности двумерного нормального распределения, как она дана в п. 2.5, либо к двумерной плотности в стандартизованных координатах, и сравним ее с произведением частных (одномерных) плотностей, которые тоже нормальны (см. п. 2.5). Сопоставляя их, мы можем убедиться, что двумерная нормальная плотность представляется в виде произведения частных плотностей тогда и только тогда, когда $\rho = 0$.

Итак, для пары признаков, имеющих совместно двумерное нормальное распределение, условие $\rho = 0$ (некоррелированность признаков) эквивалентно их независимости. Поэтому проверка гипотезы о независимости признаков, совместное распределение которых является двумерным нормальным, сводится к проверке гипотезы $H_0 : \rho = 0$.

Проверка независимости. В гауссовском случае, когда коэффициент корреляции $\rho = 0$, распределение выборочного коэффициента r известно достаточно хорошо. Это распределение симметрично и сконцентрировано около нуля (тем сильнее, чем больше n). Поэтому гипотезу H_0 следует отвергнуть, если выборочное значение r (которое отличается от гипотетического $\rho = 0$ только за счет действия случайности) слишком далеко (неправдоподобно далеко) отклоняется от нуля, т.е. $|r|$ превосходит критическое значение (для выбранного уровня значимости).

Расчет квантилей для r основан на том, что случайная величина t , получаемая из r монотонным преобразованием по формуле

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2},$$

при гипотезе H_0 подчиняется распределению Стьюдента с $m = n - 2$ степенями свободы. Поэтому квантиль уровня q распределения r (скажем, $r_{m,q}$) получится преобразованием квантили уровня q распределения Стьюдента с m степенями свободы (скажем, $t_{m,q}$) по формуле

$$r_{m,q} = \frac{t_{m,q}}{\sqrt{m + t_{m,q}^2}}.$$

Таблицы процентных точек (критических значений) для r приведены во многих сборниках таблиц по математической статистике, в частности в [19]. Однако эти процентные точки можно рассчитать и самостоятельно, имея в распоряжении таблицу квантилей или процентных точек соответствующего распределения Стьюдента.

Доверительные интервалы для ρ . Для двумерного нормального распределения коэффициент корреляции не только решает вопрос о том, зависимы признаки или нет, но и измеряет степень их связи. Поэтому в нормальном случае нужно не только уметь проверять гипотезу $H : \rho = 0$, но и указывать доверительные пределы для истинного ρ (особенно если выборка показывает, что истинное $\rho \neq 0$, т.е. признаки связаны). Для этого надо знать, каково распределение r не только при $\rho = 0$, но при произвольном ρ .

Для больших n и малых по абсолютному значению ρ выборочный коэффициент корреляции r можно считать распределенным нормально с математическим ожиданием ρ и дисперсией $(1 - \rho^2)^2 / (n - 1)$. Для указанной выше цели этот факт использовать трудно в связи с тем, что неизвестное значение ρ входит в выражение не только среднего, но и дисперсии. Р. Фишер предложил преобразовать r так, чтобы асимптотическая дисперсия преобразованной величины практически перестала зависеть от ρ . Вот это «преобразование Фишера»:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}.$$

Распределение случайной величины z хорошо аппроксимируется нормальным распределением со средним

$$\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$$

и дисперсией $1/(n-3)$. Иначе говоря, случайная величина $\sqrt{n-3}(z-\zeta)$ распределена приблизительно по закону $N(0, 1)$. Считают, что для $n \geq 20$ распределение z можно для практических целей считать нормальным (с указанными параметрами).

Величина $\rho/(2(n-1))$ мала по сравнению с $1/\sqrt{n-3}$. Поэтому ее обычно пренебрегают, когда речь идет об оценивании ρ по одной выборке. Но при соединении результатов, полученных по нескольким выборкам, это слагаемое все же может оказывать влияние.

Доверительные пределы для ρ (при данных значениях r , n и коэффициенте доверия) получают из стандартного нормального распределения путем обращения преобразования Фишера. В [19] такие доверительные пределы указаны явно.

9.6. Замечания о связи признаков, измеренных в разных шкалах

Нередки случаи, когда вопрос о независимости или связи возникает для признаков, измеряемых в шкалах различных видов, например номинальной и порядковой, порядковой и количественной и т.п. Например, соединение номинального и количественного признаков часто встречается при анализе факторных таблиц. Там вопрос о независимости истолковывается как отсутствие влияния номинального признака на количественный. Многие ситуации такого рода описаны в [49], [106].

В общем случае, к сожалению, методы анализа связи признаков становятся гораздо сложнее, чем приведенные выше. Для упрощения исследования часто приходится одну из шкал измерений понижать до уровня другой, а затем использовать стандартную методику. При подобном понижении, несомненно, происходит некоторая потеря информации, зато последующий анализ становится проще и ясней.

9.7. Анализ таблиц сопряженности и коэффициенты корреляции в пакете SPSS

Пример 9.1к. Проведем анализ таблицы сопряженности для данных о предпочтении различных видов инструкций в зависимости от типа нервной системы (табл. 9.1). Проверим гипотезу о независимости этих признаков.

Подготовка данных. Пакет SPSS содержит мощную процедуру анализа таблиц сопряженности с разветвленными возможностями. Однако эта процедура работает с исходными данными наблюдений, а не готовой таблицей кросстабуляции.

Исходные данные из табл. 9.1 должны быть записаны в виде двух переменных. В первой (**group**) для каждого индивида указывается, к какой группе он относится, а во второй (**reaction**) — тип инструкции, который он предпочитает. При вводе данных удобно использовать простые обозначения. Обозначим с помощью 1 высокореактивных индивидов и с помощью 2 — низкореактивных. Аналогично предпочтение детальной инструкции пусть будет обозначено через 1, а краткой — через 2. Тогда массив исходных наблюдений в редакторе данных пакета примет вид, показанный на рис. 9.1.

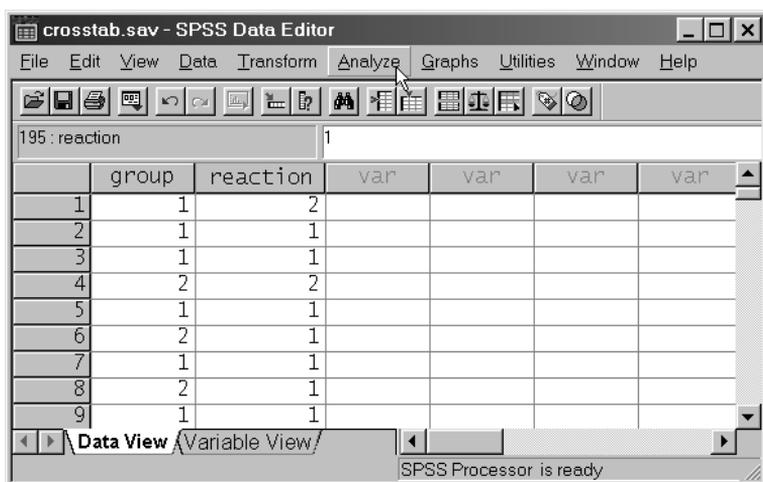


Рис. 9.1. Пакет SPSS. Подготовка данных для процедуры «Crosstabs»

Выбор процедуры. В блоке Descriptive Statistics меню Analyze выбрать процедуру Crosstabs.

Заполнение полей ввода данных. Окно ввода данных и параметров этой процедуры приведено на рис. 9.2.

Для получения таблицы сопряженности, аналогичной табл. 9.1, в нем необходимо перенести переменную **reaction** в поле **Row(s)** (строки), а переменную **group** — в поле **Column(s)** (столбцы). Затем нажать кнопку **Statistics** для настройки расчетов в процедуре.

На рис. 9.3 изображено окно выбора параметров процедуры **Crosstabs**. В этом окне, с учетом постановки задачи, указать режим расчета **Chi-**

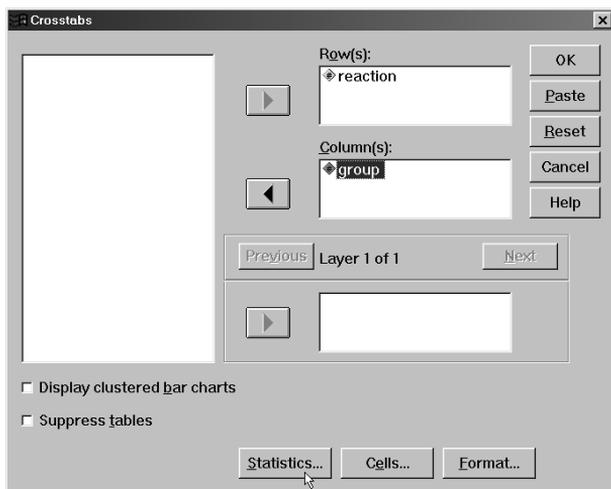


Рис. 9.2. Пакет SPSS. Окно ввода данных и параметров процедуры «Crosstabs»

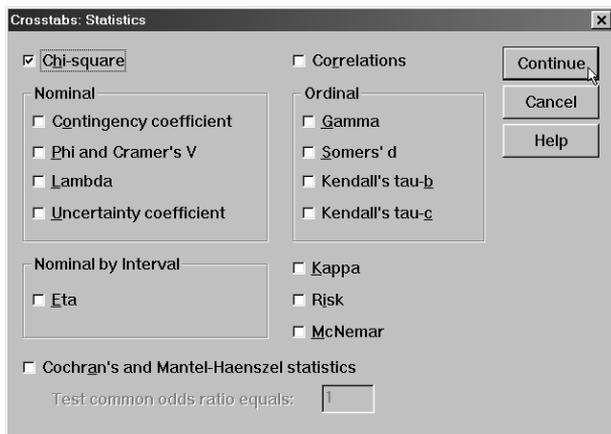


Рис. 9.3. Пакет SPSS. Окно выбора параметров процедуры «Crosstabs»

square (критерий хи-квадрат). Окно также содержит большой выбор мер связи для номинальных (**Nominal**) и порядковых (**Ordinal**) типов данных. Описание этих статистик можно найти в [7], [53], [83], [102], [106].

Для задания формы выдачи самой таблицы сопряженности нажать кнопку **Cells** в окне ввода данных и параметров процедуры **Crosstabs** (рис. 9.2). При этом появится окно, показанное на рис. 9.4. В нем можно настроить число параметров, выдаваемых процедурой в каждую клетку таблицы сопряженности. Чтобы не загромождать таблицу, укажем в группе **Counts** этого окна режимы **Observed** (наблюдения) и **Expect-**

ed (ожидаемые значения при условии независимости признаков). Это позволит нам сравнить наблюдаемое число значений в каждой клетке таблицы сопряженности с ожидаемым при независимости признаков. Нажав кнопку (Continue), вернуться в окно ввода данных и параметров и нажать кнопку (OK).

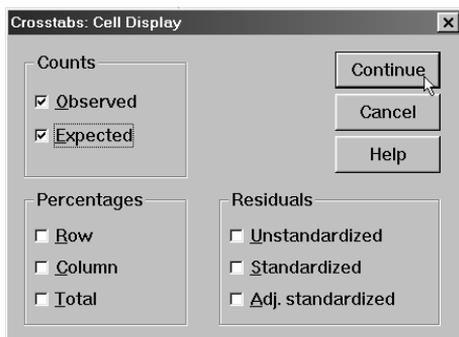


Рис. 9.4. Пакет SPSS. Окно настройки формы таблицы сопряженности

Результаты. Основные результаты работы процедуры выдаются в две таблицы. Первая (рис. 9.5) представляет саму таблицу сопряженности. В ее клетках кроме наблюдаемого числа значений (строки **Count**) приведены и ожидаемые значения (строки **Expected Count**). В этой таблице видно, что ожидаемые при условии независимости признаков значения довольно сильно отличаются от наблюдаемых.

REACTION * GROUP Crosstabulation

		GROUP		Total
		1	2	
REACTION 1	Count	63	42	105
	Expected Count	52,2	52,8	105,0
2	Count	34	56	90
	Expected Count	44,8	45,2	90,0
Total	Count	97	98	195
	Expected Count	97,0	98,0	195,0

Рис. 9.5. Пакет SPSS. Таблица сопряженности

Вторая таблица **Chi-Square Tests** (рис. 9.6) включает результаты, связанные с критерием хи-квадрат: значения различных модификаций критерия в столбце **Value**, числа степеней свободы в столбце **df**, асимптотический уровень значимости (столбец **Asymp. Sig. (2-sided)**) и точный уровень значимости, когда он известен (последние 2 столбца таблицы). Все полученные уровни значимости говорят, что гипотеза о независимости должна быть отвергнута.

Комментарии. 1. Процедура **Crosstabs** позволяет сформировать и проанализировать таблицу сопряженности из данных, имеющих несколько (более двух) факторов классификации. Для этого используется поле **Layer** в окне ввода данных

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	9,573 ^b	1	,002		
Continuity _a Correction	8,705	1	,003		
Likelihood Ratio	9,656	1	,002		
Fisher's Exact Test				,003	,002
Linear-by-Linear Association	9,524	1	,002		
N of Valid Cases	195				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 44,77.

Рис. 9.6. Пакет SPSS. Таблица результатов критерия хи-квадрат при проверке независимости признаков

и параметров этой процедуры. Скажем, если бы мы хотели получить отдельные таблицы сопряженности для юношей и девушек в этом исследовании, то в это поле следовало бы ввести переменную, в которой бы находился признак пола респондентов.

2. Критерий хи-квадрат в процедуре **Crosstabs** выдает предостережения, когда недостаточное количество наблюдений в клетках может исказить результаты проверки независимости. Более подробно об этом сказано в п. 10.4.

Следующий пример посвящен задаче выявления связи признаков, измеренных в порядковых или количественных шкалах.

Пример 9.2к. С помощью коэффициентов корреляции Спирмена, Кендэла и Пирсона выяснить связь между скоростями реакции на звук и на свет по данным табл. 3.1.

Подготовка данных. Указанные данные уже рассматривались нами в примере 3.2к. Вид экрана редактора базы данных с частью введенных данных табл. 3.1 приведен на рис. 3.6. Как и прежде, будем считать, что данные находятся в двух переменных **sound** и **light** редактора данных пакета.

Выбор процедуры. Выбрать процедуру **Bivariate** (парные корреляции) из блока **Correlate** (корреляции) меню **Analyze** редактора пакета.

Заполнение полей ввода данных. Окно ввода данных и параметров процедуры **Bivariate** приведено на рис. 9.7. В нем следует перенести переменные **light** и **sound** в поле **Variables**. В блоке **Correlation Coefficients** (коэффициенты корреляции) отметить все указанные там коэффициенты. В блоке **Test Significance** (критерии значимости) указать характер альтернатив, против которых будет проверяться нулевая гипотеза об отсутствии связи между выборками. Полезно включить опцию **Flag significant correlations** (пометка значимых коэффициентов). В этом режиме процедура будет отмечать в таблице результатов одной или двумя звездочками (* или

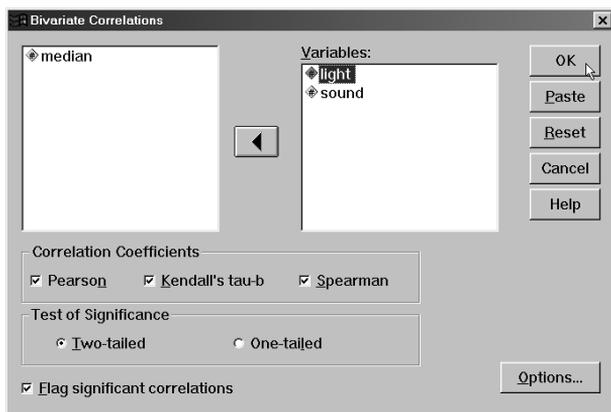


Рис. 9.7. Пакет SPSS. Окно ввода данных и параметров процедуры «Bivariate»

***) коэффициенты корреляции, значимо отличные от нуля на пяти- и однопроцентном уровнях значимости.

Результаты. В окне навигатора вывода данных процедура сформирует две таблицы. Первая из них относится к коэффициенту корреляции Пирсона и показана на рис. 9.8. В клетках этой таблицы указаны три величины: сверху — значение коэффициента корреляции, посередине — его минимальный уровень значимости против выбранных альтернатив и снизу — объем выборки. На главной диагонали этой таблицы стоят значения корреляций переменных с самими собой. Они, естественно, равны 1, а уровень значимости у них не указывается. Интересующее нас значение стоит на побочной диагонали. (Любая таблица корреляций является симметричной относительно главной диагонали.)

Correlations

		LIGHT	SOUND
LIGHT	Pearson Correlation	1,000	,213
	Sig. (2-tailed)	,17	,411
	N	17	17
SOUND	Pearson Correlation	,213	1,000
	Sig. (2-tailed)	,411	,17
	N	17	17

Рис. 9.8. Пакет SPSS. Таблица коэффициентов корреляции Пирсона процедуры «Bivariate»

Полученный коэффициент корреляции Пирсона равен — 0.213, а его уровень значимости — 0.411. То есть нельзя считать этот коэффициент значимо отличным от нуля.

Correlations

			LIGHT	SOUND
Kendall's tau_b	LIGHT	Correlation Coefficient	1,000	,222
		Sig. (2-tailed)	,	,216
	N	17	17	
	SOUND	Correlation Coefficient	,222	1,000
Sig. (2-tailed)		,216	,	
N	17	17		
Spearman's rho	LIGHT	Correlation Coefficient	1,000	,276
		Sig. (2-tailed)	,	,283
	N	17	17	
	SOUND	Correlation Coefficient	,276	1,000
Sig. (2-tailed)		,283	,	
N	17	17		

Рис. 9.9. Пакет SPSS. Таблица коэффициентов корреляции Кендэла и Спирмена процедуры «Bivariate»

Вторая таблица (рис. 9.9) содержит аналогичную информацию для ранговых коэффициентов корреляции Кендэла и Спирмена.

Значения всех трех коэффициентов примерно совпадают, и ни один из них значимо не отличается от нуля.

Комментарии. 1. Описанная процедура позволяет вычислять коэффициенты корреляции для всех возможных пар переменных, указанных в поле **Variables** окна ввода данных процедуры. Результатом подобных вычислений является корреляционная матрица.

2. Значения разбираемых нами коэффициентов корреляции далеко не всегда так хорошо совпадают, как в рассматриваемом примере. Коэффициент корреляции Пирсона гораздо более чувствителен к отдельным нехарактерным значениям, чем ранговые коэффициенты. Если в результате расчетов наблюдаются значительные расхождения в этих коэффициентах, то следует более внимательно проанализировать исходные данные. В целом коэффициенты ранговой корреляции заслуживают гораздо большего доверия, чем коэффициент корреляции Пирсона.

3. Малые выборки позволяют обнаружить связь между переменными только тогда, когда она довольно сильно выражена. Так, для выборок объема порядка 20 наблюдений, для выявления связи между ними с помощью коэффициента корреляции Спирмена на пятипроцентном уровне значимости против односторонних альтернатив, значения этого коэффициента, оцененное по выборке, должно превышать по модулю значение 0.4. Для коэффициента корреляции Кендэла аналогичное значение равно примерно 0.27. Более подробную информацию на эту тему дают табл. 9 и 10 из приложения.

Дополнительная литература

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика. Основы эконометрики. Т. 1. Теория вероятностей и прикладная статистика. — М.: Юнити-Дана, 2001. — 656 с.

2. Аптон Г. Анализ таблиц сопряженности. — М.: Финансы и статистика, 1982. — 144 с.
3. Кендэл М., Стьюарт А. Статистические выводы и связи. — М.: Наука, 1973. — 899 с.
4. Кендэл М. Ранговые корреляции. — М.: Статистика, 1975. — 212 с.
5. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. — М.: ФИЗМАТЛИТ, 2006. — 816 с.
6. Рунион Р. Справочник по непараметрической статистике. Современный подход. — М.: Финансы и статистика, 1982. — 198 с.
7. Справочник по прикладной статистике: в 2 т.; под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. — М.: Финансы и статистика, 1989, 1990.
8. Холлендер М., Вулф Д. Непараметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.

Критерии согласия

Во многих статистических задачах мы предполагаем, что некоторые случайные величины имеют заданное распределение (нормальное, экспоненциальное и т.д.) с известными или неизвестными параметрами этого распределения, и далее исходя из этого допущения мы делаем те или иные выводы. Например, мы можем предположить, что рассеяние пуль при стрельбе описывается нормальным распределением, а время службы электрической лампочки — экспоненциальным. Чем лучше мы знаем законы изменчивости данных, их распределения вероятностей, тем точнее и надежней могут быть наши статистические выводы.

Однако при этом, естественно, возникает вопрос: насколько наши предположения о распределении случайных величин соответствуют экспериментальным данным? Более реалистично поставить этот вопрос иначе: не вступает ли принятая статистическая модель в противоречие с имеющимися данными? Для решения этой задачи придуманы разные способы, иначе говоря, статистические критерии. Чтобы выделить такие критерии из остальных, их часто называют *критериями согласия*.

Определение. *Критериями согласия называют статистические критерии, предназначенные для обнаружения расхождений между гипотетической статистической моделью и реальными данными, которые эта модель призвана описать.*

В этой главе рассказано о некоторых распространенных критериях согласия — омега-квадрат, хи-квадрат, Колмогорова и Колмогорова-Смирнова. Особое внимание уделено случаю, когда необходимо проверить принадлежность распределения данных некоторому параметрическому семейству, например нормальному. Эта весьма распространенная на практике ситуация из-за своей сложности исследована не до конца и не полностью отражена в учебной и справочной литературе.

10.1. Введение

Критериями согласия называют статистические критерии, предназначенные для проверки согласия опытных данных и теоретической модели. Лучше всего этот вопрос разработан, если наблюдения представляют случайную выборку. Теоретическая модель в этом случае описывает закон распределения. В дальнейшем мы будем обсуждать

именно эту задачу, как потому, что она важна и сама по себе, так и потому, что к ней удастся свести многие другие проблемы согласия.

Теоретическое распределение. Мы будем называть теоретическим то распределение вероятностей, которое управляет случайным выбором. Представления о нем может дать не только теория. Источниками знаний здесь могут быть и традиция, и прошлый опыт, и предыдущие наблюдения. Надо лишь подчеркнуть, что это распределение должно быть выбрано независимо от тех данных, по которым мы собираемся его проверять. Иначе говоря, недопустимо сначала «подогнать» по выборке некоторый закон распределения, а потом пытаться проверить согласие с полученным законом по этой же выборке¹.

Простые и сложные гипотезы. Говоря о теоретическом законе распределения, которому гипотетически должны бы следовать элементы данной выборки, надо различать *простые* и *сложные* (т.е. составные) гипотезы об этом законе:

- простая гипотеза прямо указывает некий определенный закон вероятностей (распределение вероятностей), по которому возникли выборочные значения;
- сложная гипотеза указывает не единственное распределение, а какое-то их множество (например, параметрическое семейство).

Например, для ошибок округления при измерении расстояний с помощью линейки со шкалой 1 см мы можем предположить, что их распределение — равномерное на отрезке от -0.5 см до 0.5 см. Эта гипотеза является простой, так как она указывает единственное теоретическое распределение. А при исследовании мощности выпущенных с завода электрических лампочек мы можем предположить, что эта мощность описывается нормальным распределением с неизвестными средним и дисперсией. Эта гипотеза — сложная, она представляет собой двухпараметрическое семейство распределений.

Естественно, что методы проверки согласия с простыми и сложными гипотезами должны быть различны. Мы начнем с простых гипотез (п. 10.2–10.4), хотя на практике они встречаются реже, чем сложные: ведь в большинстве случаев теоретические соображения или традиция не идут далее указания типа распределения (нормальный, показательный, пуассоновский и т.п.), параметры которого остаются неопределенными. В п. 10.5–10.6 мы рассмотрим случай сложных гипотез.

¹ Однако можно случайным образом разбить выборку на две части, по одной «подогнать» закон распределения, а по другой — проверить его.

10.2. Критерии согласия Колмогорова и омега-квадрат в случае простой гипотезы

Простая гипотеза. Мы будем рассматривать ситуацию, когда измеряемые данные являются числами, иначе говоря, одномерными случайными величинами. Как говорилось в гл. 1, распределение одномерных случайных величин может быть полностью описано указанием их функции распределения. И многие критерии согласия основаны на проверке близости теоретической и эмпирической (выборочной) функций распределения.

Пусть мы имеем выборку размера n . Обозначим истинную функцию распределения, которой подчиняются наблюдения, $G(x)$, эмпирическую (выборочную) функцию распределения — $F_n(x)$, а гипотетическую функцию распределения — $F(x)$. Тогда гипотеза H о том, что истинная функция распределения есть $F(x)$, записывается в виде:

$$H : G(\cdot) = F(\cdot).$$

Как проверить гипотезу H ? Если H верна, то F_n и F должны проявлять определенное сходство и различие между ними должно убывать с увеличением n . Действительно, как говорилось в п. 1.8, вследствие теоремы Бернулли $F_n(x) \rightarrow F(x)$ при $n \rightarrow \infty$. Для количественного выражения сходимости функций F_n и F используют различные способы, о которых будет говориться ниже.

Статистика Колмогорова. Для выражения сходимости функций можно использовать то или иное расстояние между этими функциями. Например, можно сравнить F_n и F в равномерной метрике, т.е. рассмотреть величину:

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F(x)|. \quad (10.1)$$

Определение. Статистику D_n называют статистикой Колмогорова.

Очевидно, что D_n — случайная величина, поскольку ее значение зависит от случайного объекта F_n . Если гипотеза H справедлива и $n \rightarrow \infty$, то $F_n(x) \rightarrow F(x)$ при всяком x . Поэтому естественно, что при этих условиях $D_n \rightarrow 0$. Если же гипотеза H неверна, то $F_n \rightarrow G$ и $G \neq F$, а потому $\sup_{-\infty < x < \infty} |F_n(x) - F(x)| \rightarrow \sup_x |G(x) - F(x)|$. Эта последняя величина положительна, так как G не совпадает с F . Такое различие в поведении D_n в зависимости от того, верна H или нет, позволяет использовать D_n как статистику для проверки H .

Как всегда при проверке гипотезы, следует рассуждать так, как если бы гипотеза была верна. Ясно, что H должна быть отвергнута,

если полученное в эксперименте значение статистики D_n кажется неправдоподобно большим. Но для этого надо знать, как распределена статистика D_n при гипотезе $H : F = G$ при данных n и G .

Замечательное свойство D_n состоит в том, что если $G = F$, т.е. если гипотетическое распределение указано правильно, то закон распределения статистики D_n оказывается *одним и тем же* для всех непрерывных функций G . Он зависит только от объема выборки n .

Доказательство этого факта основано на том, что статистика (10.1) не изменяет своего значения при монотонных преобразованиях оси x . Таким преобразованием любое непрерывное распределение G можно превратить в равномерное на отрезке $[0, 1]$. При этом $F_n(\cdot)$ перейдет в функцию распределения выборки из этого равномерного распределения.

Таблицы. При малых n для статистики D_n при гипотезе H составлены таблицы процентных точек. Например, в [19], табл. 6.2, они доведены до $n = 100$. При больших n распределение D_n (при гипотезе H) указывает найденная в 1933 г. А.Н. Колмогоровым предельная теорема. Она говорит о статистике $\sqrt{n} D_n$ (поскольку сама величина $D_n \rightarrow 0$ при H , приходится умножать ее на неограниченно растущую величину, чтобы распределение стабилизировалось).

Асимптотическое приближение. Теорема Колмогорова утверждает, что при справедливости H (и если G непрерывна) величина $P(\sqrt{n} D_n < z)$ при $n \rightarrow \infty$ имеет предел, и дает его выражение:

$$\lim_{n \rightarrow \infty} P(\sqrt{n} D_n < z) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 z^2}. \quad (10.2)$$

В сборниках таблиц можно найти значения функции (10.2) (см., например, [19], табл. 6.1).

Алгоритм проверки гипотезы. Как же использовать статистику Колмогорова (10.1) для проверки простой гипотезы $H : G = F$? По исходной выборке надо вычислить значение статистики D_n . Для этого годится простая формула

$$D_n = \max_{1 \leq k \leq n} \left[\frac{k}{n} - F(x_{(k)}), F(x_{(k)}) - \frac{k-1}{n} \right]. \quad (10.3)$$

Здесь через $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ обозначены элементы вариационного ряда, построенного по исходной выборке. Полученную величину D_n затем надо сравнить с извлеченными из таблиц критическими значениями. Гипотезу H приходится отвергать (на выбранном уровне значимости), если полученное в опыте значение D_n превосходит выбранное критическое значение, соответствующее этому уровню значимости.

Критерий омега-квадрат. Другой популярный критерий согласия получим, измеряя расстояние между F_n и F в интегральной метрике. Он основан на так называемой *статистике омега-квадрат*:

$$\omega_n^2 = \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 dF(x). \quad (10.4)$$

Для вычисления ω_n^2 по реальной выборке можно использовать формулу

$$n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^n \left[F(x_{(i)}) - \frac{2i-1}{2n} \right]^2. \quad (10.5)$$

При справедливости гипотезы $H : F = G$ и непрерывности функции G распределение статистики ω_n^2 , так же, как распределение статистики D_n , зависит только от n и не зависит от G .

Таблицы. Так же, как для D_n , для ω_n^2 при малых n имеются таблицы процентных точек, а для больших значений n следует использовать предельное (при $n \rightarrow \infty$) распределение статистики $n\omega_n^2$. (Здесь снова приходится умножать на неограниченно растущий множитель: в данном случае — на n .) Предельное распределение было найдено Н.В.Смирновым в 1939 г. Приводить его здесь нет необходимости. Достаточно сказать, что для него составлены подробные таблицы и вычислительные программы (см., например, [19], табл. 6.4а).

Состоятельность. Отметим важное с теоретической точки зрения свойство критериев, основанных на D_n и ω_n^2 : они *состоятельны* против любой альтернативы $G \neq F$.

Определение. *Статистический критерий для проверки гипотезы H называют состоятельным против альтернативы H' , если вероятность с его помощью отвергнуть H , когда на самом деле верна H' , стремится к 1 при неограниченном увеличении объема наблюдений.*

Состоятельный против всех альтернатив критерий, в принципе, при большом числе наблюдений, способен обнаружить *любое* отступление от гипотезы. Таким образом, состоятельность критериев Колмогорова и омега-квадрат означает, что любое отличие распределения выборки от теоретического будет с их помощью обнаружено, если наблюдения будут продолжаться достаточно долго.

Замечание. Практическую значимость свойства состоятельности не следует преувеличивать. Во-первых, трудно рассчитывать на получение большого числа наблюдений в неизменных условиях. Во-вторых, теоретическое представление о законе распределения, которому должна подчиняться выборка, всегда

имеет характер математической модели, т.е. является в какой-то мере приближенным. Поэтому точность статистических проверок должна быть сопоставима с точностью, которую мы ожидаем от математической модели в целом и в деталях. (Скажем, представление о том, что наблюдения независимы и имеют неизменный закон распределения, является частью математической модели.) Тем не менее свойство состоятельности статистического критерия (как и статистической оценки параметра) всегда является ценным и желательным.

10.3. Практический пример (закон Менделя)

Прекрасный пример применения на деле критерия Колмогорова был дан самим А.Н. Колмогоровым спустя несколько лет после открытия этого критерия в небольшой заметке 1940 г. «Об одном новом подтверждении законов Менделя» в [56]. Мы воспроизведем изложение этой работы по брошюре В.Н. Тутубалина [96].

Законы, открытые монахом Г.И. Менделем в 1865 г. в результате восьмилетних опытов на крошечной (менее четверти сотки) делянке, являются одним из краеугольных камней современной теории наследственности. Мендель проводил опыты по гибридизации (скрещиванию) различных сортов гороха — с желтыми и зелеными зернами — и обнаружил, что в при таком скрещивании первое поколение гибридов все имеет желтые зерна, а в следующем, втором поколении снова появляются растения с зелеными зернами, причем соотношение количеств растений с желтыми и зелеными зернами — 3 : 1, а колебания этого соотношения вызываются случайными причинами. Ту же картину Мендель обнаружил и для других свойств гороха. Кроме того, он установил, что различные свойства растений передаются по наследству независимо друг от друга.

Работы Менделя намного опередили свое время. Лишь в 1900 г. его законы были заново переоткрыты, а затем были найдены публикации Менделя, описывающие эти законы. В начале XX века законы Менделя были объяснены и обобщены исходя из генетической теории наследственности. Однако в России в 30 — 50-е гг. генетика была объявлена буржуазной лженаукой, занимающиеся ею ученые преследовались, а официальная биологическая школа Т.Д. Лысенко старалась показать, что генетические законы, в частности законы Менделя, не действуют вообще. Так, Н.И. Ермолаева пыталась опровергнуть законы Менделя (журнал «Яровизация», 1939, 2(23), с. 79–86), рассматривая гибриды второго поколения не в совокупности, а по «семействам» — группам растений, выросших в одном ящике из плодов одного растения первого поколения. При обработке данных по отдельным «семействам» было

обнаружено, что отношение числа растений со слабым (рецессивным) признаком к общему числу растений-гибридов второго поколения сильно колеблется и никогда не совпадает в точности с предсказанным Менделем соотношением $1/4$. Отсюда Н.И. Ермолаева и другие сторонники Т.Д. Лысенко делали вывод, что законы Менделя не выполняются.

Однако А.Н. Колмогоров показал, что результаты опытов Н.И. Ермолаевой можно объяснить как раз на основе простейшей модели Менделя. Если для k семейств численностью n_1, n_2, \dots, n_k численности проявления рецессивного признака — $\mu_1, \mu_2, \dots, \mu_k$, то из классической теоремы Муавра—Лапласа (частного случая центральной предельной теоремы) следует, что нормированные величины

$$\mu_i^* = (\mu_i - n_i p) / \sqrt{n_i p(1 - p)}$$

имеют приблизительно нормальное распределение с параметрами $(0, 1)$. Здесь $p = 1/4$, а точность упомянутой нормальной аппроксимации вполне достаточна при n_i порядка нескольких десятков. Поэтому на совокупность $\mu_1^*, \mu_2^*, \dots, \mu_k^*$, можно смотреть (если модель Менделя верна) как на выборку, теоретическое распределение которой есть стандартный нормальный закон.

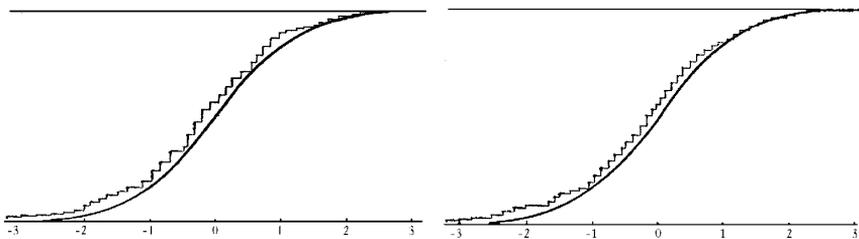


Рис. 10.1. Эмпирическая и теоретическая функции распределения: слева — для первой выборки ($k = 98$), справа — для второй выборки ($k = 123$)

А.Н. Колмогоров рассмотрел две наиболее многочисленные серии опытов Н.И. Ермолаевой, которым соответствуют две выборки размером в $k = 98$ и $k = 123$ наблюдения. Эмпирические и теоретические функции этих распределений воспроизведены на рис. 10.1 соответственно (рисунки скопированы из цитированной работы). Для количественного измерения согласия между эмпирической и теоретической функциями распределения (при числе наблюдений порядка 100) можно использовать статистику Колмогорова. Для первой выборки А.Н. Колмогоров получил $\sqrt{k} D_k = 0.82$, для второй — $\sqrt{k} D_k = 0.75$. При выполнении гипотезы о справедливости законов Менделя вероятности получить такое же или большее расхождение между выборочным и теоретическим распределением равны 0.51 для первой выборки и 0.63 для второй

выборки. Мы видим, что эти вероятности отнюдь не малы, поэтому отвергать статистическую гипотезу, а вместе с нею и закон Менделя нет никаких оснований.

Таким образом, чисто статистическое исследование превращает данные, казавшиеся опровержением законов Менделя, в их существенное подтверждение.

10.4. Критерий согласия хи-квадрат К. Пирсона для простой гипотезы

Теоретики предложили много статистических критериев, аналогичных D_n и ω_n^2 . При всей привлекательности их с математической точки зрения надо отметить, что требование непрерывности теоретического распределения $F(\cdot)$ позволяет прилагать их не ко всем выборкам. Например, вне поля их действия остаются выборки из дискретных распределений. Поэтому надо познакомиться с более универсальным критерием К. Пирсона (1900), опирающимся на теорему, также носящую имя К. Пирсона. (С обобщением этой теоремы мы встречались ранее в п. 9.3.)

Теорема К. Пирсона относится к независимым испытаниям с конечным числом исходов, т.е. к испытаниям Бернулли (в несколько расширенном смысле). Она позволяет судить о том, согласуются ли наблюдаемые в большом числе испытаний частоты этих исходов с их предполагаемыми вероятностями. Вот ее точная формулировка.

Теорема К. Пирсона. Пусть n — число независимых повторений некоего опыта, который заканчивается одним из r (r — натуральное число) элементарных исходов, скажем, A_1, \dots, A_r . Пусть p_1, \dots, p_r — вероятности этих исходов, причем $p_1 + \dots + p_r = 1$. Обозначим через m_1, \dots, m_r количества опытов, заканчивающихся соответственно исходами A_1, \dots, A_r . (Ясно, что $m_1 + \dots + m_r = n$.) Введем случайную величину

$$\chi^2 = \sum_{i=1}^r \frac{(m_i - np_i)^2}{np_i}.$$

Тогда справедливо следующее утверждение: при $n \rightarrow \infty$ случайная величина χ^2 асимптотически подчиняется распределению χ^2 (хи-квадрат) с $(r - 1)$ степенями свободы.

Гипотеза. Теорему К. Пирсона можно использовать для проверки гипотезы о том, что вероятности p_1, \dots, p_r приняли определенные

значения p_1^o, \dots, p_r^o . Далее будем называть это гипотезой H :

$$H : p_1 = p_1^o, p_2 = p_2^o, \dots, p_r = p_r^o,$$

Рассмотрим статистику:

$$X^2 = \sum_{i=1}^r \frac{(m_i - np_i^o)^2}{np_i^o} = n \sum_{i=1}^r \left(\frac{m_i}{n} - p_i^o \right)^2 / p_i^o. \quad (10.6)$$

Определение. Статистика X^2 называется статистикой хи-квадрат Пирсона для простой гипотезы.

Ясно, что $\frac{X^2}{n}$ представляет собой квадрат некоего расстояния между двумя r -мерными векторами: вектором относительных частот $(\frac{m_1}{n}, \dots, \frac{m_r}{n})$ и вектором вероятностей (p_1^o, \dots, p_r^o) . От евклидова расстояния это расстояние отличается лишь тем, что разные координаты входят в него с разными весами.

Свойства. Обсудим поведение статистики X^2 в случае, когда гипотеза H верна, и в случае, когда H неверна. Если верна H , то асимптотическое поведение X^2 при $n \rightarrow \infty$ указывает теорема К. Пирсона. Чтобы понять, что происходит с (10.6), когда H неверна, заметим, что по закону больших чисел $m_i/n \rightarrow p_i$ при $n \rightarrow \infty$, для $i = 1, \dots, r$. Поэтому при $n \rightarrow \infty$:

$$\sum_{i=1}^r \left(\frac{m_i}{n} - p_i^o \right)^2 / p_i^o \rightarrow \sum_{i=1}^r (p_i - p_i^o)^2 / p_i^o.$$

Эта величина равна 0, только если $p_i = p_i^o$ для всех i . Поэтому если H неверна, то $X^2 \rightarrow \infty$ (при $n \rightarrow \infty$).

Правило проверки гипотезы. Из сказанного следует, что H должна быть отвергнута, если полученное в опыте значение X^2 слишком велико. Здесь, как всегда, слова «слишком велико» означают, что наблюдаемое значение X^2 превосходит критическое значение, которое в данном случае можно взять из таблиц распределения хи-квадрат. Иначе говоря, вероятность $P(\chi^2 \geq X^2)$ — малая величина и, следовательно, маловероятно случайно получить такое же, как в опыте, или еще большее расхождение между вектором частот и вектором вероятностей.

Предостережение. Асимптотический характер теоремы К. Пирсона, лежащий в основе этого правила, требует осторожности при его практическом использовании. На него можно полагаться только при больших n . Судить же о том, достаточно ли n велико, надо с учетом вероятностей p_1, \dots, p_r . Поэтому нельзя сказать, к примеру, что ста наблюдений будет достаточно, поскольку не только n должно быть велико, но и произведения np_1, \dots, np_r (ожидаемые частоты) тоже не должны быть малы. Поэтому проблема применимости аппроксимации χ^2 (непрерывное распределение) к статистике X^2 , распределение

которой дискретно, оказалась сложной. Совокупность теоретических и экспериментальных доводов привела к убеждению, что эта аппроксимация применима, если все ожидаемые частоты $np_i \geq 10$. Если число r (число различных исходов) возрастает, граница для np_i может быть снижена (до 5 или даже до 3, если r порядка нескольких десятков). Чтобы соблюсти эти требования, на практике порой приходится объединять несколько исходов, т.е. переходить к схеме Бернулли с меньшим r .

Другие применения критерия хи-квадрат Пирсона. Описанный способ для проверки согласия можно прилагать не только к испытаниям Бернулли, но и к произвольным выборкам. Предварительно их наблюдения надо превратить в испытания Бернулли путем группировки. Делают это так: пространство наблюдений разбивают на конечное число непересекающихся областей, а затем для каждой области подсчитывают наблюдаемую частоту и гипотетическую вероятность.

В данном случае к перечисленным ранее трудностям аппроксимации прибавляется еще одна — выбор разумного разбиения исходного пространства. При этом надо заботиться и о том, чтобы в целом правило проверки гипотезы об исходном распределении выборки было достаточно чувствительным к возможным альтернативам. Наконец, отметим, что статистические критерии, основанные на редукции к схеме Бернулли, как правило, не являются состоятельными против всех альтернатив. Так что такой метод проверки согласия имеет ограниченную ценность.

10.5. Критерии согласия для сложной гипотезы

Постановка задачи. Более трудной, но и более важной для приложений задачей является проверка гипотезы о том, что данная выборка подчиняется определенному параметрическому закону распределения, например нормальному закону. Параметры этого закона остаются неопределенными, так что эта гипотеза сложная.

Пусть x_1, \dots, x_n — выборка из распределения с функцией распределения $F(x, \theta)$. Здесь θ — неизвестный параметр, не обязательно скалярный. Обозначим его истинное значение через θ^o . Сейчас мы не можем сравнить выборочную функцию распределения $F_n(x)$ и теоретическую, поскольку эта последняя нам не вполне известна: в ее выражение $F(x, \theta^o)$ входит неопределенный параметр θ^o . Мы, однако, можем найти для θ^o приближенное значение, основываясь на выборке x_1, \dots, x_n . Для этого можно использовать разные методы оценивания (см. гл. 4), но наиболее ясные и в определенном смысле наилучшие результаты получаются, если использовать метод наибольшего правдоподобия.

Статистики. Итак, пусть $\hat{\theta}_n$ — оценка наибольшего правдоподобия по выборке x_1, \dots, x_n для неизвестного параметра θ распределения $F(x, \theta)$. Теперь для вычисления статистики Колмогорова вместо $F(x, \theta^o)$ мы можем использовать $F(x, \hat{\theta}_n)$ и ввести *модифицированную статистику Колмогорова*:

$$\hat{D}_n = \sup_x |F(x) - F(x, \hat{\theta}_n)|. \quad (10.7)$$

Аналогично *модифицированная статистика омега-квадрат* есть:

$$\hat{\omega}_n^2 = \int_{-\infty}^{+\infty} [F_n(x) - F(x, \hat{\theta}_n)]^2 dF(x, \hat{\theta}_n). \quad (10.8)$$

Свойства. Свойства статистик \hat{D}_n и $\hat{\omega}_n^2$ во многом повторяют отмеченные ранее свойства статистик D_n и ω_n^2 . В частности, $\sqrt{n}\hat{D}_n$ и $n\hat{\omega}_n^2$ неограниченно возрастают, если проверяемая гипотеза неверна. Поэтому эту гипотезу следует отвергнуть, если наблюдаемое значение $\sqrt{n}\hat{D}_n$ (или $n\hat{\omega}_n^2$, если применяется модифицированный критерий омега-квадрат) неправдоподобно велико, например, превосходит критическое значение, о котором будет сказано ниже.

Важно отметить, что статистика \hat{D}_n распределена *иначе*, чем D_n (10.1), а статистика $\hat{\omega}_n^2$ — *иначе*, чем ω_n^2 (10.4). Причина в том, что из-за подбора $\hat{\theta}_n$ по выборке функции $F_n(x)$ и $F(x, \hat{\theta}_n)$ (в случае, если гипотеза о типе распределения верна) оказываются *ближе* друг к другу, чем $F_n(x)$ и $F(x, \theta^o)$. Поэтому при справедливости гипотезы статистика \hat{D}_n , как правило, будет принимать существенно меньшие значения, чем D_n . Аналогично соотносятся $\hat{\omega}_n^2$ и ω_n^2 .

Таблицы. Поскольку статистики (10.7), (10.8) при справедливости гипотезы имеют иные распределения, чем статистики D_n и ω_n^2 , для их применения необходимы новые таблицы распределений или хотя бы таблицы критических значений. К сожалению, модифицированные статистики (10.7), (10.8) *не обладают* столь привлекательным свойством «свободы от распределения выборки», как их прототипы, поэтому для каждого параметрического семейства распределений нужны свои таблицы. Более того, распределения (10.7), (10.8) могут зависеть и от истинного значения неизвестного параметра (параметров). К счастью, для так называемых «масштабно-сдвиговых» семейств, к которым относятся нормальное, показательное и многие другие практически важные распределения, этого последнего осложнения не возникает.

Таблицы распределений статистик (10.7), (10.8) к настоящему моменту составлены для многих семейств (см., например, [49]). Большинство из них рассчитаны методом случайных испытаний (методом Монте-Карло). Автор большинства этих расчетов М. Стефенс (M. Stephens) за-

метил, что зависимость результатов от объема выборки резко уменьшается, если вместо \hat{D}_n , $\hat{\omega}_n^2$ использовать их несколько преобразованные варианты. Стефенс утверждает, что для этих форм зависимость от n практически перестает сказываться, начиная с $n = 5$. Ниже приводятся некоторые таблицы Стефенса.

Таблица 10.1

Модифицированные критерии для проверки нормальности, оба параметра неизвестны

Статистика	Модифицированная форма	Верхние процентные точки				
		0.15	0.10	0.05	0.025	0.01
\hat{D}_n :	$\hat{D}_n \left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}} \right)$	0.775	0.819	0.895	0.955	1.035
$\hat{\omega}_n^2$:	$\hat{\omega}_n^2 \left(1 + \frac{0.5}{n} \right)$	0.091	0.104	0.126	0.148	0.178

Таблица 10.2

Модифицированные критерии для проверки экспоненциальности, параметр неизвестен

Статистика	Модифицированная форма	Верхние процентные точки				
		0.15	0.10	0.05	0.025	0.01
\hat{D}_n :	$\left(\hat{D}_n - \frac{0.2}{n} \right) \cdot \left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}} \right)$	0.926	0.990	1.094	1.190	1.308
$\hat{\omega}_n^2$:	$\hat{\omega}_n^2 \left(1 + \frac{0.16}{n} \right)$	0.149	0.177	0.224	0.273	0.337

Приближенные формулы. Предельное (при $n \rightarrow \infty$) распределение $n\hat{\omega}_n^2$ известно, но вычисляется довольно сложно. Предельное распределение для $\sqrt{n}\hat{D}_n$ найти не удалось, есть лишь приближенные формулы для критических значений, основанные на асимптотических разложениях. Сравнение расчетов по этим формулам с упомянутыми ранее таблицами показало их хорошее согласие. Как уже говорилось, для каждого параметрического семейства критические значения надо рассчитывать особо. Например, для нормального закона, оба параметра которого оцениваются по выборке,

$$\lim_{n \rightarrow \infty} P \left\{ \sqrt{n} \sup \left| F_n(x) - \Phi \left(\frac{x - \bar{x}}{s} \right) \right| > z \right\} \simeq 2 \sqrt{\frac{2\pi}{\pi - 2}} \exp \left\{ -\frac{2\pi}{\pi - 2} z^2 \right\}$$

для больших $z > 0$ (т.е. для $z \rightarrow \infty$).

Если же математическое ожидание известно и равно, скажем, a , то по выборке приходится оценивать только дисперсию. В этом случае для больших $z > 0$

$$\lim_{n \rightarrow \infty} P \left\{ \sqrt{n} \sup \left| F_n(x) - \Phi \left(\frac{x - a}{s} \right) \right| > z \right\} \simeq \frac{2\sqrt{6}}{3} e^{-2z^2}.$$

Эти приближенные формулы дают хорошие результаты для малых вероятностей и больших объемов выборок, т.е. для вероятностей, начиная примерно с 0.20 (и меньше) и для объемов n , начиная примерно со 100 (и больше).

10.6. Критерий согласия хи-квадрат Фишера для сложной гипотезы

Для проверки сложных гипотез может быть использована и соответствующая модификация критерия хи-квадрат К. Пирсона. Главные заслуги здесь принадлежат Р. Фишеру. Приведем одну из его теорем (сохраняя обозначения из теоремы К. Пирсона). Близкая к этой теорема упоминалась в п. 9.2.

Теорема Фишера. Пусть n — число независимых повторений опыта, который может заканчиваться одним из r (r — произвольное натуральное число) элементарных исходов, скажем, A_1, \dots, A_r . Пусть вероятности этих элементарных исходов известны с точностью до некоторого неопределенного, скажем, k -мерного параметра $\theta = (\theta_1, \dots, \theta_k)$. Тогда эти вероятности являются функциями от θ : $P(A_i) = p_i(\theta)$. Мы будем предполагать, что функции $p_1(\theta), \dots, p_r(\theta)$ заданы, дифференцируемы, $\sum_{i=1}^r p_i(\theta) = 1$ для всякого θ , а параметр θ изменяется в ограниченной области пространства. Тогда при $n \rightarrow \infty$ статистика:

$$X^2 = \min_{\theta} \sum_{i=1}^r \frac{[m_i - np_i(\theta)]^2}{np_i(\theta)} \quad (10.9)$$

асимптотически распределена по закону χ^2 с $r-k-1$ степенями свободы.

Существует много вариантов этой теоремы. Например, такое же, как выше, предельное распределение имеет статистика

$$X^2 = \sum_{i=1}^r \frac{[m_i - np_i(\hat{\theta}_n)]^2}{np_i(\hat{\theta}_n)}, \quad (10.10)$$

где $\hat{\theta}_n$ — оценка наибольшего правдоподобия для параметра θ , найденная по частотам m_1, \dots, m_r . Поэтому значение (10.10) в дальнейшем можно использовать вместо (10.9). Далее, знаменатели np_i в (10.9) и (10.10) можно заменить на m_i , $i = 1, \dots, r$, и это не отразится на асимптотическом распределении X^2 . Есть и другие возможности. Много интересного об этом можно узнать в книге С. Рао [82].

Определение. Статистика X^2 из (10.9) (и ее варианты) называется статистикой хи-квадрат Фишера для сложной гипотезы.

Гипотеза и ее проверка. Статистику (10.9) (и ее варианты) можно использовать для проверки описанной выше сложной гипотезы о параметрическом виде вероятностей в схеме Бернулли:

$$H : P(A_1) = p_1(\theta), \dots, P(A_r) = p_r(\theta),$$

где $p_1(\cdot), \dots, p_r(\cdot)$ — заданы, а параметр θ изменяется в заданной ограниченной области. Это можно делать так же, как мы делали с помощью статистики X^2 в случае простой гипотезы. А именно, по наблюдаемым частотам m_1, \dots, m_r надо вычислить значение X^2 (10.9) либо (10.10) и затем сравнить его с критическими значениями распределения χ^2 с числом степеней свободы $(r - k - 1)$, либо вычислить $P(\chi^2 \geq X^2)$. Однако для использования аппроксимации хи-квадрат для распределения X^2 необходимо, чтобы число наблюдений было достаточно велико, и тем самым ожидаемые частоты $np_i(\hat{\theta})$ не были малыми (см. предостережение п. 10.4).

Другие применения. Как следует из формулировки теоремы, объект ее применения — испытания с конечным числом исходов. Чтобы использовать ее в условиях другого эксперимента — например, для проверки гипотезы о типе непрерывного или дискретного распределения с бесконечным (или конечным, но большим) числом исходов — этот эксперимент надо предварительно превратить в схему Бернулли. Раньше уже говорилось, как это делается обычно — путем разбиения выборочного пространства на непересекающиеся области. Параметрический (зависящий от параметра θ) закон распределения вероятностей во всем пространстве, соответствие которого нашей выборке мы хотим проверить, превращается при этом в параметрическое распределение вероятностей между выбранными r областями.

Понятно, что результат последующего применения критерия хи-квадрат (принять гипотезу, отвергнуть гипотезу) сильно зависит от описанного перехода. К этому следует добавить условие применимости распределения χ^2 как аппроксимации для распределения X^2 , которое требует, чтобы ожидаемые частоты были достаточно большими. (Условие на ожидаемые частоты часто приходится заменять требованием, чтобы не были малы наблюдаемые частоты m_1, \dots, m_r .) Становится ясно, что подготовка к применению критерия хи-квадрат в несвойственных ему условиях составляет деликатную и не всегда простую проблему. Возникает даже опасность невольной подгонки выбираемого разбиения к желательному результату. Поэтому, строго говоря, разбиение пространства на области должно идти вне зависимости от результатов случайного эксперимента, т.е. вне влияния подлежащей обработке выборки.

Проверка нормальности. Как же после всех этих предостережений можно применить теорему Фишера к проверке гипотезы о типе выборки? Обсудим это на примере нормального распределения, параметры которого (a, σ^2) неизвестны.

Итак, есть выборка x_1, \dots, x_n большого объема, проверить нормальность которой мы хотим с помощью (10.9) или (10.10) или их модификаций. Прежде всего мы должны разбить числовую прямую на r непесекающихся областей, а еще прежде — выбрать само число r . Сейчас существует убеждение (подкрепленное асимптотическими исследованиями), что против гладкой альтернативы лучше брать r небольшим — несколько единиц. Если же конкурируют с нормальным распределением все другие возможности, число r стоит взять таким большим, какое позволяет последующее использование аппроксимации хи-квадрат.

Допустим, что r уже выбрано, и можно переходить к разбиению пространства на области. При этом надо позаботиться о том, чтобы ожидаемые частоты этих областей были достаточно велики для того, чтобы для X^2 действовала аппроксимация χ^2 . Поскольку истинное распределение вероятностей неизвестно, приходится опираться на какую-либо его оценку. В данном примере — на оценку $\Phi\left(\frac{x-\bar{x}}{s}\right)$ истинной функции распределения $\Phi\left(\frac{xa}{\sigma}\right)$.

Чтобы не ломать бесплодно голову над вопросом, какими должны быть вероятности этих областей, а точнее в данном случае — их приближенные значения, возьмем их одинаковыми. Иными словами, в качестве границ интервалов используем решения уравнений

$$\frac{k}{r} = \Phi\left(\frac{x - \bar{x}}{s}\right), \quad k = 1, \dots, r - 1.$$

Заметим, что в качестве оценки функции распределения можно использовать и выборочную функцию распределения $F_n(x)$, и другие возможности. В этом случае границами интервалов разбиения будут служить выборочные квантили (порядковые статистики).

После того, как мы определили интервалы разбиения числовой прямой, подсчитываем частоты m_1, \dots, m_r , по которым будем вычислять потом статистику X^2 (10.9) или (10.10) или какую-либо эквивалентную. Следует подчеркнуть, что, согласно теореме Фишера, для вычисления участвующих в этих формулах вероятностей $p_i(\theta)$ следует использовать частоты m_1, \dots, m_r , и только их. Никакой другой информацией пользоваться нельзя! Нельзя, например, использовать \bar{x} , s^2 в качестве оценок a и σ^2 , по которым затем вычислять $p_i(\theta)$. Причина та, что естественные оценки \bar{x} , s^2 составлены по всей выборке, а должны быть — по частотам m_i .

Можно даже сказать, какие последствия повлечет за собой нарушение этого запрета. Статистика X^2 не будет (асимптотически) следовать распределению χ^2 с $r - 3$ степенями свободы: ее функция распределения пройдет несколько ниже. Не будет она следовать и распределению χ^2 с $r - 1$ степенями свободы (как было бы при точно известных параметрах). Ее функция распределения пройдет несколько выше. В качестве иллюстрации на рис. 10.2 приведем графики функций распределения хи-квадрат с 8, 10, 18 и 20 степенями свободы. Графики, соответствующие первым двум распределениям, выделяют область, в которой будет проходить график функции распределения X^2 при $r = 11$, если для вычисления $p_i(\theta)$ использовались оценки \bar{x} , s^2 . Последние два графика задают область нахождения функции распределения X^2 при $r = 21$.

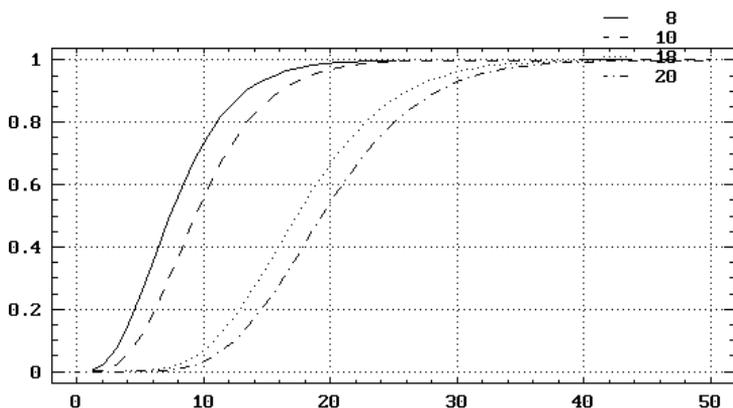


Рис. 10.2. Функции распределения хи-квадрат с 8, 10, 18 и 20 степенями свободы

При больших r относительное различие между квантилями распределений χ^2 с $(r - 3)$ и $(r - 1)$ степенями свободы невелико. Поэтому последствия такой ошибки не опасны. Но при малых r следует действовать «по теории».

Из-за всех этих сложностей, условий и оговорок можно сделать вывод, что для проверки гипотезы о нормальности выборки критерий Р. Фишера подходит плохо. Правильнее вместо этого использовать модификации критериев Колмогорова или омега-квадрат. (Начинать же проверку нормальности надо с глазомерного метода, использующего нормальную вероятностную бумагу, о чем подробно рассказывалось в гл. 5.) Но для многих распределений вероятностей (например — дискретных) другой возможности, чем обсуждаемый критерий хи-квадрат Фишера, просто нет.

10.7. Другие критерии согласия. Критерий согласия для пуассоновского распределения

Укажем, наконец, еще одну возможность для проверки согласия, которой тоже часто пользуются. Состоит она в том, что проверяют не исходную гипотезу целиком, а какое-либо ее следствие, которое считается важным. Скажем, для нормальной случайной величины ξ коэффициент асимметрии

$$\frac{M(\xi - M\xi)^3}{(D\xi)^{\frac{3}{2}}} \quad (10.11)$$

равен нулю. Поэтому коэффициент асимметрии выборки

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3} \quad (10.12)$$

тоже должен быть близок к нулю, если эта выборка — нормальная.

Чтобы судить о том, значимо ли отличается от нуля выборочное значение (10.12) и тем самым не нарушено ли обязательное для нормально-го закона соотношение (10.11), надо знать, как распределена статистика (10.12) при гипотезе. Для малых выборок исследование подобных вопросов возможно далеко не всегда и, во всяком случае, требует особого рассмотрения в каждом случае. Иное дело большие выборки.

Есть стандартная методика, которая позволяет справиться с этой задачей. Покажем ее действие на другом примере, поскольку о нормальном законе говорилось уже слишком много. Посмотрим, как можно проверить согласие выборки с распределением Пуассона (см. п. 2.2). Для случайной величины ξ , распределенной по Пуассону,

$$D\xi/M\xi = 1, \quad (10.13)$$

так как для распределения Пуассона $D\xi = M\xi = \lambda$, где λ — параметр распределения. Поэтому если выборка x_1, \dots, x_n извлечена из пуассоновской генеральной совокупности, то отношение

$$S^2/\bar{x}, \quad \text{где } S^2 = \sum_{i=1}^n (x_i - \bar{x})^2/n \quad (10.14)$$

должно быть близким к 1. Ниже пойдет речь о том, как это проверить.

Предостережение. Но сначала одно замечание общего характера: такие проверки никак не могут *доказать* соответствия выборки теоретическому закону даже при неограниченном возрастании числа наблюдений. Причина в том, что соотношения типа (10.11) и (10.13)

не являются характеристическими: даже если (10.11) справедливо, оно не означает, что ξ непременно распределено нормально. Это свойство необходимо для нормальности распределения, но недостаточно. То же самое можно сказать о (10.13): это необходимое, но недостаточное условие для того, чтобы распределение было пуассоновским. После этого обсуждения обратимся к изучению свойств статистики (10.14). Объем выборки n будем считать большим.

Распределение статистики критерия. Воспользуемся тем, что при $n \rightarrow \infty$ случайные величины $S^2 - D\xi$ и $\bar{x} - M\xi$ стремятся к 0 (закон больших чисел). Поэтому для пуассоновской выборки:

$$\frac{S^2}{\bar{x}} = \frac{D\xi + (S^2 - D\xi)}{M\xi + (\bar{x} - M\xi)} = \frac{D\xi}{M\xi} \frac{1 + \frac{S^2 - D\xi}{D\xi}}{1 + \frac{\bar{x} - M\xi}{M\xi}} = \left(1 + \frac{S^2 - D\xi}{D\xi}\right) \left(1 - \frac{\bar{x} - M\xi}{M\xi} + \dots\right).$$

Многоточие заменяет случайную величину, убывающую как n^{-1} . Раскрыв скобки, получаем, что:

$$\frac{S^2}{\bar{x}} = 1 + \frac{S^2 - D\xi}{D\xi} - \frac{\bar{x} - M\xi}{M\xi} + \dots = 1 + \frac{1}{\lambda} (S^2 - \bar{x}) + \dots$$

Исследуем при $n \rightarrow \infty$ поведение выражения $\frac{S^2 - \bar{x}}{\lambda}$, главной случайной составляющей дроби S^2/\bar{x} . Без ущерба для точности вывода вместо S^2 можно взять случайную величину:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \lambda)^2.$$

Тогда вместо $S^2 - \bar{x}$ появляется:

$$\frac{1}{n} \sum_{i=1}^n [(x_i - \lambda)^2 - x_i].$$

В силу центральной предельной теоремы эта сумма независимых и одинаково распределенных случайных величин распределена приблизительно нормально, с математическим ожиданием: $M[(\xi - \lambda)^2 - \xi] = 0$ и дисперсией $\frac{1}{n} D[(\xi - \lambda)^2 - \xi] = \frac{1}{n} M[(\xi - \lambda)^2 - \xi]^2$. Для вычисления последнего выражения надо знать, что четвертый и третий центральные моменты пуассоновского распределения равны соответственно

$$M(\xi - \lambda)^4 = 3\lambda^2 + \lambda, \quad M(\xi - \lambda)^3 = \lambda.$$

После этого подсчет дает, что $D[(\xi - \lambda)^2 - \xi] = 2\lambda^2$. Следовательно, статистика (10.14) S^2/\bar{x} распределена приблизительно по закону $N(1, 2\lambda^2/n)$.

Критерий проверки гипотезы. Зная распределение статистики (10.14) в случае справедливости нулевой гипотезы о принадлежности выборки к распределению Пуассона, можно указать пределы, в которые с вероятностью приблизительно, скажем, 0.99 должно попадать отношение S^2/\bar{x} в случае справедливости гипотезы:

$$\left| \sqrt{n} \frac{S^2/\bar{x} - 1}{\lambda\sqrt{2}} \right| < u_{0.995}, \quad (10.15)$$

где u_α обозначает квантиль уровня α стандартного нормального распределения.

Если мы хотим использовать это соотношение для практической проверки гипотезы о пуассоновском распределении выборки, надо заменить неизвестное значение λ его оценкой по выборке. Как отмечалось ранее в гл. 4, для больших выборок наилучшей является оценка наибольшего правдоподобия, которая для пуассоновского распределения равна \bar{x} . Следовательно, надо проверить по выборке, выполняется ли соотношение:

$$\left| \sqrt{n} \frac{S^2/\bar{x} - 1}{\bar{x}\sqrt{2}} \right| < u_{0.995} = 2.58, \text{ т.е. } \left| \sqrt{n} \frac{S^2 - \bar{x}}{(\bar{x})^2} \right| < 3.64. \quad (10.16)$$

Если это неравенство не выполняется, гипотезу о том, что выборка извлечена из распределения Пуассона, следует отвергнуть на уровне значимости (примерно) 0.01. Понятно, что при другом уровне значимости в правой части (10.15) будет стоять другая квантиль и поэтому правая часть (10.16) тоже будет другой.

Обсуждение и обобщения. Поскольку этот способ проверки приближенный, то чем большего объема окажется выборка в нашем распоряжении, тем точнее будет соблюден номинальный уровень значимости. К сожалению, трудно сказать определенно, начиная с какого n результат такой проверки заслуживает доверия; по-видимому, для этого требуется не менее сотни наблюдений.

Подобным образом может быть проверено любое свойство теоретического распределения, если только мы располагаем достаточно большой выборкой. Главное здесь — выбор самого свойства. Эта характеристика распределения должна быть существенна для дальнейшего. Как правило, знания о типе распределения нужны для того, чтобы на их основе сделать по выборочным данным те или иные выводы. Нередко оказывается, что для справедливости этих выводов особенно важны

лишь некоторые свойства теоретического закона распределения. Именно эти свойства и надо в первую очередь проверить.

Например, при применении критерия Стьюдента к выборкам, несколько отличающимся от нормальных, результаты будут близки к правильным (для больших выборок), если коэффициенты асимметрии и эксцесса такие же, как у нормального закона. Поэтому в проверку на нормальность в этом случае надо включить вычисление выборочных коэффициентов асимметрии и эксцесса и их значимости. Критерии проверки нормальности, опирающиеся на эти коэффициенты, подробно изложены в [19].

10.8. Критерии согласия в пакете SPSS

В этом параграфе мы покажем, как процедуры проверки согласия реализованы в пакете SPSS. Мы будем интересоваться в первую очередь типичными ситуациями, но не обойдем и те тонкости, которые отмечали в этой главе. Как будет видно, порой они бывают существенны для правильных статистических выводов. В разбираемых ниже примерах особое внимание будет обращено, во-первых, на чувствительность поведения статистик Колмогорова и хи-квадрат к «грубым» ошибкам в наблюдениях и, во-вторых, на важность правильного определения минимальных уровней значимости этих критериев для сложных гипотез.

Пример 10.1к. Проверим согласие распределения выборки диаметров головок заклепок (табл. 1.1) с нормальным распределением, используя критерий Колмогорова—Смирнова. Проведем аналогичные расчеты для «цензурированной» выборки.

Подготовка данных. Данные этого примера уже рассматривались в примерах 1.1к и 5.1к. Пусть они находятся в переменной **d** редактора данных пакета (см. рис. 1.13). В пакете нет необходимости создавать отдельную переменную для «цензурированных» данных. Достаточно задать фильтр для отбора наблюдений из исходной переменной. Для этого в меню **Data** (данные) редактора пакета следует выбрать процедуру **Select Cases** (выбор наблюдений). В окне ввода данных и параметров этой процедуры задать режим **If condition is satisfied** (если выполнено условие) и задать само условие выбора, скажем, $d < 14$. Процедура **Select Cases** сформирует в редакторе данных специальную служебную переменную **filter_\$**, в которой 1 соответствует отобранному данным, а 0 — отброшенным. Такой порядок очень удобен для формирования различных подвыборок и исключения резко выделяющихся значений.

Выбор процедуры. Для решения задачи следует выбрать процедуру **Explore** блока **Descriptive Statistics** (см. п. 1.9). Эта процедура включает критерий Колмогорова—Смирнова для нормальных выборок. При этом в ней минимальный уровень значимости критерия рассчитывается с поправкой на сложную гипотезу (поправка Лильефорса).

Комментарии. 1. В пакете существует процедура **1-Sample K-S** (одновыборочный критерий Колмогорова—Смирнова). Ее работа описана в примере 4.1к п. 4.7. Эта процедура предназначена для проверки только простой гипотезы, т.е. для случая, когда гипотетическое распределение известно нам полностью. К сожалению, процедура не позволяет задать параметры гипотетического распределения заранее. Она сама вычисляет значения этих параметров по выборке, т.е. на самом деле пытается проверить сложную гипотезу. Поэтому рассчитываемые уровни значимости этой процедуры ошибочны, а сама она не решает задачи, объявленной в документации пакета.

В этом можно убедиться на разбираемом нами примере. На рис. 10.3 в последней строке таблицы приведен минимальный уровень значимости статистики Колмогорова—Смирнова, рассчитанной для нецензурированного массива данных. Он оказался равен **0.255**. Это значит, что гипотезу о нормальном характере следует принять. Однако эту выборку нормальной считать нельзя (из-за одного грубо ошибочного значения). Причина этого ошибочного заключения ясна: статистика Колмогорова—Смирнова в том случае, когда параметры гипотетического распределения оцениваются по выборке, имеет другое распределение, чем при простой гипотезе (см. п. 10.5).

Мы не рекомендуем пользоваться этой процедурой для проверки согласия. (Процедуру можно использовать для оценки параметров четырех законов распределений, как это показано в примере 4.1к.)

2. Укажем ситуацию, когда результаты расчетов этой процедуры все же могут быть использованы для статистических выводов о согласии. Заметим, что уровень значимости статистики Колмогорова для сложной гипотезы всегда *меньше* уровня значимости этой статистики для простой гипотезы. Таким образом, если полученный уровень значимости для простой гипотезы мал, то уровень значимости для сложной гипотезы еще меньше, и эту гипотезу следует отвергать. В других случаях надо обращаться к таблицам соответствующих процентных точек или использовать возможности процедуры **Explore** для проверки нормальности, как это показано ниже.

Заполнение полей ввода данных. Для выполнения этого критерия следует в окне ввода данных процедуры **Explore** (рис. 10.4) нажать кнопку **Plot**. На экране появится окно настройки графических возможностей процедуры (рис. 10.5). В этом окне следует отметить опцию **Normality plots with test**. Вернувшись в окно ввода параметров процедуры **Explore**, запустить ее на выполнение, нажав кнопку **OK**.

Результаты. Требуемые результаты работы заданной опции процедуры приведены на рис. 10.6.

One-Sample Kolmogorov-Smirnov Test

		D
N		200
Normal Parameters ^a , Mean		13,4215
	Std. Deviation	,1344
Most Extreme Differences	Absolute	,072
	Positive	,072
	Negative	-,061
Kolmogorov-Smirnov Z		1,015
Asymp. Sig. (2-tailed)		,255

- a. Test distribution is Normal.
b. Calculated from data.

Рис. 10.3. Пакет SPSS. Результаты работы процедуры «1-Sample K-S» для «нецензурированного» массива данных

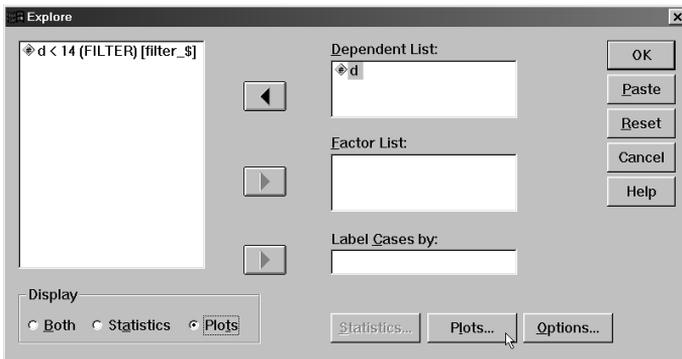


Рис. 10.4. Пакет SPSS. Окно ввода данных и параметров процедуры «Explore»

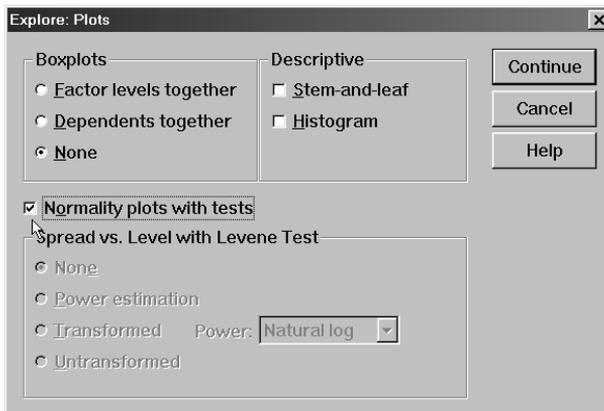


Рис. 10.5. Пакет SPSS. Окно настройки графического вывода процедуры «Explore»

Tests of Normality

	Kolmogorov-Smirnov ^a		
	Statistic	df	Sig.
D	,072	200	,014

a. Lilliefors Significance Correction

Рис. 10.6. Пакет SPSS. Результаты работы процедуры «Explore» для «цензурированного» массива данных

Они включают значение статистики Колмогорова—Смирнова, объем выборки (обозначенный в таблице **df**) и минимальный уровень значимости **Sig.** Кроме того, процедура строит график эмпирической функции распределения выборки на нормальной вероятностной бумаге **Normal Probability Plot** (см. п. 5.2).

Полученный уровень значимости критерия Колмогорова—Смирнова с поправкой Лильефорса — **0.014** — говорит, что гипотезу о принадлежности данных к нормальному семейству распределений следует отвергнуть.

Причиной отвержения гипотезы о нормальном характере (по всей выборке) данных, как это будет показано ниже, явилось одно «грубое» (аномальное) наблюдение. Механизм влияния этого наблюдения на вычисляемые характеристики критериев следующий. «Грубое» наблюдение заметно исказило значение оценки максимального правдоподобия дисперсии выборки и тем самым повлияло на значения подобранной, согласно гипотезе, функции нормального распределения $F(x, \hat{\theta})$, где вектор $\theta = (\bar{x}, s^2)$.

Результаты этой процедуры для цензурированных данных приведены на рис. 10.7.

Tests of Normality

	Kolmogorov-Smirnov ^a		
	Statistic	df	Sig.
D	,046	199	,200*

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Рис. 10.7. Пакет SPSS. Результаты работы процедуры «Explore» для «цензурированного» массива данных

В случае, когда минимальный уровень значимости достаточно велик, эта процедура вычисляет не сам уровень значимости, а его оценку снизу. Информация об этом указывается в сноске под таблицей. Оценка снизу для минимального уровня значимости критерия для цензурированных

данных достаточно велика (0.200) и не дает веских оснований отвергнуть нулевую гипотезу.

Как видно из полученных результатов, данные без резко выделяющегося значения («цензурированные») не противоречат гипотезе о нормальности распределения.

Комментарий. Заметим, что эта оценка снизу минимального уровня значимости далека от истинного значения. Но эта грубость оценки не влияет на итоговый вывод, так как вычисление нижней границы уровня значимости происходит в области значений, где заведомо нет оснований отвергать гипотезу.

Пример 10.2к. Проверим согласие распределения выборки диаметров головок заклепок (табл. 1.1) с нормальным распределением, используя критерий хи-квадрат. Проведем аналогичные расчеты для «цензурированной» выборки.

В пакете представлена процедура, реализующая критерий хи-квадрат для конечных дискретных распределений данных. Ее запуск осуществляется из пункта **Chi-Square** блока **Nonparametric Tests** меню **Analyze** (см. рис. 3.3). Обработать с помощью этой процедуры данные, имеющие непрерывное распределение, напрямую нельзя. Решение подобной задачи в SPSS возможно, но требует навыков работы с этим пакетом и хорошей статистической квалификации пользователя. Поэтому, не проводя детального разбора, укажем лишь основные этапы решения задачи. На первом этапе необходимо провести перекодировку исходных данных с целью получения таблицы частот. Для этого надо задать разбиение диапазона значений выборки на непересекающиеся интервалы (интервалы группировки). Как это лучше сделать, говорилось в п. 10.6. Затем воспользоваться процедурой **Recode** из меню **Transform** редактора пакета. С помощью этой процедуры можно создать новую переменную, в которой каждое значение исходной выборки будет заменено номером интервала группировки. Именно эту переменную следует вводить для обработки в процедуру **Chi-Square**, но предварительно необходимо для каждого из полученных интервалов группировки определить гипотетическую вероятность. Это отдельная задача, которая может решаться по-разному. Скажем, можно выбрать интервалы группировки так, чтобы вероятность попадания в каждый из них была одинаковой. Если же одинаковыми будут сами длины интервалов группировки, то придется рассчитывать гипотетические вероятности с помощью функций распределения. Полученные значения гипотетических вероятностей следует также ввести в процедуру **Chi-Square** при заполнении полей окна ввода данных и параметров этой процедуры. Результатом работы процедуры будут значения статистики хи-квадрат, число степеней свободы и минимальный уровень значимости.

Дополнительная литература

1. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика. Основы эконометрики. Т. 1. Теория вероятностей и прикладная статистика. — М.: Юнити-Дана, 2001. — 656 с.
2. *Бикел П., Доксум К.* Математическая статистика. — М.: Финансы и статистика, 1983. Вып. 1. — 280 с.; Вып. 2 — 254 с.
3. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. — М.: Наука, 1983. — 416 с.
4. *Боровков А.А.* Математическая статистика. — Новосибирск: Наука, 1997. — 772 с.
5. *Кобзарь А.И.* Прикладная математическая статистика. Для инженеров и научных работников. — М.: ФИЗМАТЛИТ, 2006. — 816 с.
6. Справочник по прикладной статистике: в 2 т.; под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. — М.: Финансы и статистика, 1989, 1990.

Выборочные обследования

11.1. Введение

Выборочные обследования и опросы широко применяются для изучения различных сторон жизни общества. В России с недавних пор они стали частью политической жизни. Накануне очередных выборов мы слышим об их результатах едва ли не ежедневно. Широкое распространение выборочных опросов, и не только политических, в ведущих странах мира относится к середине XX века. Выборочные обследования дополняют, уточняют и во многом заменяют сплошные переписи или опросы. В СССР упор делался на сплошной сбор статистических данных, а выборочным обследованиям в экономике отводилась второстепенная роль. Социологические и рыночные исследования (изучение потребления) не поощрялись и были редки.

В последнее десятилетие, уже в новой России, быстро формируется спрос на выборочные исследования, и не только с политическими целями. Заказчиками выборочных исследований выступают компании, продвигающие на рынок свои товары и услуги. Шире стали практиковаться выборочные исследования и органами государственной власти.

Укажем преимущества выборочных исследований по сравнению со сплошными обследованиями целевых групп и совокупностей. В первую очередь это оперативность получения результатов. Затем — относительно малая стоимость исследования в целом. Лучшая подготовка персонала, проводящего исследование, позволяет получить более достоверные исходные данные, а поэтому и более правильные выводы. При правильной организации выборки выборочные обследования дают высокую точность результатов. Представление о точности полученных результатов (оценку точности) можно получить в ходе обследования, основываясь на результатах выбора. Во многих случаях выборочный метод является просто единственным способом получения информации об обследуемой совокупности.

11.2. Выборки. Простой случайный выбор

При обследовании нас обычно интересуют свойства какой-либо четко очерченной совокупности однородных объектов. Эту совокупность мы будем называть *генеральной совокупностью*. При изучении генеральной совокупности выборочным методом из нее выделяют некоторую часть для дальнейшего сплошного обследования. Эту выделенную (выбранную) часть называют *выборкой*. Выборка должна быть как бы уменьшенной копией генеральной совокупности, представлять ее правильно, без искажений. Такую выборку называют *репрезентативной* (представительной). Главная опасность для выборочного метода — сформировать нерепрезентативную выборку, искаженно, неправильно представляющую генеральную совокупность. Такие выборки часто называют *смещенными* — смысл этого названия станет ясен чуть позже.

Есть много способов формирования выборки. Но только один из них несомненно обеспечивает ее репрезентативность: это случайный выбор. Именно о разных видах и формах случайного выбора и их свойствах мы будем говорить далее. К достоинствам случайного выбора добавим, что он позволяет контролировать и точность полученных с его помощью выводов. Правда, и репрезентативность, и точность при случайном выборе имеют статистический характер.

Реализация упомянутых возможностей и достоинств выборочных методов во многом определяется тщательным планированием всех этапов обследования и точным исполнением намеченного плана. «Результаты исследования никогда не могут быть лучше, чем план этого исследования», — отмечает известный американский специалист в области выборочных исследований Р. Джессен [38].

Основой выборочного метода является идея *простого случайного выбора*. Случайный выбор одного элемента из конечной генеральной совокупности называется простым, если все элементы генеральной совокупности имеют равные вероятности быть выбранными. Случайный выбор предписанного числа n элементов называют простым, если каждое множество, состоящее из n элементов генеральной совокупности, имеет равную с другими вероятность быть выбранными. Простой случайный выбор заданного количества n элементов можно произвести последовательно с помощью простого случайного выбора, формируя выборку из n элементов. (Выборку объема n , как часто говорят.)

Репрезентативность при случайном выборе. Поясним, в каком смысле простой случайный выбор является репрезентативным, избрав

для обсуждения одну из простейших задач, которую приходится решать в ходе выборочных обследований. Предположим, что образующие генеральную совокупность объекты могут обладать или не обладать некоторым определенным свойством. Обозначим его через A . В качестве подобного свойства в маркетинговых обследованиях часто выступает обладание той или иной продукцией (автомобилем, сотовым телефоном и т.д.), использование определенных косметических или гигиенических средств и т.п. В электоральных исследованиях рассматриваемое свойство может означать поддержку определенной партии или кандидата, определенной идеи или политической/экономической программы и т.д. В ходе обследования мы хотим оценить, какую долю генеральной совокупности составляют объекты со свойством A . Обозначим эту неизвестную нам долю через θ , $0 < \theta < 1$.

Ради математической простоты предположим, что численности интересующих нас генеральных совокупностей велики. (Нельзя указать определенную границу, но пусть это будут тысячи, а лучше десятки тысяч элементов или более.) Напротив, пусть объемы выборок будут малы по сравнению с численностью генеральных совокупностей. В этих условиях можно считать, что при последовательном формировании выборки объема n с помощью простого случайного выбора вероятность выбора объекта со свойством A — одна и та же на каждом шаге отбора. Эта вероятность равна доле θ всех объектов со свойством A в генеральной совокупности. В этих условиях свойства простого случайного выбора описывает вероятностная схема испытаний Бернулли.

Рассмотрим полученную выборку объема n . Обозначим через X число элементов этой выборки, которые обладают свойством A . При случайном выборе величина X тоже случайна. В качестве оценки неизвестной доли θ объектов в генеральной совокупности, обладающих свойством A , естественно взять их долю в выборке, т.е. величину X/n . Она, в силу случайности выборки, может быть как больше истинного значения θ , так и меньше его. Однако в среднем значение оценки X/n равно θ . Точный смысл этого утверждения заключается в следующем: при любых n и θ

$$M(X/n) = \theta. \quad (11.1)$$

Это равенство — одно из свойств биномиального распределения, которому подчинена случайная величина X .

Напомним (см. п. 4.5), что оценки, в среднем совпадающие с истинным значением той характеристики, приближенным выражением которой они служат, называются несмещенными. *Смещением* оценки называют разность между ее математическим ожиданием и истинным

значением оцениваемой характеристики. (Для несмещенных оценок смещения равны нулю.) Выборки называют смещенными, когда они приводят к смещению оценок. Смещение выборок — основной источник ошибок для выборочного метода. Эти смещения возникают из-за трудности осуществить простой случайный выбор.

С ростом объема простой случайной выборки доля в ней элементов со свойством A приближается к θ , так что при больших n

$$X/n \simeq \theta. \quad (11.2)$$

В том, что X/n близко к θ , можно убедиться различными способами. Проще всего выяснить средний квадрат их разности:

$$M(X/n - \theta)^2 = \frac{\theta(1 - \theta)}{n}. \quad (11.3)$$

Очевидно, что при большом n эта величина мала. (Поскольку X/n — несмещенная оценка θ , упомянутый средний квадрат разности в (11.3) — это дисперсия случайной величины X/n .)

11.3. Точность выборочной оценки

Точнее о характере приближенного равенства (11.2) можно судить с помощью доверительных интервалов. О доверительных интервалах как наиболее выразительной количественной характеристике для точности оценки нам уже приходилось говорить и ранее (см., в частности, п. 4.5, а также п. 5.3 и 6.6). Контролировать точность приближения (11.2) можно по самой выборке, зная лишь, каковы n и X . Мы расскажем о том, как это можно сделать для больших n и для значений θ , не слишком близких к 0 или 1. Эти условия характерны для большинства выборочных обследований. Но предварительно нам придется рассказать об одной теореме, которая имела важное значение для развития теории вероятностей.

Теорема Муавра—Лапласа. Рассмотрим схему испытаний Бернулли: независимые испытания с двумя исходами. Один из исходов обычно называют успехом, другой — неудачей. Вероятность успеха одинакова во всех испытаниях. Число испытаний (назначаемое заранее) обозначим через n , число успехов в них — через X . В нашей задаче n — объем выборки, успех — появление элемента со свойством A , X — число элементов со свойством A среди выбранных n , вероятность успеха — это θ (доля объектов со свойством A в генеральной совокупности). Как уже отмечалось, число n много меньше, чем численность генеральной совокупности.

Распределение вероятностей случайной величины X задает так называемая формула Бернулли:

$$P(X = k) = C_n^k \theta^k (1 - \theta)^{n-k} \quad \text{для } k = 0, 1, \dots, n.$$

Отсюда для любых целых a и b , где $0 \leq a \leq b \leq n$, получаем, что

$$P(a \leq X \leq b) = \sum_{k=a}^b C_n^k \theta^k (1 - \theta)^{n-k}.$$

В сборниках статистических таблиц можно найти значения как отдельных вероятностей $P(X = k)$, так и их накопленных сумм $P(X \leq m)$. Эти таблицы (они бывают различной степени подробности и полноты) содержат указанные вероятности для ряда значений θ и n . (Более подробное описание некоторых таких таблиц дано в п. 2.1.)

На практике нередко встречаются задачи, число испытаний в которых превосходит пределы имеющихся таблиц. В таких случаях для вычислений надо использовать приближенные формулы. Их точность неограниченно улучшается с ростом n . Мы не станем приводить формулы для $P(X = k)$, так как не собираемся ими пользоваться. Обратимся сразу к функции распределения случайной величины X :

$$F(x, n, \theta) = P(X \leq x),$$

где x — действительная переменная. Математики доказали теоретически, что при неограниченном росте n

$$P\left(\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} \leq x\right) \rightarrow \Phi(x),$$

где $\Phi(x)$ — функция Лапласа (см. п. 2.4). Этот важный результат известен как теорема Муавра—Лапласа.

Практически в этом можно убедиться следующим образом. Сравнить графики функций $y = F(x, n, \theta)$ и $y = \Phi\left(\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}\right)$ для некоторых разных n и θ . Построить их можно по-разному, например на нормальной вероятностной бумаге (см. п. 5.2). График функции $y = \Phi\left(\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}\right)$ на этой бумаге — прямая линия. График $y = F(x, n, \theta)$ на нормальной вероятностной бумаге выглядит как лестница со ступенями почти постоянной высоты и ширины. Упомянутая прямая пересекает эти ступени почти посередине. С ростом n эти графики сближаются. Характер этого сближения следующий:

- 1) функция распределения случайной величины X (число успехов в n испытаниях Бернулли) при увеличении n становится все более

похожей на функцию нормального распределения $N(n\theta, n\theta(1-\theta))$;

- 2) при данном n сходство тем больше, чем ближе θ к значению $\theta = 0.5$. (При малом $n\theta(1-\theta)$ нормальное приближение для биномиального распределения действует плохо.);
- 3) при вычислении $P(X \leq m)$ (для целых значений m) можно пользоваться приближением

$$P(X \leq m) \approx \Phi\left(\frac{m - n\theta}{\sqrt{n\theta(1-\theta)}}\right),$$

но более точный результат получается, если в правой части m увеличить на 0.5:

$$P(X \leq m) \approx \Phi\left(\frac{m + 0.5 - n\theta}{\sqrt{n\theta(1-\theta)}}\right). \quad (11.4)$$

Последнюю формулу называют нормальным приближением биномиального распределения с «поправкой на непрерывность» (К подобному приему часто приходится прибегать при использовании непрерывного закона распределения для приближенной замены им дискретного распределения.);

- 4) из формулы (11.4) следует, что для целых значений m

$$P(X \geq m) \approx 1 - \Phi\left(\frac{m - 0.5 - n\theta}{\sqrt{n\theta(1-\theta)}}\right).$$

Соответственно, для целых значений a, b (где $0 \leq a \leq b \leq n$)

$$P(a \leq X \leq b) \approx \Phi\left(\frac{b + 0.5 - n\theta}{\sqrt{n\theta(1-\theta)}}\right) - \Phi\left(\frac{a - 0.5 - n\theta}{\sqrt{n\theta(1-\theta)}}\right).$$

Традиционно эти результаты формулируют в виде предельной теоремы, называемой теоремой Муавра—Лапласа:

Для фиксированных x_1 и x_2 , где $x_1 < x_2$, при $n \rightarrow \infty$ справедливо соотношение

$$P(n\theta + x_1\sqrt{n\theta(1-\theta)} \leq X \leq n\theta + x_2\sqrt{n\theta(1-\theta)}) \rightarrow \Phi(x_2) - \Phi(x_1),$$

или

$$P\left(x_1 \leq \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} \leq x_2\right) \rightarrow \Phi(x_2) - \Phi(x_1). \quad (11.5)$$

Помимо большой исторической важности (эта теорема была исторически первой центральной предельной теоремой), эта теорема оправдывает использование правой части (11.5) как приближения для левой части

(11.5) при больших n . Как отмечалось выше, поправки на непрерывность улучшают точность приближения. Впрочем, для действительно больших n (порядка сотен) удовлетворительную точность приближения можно получить и без них.

Доверительные интервалы. Желая оценить близость X/n к неизвестному θ , естественно рассмотреть их разность

$$\frac{X}{n} - \theta.$$

К сожалению, говорить о малости этой величины (по модулю) мы можем только с некоторой вероятностью, так как в силу случайности оценка X/n может отклоняться от θ . (Примером такого редкого, но не невозможного события является, скажем, выпадение десяти гербов при десяти бросаниях правильной монеты. Доля гербов в такой выборке составит 1, хотя вероятность выпадения герба для правильной монеты равна 0.5. Впрочем, вероятность этого события меньше 0.001.)

Приближенные, но достаточно точные для практики доверительные интервалы для θ можно построить по X и n , опираясь на теорему Муавра—Лапласа. В силу этой теоремы случайная величина $\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$ из п. 11.6, которую мы запишем сейчас в виде

$$\left(\frac{X}{n} - \theta\right) \times \sqrt{\frac{n}{\theta(1-\theta)}}, \quad (11.6)$$

с достаточной точностью следует стандартному нормальному распределению $N(0, 1)$. Это позволит нам действовать примерно по тому же плану, что в п. 5.3.

Выберем близкое к единице значение доверительной вероятности. Обозначим ее через $1 - 2\alpha$, где α — число малое. Обычно одновременно вычисляют сразу несколько доверительных интервалов; следовательно, действуют с несколькими значениями доверительных вероятностей. Так, традиционны для $1 - 2\alpha$ значения 0.90, 0.95 и 0.99. (Значения α при этом суть 0.05, 0.025 и 0.005.)

С помощью таблиц или специальных процедур в статистических пакетах найдем $(1 - \alpha)$ — квантили стандартного нормального распределения. Как и ранее в п. 5.3, обозначим их через $z_{1-\alpha}$. Если через η обозначить на минуту стандартную нормальную случайную величину, то с ее помощью соотношения между вероятностью $1 - 2\alpha$ и квантилью $z_{1-\alpha}$ можно выразить так:

$$P(|\eta| < z_{1-\alpha}) = 1 - 2\alpha. \quad (11.7)$$

Далее в это равенство вместо η подставим случайную величину (11.6). При такой замене равенство (11.7) становится не вполне точным. Для объемов выборок n , с которыми мы обычно имеем дело в выборочных опросах и обследованиях, упомянутой неточностью вполне можно пренебречь. Все же ради аккуратности поставим знак приближенного равенства:

$$P\left(\left|\frac{X}{n} - \theta\right| \times \sqrt{\frac{n}{\theta(1-\theta)}} < z_{1-\alpha}\right) \approx 1 - 2\alpha.$$

Из этого заключаем, что с (приближенной) вероятностью $1 - 2\alpha$ выполняется неравенство

$$\left|\frac{X}{n} - \theta\right| < z_{1-\alpha} \sqrt{\frac{\theta(1-\theta)}{n}}. \quad (11.8)$$

При разных α эти неравенства говорят нам о том, как далеко выборочная оценка X/n может из-за случайностей выбора отступить от интересующего нас числа θ .

Непосредственно воспользоваться неравенством (11.8) нельзя, так как его правая часть содержит неизвестную нам величину $v = \theta(1-\theta)$. Есть несколько способов обойти это неудобство.

Можно, например, превратить (11.8) в квадратное неравенство

$$\left(\frac{X}{n} - \theta\right)^2 < z_{1-\alpha}^2 \frac{\theta(1-\theta)}{n},$$

которое затем решить относительно θ .

Можно воспользоваться тем, что $\theta(1-\theta) \leq \frac{1}{4}$ для θ из интервала $(0, 1)$. Если мы теперь заменим $v = \theta(1-\theta)$ в правой части (11.8) его максимально возможным значением $\frac{1}{4}$, мы получим для $\left|\frac{X}{n} - \theta\right|$ оценку

$$\left|\frac{X}{n} - \theta\right| < \frac{z_{1-\alpha}}{2\sqrt{n}}. \quad (11.9)$$

Самый же простой (и достаточно надежный) способ состоит в том, чтобы заменить неизвестное $v = \theta(1-\theta)$ его выборочной оценкой $\hat{v} = \frac{X}{n} \left(1 - \frac{X}{n}\right)$. В этом случае приближенный $(1 - 2\alpha)$ - доверительный интервал для θ имеет вид:

$$\frac{X}{n} - \frac{z_{1-\alpha}}{\sqrt{n}} \times \sqrt{\frac{X}{n} \left(1 - \frac{X}{n}\right)} < \theta < \frac{X}{n} + \frac{z_{1-\alpha}}{\sqrt{n}} \times \sqrt{\frac{X}{n} \left(1 - \frac{X}{n}\right)}.$$

Что влияет на точность оценки? Выражение (11.8) позволяет указать факторы, непосредственно влияющие на точность оценки:

- 1) Чем больше число наблюдений в выборке, тем точнее выборочная оценка.
- 2) Для увеличения точности оценки вдвое необходимо увеличить объем выборки вчетверо, так как точность оценки обратно пропорциональна квадратному корню из объема выборки.
- 3) Чем больше величина $\theta(1 - \theta)$, тем ниже точность. Чем ближе θ к нулю или к 1, тем меньшим может быть объем выборки, необходимый для оценки θ с заданной точностью. Это обстоятельство часто используют для сокращения объема выборок, путем предварительного «расслаивания» генеральной совокупности на непересекающиеся слои, в которых исследуемый признак выражен либо очень сильно (θ близко к единице), либо, наоборот, очень слабо (θ близко к нулю). Подробнее об этом будет сказано ниже.
- 4) В общем случае, когда на базе простой случайной выборки оценивается не доля совокупности, а некоторые произвольные параметры, например величина среднего дохода, средние затраты на определенный тип продукции или услуг и т.п., точность оценки обратно пропорциональна квадратному корню из дисперсии оценки.
- 5) Точность оценивания при простом случайном отборе можно оценить по самой выборке.
- 6) Простой случайный выбор при достаточном объеме выборки обеспечивает ее репрезентативность, т.е. доставляет выборку, правильно представляющую генеральную совокупность.
- 7) Если механизм случайного выбора не обеспечивает равных шансов быть выбранными для всех элементов генеральной совокупности, возникающие из-за этого смещения невозможно устранить никаким увеличением объема выборки.

Необходимый объем выборки. Мы уже знаем, что простой случайный выбор обеспечивает репрезентативность выборок лишь в статистическом смысле. Поэтому о достигаемой при этом точности выводов следует говорить тоже в статистических терминах, задавшись, в частности, доверительной вероятностью. Положим, например, доверительную вероятность равной 0.95. Такой уровень доверия (статистическая надежность вывода) считается умеренным. Посмотрим, каков должен быть объем выборки, который обеспечивает с доверительной вероятностью 0.95 точность в оценивании θ не ниже, скажем, 0.03. (То есть

точность не ниже 3%, если в процентах исчислять долю объектов с интересующим нас свойством.)

Мы отмечали выше, что достигаемая точность в оценивании θ зависит от величины θ . Так как величина θ нам не известна, задача на первый взгляд кажется неразрешимой. Мы, впрочем, знаем, что наибольший объем выборки требуется для $\theta = \frac{1}{2}$. Это наиболее трудная ситуация. Объем выборки, который обеспечивает требуемую точность и надежность при $\theta = \frac{1}{2}$, обеспечивает эти требования и при других значениях θ . Сказанное означает, что при расчетах объемов выборок, т.е. при планировании выборочного обследования, нам следует обратиться к формуле (11.9).

При $1 - 2\alpha = 0.95$ квантиль $z_{1-\alpha} = z_{0.975} = 1.96$. Согласно (11.9), точность в 3% будет обеспечена при n таких, что

$$\frac{z_{1-\alpha}}{2\sqrt{n}} \leq 0.03.$$

Отсюда

$$\sqrt{n} \geq \frac{z_{1-\alpha}}{2} * \frac{1}{0.03},$$

поэтому наименьший необходимый объем выборки оказывается равным $n = 107$.

Для достижения точности в 1%, т.е. втрое лучшей, объем выборки должен быть увеличен в девять раз (квадрат числа три). Поэтому необходимое $n = 963$.

Объем простой случайной выборки $n = 2000$ (примерно с такими объемами работают социологические агентства) обеспечивает:

точность 3% с надежностью 0.996,

точность 1% с надежностью 0.814, и т.д.

Расслоенный выбор, который будет описан ниже, обеспечивает при тех же объемах выборок еще более высокую точность.

Требование к точности оценивания в пределах 0.01 – 0.02 вполне разумно, когда речь идет о доле популяции, обладающей определенным признаком. Например, в текущих электоральных исследованиях шансов различных претендентов на пост президента, проводимых фондом «Общественное мнение», доли избирателей, поддерживающих того или иного политика, часто отличаются на 2–3%. Поэтому для определения более популярного кандидата требуется точность не менее 1%. В маркетинговых исследованиях требования к точности оценок могут быть и не такими высокими и находиться в пределах 5 или даже 10%. Подставляя выбранные значения точности в выражение (11.8), нетрудно получить представление о необходимом объеме случайной выборки.

Таблица 11.1

Примерные объемы выборок для обеспечения точности оценки в пределах 0.01 с доверительной вероятностью $\alpha = 0.95$ и $\alpha = 0.99$ в зависимости от выраженности признака θ в генеральной совокупности

$\alpha \backslash \theta$	0.05	0.1	0.2	0.3	0.4	0.5
0.95	1825	3457	6147	8067	9220	9604
0.99	3162	5991	10650	13978	15975	16641

Таблица 11.2

Примерные объемы выборок для обеспечения точности оценки в пределах 0.02 с доверительной вероятностью $\alpha = 0.95$ и $\alpha = 0.99$ в зависимости от выраженности признака θ в генеральной совокупности

$\alpha \backslash \theta$	0.05	0.1	0.2	0.3	0.4	0.5
0.95	456	864	1537	2017	2305	2401
0.99	790	1498	2663	3495	3994	4160

В тех случаях, когда θ превышает 0.5, (например, равно 0.6) объем выборки также можно определить из приведенных выше таблиц. Для этого надо обратиться к данным таблицам для $\theta' = (1 - \theta)$, т.е. для $\theta' = 0.4$ в упомянутом примере. Необходимые объемы выборок для оценки параметров θ и $(1 - \theta)$ совпадают.

Числа в таблицах показывают, сколь существенно можно сократить объем выборки в зависимости от меры выраженности признака в совокупности. Эту особенность случайного выбора можно использовать для расслоенных генеральных совокупностей. Так говорят о популяции, разделенной на части. В некоторых из этих частей признак может быть выражен более, а в других – менее ярко, чем в совокупности в целом. Проводя случайный выбор необходимого объема из каждой части отдельно, можно либо уменьшить суммарный необходимый объем, либо увеличить точность конечного результата. Об этом будет сказано в следующем разделе.

11.4. Выборки. Сложные планы

Расслоенные совокупности. Предположим, что помимо признака A , каждый объект генеральной совокупности обладает еще и другим признаком, который мы назовем признаком B . Предположим, что этот признак имеет несколько различных уровней. Например, если элементы генеральной совокупности – это люди, то признаком B может быть, например, цвет глаз. Тогда его уровни – это те цвета глаз, которые мы различаем: серые, карие, голубые и т.д. Другой пример: генеральную

совокупность составляет все взрослое население Москвы; признак B – годовой доход взрослого москвича в денежном выражении. Уровнями этого признака могут быть, например, такие:

- B_1 – годовой доход до 30 тысяч рублей;
 - B_2 – годовой доход от 30 до 50 тысяч рублей;
 - B_3 – годовой доход от 50 до 100 тысяч рублей
- и т.д.

Признак B может быть и составным; его уровнями могут быть комбинации уровней нескольких признаков. Произвольно занумеруем уровни признака. В дальнейших обсуждениях удобнее говорить о номерах уровней, чем об их названиях. Пусть $B_1, B_2, \dots, B_l, \dots, B_L$ – уровни признака B , L – их общее число, l – текущий номер уровня. Признак B разделяет генеральную совокупность на непересекающиеся множества. Каждое такое множество составляют элементы генеральной совокупности, обладающие признаком B на определенном уровне. Эти множества называются *слоями*, или *стратами*. О признаке B говорят, что своими уровнями он *расслаивает* (стратифицирует) генеральную совокупность. Мы будем называть слоем l совокупность элементов, обладающих признаком B на уровне l , $l = 1, \dots, L$.

Обозначим через w_l ту долю, которую в генеральной совокупности составляют элементы из слоя l , т.е. те элементы, которые обладают свойством B_l . Иногда w_l называют весом слоя l . Ясно, что

$$\sum_{l=1}^L w_l = 1. \quad (11.9)$$

При простом случайном выборе w_l – это вероятность того, что мы выберем элемент со свойством B_l .

В каждом слое некоторая часть элементов обладает признаком A . Обозначим долю таких элементов в слое l через θ_l , $0 \leq \theta_l \leq 1$. При простом случайном выборе θ_l – это условная вероятность того, что выбранный элемент обладает свойством A при условии, что он обладает свойством B_l . Ясно, что

$$\theta = \sum_{l=1}^L w_l * \theta_l. \quad (11.10)$$

При простом случайном выборе эта формула выражает вероятность выбора элемента со свойством A через условные вероятности A при условии B_l и вероятности выбора B_l . В теории вероятностей равенство (11.10) известно как формула полной вероятности.

Стратифицированный выбор. Предположим, что веса слоев $w_1, \dots, w_l, \dots, w_L$ нам известны. В социологии, где расслоение ча-

сто идет по возрастным, географическим, профессиональным и т.д. признакам, веса этих групп могут быть извлечены из официальных статистических данных.

Предположим далее, что из каждого слоя мы можем произвести простой случайный выбор элементов в нужном числе, независимый от результатов других выборов. Как и в предыдущих разделах, этот выбор мы будем описывать с помощью испытаний Бернулли. Как мы отмечали, эта схема дает правильные результаты для больших совокупностей (в данном случае слоев) и объемов выборок, малых по сравнению с численностью совокупностей.

Пусть n_l обозначает объем простой случайной выборки, извлеченной из слоя l ; пусть $N = \sum_l n_l$ — суммарный объем всех выборок. В дальнейшем мы положим

$$n_l = N * w_l,$$

т.е. положим объем выборки из каждого слоя пропорциональным его численности (его весу). Такой выбор называется пропорциональным.

Пусть X_l обозначает число элементов со свойством A в выборке из слоя l . Как мы уже знаем, частота (относительная) события A в известной выборке несмещенно оценивает θ_l , т.е.:

$$\hat{\theta}_l = X_l/n_l \quad M \frac{X_l}{n_l} = \theta_l.$$

Дисперсия этой оценки

$$D\hat{\theta}_l = \frac{\theta_l * (1 - \theta_l)}{n_l}.$$

Из оценок $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_L$ можно составить несмещенную оценку для θ :

$$\hat{\theta} = \sum_l w_l * \hat{\theta}_l.$$

Дисперсия этой оценки равна

$$D\hat{\theta} = \sum_l w_l^2 \frac{\theta_l * (1 - \theta_l)}{n_l}.$$

Для пропорционального выбора

$$D\hat{\theta} = \frac{1}{N} \sum_l w_l * \theta_l * (1 - \theta_l). \quad (11.11)$$

Доверительные интервалы для θ можно строить так же, как мы действовали в разделе 11.7 («Доверительные интервалы»).

Преимущества стратифицированного выбора. Остается ответить на вопрос, имеет ли стратифицированный выбор какие-либо преимущества перед простым случайным выбором, и если да, то когда эти преимущества проявляются? Понятно, что сравнение надо проводить в равных условиях – в данном случае при одинаковом количестве обследованных элементов N . Так как обе оценки для θ – несмещенные, для их сравнения надо обратиться к дисперсиям. Та оценка окажется лучше (точнее), чья дисперсия меньше.

Для простой случайной выборки объема N дисперсия несмещенной оценки θ равна

$$\frac{\theta * (1 - \theta)}{N}.$$

Эту величину надо сравнить с (11.11). Заметим, что выражение $\theta * (1 - \theta)$ можно представить как:

$$\theta * (1 - \theta) = \sum_l w_l * \theta_l * (1 - \theta_l) + \sum_l w_l * (\theta_l - \theta)^2. \quad (11.12)$$

(Проверьте!) Отсюда следует, что дисперсия оценки при простом случайном выборе превосходит дисперсию оценки при стратифицированном выборе на неотрицательную величину, пропорциональную

$$\sum_l w_l * (\theta_l - \theta)^2.$$

Величина эта положительна, исключая случай

$$\theta_1 = \theta_2 = \dots = \theta_L = \theta.$$

Если вспомнить вероятностный смысл чисел $\theta, \theta_1, \dots, \theta_L$, мы увидим, что пропорциональный выбор лишь тогда не имеет преимуществ перед простым случайным выбором, когда

$$P(A) = P(A/B_1) = P(A/B_2) = \dots = P(A/B_L).$$

Эти равенства, когда условные вероятности события A не зависят от условий, означает независимость признаков A и B . Лишь в том случае, когда расслоение совокупности произведено с помощью признака, статистически независимого от признака A , расслоенный выбор не дает нам преимуществ в статистической точности. Во всех других случаях такой способ выбора лучше. Притом тем лучше, чем больше зависимость между признаками A и B . К слову, выражение

$$\sum_{l=1}^L [P(A/B_l) - P(A)]^2 * P(B_l)$$

может служить мерой, выражающей зависимость A от признака B (предсказуемость A по наблюдению над B).

Суммируя полученные результаты, заметим, что:

- 1) расслоенный (стратифицированный) выбор дает лучшие результаты, чем простой случайный выбор;
- 2) этот план обследования возможен, если известны доли w_i слоев в генеральной совокупности;
- 3) для разделения на слои следует выбирать признаки, наиболее тесно связанные с исследуемым признаком;
- 4) эффективно осуществлять расслоенный отбор тем легче, чем лучше предварительно изучена генеральная совокупность.

Квотируемый выбор. Квотируемый выбор представляет собой упрощенную форму стратифицированного выбора, в которой контроль за соблюдением принципов случайного выбора несколько ослаблен. При квотируемом выборе интервьюеры получают задания опросить (обследовать) определенное число (квоту) респондентов из определенных слоев. В отличие от стратифицированного выбора, в котором из каждого слоя производится простой случайный выбор, здесь интервьюерам предоставлена свобода в выборе респондентов. Поэтому квотируемый выбор не обеспечивает полной репрезентативности выборок и не свободен от неконтролируемого смещения результатов. Эта опасность тем меньше, чем более однородны слои по отношению к признаку A .

Кластерный выбор. Предположим, что генеральная совокупность разделена на непересекающиеся группы. Будем называть их кластерами. С формальной точки зрения кластеры не отличаются от слоев, о которых мы говорили при выборе из расслоенной совокупности. Однако признаки, по которым генеральная совокупность разделена на кластеры, слабо (либо вовсе не) связаны с интересующим нас признаком A . Поэтому такое разделение не подходит для стратифицированного выбора. Неформальное, но существенное различие между слоями и кластерами: слоев обычно немного, и их объемы велики, кластеры же многочисленны, но состоят из небольшого числа элементов. Разделение генеральной совокупности на кластеры нередко возникает естественным образом. Например, генеральная совокупность жителей города распадается на группы жителей отдельных улиц или домов; каждый человек является членом определенного семейного хозяйства и т.д. Такая кластерная структура может быть существенной для организации выборки, даже если она и не связана с интересующими нас характеристиками генеральной совокупности. Скажем, по плану выбора следует опросить некоего

жителя некоего села. Добраться к нему требует времени и денег. И уж если все это затрачено, то как не воспользоваться возможностью и без больших дополнительных усилий не обследовать его/ее соседей или членов его/ее семьи?

При одноступенчатом кластерном выборе мы производим простой случайный выбор определенного числа кластеров. (Далее — m кластеров.) Затем каждый кластер обследуем сплошь. При многоступенчатом выборе вместо сплошного обследования выбранных кластеров мы разделяем каждый из них на более мелкие кластеры и еще раз прибегаем к кластерному выбору. Употребительны и более сложно устроенные выборки (см. об этом ниже).

Коротко обсудим свойства одноступенчатого кластерного выбора. Естественной оценкой $\hat{\theta}$ для неизвестного θ может служить доля элементов с признаком A среди всех обследованных элементов. Если обозначить через N объем всей выборки, а через X — число элементов с признаком A , получим для $\hat{\theta}$ формулу, внешне не отличающуюся от оценки при простом случайном выборе:

$$\hat{\theta} = \frac{X}{N}. \quad (11.13)$$

Отличие, однако, имеется, и притом серьезное: в отличие от простого случайного выбора (выбора элементов) при кластерном выборе объем выборочной совокупности N оказывается случайным. Поэтому оценка $\hat{\theta}$ при кластерном выборе уже не является несмещенной. Точно указать ее математическое ожидание в общем случае невозможно, так как оно зависит от статистических свойств кластеров. Можно лишь утверждать, что смещение оценки $\hat{\theta}$ имеет порядок $\frac{1}{m}$. Поэтому при большом числе выбранных кластеров оценки получаются почти несмещенными.

Лишь в одном исключительном случае $\hat{\theta}$ оказывается несмещенной оценкой: когда объемы кластеров одинаковы. Знаменатель дроби (11.13) в этом случае не зависит от результатов выбора, и оценка получается несмещенной. Помня об этом свойстве, при разбиении генеральной совокупности на кластеры для дальнейшего кластерного выбора эти кластеры стараются сделать одинаковой численности.

Многоступенчатый отбор. На примерах простого и расслоенного случайного выбора мы рассказали об основных принципах выборочных обследований. Однако реализовать описанные методы на практике в неименном виде удастся довольно редко. Главная причина — трудность в организации простого случайного выбора из-за отсутствия полного списка элементов генеральной совокупности.

Из-за этого приходится прибегать к более сложным и многоступенчатым формам выбора. На первом этапе исследователь разбивает ге-

неральную совокупность на непересекающиеся группы, полный список которых (без повторений и без пропусков) ему известен. Так, в электоральных исследованиях, где генеральной совокупностью выступает взрослое население России, на первом шаге многоступенчатого отбора можно использовать списки избирательных участков, списки почтовых отделений и т.п.

Составляющие этот список единицы именуются *первичными единицами отбора* (ПЕО). Заметим, что число ПЕО не должно быть ни слишком малым, ни слишком большим. В первом случае снижается точность оценок и есть риск пропустить что-то существенное, а во втором — значительно увеличиваются затраты на обследование.

На первом этапе многоступенчатого отбора осуществляется случайный выбор определенного числа ПЕО, с учетом их долей охвата генеральной совокупности. На втором этапе в каждой отобранной единице можно проводить простой либо стратифицированный случайный отбор (если это возможно) либо же опять выделять список единиц для отбора. Так, если ПЕО являлись населенными пунктами или сельскими административными районами, то на втором этапе в качестве единиц отбора можно использовать списки избирательных участков в городах и деревни и села в сельских районах. Иногда единицы отбора на втором этапе именуют *вторичными единицами отбора* (ВЕО). Случайный отбор ВЕО осуществляется исходя из тех же принципов, что и на первом этапе. Продвигаясь таким образом шаг за шагом, на последнем этапе осуществляется выбор опрашиваемых респондентов.

Мы ограничимся этим общим описанием принципов многоступенчатого отбора. Не станем касаться ни его разновидностей, ни точных формул для оценок и их статистических характеристик. Эту информацию можно найти в специальной литературе [38], [54]. Заметим только, что реальные выборки, используемые в исследованиях общественного мнения, устроены еще более сложно, так как они обычно сочетают в себе принципы многоступенчатости и стратификации.

Другие способы формирования выборок. Организационные трудности, денежные затраты, стремление провести обследование быстро и прочие подобные причины часто заставляют отступать от классических теоретических схем и проводить исследования по какому-то другому плану. Таких планов известно много. Более того, каждое конкретное исследование, с учетом его особенностей, может идти по собственному плану. Иногда особенности этих планов диктуются выбранной методикой проведения опроса. Так, проводя опрос по телефону, необходимо использовать специальные планы опроса, учитывающие специфику метода.

11.5. Основные выводы

- 1) Построение реальной репрезентативной выборки при выборочных исследованиях является необходимой, но, возможно, весьма трудоемкой задачей, требующей как высокой квалификации специалистов, так и довольно обширной социально-демографической статистической информации. В то же время построенную выборку можно использовать многократно в различных исследованиях. Именно так и поступают ведущие социологические службы и органы Госкомстата.
- 2) В основе построения репрезентативных выборок лежат принципы случайного отбора. Другие способы формирования выборок не позволяют сделать обоснованных заключений о качестве полученных результатов.
- 3) Точность оценок, получаемых по случайной выборке, зависит от плана построения (дизайна) выборки и рассчитывается на основе самой выборки. Она никак не связана с объемом генеральной совокупности (в тех случаях, когда объем выборки пренебрежимо мал по сравнению с объемом совокупности). Поэтому «процент охвата» генеральной совокупности не является разумной характеристикой выборочного обследования и не должен участвовать в его планировании.
- 4) Необходимый объем выборки зависит от требуемой точности и достоверности результатов и плана построения выборки. Выше были приведены методы расчета и значения объемов для простейших планов выборок. Как показывает опыт ведущих социологических служб, при всероссийских опросах общественного мнения объемы многоступенчатых, стратифицированных выборок обычно колеблются около 1500 – 2500 респондентов. В связи со сказанным примерно такими же должны быть выборки при обследовании одного субъекта Федерации или просто крупного населенного пункта. В более детальных выборочных исследованиях могут фигурировать выборки большего объема. Например, общероссийская выборка Госкомстата для обследования социально-экономического положения семей включает в себя порядка 50 000 респондентов.
- 5) В данном обзоре мы не обсуждали, как следует составлять анкеты или опросные листы. Это отдельная и весьма важная тема. Даже при хорошо составленной выборке недостаточно продуманные формулировки вопросов могут привести к результатам, сильно искажающим реальное положение дел.

Дополнительная литература

1. *Джессен Р.* Методы статистических обследований / пер. с англ.; под ред. и с предисл. Е.М. Четыркина. — М.: Финансы и статистика, 1985. — 478 с.: ил.
2. *Кокрен У.* Методы выборочного исследования. — М.: Статистика, 1976. — 440 с.
3. *Крыштановский А.О.* Анализ социологических данных с помощью пакета SPSS: учеб. пособие. — М.: ГУ ВШЭ, 2006. — 281 с.
4. *Сигел Э.* Практическая бизнес-статистика: пер. с англ. — М.: Вильямс, 2002. — 1056 с.
5. Справочник по прикладной статистике: в 2 т.; под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. — М.: Финансы и статистика, 1989, 1990.
6. *Шварц Г.* Выборочный метод. Руководство по применению статистических методов оценивания: пер. с нем. — М.: Статистика, 1978. — 213 с.

Многомерный анализ и другие статистические методы

12.1. Введение

Арсенал методов анализа данных, предлагаемых современной статистикой, разумеется, далеко не ограничивается тем, что было изложено в предыдущих главах этой книги. Так, за рамками рассмотрения остались широко используемые на практике методы многомерного статистического анализа (т.е. анализа многомерных статистических данных), а также всевозможные специализированные статистические методы, предназначенные для анализа специфических данных в конкретных предметных областях. В настоящей главе мы дадим очень краткий обзор таких методов, выбрав из них наиболее широко используемые и включенные в статистические пакеты для ЭВМ.

Замечание для профессиональных математиков и статистиков. Цель этой главы — всего лишь дать знакомящимся со статистикой читателям самое общее представление о назначении некоторых из тех областей статистики, которые не были затронуты в этой книге, а также указать список книг для дальнейшего чтения. Поэтому просим быть снисходительными к упрощениям и неточностям, неизбежным при описании сути сложных научных проблем в двух-трех абзацах.

12.2. Многомерный статистический анализ

В предыдущих главах книги мы обсуждали в основном такие проблемы, в которых случайная изменчивость была представлена одной (случайной) переменной. Например, у каждого наудачу выбранного объекта мы измеряли какой-то один признак; либо при каждой комбинации управляющих факторов измеряли одномерный отклик и т.д. Исключение составила гл. 9, в которой мы рассматривали вопросы связи двух (случайных) признаков. Там мы встретились с ситуацией, когда в одном эксперименте — например, при обследовании одного объекта — измеряются сразу несколько характеристик. В таких опытах каждое наблюдение представляется не одним-единственным числом, а некоторым конечным набором чисел, в котором в заданном порядке записаны все

измеренные характеристики объекта. Та часть математической статистики, которая исследует эксперименты с такими многомерными наблюдениями, называется *многомерным статистическим анализом*.

Измерение сразу нескольких признаков (свойств объектов) в одном эксперименте в общем более естественно, чем измерение лишь какого-то одного. Поэтому потенциально многомерный статистический анализ имеет обширное поле для применений. К тому же с формальной точки зрения одномерный статистический анализ (который мы и обсуждали ранее) представляет частный случай многомерного.

В настоящее время хорошо разработана математическая теория для многомерных гауссовских наблюдений, т.е. для случайных величин, подчиняющихся многомерному нормальному распределению. Здесь почти для каждого одномерного гауссовского метода существует соответствующий многомерный вариант. Кроме того, имеются решения и для некоторых специфически многомерных статистических проблем. О многомерном гауссовском статистическом анализе написаны книги, из которых мы особо отметим [5] и [8]. Этому вопросу обычно отводится место и в учебниках общего назначения.

К сожалению, построение теории для многомерных статистических данных оказалось делом весьма трудным. Такая теория до сих пор еще далеко не достигает той полноты и законченности, которая свойственна ее одномерной версии. Хорошо разработана лишь теория для гауссовских (имеющих многомерное нормальное распределение) данных. Здесь почти для каждого одномерного гауссовского статистического метода имеется соответствующий многомерный вариант. Кроме того, естественно, имеются и методы для решения некоторых специфически многомерных задач.

Построение многомерных версий для других статистических методов удается далеко не так гладко. В частности, непараметрические методы, такие важные и эффективные в одномерном случае, все еще не имеют своего законченного многомерного аналога (соответствующая теория находится в процессе разработки). Поэтому для аккуратного статистического анализа имеющихся данных нередко не находится адекватных статистических средств. Из-за этого, в частности, рассчитанные на гауссовские данные правила нередко приходится применять и там, где для этого нет достаточных оснований. Конечные выводы в таких случаях бывает нелегко интерпретировать. Более того, при анализе многомерных данных часто используют и методы, вообще не имеющие четкой статистической трактовки в духе рассмотренных ранее концепций проверки гипотез, построения доверительных интервалов и т.д. Поэтому мы не будем пытаться изложить здесь хоть сколько-нибудь

цельную картину многомерного анализа, а ограничимся упоминанием и кратким пояснениями нескольких наиболее популярных методов — тех, которые уже нашли отражение в статистических пакетах. Подробное изложение этих и других методов можно найти в [5], [87], [103].

12.3. Факторный анализ

При исследовании сложных объектов и систем (например, в психологии, биологии, социологии и т.д.) часто мы не можем непосредственно измерить величины, определяющие свойства этих объектов (так называемые *факторы*), а иногда нам не известны даже число и содержательный смысл факторов. Для измерений могут быть доступны иные величины, тем или иным способом зависящие от этих факторов. При этом, когда влияние неизвестного фактора проявляется в нескольких измеряемых признаках, эти признаки могут обнаруживать тесную связь между собой (например, коррелированность), поэтому общее число факторов может быть гораздо меньше, чем число измеряемых переменных, которое обычно выбирается исследователем в той или иной мере произвольно. Для обнаружения влияющих на измеряемые переменные факторов используются методы *факторного анализа*¹.

В качестве примера применения факторного анализа приведем изучение свойств личности с помощью психологических тестов. Свойства личности не поддаются прямому измерению, о них можно судить только на основании поведения человека, ответа на те или иные вопросы и т.д. Для объяснения результатов проведенных опытов их результаты подвергаются факторному анализу, который и позволяет выявить те личностные свойства, которые оказывали влияние на поведение испытуемых в проведенных опытах.

Первым этапом факторного анализа, как правило, является выбор новых признаков, которые являются линейными комбинациями прежних и «вбирают» в себя большую часть общей изменчивости наблюдаемых данных, а поэтому передают большую часть информации, заключенной в первоначальных наблюдениях. Обычно это осуществляют с помощью *метода главных компонент*, хотя иногда используют и другие приемы (скажем, метод максимального правдоподобия). Метод главных

¹ Обратите внимание, что факторный анализ — это метод совсем другого назначения, чем одно-, двух- и многофакторный анализ, которые рассматривались нами в гл. 6 и 7. В однофакторном, двухфакторном и т.д. анализе (по-английски: One-way, Two-way и т.д. Analysis of Variance) влияющие на результат факторы считаются известными и речь идет только о выяснении существенности или оценке этого влияния. А в факторном анализе (по-английски: Factor Analysis) речь идет о выделении из множества измеряемых характеристик объекта новых факторов, более адекватно отражающих свойства объекта.

компонент по существу сводится к выбору новой ортогональной системы координат в пространстве наблюдений. В качестве первой главной компоненты избирают направление, вдоль которого массив наблюдений имеет наибольший разброс, выбор каждой последующей главной компоненты происходит так, чтобы разброс наблюдений вдоль нее был максимальным и чтобы эта главная компонента была ортогональна другим главным компонентам, выбранным прежде.

Однако обычно факторы, полученные методом главных компонент, не поддаются достаточно наглядной интерпретации. Поэтому следующим шагом факторного анализа служит преобразование (вращение) факторов таким образом, чтобы облегчить их интерпретацию.

Более подробно о методах факторного анализа можно прочесть в книгах [10], [103], [110].

12.4. Дискриминантный анализ

Предположим, что мы имеем совокупность объектов, разбитую на несколько групп (т.е. для каждого объекта мы можем сказать, к какой группе он относится). Пусть для каждого объекта имеются изменения нескольких количественных характеристик. Мы хотим найти способ, как на основании этих характеристик можно узнать группу, к которой принадлежит объект. Это позволит нам для новых объектов из той же совокупности предсказывать группы, к которой они относятся.

Например, исследуемыми объектами могут быть пациенты — здоровые или больные той или иной болезнью, а характеристиками — результаты медицинских анализов. Если мы научимся по этим характеристикам узнавать, здоров ли пациент либо болен той или иной болезнью, это позволит значительно повысить эффективность медицинских обследований.

Для решения этой задачи применяются методы *дискриминантного анализа*, они позволяют строить функции измеряемых характеристик, значения которых и объясняют разбиение объектов на группы. Желательно, чтобы этих функций (дискриминирующих признаков) было немного — в этом случае результаты анализа легче содержательно истолковать. Особую роль, благодаря своей простоте, играет *линейный дискриминантный анализ*, в котором классифицирующие признаки выбираются как линейные функции от первичных признаков. В случае разделения нескольких нормальных (гауссовских) совокупностей линейный дискриминантный анализ имеет ясные статистические свойства.

Более подробно о дискриминантном анализе говорится в книгах [10], [103].

12.5. Кластерный анализ

Методы кластерного анализа позволяют разбить изучаемую совокупность объектов на группы «схожих» объектов, называемых *кластерами*.

Большинство методов кластеризации (иерархической группировки) являются *агломеративными* (объединительными) — они начинают с создания элементарных кластеров, каждый из которых состоит ровно из одного исходного наблюдения (одной точки), а на каждом последующем шаге происходит объединение двух наиболее близких кластеров в один. Момент остановки этого процесса может задаваться исследователем (например, указанием требуемого числа кластеров или максимального расстояния, при котором допустимо объединение). Графическое изображение процесса объединения кластеров может быть получено с помощью *дендрограммы* — дерева объединения кластеров. Другие методы кластерного анализа являются *дивизивными* — они пытаются разбивать объекты на кластеры непосредственно.

Методы кластеризации довольно разнообразны, в них по-разному выбирается способ определения близости между кластерами (и между объектами), а также используются различные алгоритмы вычислений. Заметим, что результаты кластеризации зависят от выбранного метода, и эта зависимость тем сильнее, чем менее явно изучаемая совокупность разделяется на группы объектов. Поэтому результаты вычислительной кластеризации могут быть дискуссионными, и часто они служат лишь подспорьем для содержательного анализа.

Заметим также, что методы кластерного анализа не дают какого-либо способа для проверки статистической гипотезы об адекватности полученных классификаций. Иногда результаты кластеризации можно обосновать с помощью методов дискриминантного анализа.

Более подробно с методами кластерного анализа можно познакомиться в [47], [87], [103].

12.6. Многомерное шкалирование

Во многих областях исследования (например, в психологии, биологии, социологии, лингвистике и т.д.) бывает затруднительно или невозможно проводить непосредственное измерение интересующих исследователя характеристик объектов из изучаемой совокупности, зато можно экспертным или каким-то другим путем оценивать степень сходства или различия между парами объектов. В этом случае для интерпретации получаемых данных используются методы многомерного шкалирования.

Они позволяют представить совокупность интересующих исследователя объектов в виде некоторого набора точек многомерного пространства некоторой небольшой размерности, при этом каждому объекту соответствует одна точка. Координаты точек истолковываются как значения неких характеристик исходных объектов, которые и объясняют их свойства или взаимоотношения.

В случае удачного шкалирования, когда точки полученного пространства представляют объекты без серьезных погрешностей и размерность этого пространства невелика (равна, скажем, двум или трем), исследователь получает возможность представить изучаемую совокупность объектов наглядно. Часто это помогает по-новому осознать проблему, увидеть ее новые черты и особенности либо осознать те скрытые признаки, которые и определяют видимые свойства объектов или их взаимоотношения.

Типичный пример использования методов многомерного шкалирования — изучение политических деятелей. Здесь исходными данными для анализа могут служить экспертные оценки сходства или различия взглядов политических деятелей по некоторому набору вопросов. Для депутатов парламента такими данными могут служить результаты голосований. И очень часто с помощью методов многомерного шкалирования удается объяснить исходные данные с помощью нескольких характеристик взглядов политических деятелей, которые и описывают (в основном) их поведение. Например, может оказаться, что результаты голосований депутатов в парламенте в основном объясняются всего двумя-тремя характеристиками. Исследователь может условно их назвать, скажем, «приверженность к либеральной или к государственной модели экономики» и «прозападная или почвенническая ориентированность» или как-то еще. Результаты подобных исследований иногда публикуются в газетах.

Часто в качестве исходных данных для шкалирования используются не сами оценки степени сходства объектов, а результаты их ранжирования. Соответствующие методы шкалирования называются *неметрическими*. Они были разработаны для решения проблем психологии: здесь исходными данными часто служат суждения человека (как испытуемого либо как эксперта), поэтому их количественные значения носят в значительной мере условный характер. Чтобы избавиться от этой условности, и прибегают к ранжированию. Сейчас неметрическое многомерное шкалирование широко применяется и для других данных. Подробнее о методах многомерного шкалирования можно прочесть в книгах [59], [87], [89].

12.7. Методы контроля качества

Из многочисленных специализированных разделов статистики мы рассмотрим один — методы контроля качества. Эти методы, как следует из их названия, предназначены для контроля качества выпускаемой продукции с целью выявления нарушений и «узких мест» в организации производства и в технологических процессах, ведущих к снижению качества продукции. Повсеместное применение научно обоснованных методов контроля качества явилось немаловажным фактором успехов стран — лидеров мировой экономики, в особенности Японии.

В отличие от большинства описанных выше многомерных методов методы контроля качества не требуют трудоемких вычислений — они исключительно просты и наглядны. Целью этих методов может быть:

- получение наглядного представления о выборочном распределении значения некоторого параметра в выпускаемой продукции и сравнение этого распределения с границами допуска (*гистограмма качества*);
- наглядное выделение наиболее важных факторов, влияющих на качество продукции (*диаграмма Парето*);
- выявление необычных отклонений в параметрах выпускаемой продукции и отделения случайных отклонений от неслучайных и требующих вмешательства тенденций (*контрольные карты*).

Простота, наглядность и эффективность статистических методов контроля качества сделали возможным и оправданным их повсеместное (вплоть до мастеров, а иногда и отдельных рабочих) применение в передовых странах. Более подробно об этих методах можно прочесть в книгах [73], [90].

12.8. Использование статистических пакетов

В пакете SPSS представлены все перечисленные выше методы. При этом есть возможность гибкого выбора и настройки параметров соответствующих процедур. Например, иерархическая кластеризация в пакете предусматривает задание различных расстояний между объектами, различных методов объединения объектов в кластеры и т.п.

В документации и во встроенном справочнике SPSS читатель может найти дополнительные пояснения по назначению и методике применения статистических методов, описанных в этой главе.

Дополнительная литература

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности: справочное издание / под ред. С.А. Айвазяна. — М.: Финансы и статистика, 1989. — 607 с.
2. Андерсен Т. Введение в многомерный статистический анализ. — М.: Физматгиз, 1963. — 500 с.
3. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: учебник. — М.: Финансы и статистика, 1998. — 352 с.: ил.
4. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа. — М.: Финансы и статистика, 1986.
5. Факторный, дискриминантный и кластерный анализ. — М.: Финансы и статистика, 1989. — 215 с.
6. Хартман Г. Современный факторный анализ. — М.: Статистика, 1972.

Таблицы математической статистики

Здесь представлено несколько таблиц математической статистики. Овладевая принципами обработки данных, необходимо «вручную» выполнить десяток-другой упражнений с числами и вычислениями. В статистике такая работа чаще всего заканчивается обращением к таблицам. Их удобно иметь под рукой.

При обработке данных с помощью пакетов статистических программ все необходимые числа, как правило, предоставляет компьютер. Для классических методов статистики и распределений, связанных с нормальным (Стьюдента, хи-квадрат и т.д.), с этим проблем обычно не возникает. Для непараметрических статистик пакеты чаще всего указывают критические значения, основанные на аппроксимациях для больших выборок. Для малых выборок эти приближенные значения могут оказаться неудовлетворительными. В этих случаях необходимы таблицы.

Чтобы уложиться в разумный объем, мы даем таблицы в упрощенном виде и не для всех объемов выборок. Это только таблицы процентных точек и критических значений. В описании таблиц мы не касаемся вопросов интерполяции и экстраполяции, которые часто возникают при пользовании таблицами. Частично эти вопросы рассматриваются в основном тексте книги.

Возможно, что для некоторых статистических исследований эти таблицы окажутся недостаточными. Тогда следует обратиться к более обширным таблицам математической статистики, например [19], [32], [65], либо к пакетам статистических программ.

В настоящем приложении представлены десять наиболее употребительных таблиц математической статистики. Часть из них (нормальное, биномиальное распределения, распределения Стьюдента, хи-квадрат и F-распределение) более доступна для широкого читателя, так как эти таблицы часто публикуются в приложениях к учебным пособиям, в общестатистической литературе и в справочниках. Другая часть — это таблицы непараметрической статистики. В приложении представлены таблицы процентных точек для статистик Уилкоксона, Краскела—Уоллиса, Фридмана и двух коэффициентов ранговой корреляции — Кендэла и Спирмена. Непараметрические методы статистики позже вошли в программы учебных курсов, они хуже отражены в научной и учебной

литературе. Между тем непараметрические методы очень нужны для экономических, социальных, медицинских, биологических, экологических и других исследований. Сейчас эти методы очень признаны и популярны. К сожалению, таблицы по непараметрической статистике публиковались на русском языке малыми тиражами, в основном — в специальной литературе и давно стали библиографическими редкостями. Впрочем, сетования на недоступность таблиц относятся ко всей математической статистике, в последний раз сборник таких таблиц был издан в СССР в 1985 г. [65].

Основой для настоящего приложения послужили таблицы из ГОСТ 23554.2—81 [32] и New Cambridge Elementary Statistical Tables [133], а также из книги М. Холлелера и Д. Вулфа [115]. Часть таблиц приведена в переработанном виде. Каждая таблица снабжена кратким описанием; к каждой дан пример на считывание табличных значений. Примеры на считывание таблиц соотнесены с теми статистическими критериями (с указаниями нулевых гипотез и альтернатив), для которых нужна соответствующая таблица. Подробное описание этих критериев дано в основной части книги.

В названных таблицах мы употребляем два выражения: процентные точки и критические значения. Мы говорим о процентных точках распределений вероятностей и о критических значениях выборочных статистик. Для нормального, биномиального, студентовского, хи-квадрат и F-распределений приводятся процентные точки, а для статистик Уилкоксона, Краскела—Уоллиса, Фридмана и коэффициентов ранговой корреляции — критические значения.

Определение 1. *Рассмотрим случайную величину ξ с заданным распределением вероятностей. Верхней альфа-процентной точкой x данного распределения называется решение уравнения*

$$P(\xi \geq x) = \alpha,$$

где α — заданное число, $0 < \alpha < 1$.

Если, например, $\alpha = 0.10$, то соответствующее x — это верхняя десятипроцентная точка. Название «процентной» точки больше подходит, когда вероятности выражаются в процентах. Мы употребляем его и тогда, когда вероятности выражаются в долях единицы.

О критических значениях статистик мы говорим тогда, когда их распределения (обычно — при нулевых гипотезах) еще не обрели в языке самостоятельного существования. Вот разница: говорят, например, о распределении статистики ранговых сумм Уилкоксона, но о распределении Стьюдента, уже не связывая его с соответствующей статистикой (студентовой дробью, или студентовым отношением).

Определение 2. Если S — статистика критерия, ее верхним критическим значением x уровня альфа называют решение уравнения: $P(S \geq x) = \alpha$, где α — заданное число, $0 < \alpha < 1$.

При этом приходится дополнительно говорить о предположениях и условиях, в которых мы рассматриваем статистику S . Скажем, для упомянутой выше статистики сумм Уилкоксона приходится уточнять, что мы рассматриваем ее для независимых, непрерывных и однородных выборок. Название верхние критические значения прямо сообщает его статистический смысл: надо отвергать гипотезу (на указанном уровне значимости), если выборочное значение статистики превосходит это критическое значение или равно ему.

Многие распределения и статистики зависят от одного или нескольких целочисленных параметров. Обычно эти параметры — число степеней свободы или объема выборок. Чтобы указать, какие значения этих параметров отражены в таблицах, мы употребляем обозначения типа $n = 4(1)10(2)20$ и т.д. Читать эту запись надо так: n изменяется от $n = 4$ с шагом единица (т.е. увеличиваясь последовательно на единицу) до $n = 10$; далее n изменяется от $n = 10$ с шагом 2 до $n = 20$ и т.д.

Благодарим Н.П. Нискину и Д.С. Шмерлинга за большую помощь в подготовке таблиц.

П1. Верхние процентные точки стандартного нормального распределения

Описание таблицы. В таблице для различных значений уровня значимости α приведены процентные точки z_α стандартного нормального распределения $N(0, 1)$ (см. п. 2.4).

По определению, z_α есть решение уравнения:

$$\Phi(z_\alpha) = 1 - \alpha,$$

где $\Phi(\cdot)$ — функция стандартного нормального распределения (функция Лапласа). Если через z обозначить случайную величину, распределенную по стандартному нормальному закону, то z_α можно определить и с помощью уравнения $P(z \geq z_\alpha) = \alpha$.

В таблице даны значения z_α для некоторых значений $\alpha \geq 0.500$. Для $\alpha < 0.5$ следует использовать соотношение $z_\alpha = -z_{1-\alpha}$.

Таблица взята в переработанном виде из [133]. Другие таблицы см. в [19], [77], [115] и др.

Пример на считывание таблицы. Случайная величина z , имеющая стандартное нормальное распределение, с вероятностью $\alpha = 0.050$

превышает значение $z_\alpha = 1.6449$. Другими словами, мы должны отвергнуть гипотезу, приводящую к стандартному нормальному распределению статистики z против односторонних альтернатив на уровне значимости $\alpha = 0.050$, если значение z статистики превысило $z_\alpha = 1.6449$. В случае двусторонних альтернатив для отвержения гипотезы на уровне значимости α следует выбирать $\alpha/2$ процентную точку $z_{\alpha/2}$. Так, при $\alpha = 0.050$ $z_{\alpha/2} = 1.9600$.

Таблица 1

Верхние процентные точки стандартного нормального распределения

α	0.500	0.450	0.400	0.350	0.300	0.250
z_α	0.0000	0.1257	0.2533	0.3853	0.5244	0.7645
α	0.200	0.150	0.100	0.050	0.025	0.010
z_α	0.8416	1.0364	1.2816	1.6449	1.9600	2.3263
α	0.005	0.0025	0.001	0.0005	0.0001	
z_α	2.5758	2.8070	3.0902	3.2905	3.7190	

П2. Верхние процентные точки распределения Стьюдента

Описание таблицы. В таблице для числа степеней свободы $n = 1(1)30(2)50(5)70(10)100$ и некоторых других приведены процентные точки $t(n, \alpha)$ распределения Стьюдента для различных значений α (см. п. 2.6.2).

Если через t_n обозначить случайную величину, распределенную по Стьюденту с n степенями свободы, то $t(n, \alpha)$ можно определить как решение уравнения

$$P(t_n \geq t(n, \alpha)) = \alpha.$$

В таблице даны значения $t(n, \alpha)$ для некоторых значений $\alpha \geq 0.500$. Для $\alpha < 0.500$ следует использовать соотношение $t(n, \alpha) = -t(n, 1 - \alpha)$. В строке, начинающейся символом ∞ (в столбце значений n), приведены процентные точки стандартного нормального распределения — это предельные значения $t(n, \alpha)$ при $n \rightarrow \infty$.

Таблица взята из [133]. Другие таблицы см. в [65], [74], [93].

Пример на считывание таблицы. Случайная величина t_n , имеющая распределение Стьюдента с числом степеней свободы $n = 13$, с вероятностью $\alpha = 0.05$ превышает значение $t(n, \alpha) = 1.7709$. Другими словами, мы должны отвергнуть гипотезу, приводящую к распределению Стьюдента с n степенями свободы для статистики t_n , против односторонних альтернатив на уровне значимости $\alpha = 0.05$, если значение t_n превысило $t(n, \alpha) = 1.7709$. В случае двусторонних альтернатив для отвержения гипотезы на уровне значимости α следует выбирать процентную точку $t(n, \alpha/2)$. Так, при $\alpha = 0.05$ и $n = 13$ $t(n, \alpha/2) = 2.1604$.

Таблица 2

Верхние процентные точки распределения Стьюдента

n	$\alpha = 0.10$	$\alpha = 0.050$	$\alpha = 0.025$	$\alpha = 0.010$	$\alpha = 0.005$	$\alpha = 0.0025$
1	3.0777	6.314	12.706	31.820	63.657	127.32
2	1.8856	2.9200	4.3027	6.9646	9.9248	14.089
3	1.6377	2.3534	3.1824	4.5407	5.8409	7.4533
4	1.5332	2.1318	2.7764	3.7469	4.6041	5.5976
5	1.4759	2.0150	2.5706	3.3649	4.0321	4.7733
6	1.4398	1.9432	2.4469	3.1427	3.7074	4.3168
7	1.4149	1.8946	2.3646	2.9980	3.4995	4.0293
8	1.3968	1.8595	2.3060	2.8965	3.3554	3.8325
9	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897
10	1.3722	1.8125	2.2281	2.7638	3.1693	3.5814
11	1.3634	1.7959	2.2010	2.7181	3.1058	3.4966
12	1.3562	1.7823	2.1788	2.6810	3.0545	3.4284
13	1.3502	1.7709	2.1604	2.6503	3.0123	3.3725
14	1.3450	1.7613	2.1448	2.6245	2.9768	3.3257
15	1.3406	1.7530	2.1314	2.6025	2.9467	3.2860
16	1.3368	1.7459	2.1199	2.5835	2.9208	3.2520
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.1534
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.1352
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.1188
23	1.3195	1.7139	2.7139	2.0687	2.4999	3.1040
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.0905
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.0782
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.0669
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.0565
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.0469
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.0380
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.0298
32	1.3086	1.6939	2.0369	2.4487	2.7385	3.0149
34	1.3070	1.6909	2.0322	2.4411	2.7284	3.0020
36	1.3055	1.6883	2.0281	2.4345	2.7195	2.9905
38	1.3042	1.6860	2.0244	2.4286	2.7116	2.9803
40	1.3031	1.6839	2.1211	2.4233	2.7045	2.9712
42	1.3020	1.6820	2.0181	2.4185	2.6981	2.9630
44	1.3011	1.6802	2.0154	2.4141	2.6923	2.9555
46	1.3002	1.6787	2.0129	2.4102	2.6870	2.9488
48	1.2994	1.6772	2.01060	2.4066	2.6822	2.9426
50	1.2987	1.6759	2.0086	2.4033	2.6778	2.9370
55	1.2971	1.6730	2.0040	2.3961	2.6682	2.9247
60	1.2958	1.6706	2.0003	2.3901	2.6603	2.9146
65	1.2947	1.6686	1.9971	2.3851	2.6536	2.9060
70	1.2938	1.6669	1.9944	2.3808	2.6479	2.8987

Таблица 2

Верхние процентные точки распределения Стьюдента
(окончание)

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.010$	$\alpha = 0.005$	$\alpha = 0.0025$
80	1.2922	1.6641	1.9901	2.3739	2.6387	2.8870
90	1.2910	1.6620	1.9867	2.3685	2.6316	2.8779
100	1.2901	1.6602	1.9840	2.3642	2.6259	2.8707
120	1.2886	1.6577	1.9799	2.3578	2.6174	2.8599
150	1.2872	1.6551	1.9759	2.3515	2.6090	2.8492
200	1.2858	1.6525	1.9719	2.3451	2.6006	2.8385
250	1.2849	1.6510	1.9695	2.3414	2.5956	2.8322
300	1.2844	1.6499	1.9679	2.3388	2.5923	2.8279
400	1.2837	1.6487	1.9659	2.3357	2.5882	2.8227
500	1.2832	1.6479	1.9647	2.3338	2.5857	2.8195
∞	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070

ПЗ. Верхние процентные точки распределения хи-квадрат

Описание таблицы. В таблице для числа степеней свободы $n = 1(1)120$ приведены верхние процентные точки $\chi^2(n, \alpha)$ распределения хи-квадрат с n степенями свободы (см. п. 2.6.1) для значений $\alpha = 0.1, 0.05$ и 0.01 .

Если через χ_n^2 обозначить случайную величину, распределенную по закону хи-квадрат с n степенями свободы, то $\chi^2(n, \alpha)$ можно определить как решение уравнения

$$P(\chi_n^2 \geq \chi^2(n, \alpha)) = \alpha.$$

Таблица взята из [32]. Другие таблицы см. в [19], [65], [74], [125].

Пример на считывание таблицы. Случайная величина χ_n^2 , имеющая распределение хи-квадрат с числом степеней свободы $n = 13$, с вероятностью $\alpha = 0.05$ превышает значение $\chi^2(n, \alpha) = 22.362$. Другими словами, следует отвергнуть гипотезу, приводящую к распределению хи-квадрат с n степенями свободы для статистики χ_n^2 , на уровне значимости $\alpha = 0.05$, если значение χ_n^2 превысило $\chi^2(n, \alpha) = 22.362$.

Таблица 3

Верхние процентные точки распределения хи-квадрат

n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	2.7055	3.8415	6.6349
2	4.6052	5.9915	9.2103
3	6.2514	7.8147	11.345
4	7.7794	9.4877	13.277
5	9.2364	11.071	15.086
6	10.645	12.592	16.812
7	12.017	14.067	18.475
8	13.362	15.507	20.090
9	14.684	16.919	21.666

Таблица 3

Верхние процентные точки распределения хи-квадрат
(продолжение)

п	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
10	15.987	18.307	23.209
11	17.275	19.675	24.725
12	18.549	21.026	26.217
13	19.812	22.362	27.688
14	21.064	23.685	29.141
15	22.307	24.996	30.578
16	23.542	26.296	32.000
17	24.769	27.587	33.409
18	25.989	28.869	34.805
19	27.204	30.144	36.191
20	28.412	31.410	37.566
21	29.615	32.671	38.932
22	30.813	33.924	40.289
23	32.007	35.172	41.638
24	33.196	36.415	41.980
25	34.382	37.652	44.314
26	35.563	38.885	45.642
27	36.741	40.113	46.963
28	37.916	41.337	48.278
29	39.087	42.557	49.588
30	40.256	43.773	50.892
31	41.422	44.985	52.191
32	42.585	46.194	53.486
33	43.745	47.400	54.776
34	44.903	48.602	56.061
35	46.059	49.802	57.342
36	47.212	50.998	58.619
37	48.363	52.192	59.892
38	49.513	53.384	61.162
39	50.660	54.572	62.428
40	51.805	55.758	63.691
41	52.949	56.942	64.950
42	54.090	58.124	66.206
43	55.230	59.304	67.459
44	56.369	60.481	68.710
45	57.505	61.656	69.957
46	58.641	62.830	71.201
47	59.774	64.001	72.443
48	60.907	65.023	73.683
49	62.038	66.339	74.919
50	63.167	67.505	76.154
51	64.295	68.669	77.386
52	65.422	69.832	78.616
53	66.548	70.993	79.843
54	67.673	72.153	81.069

Таблица 3

Верхние процентные точки распределения хи-квадрат
(продолжение)

n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
55	68.796	73.311	82.292
56	69.919	74.468	83.513
57	71.040	75.624	84.733
58	72.160	76.778	85.950
59	73.279	77.931	87.166
60	74.397	79.082	88.379
61	75.514	80.232	89.591
62	76.630	81.381	90.802
63	77.745	82.529	92.010
64	78.860	83.675	93.217
65	79.973	84.821	94.422
66	81.085	85.965	95.626
67	82.197	87.108	96.828
68	83.308	88.250	98.028
69	84.418	89.391	99.228
70	85.527	90.531	100.425
71	86.635	91.670	101.621
72	87.743	92.808	102.816
73	88.850	93.945	107.862
74	89.956	95.081	109.074
75	91.061	96.217	106.393
76	92.166	97.351	107.583
77	93.270	98.484	108.771
78	94.374	99.617	109.958
79	95.476	100.749	111.144
80	96.579	101.879	112.329
81	97.680	103.010	113.512
82	98.780	104.139	114.695
83	99.880	105.267	115.876
84	100.980	106.395	117.057
85	102.079	107.522	118.236
86	103.177	108.648	119.414
87	104.275	109.773	120.591
88	105.372	110.898	121.767
89	106.469	112.022	122.942
90	107.565	113.145	124.116
91	108.661	114.268	125.289
92	108.756	115.390	126.462
93	110.850	116.511	127.633
94	111.944	117.632	128.803
95	113.038	118.752	129.973
96	114.131	119.871	131.141
97	115.223	120.990	132.309
98	116.315	122.108	133.476
99	117.407	123.225	134.642

Верхние процентные точки распределения хи-квадрат
(окончание)

n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
100	118.498	124.342	135.807
101	119.590	125.460	136.970
102	120.679	126.574	138.134
103	121.770	127.690	139.300
104	122.858	128.804	140.459
105	123.950	129.920	141.620
106	125.035	131.031	142.780
107	126.120	132.140	143.940
108	127.211	133.257	145.099
109	128.300	134.370	146.260
110	129.385	135.480	147.414
111	130.470	136.590	148.570
112	131.558	137.701	149.727
113	132.640	138.810	150.880
114	133.729	139.921	152.037
115	134.810	141.030	153.190
116	135.898	142.138	154.344
117	136.980	143.250	155.500
118	138.066	144.354	156.648
119	139.150	145.460	157.800
120	140.233	146.657	158.950

П4. Верхние процентные точки F-распределения

Описание таблиц. В последующих трех таблицах для $\alpha = 0.10, 0.05$ и 0.01 приведены верхние процентные точки $F(n_1, n_2, \alpha)$ F-распределений со степенями свободы (n_1, n_2) для $n_1 = 1(1)10(5)20$, $n_2 = 1(1)30(2)40(20)200$ и некоторых других (см. п. 2.6.3). В строках и столбцах, начинающихся символом ∞ (как значением n_2 или значением n_1), приведены предельные значения $F(n_1, n_2, \alpha)$ при $n_2 \rightarrow \infty$ и $n_1 \rightarrow \infty$.

Если через F_{n_1, n_2} обозначить случайную величину, имеющую F-распределение с (n_1, n_2) степенями свободы, то $F(n_1, n_2, \alpha)$ можно определить как решение уравнения

$$P(F_{n_1, n_2} \geq F(n_1, n_2, \alpha)) = \alpha.$$

Таблицы взяты из [133]. Другие таблицы см. в [19], [65], [74].

Пример на считывание таблицы. Случайная величина, имеющая F-распределение с числом степеней свободы $n_1 = 4$ и $n_2 = 9$ с вероятностью 0.10 превышает значение $F(n_1, n_2, \alpha) = 2.693$ (см. табл. 4.1). Другими словами, следует отвергнуть нулевую гипотезу, приводящую к F-распределению статистики критерия на 10% уровне значимости (то есть при $\alpha = 0.10$), если рассчитанное значение этой статистики равно или превысило 2.693 .

Таблица 4.1

Верхние 10% точки $F(n_1, n_2, 0.10)$ F-распределения F_{n_1, n_2}

$n_2 \setminus n_1$	1	2	3	4	5	6
1	39.86	49.50	53.59	55.83	57.24	58.20
2	8.526	9.000	9.162	9.243	9.293	9.326
3	5.538	5.462	5.391	5.343	5.309	5.285
4	4.545	4.325	4.191	4.107	4.051	4.010
5	4.060	3.780	3.619	3.520	3.453	3.405
6	3.776	3.463	3.289	3.181	3.108	3.055
7	3.589	3.257	3.074	2.961	2.883	2.827
8	3.458	3.113	2.924	2.806	2.726	2.668
9	3.360	3.006	2.813	2.693	2.611	2.551
10	3.285	2.924	2.728	2.605	2.522	2.461
11	3.225	2.860	2.660	2.536	2.451	2.389
12	3.177	2.807	2.606	2.480	2.394	2.331
13	3.136	2.763	2.560	2.434	2.347	2.283
14	3.102	2.726	2.522	2.395	2.307	2.243
15	3.073	2.695	2.490	2.361	2.273	2.208
16	3.048	2.668	2.462	2.333	2.244	2.178
17	3.026	2.645	2.437	2.308	2.218	2.152
18	3.007	2.624	2.416	2.286	2.196	2.130
19	2.990	2.606	2.397	2.266	2.176	2.109
20	2.975	2.589	2.380	2.249	2.158	2.091
21	2.961	2.575	2.365	2.233	2.142	2.075
22	2.949	2.561	2.351	2.219	2.128	2.060
23	2.973	2.549	2.339	2.207	2.115	2.047
24	2.927	2.538	2.327	2.195	2.103	2.035
25	2.918	2.528	2.317	2.184	2.092	2.024
26	2.909	2.519	2.307	2.174	2.082	2.014
27	2.901	2.511	2.299	2.165	2.073	2.005
28	2.894	2.503	2.291	2.157	2.064	1.996
29	2.887	2.495	2.283	2.149	2.057	1.988
30	2.881	2.489	2.276	2.142	2.049	1.980
32	2.869	2.477	2.263	2.129	2.036	1.967
34	2.869	2.466	2.252	2.118	2.024	1.955
36	2.850	2.456	2.243	2.108	2.014	1.945
38	2.842	2.448	2.234	2.099	2.005	1.935
40	2.835	2.440	2.226	2.091	1.997	1.927
60	2.791	2.393	2.177	2.041	1.946	1.875
80	2.769	2.370	2.154	2.017	1.921	1.849
100	2.756	2.356	2.139	2.002	1.906	1.834
120	2.748	2.347	2.130	1.992	1.896	1.824
140	2.742	2.341	2.123	1.985	1.889	1.817
160	2.737	2.336	2.118	1.980	1.884	1.811
180	2.734	2.332	2.114	1.976	1.880	1.807
200	2.731	2.329	2.111	1.973	1.876	1.804
∞	2.711	2.308	2.089	1.951	1.853	1.780

Таблица 4.1

Верхние 10% точки $F(n_1, n_2, 0.10)$ F-распределения F_{n_1, n_2}
(продолжение)

$n_2 \setminus n_1$	7	8	9	10	15	20
1	58.91	59.44	59.86	60.20	61.22	61.74
2	9.349	9.367	9.381	9.392	9.425	9.441
3	5.266	5.252	5.240	5.230	5.200	5.184
4	3.979	3.955	3.936	3.920	3.870	3.844
5	3.368	3.339	3.316	3.297	3.238	3.207
6	3.014	2.983	2.958	2.937	2.871	2.836
7	2.785	2.752	2.725	2.703	2.632	2.595
8	2.624	2.589	2.561	2.538	2.464	2.425
9	2.505	2.469	2.440	2.416	2.340	2.298
10	2.414	2.377	2.347	2.323	2.244	2.201
11	2.342	2.304	2.274	2.248	2.167	2.123
12	2.283	2.245	2.214	2.188	2.105	2.060
13	2.234	2.195	2.164	2.138	2.053	2.007
14	2.193	2.154	2.122	2.095	2.010	1.962
15	2.158	2.119	2.086	2.059	1.972	1.924
16	2.128	2.088	2.055	2.028	1.940	1.891
17	2.102	2.061	2.128	2.001	1.912	1.862
18	2.079	2.038	2.005	1.977	1.887	1.837
19	2.058	2.017	1.984	1.956	1.865	1.814
20	2.040	1.999	1.965	1.937	1.845	1.794
21	2.023	1.982	1.948	1.920	1.827	1.776
22	2.008	1.967	1.933	1.904	1.811	1.759
23	1.995	1.953	1.919	1.890	1.796	1.744
24	1.983	1.941	1.906	1.877	1.783	1.730
25	1.971	1.929	1.895	1.866	1.771	1.718
26	1.961	1.919	1.884	1.855	1.760	1.706
27	1.952	1.909	1.874	1.845	1.749	1.695
28	1.943	1.900	1.865	1.836	1.740	1.685
29	1.935	1.892	1.857	1.827	1.731	1.676
30	1.927	1.884	1.849	1.819	1.722	1.667
32	1.913	1.870	1.835	1.805	1.707	1.652
34	1.901	1.858	1.822	1.793	1.694	1.638
36	1.891	1.847	1.812	1.781	1.682	1.626
38	1.881	1.838	1.802	1.772	1.672	1.615
40	1.873	1.829	1.793	1.763	1.662	1.605
60	1.819	1.775	1.738	1.707	1.603	1.544
80	1.793	1.748	1.711	1.680	1.574	1.513
100	1.778	1.732	1.695	1.663	1.557	1.494
120	1.767	1.722	1.684	1.652	1.545	1.482
140	1.760	1.715	1.677	1.645	1.537	1.473
160	1.755	1.709	1.671	1.639	1.531	1.467
180	1.750	1.705	1.667	1.634	1.526	1.462
200	1.747	1.701	1.663	1.631	1.522	1.458
∞	1.723	1.676	1.638	1.605	1.494	1.428

Таблица 4.1

Верхние 10% точки $F(n_1, n_2, 0.10)$ F-распределения F_{n_1, n_2}
(окончание)

$n_2 \setminus n_1$	24	30	40	60	120	∞
1	62.00	62.26	62.53	62.79	63.06	63.33
2	9.450	9.458	9.466	9.475	9.483	9.491
3	5.176	5.168	5.160	5.151	5.142	5.134
4	3.831	3.817	3.804	3.790	3.775	3.761
5	3.191	3.174	3.157	3.140	3.123	3.105
6	2.818	2.800	2.781	2.762	2.742	2.722
7	2.575	2.556	2.535	2.514	2.493	2.471
8	2.404	2.383	2.361	2.339	2.316	2.293
9	2.277	2.255	2.232	2.208	2.184	2.159
10	2.178	2.155	2.132	2.107	2.082	2.055
11	2.100	2.076	2.052	2.026	2.000	1.972
12	2.036	2.011	1.986	1.960	1.932	1.904
13	1.983	1.958	1.932	1.904	1.876	1.846
14	1.938	1.912	1.885	1.857	1.828	1.797
15	1.899	1.873	1.845	1.817	1.787	1.755
16	1.866	1.839	1.811	1.782	1.751	1.718
17	1.836	1.809	1.781	1.751	1.719	1.686
18	1.810	1.783	1.754	1.723	1.692	1.657
19	1.787	1.759	1.730	1.699	1.666	1.631
20	1.767	1.738	1.708	1.677	1.643	1.607
21	1.748	1.719	1.689	1.657	1.623	1.586
22	1.731	1.702	1.671	1.639	1.604	1.567
23	1.716	1.686	1.655	1.622	1.587	1.549
24	1.702	1.672	1.641	1.607	1.572	1.533
25	1.689	1.659	1.627	1.593	1.557	1.518
26	1.677	1.648	1.615	1.581	1.544	1.504
27	1.666	1.636	1.603	1.569	1.531	1.491
28	1.656	1.625	1.592	1.558	1.520	1.478
29	1.646	1.616	1.582	1.547	1.509	1.467
30	1.638	1.607	1.573	1.538	1.499	1.456
32	1.622	1.590	1.556	1.520	1.481	1.437
34	1.608	1.576	1.542	1.505	1.464	1.420
36	1.595	1.563	1.528	1.491	1.450	1.404
38	1.584	1.551	1.516	1.478	1.437	1.390
40	1.574	1.541	1.506	1.467	1.425	1.377
60	1.511	1.476	1.437	1.395	1.348	1.292
80	1.479	1.443	1.403	1.358	1.307	1.245
100	1.460	1.423	1.382	1.336	1.282	1.214
120	1.447	1.409	1.368	1.320	1.265	1.193
140	1.438	1.400	1.358	1.309	1.252	1.176
160	1.431	1.393	1.350	1.301	1.242	1.163
180	1.426	1.387	1.344	1.294	1.235	1.153
200	1.422	1.383	1.339	1.289	1.228	1.144
∞	1.383	1.350	1.304	1.240	1.169	1.000

Таблица 4.2

Верхние 5% точки $F(n_1, n_2, 0.05)$ F-распределения F_{n_1, n_2}

$n_2 \setminus n_1$	1	2	3	4	5	6
1	161.4	199.5	215.7	224.6	230.2	234.0
2	18.51	19.00	19.16	19.25	19.30	19.33
3	10.13	9.552	9.277	9.117	9.013	8.941
4	7.709	6.944	6.591	6.388	6.256	6.163
5	6.608	5.786	5.409	5.192	5.050	4.950
6	5.987	5.143	4.757	4.534	4.387	4.284
7	5.591	4.737	4.347	4.120	3.972	3.866
8	5.318	4.459	4.066	3.838	3.687	3.581
9	5.117	4.256	3.863	3.633	3.482	3.374
10	4.965	4.103	3.708	3.478	3.326	3.217
11	4.844	3.982	3.587	3.357	3.204	3.095
12	4.747	3.885	3.490	3.259	3.106	2.996
13	4.667	3.806	3.411	3.179	3.025	2.915
14	4.600	3.739	3.344	3.112	2.958	2.848
15	4.543	3.682	3.287	3.056	2.901	2.790
16	4.494	3.634	3.239	3.007	2.852	2.741
17	4.451	3.592	3.197	2.965	2.810	2.699
18	4.414	3.555	3.160	2.928	2.773	2.661
19	4.381	3.522	3.127	2.895	2.740	2.628
20	4.351	3.493	3.098	2.866	2.711	2.599
21	4.325	3.467	3.072	2.840	2.685	2.573
22	4.301	3.443	3.049	2.817	2.661	2.549
23	4.279	3.422	3.028	2.796	2.640	2.528
24	4.260	3.413	3.009	2.776	2.621	2.508
25	4.242	3.385	2.991	2.759	2.603	2.490
26	4.225	3.369	2.975	2.743	2.587	2.474
27	4.210	3.354	2.960	2.728	2.572	2.459
28	4.196	3.340	2.947	2.714	2.558	2.445
29	4.183	3.328	2.934	2.701	2.545	2.432
30	4.171	3.316	2.922	2.690	2.534	2.421
32	4.149	3.295	2.901	2.668	2.512	2.399
34	4.130	3.276	2.883	2.650	2.494	2.380
36	4.113	3.259	2.866	2.634	2.477	2.364
38	4.098	3.245	2.852	2.619	2.463	2.349
40	4.085	3.232	2.839	2.606	2.449	2.336
60	4.001	3.150	2.758	2.525	2.368	2.254
80	3.960	3.111	2.719	2.486	2.329	2.214
100	3.936	3.087	2.696	2.463	2.305	2.191
120	3.920	3.072	2.680	2.447	2.290	2.175
140	3.909	3.061	2.669	2.436	2.279	2.164
160	3.900	3.052	2.661	2.428	2.271	2.156
180	3.894	3.046	2.655	2.422	2.264	2.149
200	3.888	3.041	2.650	2.417	2.259	2.144
∞	3.851	3.005	2.614	2.381	2.223	2.108

Таблица 4.2

Верхние 5% точки $F(n_1, n_2, 0.05)$ F-распределения F_{n_1, n_2}
(продолжение)

$n_2 \setminus n_1$	7	8	9	10	15	20
1	236.8	238.9	240.5	241.9	245.9	248.0
2	19.35	19.37	19.38	19.40	19.43	19.45
3	8.887	8.845	8.812	8.786	8.703	8.660
4	6.094	6.041	5.999	5.964	5.858	5.802
5	4.876	4.818	4.772	4.735	4.619	4.558
6	4.207	4.147	4.099	4.060	3.938	3.874
7	3.787	3.726	3.677	3.637	3.511	3.444
8	3.500	3.438	3.388	3.347	3.218	3.150
9	3.293	3.230	3.179	3.137	3.006	2.936
10	3.135	3.072	3.020	2.978	2.845	2.774
11	3.012	2.948	2.896	2.854	2.718	2.646
12	2.913	2.849	2.796	2.753	2.617	2.544
13	2.832	2.767	2.714	2.671	2.533	2.459
14	2.764	2.699	2.646	2.602	2.463	2.388
15	2.707	2.641	2.588	2.544	2.403	2.328
16	2.657	2.591	2.538	2.494	2.352	2.276
17	2.614	2.548	2.494	2.450	2.308	2.230
18	2.577	2.510	2.456	2.412	2.269	2.191
19	2.544	2.477	2.423	2.378	2.234	2.156
20	2.514	2.447	2.393	2.348	2.203	2.124
21	2.488	2.420	2.366	2.321	2.176	2.096
22	2.464	2.397	2.342	2.297	2.151	2.071
23	2.442	2.375	2.320	2.275	2.128	2.048
24	2.423	2.355	2.300	2.255	2.108	2.027
25	2.405	2.337	2.282	2.236	2.089	2.008
26	2.388	2.321	2.266	2.220	2.072	1.990
27	2.373	2.305	2.250	2.204	2.056	1.974
28	2.359	2.291	2.236	2.190	2.041	1.959
29	2.346	2.278	2.223	2.177	2.028	1.945
30	2.334	2.266	2.211	2.165	2.015	1.932
32	2.313	2.244	2.189	2.142	1.992	1.908
34	2.294	2.225	2.170	2.123	1.972	1.888
36	2.277	2.209	2.153	2.106	1.954	1.870
38	2.262	2.194	2.138	2.091	1.939	1.853
40	2.249	2.180	2.124	2.077	1.924	1.839
60	2.167	2.097	2.040	1.993	1.836	1.748
80	2.126	2.056	1.999	1.951	1.793	1.703
100	2.102	2.032	1.975	1.927	1.768	1.676
120	2.087	2.016	1.959	1.910	1.751	1.659
140	2.076	2.005	1.947	1.899	1.738	1.646
160	2.067	1.997	1.939	1.890	1.729	1.637
180	2.061	1.990	1.932	1.884	1.722	1.629
200	2.056	1.985	1.927	1.878	1.717	1.623
∞	2.019	1.948	1.889	1.840	1.676	1.581

Таблица 4.2

Верхние 5% точки $F(n_1, n_2, 0.05)$ F-распределения F_{n_1, n_2}
(окончание)

$n_2 \setminus n_1$	24	30	40	60	120	∞
1	249.1	250.1	251.1	252.2	253.2	254.3
2	19.45	19.46	19.47	19.48	19.49	19.50
3	8.638	8.617	8.594	8.572	8.549	8.526
4	5.774	5.746	5.717	5.688	5.658	5.628
5	4.527	4.496	4.464	4.431	4.399	4.365
6	3.842	3.808	3.774	3.740	3.705	3.669
7	3.411	3.376	3.340	3.304	3.267	3.230
8	3.115	3.079	3.043	3.005	2.967	2.928
9	2.901	2.864	2.826	2.787	2.748	2.707
10	2.737	2.700	2.661	2.621	2.580	2.538
11	2.609	2.571	2.531	2.490	2.448	2.404
12	2.506	2.466	2.426	2.384	2.341	2.296
13	2.420	2.380	2.339	2.297	2.252	2.206
14	2.349	2.308	2.266	2.223	2.178	2.131
15	2.288	2.247	2.204	2.160	2.114	2.066
16	2.235	2.194	2.151	2.106	2.059	2.010
17	2.190	2.148	2.104	2.058	2.011	1.964
18	2.150	2.107	2.063	2.017	1.968	1.917
19	2.114	2.071	2.026	1.980	1.930	1.878
20	2.083	2.039	1.994	1.946	1.896	1.843
21	2.054	2.010	1.965	1.916	1.866	1.812
22	2.028	1.984	1.938	1.889	1.838	1.783
23	2.005	1.961	1.914	1.865	1.813	1.757
24	1.984	1.939	1.892	1.842	1.890	1.733
25	1.964	1.919	1.872	1.822	1.768	1.711
26	1.946	1.901	1.853	1.803	1.749	1.691
27	1.930	1.884	1.836	1.785	1.731	1.672
28	1.915	1.869	1.820	1.769	1.714	1.654
29	1.901	1.854	1.806	1.754	1.698	1.638
30	1.887	1.841	1.792	1.740	1.684	1.622
32	1.864	1.817	1.767	1.714	1.657	1.594
34	1.843	1.795	1.745	1.691	1.633	1.569
36	1.824	1.776	1.726	1.671	1.612	1.547
38	1.808	1.760	1.708	1.653	1.594	1.527
40	1.793	1.744	1.693	1.637	1.577	1.509
60	1.700	1.649	1.594	1.534	1.467	1.389
80	1.654	1.602	1.545	1.482	1.411	1.325
100	1.627	1.573	1.515	1.450	1.376	1.283
120	1.608	1.554	1.495	1.429	1.352	1.254
140	1.595	1.541	1.481	1.414	1.335	1.232
160	1.586	1.531	1.470	1.402	1.321	1.214
180	1.578	1.523	1.462	1.393	1.311	1.200
200	1.572	1.516	1.455	1.386	1.302	1.189
∞	1.517	1.471	1.406	1.332	1.221	1.000

Таблица 4.3

Верхние 1% точки $F(n_1, n_2, 0.01)$ F-распределения F_{n_1, n_2}

$n_2 \setminus n_1$	1	2	3	4	5	6
1	4052.	4999.	5403.	5625.	5764.	5859.
2	98.50	99.00	99.17	99.25	99.30	99.33
3	34.12	30.82	29.46	28.71	28.24	27.91
4	21.20	18.00	16.69	15.98	15.52	15.21
5	16.26	13.27	12.06	11.39	10.97	10.67
6	13.75	10.92	9.780	9.148	8.746	8.466
7	12.25	9.547	8.451	7.847	7.460	7.191
8	11.26	8.649	7.591	7.006	6.632	6.371
9	10.56	8.022	6.992	6.422	6.057	5.802
10	10.04	7.559	6.552	5.994	5.636	5.386
11	9.646	7.206	6.217	5.668	5.316	5.069
12	9.330	6.927	5.953	5.412	5.064	4.821
13	9.074	6.701	5.739	5.205	4.862	4.620
14	8.862	6.515	5.564	5.035	4.693	4.456
15	8.683	6.359	5.417	4.893	4.556	4.318
16	8.531	6.226	5.292	4.773	4.437	4.202
17	8.400	6.112	5.185	4.669	4.336	4.102
18	8.285	6.013	5.092	4.579	4.248	4.015
19	8.185	5.926	5.010	4.500	4.171	3.939
20	8.096	5.849	4.938	4.431	4.103	3.871
21	8.017	5.780	4.874	4.369	4.042	3.812
22	7.945	5.719	4.817	4.313	3.988	3.758
23	7.881	5.664	4.765	4.264	3.939	3.710
24	7.823	5.614	4.718	4.218	3.895	3.667
25	7.770	5.568	4.675	4.177	3.855	3.627
26	7.721	5.526	4.637	4.140	3.818	3.591
27	7.677	5.488	4.601	4.106	3.785	3.558
28	7.636	5.453	4.568	4.074	3.754	3.528
29	7.598	5.420	4.538	4.045	3.725	3.499
30	7.562	5.390	4.510	4.018	3.699	3.473
32	7.499	5.336	4.459	3.969	3.652	3.427
34	7.444	5.289	4.416	3.927	3.611	3.386
36	7.396	5.248	4.377	3.890	3.574	3.351
38	7.353	5.211	4.343	3.858	3.542	3.319
40	7.314	5.179	4.313	3.828	3.514	3.291
60	7.077	4.977	4.126	3.649	3.339	3.119
80	6.963	4.881	4.036	3.563	3.255	3.036
100	6.895	4.824	3.984	3.513	3.206	2.988
120	6.851	4.787	3.949	3.480	3.174	2.956
140	6.819	4.760	3.925	3.456	3.151	2.933
160	6.796	4.740	3.906	3.439	3.134	2.917
180	6.778	4.725	3.892	3.425	3.121	2.904
200	6.763	4.713	3.881	3.414	3.110	2.894
∞	6.660	4.626	3.801	3.338	3.036	2.820

Таблица 4.3

Верхние 1% точки $F(n_1, n_2, 0.01)$ F-распределения F_{n_1, n_2}
(продолжение)

$n_2 \setminus n_1$	7	8	9	10	15	20
1	5928.	5981.	6022.	6056.	6157.	6209.
2	99.36	99.37	99.39	99.40	99.43	99.45
3	27.67	27.49	27.34	27.23	26.87	26.69
4	14.98	14.80	14.66	14.55	14.20	14.02
5	10.46	10.29	10.16	10.05	9.722	9.553
6	8.260	8.102	7.976	7.874	7.559	7.396
7	6.993	6.840	6.719	6.620	6.314	6.155
8	6.178	6.029	5.911	5.814	5.515	5.359
9	5.613	5.467	5.351	5.257	4.962	4.808
10	5.200	5.057	4.942	4.849	4.558	4.405
11	4.886	4.744	4.632	4.539	4.251	4.099
12	4.640	4.499	4.388	4.296	4.010	3.858
13	4.441	4.302	4.191	4.100	3.815	3.665
14	4.278	4.140	4.030	3.939	3.656	3.505
15	4.142	4.004	3.895	3.805	3.522	3.372
16	4.026	3.890	3.780	3.691	3.409	3.259
17	3.927	3.791	3.682	3.593	3.312	3.162
18	3.841	3.705	3.597	3.508	3.227	3.077
19	3.765	3.631	3.522	3.434	3.153	3.003
20	3.699	3.564	3.457	3.368	3.088	2.938
21	3.640	3.506	3.398	3.310	3.030	2.880
22	3.587	3.453	3.346	3.258	2.978	2.827
23	3.539	3.406	3.299	3.211	2.931	2.781
24	3.496	3.363	3.256	3.168	2.889	2.738
25	3.457	3.324	3.217	3.129	2.850	2.699
26	3.421	3.288	3.182	3.094	2.815	2.664
27	3.388	3.256	3.149	3.062	2.783	2.632
28	3.358	3.226	3.120	3.032	2.753	2.602
29	3.330	3.198	3.092	3.005	2.726	2.574
30	3.304	3.173	3.066	2.979	2.700	2.549
32	3.258	3.127	3.021	2.934	2.655	2.503
34	3.218	3.087	2.981	2.894	2.615	2.463
36	3.183	3.052	2.946	2.859	2.580	2.428
38	3.152	3.021	2.915	2.828	2.549	2.397
40	3.124	2.993	2.888	2.801	2.522	2.369
60	2.953	2.823	2.718	2.632	2.352	2.198
80	2.871	2.742	2.637	2.551	2.271	2.115
100	2.823	2.694	2.590	2.503	2.223	2.067
120	2.792	2.663	2.559	2.472	2.336	2.034
140	2.770	2.641	2.536	2.536	2.450	2.012
160	2.753	2.624	2.520	2.433	2.153	1.995
180	2.740	2.611	2.507	2.421	2.140	1.982
200	2.730	2.601	2.497	2.411	2.129	1.971
∞	2.657	2.529	2.425	2.339	2.056	1.897

Таблица 4.3

Верхние 1% точки $F(n_1, n_2, 0.01)$ F-распределения F_{n_1, n_2}
(окончание)

$n_2 \setminus n_1$	24	30	40	60	120	∞
1	6234.	6261.	6287.	6313.	6339.	6366.
2	99.46	99.47	99.47	99.48	99.49	99.50
3	26.60	26.51	26.41	26.32	26.22	26.12
4	13.93	13.84	13.74	13.65	13.56	13.46
5	9.466	9.379	9.291	9.202	9.112	9.020
6	7.313	7.228	7.143	7.057	6.969	6.880
7	6.074	5.992	5.908	5.824	5.737	5.650
8	5.279	5.198	5.116	5.032	4.946	4.859
9	4.729	4.649	4.567	4.483	4.398	4.311
10	4.327	4.247	4.165	4.082	3.997	3.909
11	4.021	3.941	3.860	3.776	3.690	3.602
12	3.781	3.701	3.619	3.536	3.449	3.361
13	3.587	3.507	3.425	3.341	3.255	3.165
14	3.427	3.348	3.266	3.181	3.094	3.004
15	3.294	3.214	3.132	3.047	2.960	2.868
16	3.181	3.101	3.018	2.933	2.845	2.753
17	3.084	3.003	2.921	2.834	2.746	2.653
18	2.999	2.919	2.835	2.749	2.660	2.566
19	2.925	2.844	2.761	2.674	2.584	2.489
20	2.859	2.778	2.695	2.608	2.517	2.421
21	2.801	2.720	2.636	2.548	2.457	2.360
22	2.749	2.668	2.583	2.495	2.403	2.306
23	2.702	2.620	2.536	2.447	2.354	2.256
24	2.659	2.577	2.492	2.403	2.310	2.211
25	2.620	2.538	2.453	2.364	2.270	2.169
26	2.585	2.503	2.417	2.327	2.233	2.132
27	2.552	2.470	2.384	2.294	2.199	2.096
28	2.522	2.440	2.354	2.263	2.167	2.064
29	2.495	2.412	2.325	2.234	2.138	2.034
30	2.469	2.386	2.299	2.208	2.111	2.006
32	2.423	2.340	2.252	2.160	2.062	1.956
34	2.383	2.299	2.211	2.118	2.019	1.911
36	2.347	2.263	2.175	2.182	1.982	1.872
38	2.316	2.232	2.143	2.049	1.948	1.836
40	2.288	2.203	2.114	2.019	1.917	1.805
60	2.115	2.228	1.936	1.836	1.726	1.601
80	2.032	1.944	1.849	1.746	1.631	1.494
100	1.983	1.893	1.797	1.692	1.572	1.427
120	1.950	1.860	1.763	1.656	1.533	1.381
140	1.927	1.836	1.738	1.630	1.505	1.346
160	1.910	1.819	1.720	1.610	1.483	1.318
180	1.896	1.805	1.706	1.595	1.466	1.297
200	1.886	1.794	1.694	1.583	1.453	1.278
∞	1.791	1.716	1.613	1.473	1.325	1.000

П5. Верхние процентные точки биномиального распределения вероятностей

Описание таблицы. В таблице для числа испытаний Бернулли $n = 10(1)25$ и вероятности «успеха» 0.5 даны верхние процентные точки $S(\alpha, n)$ для числа «успехов». Если упомянутое число «успехов» обозначить через S_n , то $S(\alpha, n)$ для данного α можно определить как решение уравнения:

$$P(S_n \geq S(\alpha, n)) = \alpha.$$

Распределение случайной величины S_n дискретно, и решение указанного уравнения существует лишь для некоторых α . Поэтому в следующей таблице приводятся решения этого уравнения для тех α , которые близки к назначенным уровням значимости $\alpha = 0.10, 0.05$ и 0.01 . Числа в таблице расположены в три столбца, соответствующие упомянутым значения α . Внутри каждого столбца слева в строке указаны значения $S(\alpha, n)$, где α близко к назначенному уровню; справа даны точные значения α .

Пример на считывание таблицы. Для $n = 15$ испытаний Бернулли следует на уровне $\alpha = 0.059$ отвергнуть гипотезу о том, что вероятность «успеха» $p = 0.5$ (нулевая гипотеза) против правосторонней альтернативы $p > 0.5$, если выборочное значение S_n превосходит или равно $S(\alpha, n) = 11$. В случае левосторонней альтернативы (т.е. против предположения, что $p < 0.5$) нулевую гипотезу следует отвергнуть на том же уровне значимости, если S_n меньше или равно $n - S(\alpha, n) = 15 - 11 = 4$. Против двусторонней альтернативы (т.е. предположения, что p отлично от 0.5) нулевую гипотезу следует отвергнуть, если $S_n \leq 4$ или $S_n \geq 11$, но уже на уровне значимости $2\alpha = 0.118$.

Таблица 5

Верхние процентные точки биномиального распределения
(для вероятности «успеха» $p = 0.5$)

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
10	7 .1719	8 .0547	9 .0107
11	8 .1113	9 .0327	10 .0059
12	9 .0730	10 .0193	11 .0032
13	9 .1334	10 .0461	11 .0112
14	10 .0898	11 .0287	12 .0065
15	10 .1509	11 .0592	12 .0176
16	11 .1051	12 .0384	13 .0106
17	12 .0717	13 .0245	14 .0064
18	12 .1189	13 .0481	14 .0154
19	13 .0835	14 .0318	15 .0096
20	13 .1316	14 .0577	15 .0207
21	14 .0946	15 .0392	16 .0133
22	14 .1431	15 .0669	16 .0233
23	15 .1050	16 .0466	17 .0173
24	15 .1537	16 .0758	18 .0320
25	16 .1148	17 .0539	19 .0216

П6. Верхние критические значения для статистики Уилкоксона

Описание таблицы. В таблице для объемов первой выборки m и второй выборки n , $2 \leq n \leq m \leq 10$ и $m = 11(1)20$, $n = 2, 3, 4$ приведены верхние критические значения $W(\alpha, m, n)$ для двухвыборочной статистики ранговых сумм Уилкоксона при проверке гипотезы однородности против правосторонней альтернативы (см. п. 3.5.2).

Если для данных $m \geq n$ через W обозначить упомянутую статистику и предположить, что выборки однородны, то $W(\alpha, m, n)$ для данного α можно определить как решение уравнения

$$P(W \geq W(\alpha, m, n)) = \alpha.$$

Нижние критические значения уровня α вычисляются по формуле $n(n + m + 1) - W(\alpha, m, n)$.

Так как распределение случайной величины W дискретно, то решения указанного уравнения существуют лишь для некоторых α , $0 < \alpha < 1$.

Данные в таблице сгруппированы в три столбца. Внутри каждого столбца слева в строке приведены значения $W(\alpha, m, n)$, где α близко к назначенному уровню (т.е. к 0.10, 0.05 или 0.01 соответственно столбцу); справа даны точные значения α . Пропуски в таблице означают, что для данных α , m , n указанное уравнение не имеет даже приближенного решения.

В переработанном виде таблица взята из [115]. Другие таблицы см. в [19], [48] [142].

Пример на считывание таблицы. При проверке гипотезы однородности (нулевой гипотезы) двух выборок объемов $m = 13$, $n = 4$ против правосторонней альтернативы, гипотезу следует отвергнуть на уровне значимости 0.011 (или, приближенно, на уровне 0.01), если статистика W превышает или равна 56. Против левосторонней альтернативы нулевую гипотезу следует отвергнуть (на том же уровне), если W меньше или равно $n(n + m + 1) - W(\alpha, n, m) = 4 * (4 + 13 + 1) - 56 = 16$.

Таблица 6

Верхние критические значения $W(\alpha, n, m)$ статистики Уилкоксона W

n	m	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	
2	2				
2	3	9	.100		
2	4	10	.133	11 .067	
2	5	12	.095	13 .048	
2	6	14	.071	15 .036	
2	7	15	.111	16 .056	17 .028
2	8	17	.089	18 .044	19 .022
2	9	18	.109	20 .036	21 .018
2	10	20	.091	21 .061	23 .015
2	11	21	.115	23 .051	25 .013
2	12	23	.099	25 .044	27 .011
2	13	24	.114	26 .057	29 .010
2	14	26	.100	28 .050	31 .008

Таблица 6

Верхние критические значения $W(\alpha, n, m)$ статистики Уилкоксона W
(продолжение)

n	m	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
2	15	28	.088	30	.044	33	.007
2	16	29	.105	31	.059	35	.007
2	17	31	.094	33	.053	36	.012
2	18	32	.105	35	.047	38	.011
2	19	34	.095	37	.043	40	.010
2	20	35	.108	38	.052	42	.009
3	3	14	.100	15	.050		
3	4	16	.114	17	.057	18	.029
3	5	18	.125	20	.036	21	.018
3	6	21	.083	22	.048	24	.012
3	7	23	.092	24	.058	27	.008
3	8	25	.097	27	.042	29	.012
3	9	27	.105	29	.050	32	.009
3	10	29	.108	31	.056	35	.007
3	11	31	.113	34	.044	37	.011
3	12	34	.090	36	.051	40	.009
3	13	36	.095	38	.055	42	.012
3	14	38	.099	41	.046	45	.010
3	15	40	.102	43	.050	48	.009
3	16	42	.105	45	.055	50	.011
3	17	44	.108	48	.046	53	.010
3	18	47	.092	50	.050	56	.008
3	19	49	.095	52	.054	58	.010
3	20	51	.098	55	.047	61	.009
4	4	23	.100	24	.057	26	.014
4	5	26	.095	27	.056	30	.008
4	6	29	.086	30	.057	33	.010
4	7	31	.115	33	.055	36	.012
4	8	34	.107	36	.055	40	.008
4	9	37	.099	39	.053	43	.010
4	10	40	.094	42	.053	46	.012
4	11	43	.089	45	.052	50	.009
4	12	45	.106	48	.052	53	.010
4	13	48	.101	51	.051	56	.011
4	14	51	.096	54	.051	60	.009
4	15	54	.092	57	.050	63	.010
4	16	56	.106	60	.050	66	.011
4	17	59	.101	63	.049	70	.009
4	18	62	.098	66	.049	73	.010
4	19	65	.094	69	.049	76	.011
4	20	67	.105	72	.048	80	.009
5	5	34	.111	36	.048	39	.008
5	6	38	.089	40	.041	43	.009
5	7	41	.101	43	.053	47	.009
5	8	44	.111	47	.047	51	.009

Верхние критические значения $W(\alpha, n, m)$ статистики Уилкоксона W
(окончание)

n	m	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
5	9	48 .095	50 .056	55 .009
5	10	51 .103	54 .050	59 .010
6	6	48 .090	50 .047	54 .008
6	7	52 .090	54 .051	58 .011
6	8	56 .091	58 .054	63 .010
6	9	60 .091	53 .044	68 .009
6	10	64 .090	67 .047	72 .011
7	7	63 .104	66 .049	71 .009
7	8	68 .095	71 .047	76 .010
7	9	72 .105	76 .045	81 .011
7	10	77 .097	80 .054	87 .009
8	8	81 .097	84 .052	90 .010
8	9	86 .100	90 .046	96 .010
8	10	91 .102	95 .051	102 .010
9	9	101 .095	105 .047	112 .009
9	10	106 .106	111 .047	119 .009
10	10	123 .095	127 .053	136 .009

П7. Верхние критические значения статистики Краскела—Уоллиса для различных планов эксперимента

Описание таблицы. В таблице для заданного числа способов обработки $k = 3(1)6$ и заданного числа наблюдений для каждого из способов обработки n_i , $i = 1, \dots, k$ ($n_1 = 2(1)8, n_2 = 2(1)8, n_3 = 2(1)8$ при $k = 3$; $n_1 = 2(1)4, n_2 = 2(1)4, n_3 = 1(1)4, n_4 = 1(1)4$ при $k = 4$; $n_1 = 2(1)3, n_2 = 2(1)3, n_3 = 1(1)3, n_4 = 1(1)3, n_5 = 1(1)3$ при $k = 5$; $n_1 = 2(1)3, n_2 = 2(1)3, n_3 = 1(1)2, n_4 = 1(1)2, n_5 = 1(1)2, n_6 = 1(1)2$ при $k = 6$) указаны верхние критические значения $H(\alpha, k, n_1, \dots, n_k)$ для статистики Краскела—Уоллиса H при проверке гипотезы однородности.

Если для данных k, n_1, \dots, n_k через H обозначить упомянутую статистику и предположить, что выборки однородны, то $H(\alpha, k, n_1, \dots, n_k)$ для данного α можно определить как решение уравнения

$$P(H \geq H(\alpha, k, n_1, \dots, n_k)) = \alpha.$$

Так как распределение случайной величины H дискретно, то решение указанного уравнения существует лишь для некоторых α , $0 < \alpha < 1$. Поэтому в следующей ниже таблице приводятся решения этого уравнения для тех α , которые близки к назначенным уровням $\alpha = 0.10, 0.05, 0.025$ и 0.01 . Таким образом, проверки нулевых гипотез с помощью этих таблиц проводятся на уровнях значимости, лишь приближенных к номинальным. Для некоторых планов эксперимента и уровней значимости α решения, приближенные к номинальным

уровням значимости, не существуют. В соответствующих случаях в таблице стоят пропуски.

Для более обширных планов эксперимента, помеченных в таблице выражениями $n_1 = 99, n_2 = 99, \dots, n_k = 99$, при $k = 3(1)5$ приведены приближенные значения, основанные на аппроксимации распределения H распределением хи-квадрат с $(k - 1)$ степенями свободы.

Таблица взята из [133]. Другие таблицы см. в [115].

Пример на считывание таблицы. Для плана эксперимента $k = 3, n_1 = 5, n_2 = 4, n_3 = 4$ мы должны отвергнуть гипотезу об отсутствии эффектов обработки (нулевую гипотезу) на приближенном уровне значимости $\alpha = 0.05$, если вычисленное значение статистики Краскела—Уоллиса H равно или превышает значение $H(\alpha, k, n_1, \dots, n_k) = 5.657$.

Таблица 7

Верхние критические значения $H(\alpha, kn_1, \dots, n_k)$ статистики Краскела—Уоллиса H

n_1	n_2	n_3	n_4	n_5	n_6	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
2	2	2	0	0	0	4.571			
3	2	1	0	0	0	4.286			
3	2	2	0	0	0	4.500	4.714		
3	3	1	0	0	0	4.571	5.143		
3	3	2	0	0	0	4.556	5.361	5.556	
3	3	3	0	0	0	4.622	5.600	5.956	7.200
4	2	1	0	0	0	4.500			
4	2	2	0	0	0	4.458	5.333	5.500	
4	3	1	0	0	0	4.056	5.208	5.833	
4	3	2	0	0	0	4.511	5.444	6.000	6.444
4	3	3	0	0	0	4.709	5.791	6.155	6.745
4	4	2	0	0	0	4.555	5.455	6.327	7.036
4	4	3	0	0	0	4.545	5.598	6.394	7.144
4	4	4	0	0	0	4.654	5.692	6.615	7.654
5	2	1	0	0	0	4.200	5.000		
5	2	2	0	0	0	4.373	5.160	6.000	6.533
5	3	1	0	0	0	4.018	4.960	6.044	
5	3	2	0	0	0	4.651	5.251	6.004	6.909
5	3	3	0	0	0	4.533	5.648	6.315	7.079
5	4	1	0	0	0	3.987	4.985	5.858	6.955
5	4	2	0	0	0	4.541	5.273	6.068	7.205
5	4	3	0	0	0	4.549	5.656	6.410	7.445
5	4	4	0	0	0	4.668	5.657	6.673	7.760
5	5	1	0	0	0	4.109	5.127	6.000	7.309
5	5	2	0	0	0	4.623	5.338	6.346	7.338
5	5	3	0	0	0	4.545	5.705	6.549	7.578
5	5	4	0	0	0	4.523	5.666	6.760	7.823
5	5	5	0	0	0	4.560	5.780	6.740	8.000
6	2	1	0	0	0	4.200	4.822	5.600	
6	2	2	0	0	0	4.545	5.345	5.745	6.655
6	3	2	0	0	0	3.909	4.855	5.945	6.873
6	3	2	0	0	0	4.682	5.348	6.136	6.970

Таблица 7

Верхние критические значения $H(\alpha, kn_1, \dots, n_k)$ статистики Краскела–Уоллиса H
(продолжение)

n_1	n_2	n_3	n_4	n_5	n_6	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
6	3	3	0	0	0	4.590	5.615	6.436	7.410
6	4	1	0	0	0	4.038	4.947	5.856	7.106
6	4	2	0	0	0	4.494	5.340	6.186	7.340
6	4	3	0	0	0	4.604	5.610	6.538	7.500
6	4	4	0	0	0	4.595	5.681	6.667	7.795
6	5	1	0	0	0	4.128	4.990	5.951	7.182
6	5	2	0	0	0	4.596	5.338	6.196	7.376
6	5	3	0	0	0	4.535	5.602	6.667	7.590
6	5	4	0	0	0	4.522	5.661	6.750	7.936
6	5	5	0	0	0	4.547	5.729	6.788	8.028
6	6	1	0	0	0	4.000	4.945	5.923	7.121
6	6	2	0	0	0	4.438	5.410	6.210	7.467
6	6	3	0	0	0	4.558	5.625	6.725	7.725
6	6	4	0	0	0	4.548	5.724	6.812	8.000
6	6	5	0	0	0	4.542	5.765	6.848	8.124
6	6	6	0	0	0	4.643	5.801	6.889	8.222
7	1	1	0	0	0	4.267			
7	2	1	0	0	0	4.200	4.706	5.727	
7	2	2	0	0	0	4.526	5.143	5.818	7.000
7	3	1	0	0	0	4.173	4.952	5.758	7.030
7	3	2	0	0	0	4.502	5.357	6.201	6.839
7	3	3	0	0	0	4.603	5.620	6.449	7.228
7	4	1	0	0	0	4.121	4.986	5.791	6.986
7	4	2	0	0	0	4.549	5.376	6.184	7.321
7	4	3	0	0	0	4.527	5.623	6.578	7.550
7	4	4	0	0	0	4.562	5.650	6.707	7.814
7	5	1	0	0	0	4.035	5.064	5.953	7.061
7	5	2	0	0	0	4.485	5.393	6.221	7.450
7	5	3	0	0	0	4.535	5.607	6.627	7.697
7	5	4	0	0	0	4.542	5.733	6.738	7.931
7	5	5	0	0	0	4.571	5.708	6.835	8.108
7	6	1	0	0	0	4.033	5.067	6.067	7.254
7	6	2	0	0	0	4.500	5.357	6.223	7.490
7	6	3	0	0	0	4.550	5.689	6.694	7.756
7	6	4	0	0	0	4.562	5.706	6.787	8.039
7	6	5	0	0	0	4.560	5.770	6.857	8.157
7	6	6	0	0	0	4.530	5.730	6.897	8.257
7	7	1	0	0	0	3.986	4.986	6.057	7.157
7	7	2	0	0	0	4.491	5.398	6.328	7.491
7	7	3	0	0	0	4.613	5.688	6.708	7.810
7	7	4	0	0	0	4.563	5.766	6.788	8.142
7	7	5	0	0	0	4.546	5.746	6.886	8.257
7	7	6	0	0	0	4.568	5.793	6.927	8.345
7	7	7	0	0	0	4.594	5.818	6.954	8.378

Таблица 7

Верхние критические значения $H(\alpha, kn_1, \dots, n_k)$ статистики Краскела–Уоллиса H
(продолжение)

n_1	n_2	n_3	n_4	n_5	n_6	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
8	1	1	0	0	0	4.418			
8	2	1	0	0	0	4.011	4.909	5.420	
8	2	2	0	0	0	4.587	5.356	5.817	6.663
8	3	1	0	0	0	4.010	4.881	6.064	6.804
8	3	2	0	0	0	4.451	5.316	6.195	7.022
8	3	3	0	0	0	4.543	5.617	6.588	7.350
8	4	1	0	0	0	4.038	5.044	5.885	6.973
8	4	2	0	0	0	4.500	5.393	6.193	7.350
8	4	3	0	0	0	4.529	5.623	6.562	7.585
8	4	4	0	0	0	4.561	5.779	6.750	7.853
8	5	1	0	0	0	3.967	4.869	5.864	7.110
8	5	2	0	0	0	4.466	5.415	6.260	7.440
8	5	3	0	0	0	4.514	5.614	6.614	7.706
8	5	4	0	0	0	4.549	5.718	6.782	7.992
8	5	5	0	0	0	4.555	5.769	6.843	8.116
8	6	1	0	0	0	4.015	5.015	5.933	7.256
8	6	2	0	0	0	4.463	5.404	6.294	7.522
8	6	3	0	0	0	4.575	5.678	6.658	7.796
8	6	4	0	0	0	4.563	5.743	6.795	8.045
8	6	5	0	0	0	4.550	5.750	6.867	8.226
8	6	6	0	0	0	4.599	5.770	6.932	8.313
8	7	1	0	0	0	4.045	5.041	6.047	7.308
8	7	2	0	0	0	4.451	5.403	6.339	7.571
8	7	3	0	0	0	4.556	5.698	6.671	7.872
8	7	4	0	0	0	4.548	5.759	6.837	8.118
8	7	5	0	0	0	4.551	5.782	6.884	8.242
8	7	6	0	0	0	4.553	5.781	6.917	8.333
8	7	7	0	0	0	4.585	5.802	6.980	8.363
8	8	1	0	0	0	4.044	5.039	6.005	7.314
8	8	2	0	0	0	4.509	5.408	6.351	7.654
8	8	3	0	0	0	4.555	5.734	6.682	7.889
8	8	4	0	0	0	4.579	5.743	6.886	8.168
8	8	5	0	0	0	4.573	5.761	6.920	8.297
8	8	6	0	0	0	4.572	5.779	6.953	8.367
8	8	7	0	0	0	4.571	5.791	6.980	8.419
8	8	8	0	0	0	4.595	5.805	6.995	8.465
99	99	99	0	0	0	4.605	5.991	7.378	9.210
2	2	2	1	0	0	5.357	5.679		
2	2	2	2	0	0	5.667	6.167	6.667	6.667
3	2	1	1	0	0	5.143			
3	2	2	1	0	0	5.556	5.833	6.250	
3	2	2	2	0	0	5.644	6.333	6.978	7.133
3	3	1	1	0	0	5.333	6.333	6.333	
3	3	2	1	0	0	5.689	6.244	6.689	7.200
3	3	2	2	0	0	5.745	6.527	7.055	7.636

Таблица 7

Верхние критические значения $H(\alpha, kn_1, \dots, n_k)$ статистики Краскела–Уоллиса H
(продолжение)

n_1	n_2	n_3	n_4	n_5	n_6	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
3	3	3	1	0	0	5.655	6.600	7.036	7.400
3	3	3	2	0	0	5.879	6.727	7.515	8.015
3	3	3	3	0	0	6.026	7.000	7.667	8.538
4	2	1	1	0	0	5.250	5.833		
4	2	2	1	0	0	5.533	6.133	6.533	7.000
4	2	2	2	0	0	5.755	6.545	7.064	7.391
4	3	1	1	0	0	5.067	6.178	6.711	7.067
4	3	2	1	0	0	5.591	6.309	6.955	7.455
4	3	2	2	0	0	5.750	6.621	7.326	7.871
4	3	3	1	0	0	5.689	6.545	7.326	7.758
4	3	3	2	0	0	5.872	6.795	7.564	8.333
4	3	3	3	0	0	6.016	6.984	7.775	8.659
4	4	1	1	0	0	5.182	5.945	6.955	7.909
4	4	2	1	0	0	5.568	6.386	7.159	7.909
4	4	2	2	0	0	5.808	6.731	7.538	8.346
4	4	3	1	0	0	5.692	6.635	7.500	8.231
4	4	3	2	0	0	5.901	6.874	7.747	8.621
4	4	3	3	0	0	6.019	7.038	7.929	8.876
4	4	4	1	0	0	5.654	6.725	7.648	8.588
4	4	4	2	0	0	5.914	6.957	7.914	8.871
4	4	4	3	0	0	6.042	7.142	8.079	9.075
4	4	4	4	0	0	6.088	7.235	8.228	9.287
99	99	99	99	0	0	6.251	7.815	9.348	11.34
2	2	1	1	1	0	5.786			
2	2	2	1	1	0	6.250	6.750	6.750	
2	2	2	2	1	0	6.600	7.133	7.333	7.533
2	2	2	2	2	0	6.982	7.418	7.964	8.291
3	2	1	1	1	0	6.139	6.583		
3	2	2	1	1	0	6.511	6.800	7.200	7.600
3	2	2	2	1	0	6.709	7.309	7.745	8.127
3	2	2	2	2	0	6.955	7.682	8.182	8.682
3	3	1	1	1	0	6.311	7.111	7.467	
3	3	2	1	1	0	6.600	7.200	7.618	
3	3	2	2	1	0	6.788	7.591	8.121	8.576
3	3	2	2	2	0	7.026	7.910	8.538	9.115
3	3	3	1	1	0	6.788	7.576	8.061	8.424
3	3	3	2	1	0	6.910	7.769	8.449	9.051
3	3	3	2	2	0	7.121	8.044	8.813	9.505
3	3	3	3	1	0	7.077	8.000	8.703	9.451
3	3	3	3	2	0	7.210	8.200	9.038	9.876
3	3	3	3	3	0	7.333	8.333	9.233	10.20
99	99	99	99	99	0	7.779	9.488	11.14	13.28
2	2	1	1	1	1	6.833			
2	2	2	1	1	1	7.267	7.800		
2	2	2	2	1	1	7.527	8.018	8.345	8.618

Таблица 7

Верхние критические значения $H(\alpha, kn_1, \dots, n_k)$ статистики Краскела–Уоллиса H (окончание)

n_1	n_2	n_3	n_4	n_5	n_6	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
2	2	2	2	2	1	7.909	8.455	8.864	9.227
2	2	2	2	2	2	8.154	8.846	9.385	9.846
3	2	1	1	1	1	7.133	7.467	7.667	
3	2	2	1	1	1	7.418	7.945	8.236	8.509
3	2	2	2	1	1	7.727	8.348	8.727	9.136
3	2	2	2	2	1	7.987	8.731	9.218	9.692
3	2	2	2	2	2	8.198	9.033	9.648	10.22
3	3	1	1	1	1	7.400	7.909	8.564	8.564
3	3	2	1	1	1	7.697	8.303	8.667	9.045
3	3	2	2	1	1	7.872	8.615	9.128	9.628
3	3	2	2	2	1	8.077	8.923	9.549	10.15
3	3	2	2	2	2	8.305	9.190	9.914	10.61

П8. Верхние критические значения для статистики Фридмана

Описание таблицы. В таблице для заданного числа способов обработки $k = 3(17)$ и некоторого числа блоков n приведены верхние критические значения $S(\alpha, k, n)$ статистики Фридмана (см. п. 7.4.1).

Если для данных k и n обозначить через S упомянутую статистику Фридмана и предположить, что эффекты обработки отсутствуют (нулевая гипотеза), то $S(\alpha, k, n)$ для данного α можно определить как решение уравнения

$$P(S \geq S(\alpha, k, n)) = \alpha.$$

Так как распределение случайной величины S дискретно, то решение указанного уравнения существует лишь для некоторых α , $0 < \alpha < 1$. Поэтому в следующей ниже таблице приводятся решения этого уравнения для тех α , которые близки к назначенным уровням значимости $\alpha = 0.10, 0.05$ и 0.01 .

Числа в таблице расположены в три столбца, соответствующие упомянутым значениям α . Внутри каждого столбца слева в строке указаны значения $S(\alpha, k, n)$, где α близко к назначенному уровню; справа даны точные значения α . В том случае, когда решение указанного выше уравнения отсутствует для α , близких к номинальным уровням значимости, в таблице стоят пропуски.

В переработанном виде таблица взята из [32]. Другие таблицы см. в [115], [136], [137].

Пример на считывание таблицы. Для плана эксперимента с $k = 3$, $n = 13$ следует отвергнуть предположение об отсутствии эффектов обработки (нулевую гипотезу) на уровне значимости $\alpha = 0.05$, если вычисленное значение статистики Фридмана S равно или превышает значение $S(\alpha, k, n) = 6$.

Таблица 8

Верхние критические значения $S(\alpha, k, n)$ статистики Фридмана S

k	n	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
3	2	4.000	.167				
3	3	4.667	.194	6.000	.028	6.000	.028
3	4	4.500	.125	6.500	.042	8.000	.005
3	5	5.200	.093	6.400	.039	8.400	.008
3	6	5.333	.072	6.333	.052	9.000	.008
3	7	4.571	.112	6.000	.051	8.857	.008
3	8	4.750	.120	6.250	.047	9.000	.010
3	9	4.667	.107	6.222	.048	8.667	.010
3	10	5.000	.092	6.200	.046	8.600	.012
3	11	4.909	.100	6.545	.043	8.909	.011
3	12	4.667	.108	6.167	.051	8.667	.011
3	13	4.769	.098	6.000	.050	9.385	.009
3	14	5.143	.089	6.143	.049	9.000	.010
3	15	4.933	.096	6.400	.047	8.933	.010
3	16	4.875	.091	6.125	.052	9.125	.010
3	17	4.588	.105	6.118	.046	8.941	.010
3	18	4.778	.098	6.333	.045	9.000	.009
3	19	5.053	.092	6.000	.044	8.842	.011
3	20	4.900	.097	6.100	.052	9.100	.011
3	21	4.667	.108	6.000	.052	8.857	.011
3	22	4.727	.091	5.818	.052	9.091	.009
3	23	4.522	.106	5.826	.054	9.391	.009
3	24	4.750	.100	6.083	.053	9.083	.011
3	25	4.880	.097	6.080	.050	9.333	.010
4	2	5.400	.167	6.000	.042		
4	3	6.600	.075	7.000	.054	8.200	.017
4	4	6.000	.105	7.500	.052	9.300	.012
4	5	6.360	.093	7.320	.055	9.960	.009
4	6	6.400	.098	7.400	.056	10.200	.010
4	7	6.257	.100	7.629	.052	10.371	.010
4	8	6.300	.100	7.650	.049	10.500	.009
4	9	6.200	.098	7.667	.049	10.467	.010
4	10	6.240	.101	7.680	.047	10.680	.010
5	2	6.800	.117	7.600	.042	8.000	.008
5	3	7.467	.096	8.533	.045	10.133	.008
5	4	7.600	.095	8.800	.049	11.000	.010
5	5	7.680	.094	8.960	.049	11.680	.010
5	6	7.600	.102	9.067	.049	11.867	.010
5	7	7.057	.103	9.114	.049	12.114	.010
5	8	7.700	.100	9.200	.050	12.300	.010
6	2	8.286	.087	8.857	.051	9.714	.008
6	3	8.714	.095	9.857	.046	11.762	.009
6	4	8.857	.102	10.143	.052	12.714	.010
6	5	9.000	.099	10.371	.051	13.229	.010
6	6	9.048	.099	10.517	.049	13.619	.010
6	7	9.200	.098	10.476	.052	14.100	.009

Верхние критические значения $S(\alpha, k, n)$ статистики Фридмана S
(окончание)

k	n	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
6	8	9.000	.098	10.790	.050	13.860	.009
7	7	7.710	.098	9.550	.050	13.810	.010
7	8	7.850	.098	9.780	.049	13.690	.010

П9. Верхние критические значения для коэффициента ранговой корреляции Кендэла

Описание таблицы. В таблице для различных объемов выборок $n = 4(1)40$ приведены верхние критические значения $\tau(\alpha, n)$ для коэффициента ранговой корреляции Кендэла при проверке гипотезы о независимости двух признаков (нулевой гипотезы) против правосторонней альтернативы (т.е. предположения о положительной связи признаков) (см. п. 9.4).

Если для данного n через τ обозначить упомянутый коэффициент ранговой корреляции и предположить, что признаки независимы, то $\tau(\alpha, n)$ для данного α можно определить как решение уравнения $P(\tau \geq \tau(\alpha, n)) = \alpha$.

Нижние критические значения уровня α равны $-\tau(\alpha, n)$. При проверке нулевой гипотезы против двусторонней альтернативы (т.е. против предположения о какой-либо статистической связи между признаками) величина $\tau(\alpha, n)$ служит критическим значением уровня 2α для статистики $|\tau|$.

Так как распределение случайной величины τ дискретно, то решение указанного уравнения существует лишь для некоторых α , $0 < \alpha < 1$. Поэтому в следующей ниже таблице приводятся решения этого уравнения для тех α , которые близки к назначенным уровням значимости $\alpha = 0.10, 0.05$ и 0.01 . Числа в таблице расположены в три столбца, соответствующие упомянутым значениям α . Внутри каждого столбца слева в строке указаны значения $\tau(\alpha, n)$, где α близко к назначенному уровню; справа даны точные значения α .

Таблица в переработанном виде взята из [115]. Другие таблицы см. в [53].

Пример на считывание таблицы. Для выборок объема $n = 13$ следует отвергнуть гипотезу о независимости двух признаков против предположения об их положительной зависимости, на уровне значимости $\alpha = 0.05$, если вычисленное значение коэффициента ранговой корреляции Кендэла равно или превышает значение $\tau(\alpha, n) = 0.359$.

Таблица 9

Верхние критические значения $\tau(\alpha, n)$
коэффициента ранговой корреляции Кендэла τ

n	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
4	0.667	.167	1.000	.042		
5	0.600	.117	0.800	.042	1.000	.008
6	0.467	.136	0.600	.068	0.867	.010
7	0.429	.119	0.619	.035	0.810	.005

Таблица 9

Верхние критические значения $\tau(\alpha, n)$
коэффициента ранговой корреляции Кендэла τ (окончание)

n	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
8	0.429	.089	0.500	.054	0.714	.007
9	0.389	.090	0.444	.060	0.611	.010
10	0.333	.108	0.422	.054	0.600	.008
11	0.309	.109	0.418	.043	0.564	.008
12	0.303	.098	0.394	.043	0.515	.010
13	0.282	.102	0.359	.050	0.487	.010
14	0.275	.096	0.341	.050	0.473	.010
15	0.257	.101	0.333	.046	0.448	.010
16	0.250	.097	0.317	.048	0.433	.010
17	0.235	.102	0.309	.046	0.426	.010
18	0.229	.100	0.294	.048	0.412	.010
19	0.216	.100	0.287	.097	0.392	.010
20	0.211	.104	0.274	.049	0.379	.010
21	0.210	.098	0.267	.049	0.371	.010
22	0.203	.099	0.255	.051	0.359	.010
23	0.194	.104	0.249	.051	0.352	.010
24	0.188	.104	0.246	.048	0.341	.010
25	0.187	.101	0.240	.049	0.333	.010
26	0.182	.102	0.231	.052	0.323	.010
27	0.179	.099	0.225	.052	0.316	.010
28	0.175	.101	0.222	.051	0.312	.010
29	0.172	.099	0.222	.048	0.305	.010
30	0.168	.100	0.214	.051	0.301	.010
31	0.166	.099	0.211	.050	0.295	.010
32	0.161	.101	0.206	.051	0.290	.010
33	0.159	.100	0.205	.049	0.284	.010
34	0.155	.102	0.201	.049	0.280	.010
35	0.153	.101	0.197	.050	0.277	.010
36	0.152	.099	0.194	.050	0.273	.010
37	0.150	.098	0.189	.051	0.267	.010
38	0.147	.101	0.189	.049	0.263	.010
39	0.144	.101	0.185	.050	0.258	.010
40	0.144	.099	0.182	.050	0.256	.010

П10. Верхние критические значения для коэффициента ранговой корреляции Спирмена

Описание таблицы. В таблице для различных объемов выборок $n = 4(1)50(2)70$ приведены верхние критические значения $\rho(\alpha, n)$ коэффициента ранговой корреляции Спирмена при проверке гипотезы о независимости двух признаков (нулевой гипотезы) против правосторонней альтернативы (т.е. против предположения о положительной связи признаков) (см. п. 9.4).

Если для данного n через ρ обозначить упомянутый коэффициент ранговой корреляции и предположить, что признаки независимы, то $\rho(\alpha, n)$ для данного α можно определить как решение уравнения $P(\rho \geq \rho(\alpha, n)) = \alpha$.

Нижние критические значения уровня α равны $-\rho(\alpha, n)$. При проверке нулевой гипотезы против двусторонней альтернативы (т.е. против предположения о какой-либо связи между признаками) величина $\rho(\alpha, n)$ служит критическим значением уровня 2α для статистики $|\rho|$.

Так как распределение случайной величины ρ дискретно, то решение указанного уравнения существует лишь для некоторых α , $0 < \alpha < 1$. Поэтому в таблице в столбцах для $\rho(\alpha, n)$ представлены приближенные решения указанного уравнения.

Таблица взята из [115]. Другие таблицы см. в [79], [144].

Пример на считывание таблицы. Для выборок объема $n = 13$ следует отвергнуть гипотезу о независимости двух признаков против предположения об их положительной зависимости на уровне значимости $\alpha = 0.05$, если вычисленное значение коэффициента ранговой корреляции Спирмена ρ равно или превышает значение $\rho(\alpha, n) = 0.484$.

Таблица 10

Критические значения $\rho(\alpha, n)$ коэффициента ранговой корреляции Спирмена ρ

n	$\alpha = 0.25$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
4	0.600	1.000	1.000			
5	0.500	0.800	0.900	1.000	1.000	
6	0.371	0.657	0.829	0.886	0.943	1.000
7	0.321	0.571	0.714	0.786	0.893	0.929
8	0.310	0.524	0.643	0.738	0.833	0.881
9	0.267	0.483	0.600	0.700	0.783	0.833
10	0.248	0.455	0.564	0.648	0.745	0.794
11	0.236	0.427	0.536	0.618	0.709	0.755
12	0.224	0.406	0.503	0.587	0.671	0.727
13	0.209	0.385	0.484	0.560	0.648	0.703
14	0.200	0.367	0.464	0.538	0.622	0.675
15	0.189	0.354	0.443	0.521	0.604	0.654
16	0.182	0.341	0.429	0.503	0.582	0.635
17	0.176	0.328	0.414	0.485	0.566	0.615
18	0.170	0.317	0.401	0.472	0.550	0.600
19	0.165	0.309	0.391	0.460	0.535	0.584
20	0.161	0.299	0.380	0.447	0.520	0.570
21	0.156	0.292	0.370	0.435	0.508	0.556
22	0.152	0.284	0.361	0.425	0.496	0.544
23	0.148	0.278	0.353	0.415	0.486	0.532
24	0.144	0.271	0.344	0.406	0.475	0.521
25	0.142	0.265	0.337	0.398	0.466	0.511
26	0.138	0.259	0.331	0.390	0.457	0.501
27	0.136	0.255	0.324	0.382	0.448	0.491
28	0.133	0.250	0.317	0.375	0.440	0.483
29	0.130	0.245	0.312	0.368	0.433	0.475
30	0.128	0.240	0.306	0.362	0.425	0.467
31	0.126	0.236	0.301	0.356	0.418	0.459

Таблица 10

Критические значения $\rho(\alpha, n)$ коэффициента ранговой корреляции Спирмена ρ
(окончание)

n	$\alpha = 0.25$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
32	0.124	0.232	0.296	0.350	0.412	0.452
33	0.121	0.229	0.291	0.345	0.405	0.446
34	0.120	0.225	0.287	0.340	0.399	0.439
35	0.118	0.222	0.283	0.335	0.394	0.433
36	0.116	0.219	0.279	0.330	0.388	0.427
37	0.114	0.216	0.275	0.325	0.383	0.421
38	0.113	0.212	0.271	0.321	0.378	0.415
39	0.111	0.210	0.267	0.317	0.373	0.410
40	0.110	0.207	0.264	0.313	0.368	0.405
41	0.108	0.204	0.261	0.309	0.364	0.400
42	0.107	0.202	0.257	0.305	0.359	0.395
43	0.105	0.199	0.254	0.301	0.355	0.391
44	0.104	0.197	0.251	0.298	0.351	0.386
45	0.103	0.194	0.248	0.294	0.347	0.382
46	0.102	0.192	0.246	0.291	0.343	0.378
47	0.101	0.190	0.243	0.288	0.340	0.374
48	0.100	0.188	0.240	0.285	0.336	0.370
49	0.098	0.186	0.238	0.282	0.333	0.366
50	0.097	0.184	0.235	0.279	0.329	0.363
52	0.095	0.180	0.231	0.274	0.323	0.356
54	0.094	0.177	0.226	0.268	0.317	0.349
56	0.092	0.174	0.222	0.264	0.311	0.343
58	0.090	0.171	0.218	0.259	0.306	0.337
60	0.089	0.168	0.214	0.255	0.300	0.331
62	0.087	0.165	0.211	0.250	0.296	0.326
64	0.086	0.162	0.207	0.246	0.291	0.321
66	0.084	0.160	0.204	0.243	0.287	0.316
68	0.083	0.157	0.201	0.239	0.282	0.311
70	0.082	0.155	0.198	0.235	0.278	0.307
72	0.081	0.153	0.195	0.232	0.274	0.303
74	0.080	0.151	0.193	0.229	0.271	0.299
76	0.078	0.149	0.190	0.226	0.267	0.295
78	0.077	0.147	0.188	0.223	0.264	0.291
80	0.076	0.145	0.185	0.220	0.260	0.287
82	0.075	0.143	0.183	0.217	0.257	0.284
84	0.074	0.141	0.181	0.215	0.254	0.280
86	0.074	0.139	0.179	0.212	0.251	0.277
88	0.073	0.138	0.176	0.210	0.248	0.274
90	0.072	0.136	0.174	0.207	0.245	0.271
92	0.071	0.135	0.173	0.205	0.243	0.268
94	0.070	0.133	0.171	0.203	0.240	0.265
96	0.070	0.132	0.169	0.201	0.238	0.262
98	0.069	0.130	0.167	0.199	0.235	0.260
100	0.068	0.129	0.165	0.197	0.233	0.257

Литература

1. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика. Основы эконометрики. Т. 1. Теория вероятностей и прикладная статистика. — М.: Юнити-Дана, 2001 — 656 с.
2. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Классификация и снижение размерности: справочное издание / под ред. С.А. Айвазяна. — М.: Финансы и статистика, 1989. — 607 с.
3. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Основы моделирования и первичная обработка данных: справочное издание / под ред. С.А. Айвазяна. — М.: Финансы и статистика, 1983. — 471 с.
4. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Исследование зависимостей: справочное издание / под ред. С.А. Айвазяна. — М.: Финансы и статистика, 1985. — 471 с.
5. *Андерсен Т.* Введение в многомерный статистический анализ. — М.: Физматгиз, 1963. — 500 с.
6. *Андерсен Т.* Статистический анализ временных рядов. — М.: Мир, 1976. — 756 с.
7. *Аптон Г.* Анализ таблиц сопряженности. — М.: Финансы и статистика, 1982. — 144 с.
8. *Аренс Х., Лейтер Ю.* Многомерный дисперсионный анализ — М.: Финансы и статистика, 1985. — 230 с.
9. *Афанасьев В.Н., Юзбашев М.М.* Анализ временных рядов: учебник. — М.: Финансы и статистика, 2001. — 228 с.: ил.
10. *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ. — М.: Мир, 1982. — 488 с.
11. *Баласанов Ю.Г., Дойников А.Н., Королев М.Ф., Юровский А.Ю.* Прикладной анализ временных рядов с программой ЭВРИСТА. — Центр СП «Диалог» МГУ, 1991. — 328 с.
12. *Бард Й.* Нелинейное оценивание параметров. — М.: Финансы и статистика, 1979. — 349 с.
13. *Бендат Дж., Пирсол А.* Прикладной анализ случайных данных. — М.: Мир, 1989. — 540 с.
14. *Бендат Дж., Пирсол А.* Применение корреляционного и спектрального анализа. — М.: Мир, 1979. — 311 с.
15. *Бернулли Я.* О законе больших чисел / под общ. ред. Ю.В. Прохорова. — М.: Наука, 1986. — 176 с.
16. *Бикел П., Доксум К.* Математическая статистика. — М.: Финансы и статистика, 1983. Вып. 1 — 280 с.; Вып. 2 — 254 с.
17. *Бокс Дж., Дженкинс Г.* Анализ временных рядов. Прогноз и управление. — М.: Мир, 1974. Вып. 1 — 288 с.; Вып. 2 — 197 с.
18. *Болдин М.В., Симонова Г.И., Тюрин Ю.Н.* Знаковый статистический анализ линейных моделей. — М.: Наука: Физматлит, 1997. — 288 с.
19. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. — М.: Наука, 1983. — 416 с.

20. *Боровиков В.П.* Популярное введение в программу STATISTICA. — М.: КомпьютерПресс, 1998. — 267 с.: ил.
21. *Боровиков В.* STATISTICA: искусство анализа данных на компьютере. Для профессионалов. — СПб.: Питер, 2001. — 656 с.: ил.
22. *Боровиков В.П., Боровиков И.П.* STATISTICA. — Статистический анализ и обработка данных в среде Windows. — М.: Информационно-издательский дом «Филин», 1997. — 608 с.
23. *Боровиков В.П., Ивченко Г.И.* Прогнозирование в системе STATISTICA в среде Windows. Основы теории и интенсивная практика на компьютере: учеб. пособие. — М.: Финансы и статистика, 1999. — 384 с.: ил.
24. *Боровков А.А.* Теория вероятностей. 2-е изд., доп. — М.: Наука, 1986. — 431 с.
25. *Бриллинджер Д.* Временные ряды. — М.: Мир, 1980. — 536 с.
26. *Векслер Л.С.* Статистический анализ на персональном компьютере // МИР ПК. 1992. № 2. — С. 89—97.
27. *Вучков И., Бояджиева Л., Солаков Е.* Прикладной линейный регрессионный анализ. — М.: Финансы и статистика, 1987. — 239 с.
28. *Гаек Я., Шидак З.* Теория ранговых критериев. — М.: Наука, 1971. — 376 с.
29. *Гитис Э.И., Пискулов Е.А.* Аналого-цифровые преобразователи. — М.: Энергоиздат, 1981. — 360 с.
30. *Гнеденко Б.В.* Курс теории вероятностей: учебник. 8-е изд., испр. и доп. — М.: Едиториал УРСС, 2005. — 448 с.
31. *Гоноровский И.С.* Радиотехнические цепи и сигналы. — М.: Сов. радио, 1977. — 608 с.
32. ГОСТ 23554.2–81. Система управления качеством продукции. Экспертные методы оценки качества промышленной продукции. Обработка значений экспертных оценок качества продукции. — М.: Изд-во Стандартов, 1982. — 66 с.
33. *Готтсданкер Р.* Основы психологического эксперимента. — М.: МГУ, 1982. — 463 с.
34. *Григорьев С.Г., Перфилов А.М., Левандовский В.В., Юнкеров В.И.* Пакет прикладных программ STATGRAPHICS на персональном компьютере (практическое пособие по обработке результатов медико-биологических исследований). СПб., 1992. — 104 с.
35. *Гусев А.Н.* Дисперсионный анализ в экспериментальной психологии: Учеб. пособие. — М.: Учебно-методический коллектор «Психология», 2000. — 136 с.
36. *Демиденко Е.З.* Линейная и нелинейная регрессия. — М.: Финансы и статистика, 1981. — 302 с.
37. *Дженкинс Г., Ваттс Д.* Спектральный анализ и его приложения. — М.: Мир. Вып. 1, 1971. — 316 с.; Вып. 2, 1972. — 288 с.
38. *Джессен Р.* Методы статистических обследований: пер. с англ.; под ред. и с предисл. Е.М. Четыркина. — М.: Финансы и статистика, 1985. — 478 с.: ил.
39. *Джонсон Н., Лион Ф.* Статистика и планирование эксперимента в технике и науке. — М.: Мир. Т. 1, 1980, — 610 с., Т. 2, 1981, — 520 с.
40. *Дугерти К.* Введение в эконометрику: пер. с англ. — М.: ИНФРА-М, 1999. — XIV, 402 с.

41. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. 3-е изд.: пер. с англ. — М.: Вильямс, 2007. — 912 с.
42. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: учебник. — М.: Финансы и статистика, 1998. — 352 с.: ил.
43. Дэниел К. Применение статистики в промышленном эксперименте. — М.: Мир 1979. — 299 с.
44. Дюк В. Обработка данных на ПК в примерах — СПб.: Питер, 1997. — 240 с.: ил.
45. Дюк В.А., Мирошников А.И. STATGRAPHICS Plus for Windows — учебное пособие по прикладной статистике //Тезисы доклада на международной конференции «Статистическое образование в современном мире: идеи, ориентации, технологии». СПб., 1996. — С. 193–196
46. Дюк В.А., Мирошников А.И. Эволюция STATGRAPHICS //МИР ПК. 1995. № 12.
47. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа. — М.: Финансы и статистика, 1986.
48. Закс Л. Статистическое оценивание / пер. с нем.; науч. ред. Ю.П. Адлера и В.Г. Горского. — М.: Статистика, 1976. — 598 с.
49. Кендэл М., Стьюарт А. Статистические выводы и связи. — М.: Наука, 1973. — 899 с.
50. Кендэл М., Стьюарт А. Теория распределений. — М.: Наука, 1966.
51. Кендэл М., Стьюарт А. Многомерный статистический анализ и временные ряды. — М.: Наука, 1976. — 736 с.
52. Кендэл М. Временные ряды. — М.: Финансы и статистика, 1981. — 199 с.
53. Кендэл М. Ранговые корреляции. — М.: Статистика, 1975. — 212 с.
54. Кокрен У. Методы выборочного исследования. — М.: Статистика, 1976. — 440 с.
55. Кокс Д.Р., Оукс Д. Анализ данных типа времени жизни. — М.: Финансы и статистика, 1988. — 192 с.
56. Колмогоров А.Н. Об одном новом подтверждении законов Менделя //ДАН СССР. 1940. Т. 27. № 1. — С. 38—42.
57. Крамер Г. Математические методы статистики. — М.: Мир, 1975. — 648 с.
58. Кремер Н.Ш. Теория вероятностей и математическая статистика. — М.: ЮНИТИ, 2000.
59. Крылов В.Ю. Геометрическое представление данных в психологических исследованиях. — М.: Наука, 1990. — 117 с.
60. Кулаицев А.П. Пакеты для анализа данных //МИР ПК. 1995. № 1.
61. Кулаицев А.П. Методы и средства анализа данных в среде Windows. STADIA. 3-е изд., перераб. и доп. — М.: Информатика и компьютеры, 1999. — 341 с.: ил.
62. Левин Б.Р. Теоретические основы статистической радиотехники: в 3 кн. — М.: Сов. радио, 1975.
63. Леман Э. Проверка статистических гипотез. — М.: Наука, 1964. — 498 с.
64. Леман Э. Теория точечного оценивания. — М.: Наука, 1991. — 448 с.
65. Ликеш И., Ляга И. Основные таблицы математической статистики. — М.: Финансы и статистика, 1985. — 356 с.

66. *Литтл Р.Дж., Рубин Д.Б.* Статистический анализ данных с пропусками. — М.: Финансы и статистика, 1991. — 336 с.
67. *Лукашин Ю.П.* Адаптивные методы краткосрочного прогнозирования. — М.: Статистика, 1979. — 254 с.
68. *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика. Начальный курс: учебник. 4-е изд. — М.: Дело, 2000. — 400 с.
69. *Макаров А.А.* STADIA против STATGRAPHICS, или Кто ваш «лоцман» в море статистических данных //МИР ПК. 1992. № 3. — С. 58—66.
70. *Макаров А.А.* Роль и место статистических пакетов программ в курсах математической и прикладной статистики //Тезисы доклада на международной конференции «Информационные технологии в непрерывном образовании». Петрозаводск: 1995. — С. 127—128.
71. *Макаров А.А.* Статистические пакеты в обучении математической и прикладной статистике //Тезисы доклада на международной конференции «Статистическое образование в современном мире: идеи, ориентации, технологии». СПб., 1996. — С. 193—196.
72. *Макаров А.А., Кулаишев А.П., Синева И.С.* Использование программ обработки данных в преподавании курсов теории вероятностей, математической и прикладной статистики и информатики. Метод. рекомендации (выпуск 1) — М.: МГУ, 2002. — 39 с.: ил.
73. *Макино Т., Охаси М., Доке Х., Макино К.* Контроль качества с помощью персональных компьютеров. — М.: Машиностроение, 1991.
74. *Мардиа К., Земрош П.* Таблицы F-распределений. — М.: Наука, 1984. — 255 с.
75. *Марпл-мл. С.Л.* Цифровой спектральный анализ и его приложения. — М.: Мир, 1990. — 584 с.
76. *Мэйндоналд Дж.* Вычислительные алгоритмы в прикладной статистике. — М.: Финансы и статистика, 1988. — 350 с.
77. *Мюллер П., Нойман П., Шторм Р.* Таблицы по математической статистике. — М.: Финансы и статистика, 1982.— 278 с.
78. *Никитин Я.Ю.* Асимптотическая эффективность непараметрических критериев. — М.: Физматлит, 1995. — 240 с.
79. *Оуэн Д.Б.* Сборник статистических таблиц. 2-е изд., испр. — М.: ВЦ АН СССР, 1973. — 586 с.
80. *Плис А.И., Сливина Н.А.* Mathcad: математический практикум для экономистов и инженеров: учеб. пособие. — М.: Финансы и статистика, 1999. — 656 с.: ил.
81. *Поллард Дж.* Справочник по вычислительным методам статистики. — М.: Финансы и статистика, 1982. — 344 с.
82. *Рао С.Р.* Линейные статистические методы и их применение. — М.: Наука, 1968. — 548 с.
83. *Рунион Р.* Справочник по непараметрической статистике. Современный подход. — М.: Финансы и статистика, 1982. — 198 с.
84. *Семенов Н.А.* Программы регрессионного анализа и прогнозирования временных рядов. Пакеты ПАРИС и МАВР. — М.: Финансы и статистика, 1990. — 111 с.
85. *Смирнов Н.В., Дунин-Барковский И.В.* Курс теории вероятностей и математической статистики для технических приложений. 2-е изд., испр. и доп. — М.: Наука, 1965. — 511 с.

86. *Смоляк С.А., Титаренко Б.П.* Устойчивые методы оценивания. — М.: Статистика, 1980. — 206 с.
87. Справочник по прикладной статистике: в 2 т. / под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. — М.: Финансы и статистика, 1989, 1990.
88. Справочник по специальным функциям с формулами, графиками и таблицами / под ред. М.А. Абрамовица, И. Стиган. — М.: Наука, 1979. — 830 с.
89. Статистические методы для ЭВМ / под ред. К. Эйнслеяна, Э. Рэлстона, Г.С. Уолфа. — М.: Наука, 1986. — 459 с.
90. Статистические методы повышения качества / под ред. Хитоси Куме. — М.: Финансы и статистика, 1991.
91. *Стрелю Я.* Роль темперамента в психическом развитии. — М.: Прогресс, 1982. — 231 с.
92. Таблицы вероятностных функций. Т. 2. — М.: ВЦ АН СССР, 1959. — 344 с.
93. Таблицы функций распределения и плотностей распределения Стъдента / под ред. Н.В. Смирнова. — М.: АН СССР, 1960.
94. *Теннант-Смит Дж.* Бейсик для статистиков. — М.: Мир, 1988. — 207 с.
95. *Томас Р.* Количественные методы анализа хозяйственной деятельности / Пер. с англ. — М.: Изд-во «Дело и Сервис», 1999. — 432 с.
96. *Тутубалин В.Н.* Границы применимости (вероятностно-статистические методы и их возможности). — М.: Знание, 1977. — 61 с.
97. *Тутубалин В.Н.* Теория вероятностей и случайных процессов. — М.: Изд-во МГУ, 1992. — 400 с.
98. *Тьюки Дж.* Анализ результатов наблюдений. Разведочный анализ. — М.: Мир, 1981. — 693 с.
99. *Тюрин Ю.Н., Макаров А.А.* Анализ данных на компьютере / под ред. В.Э. Фигурнова. — М.: ИНФРА-М: Финансы и статистика, 1995. — 384 с.: ил.
100. *Тюрин Ю.Н., Макаров А.А.* Статистический анализ данных на компьютере / под ред. В.Э. Фигурнова. — М.: ИНФРА-М, 1998. — 528 с.: ил.
101. *Тюрин Ю.Н., Симонова Г.И.* Знаковый анализ линейных моделей //Обозрение прикладной и промышленной математики. Т. 1. Вып. 2. 1994. — С. 214—278.
102. *Урбах В.Ю.* Математическая статистика для биологов и медиков. — М.: Изд-во АН СССР, 1963. — 323 с.
103. Факторный, дискриминантный и кластерный анализ. — М.: Финансы и статистика, 1989. — 215 с.
104. *Феллер В.* Введение в теорию вероятностей и её приложения: в 2 т. Т. 1: пер. с англ. — М.: Мир, 1984. — 528 с.
105. *Фигурнов В.Э.* IBM PC для пользователя. Краткий курс. — М.: ИНФА-М, 1997. — 480 с.
106. *Флейс Дж.* Статистические методы для изучения таблиц долей и пропорций. — М.: Финансы и статистика, 1989. — 319 с.
107. *Хальд А.* Математическая статистика с техническими приложениями. — М.: Изд-во иностранной литературы, 1956. — 664 с.
108. *Хампель Ф., Рончетти Э., Рауссей П., Штаэль В.* Робастность в статистике. Подход на основе функций влияния. — М.: Мир, 1989. — 512 с.
109. *Хан Г., Шапиро С.* Статистические модели в инженерных задачах. — М.: Статистика, 1980. — 444 с.

110. *Хартман Г.* Современный факторный анализ. — М.: Статистика, 1972.
111. *Хастингс Н., Пикок Дж.* Справочник по статистическим распределениям. М.: Статистика, 1980. — 95 с.
112. *Хенинен А.Я., Павлов Ю.Л.* Статистик-Консультант, или Еще один довод в пользу неизбежного // МИР ПК. 1994. № 6.
113. *Хеттсманпергер Т.* Статистические выводы, основанные на рангах. — М.: Финансы и статистика, 1987. — 334 с.
114. *Хикс Ч.* Основные принципы планирования эксперимента. — М.: Мир, 1967. — 406 с.
115. *Холлендер М., Вулф Д.* Непараметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.
116. *Хьюбер П.* Робастность в статистике. — М.: Мир, 1984. — 304 с.
117. *Шиндовский Э., Шюрц О.* Статистические методы контроля производства. М.: Госкомстандарт, 1969. — 542 с.
118. *Ширяев А.Н.* Вероятность. — М.: Наука, 1980. — 574 с.
119. *Ширяев А.Н.* Основы стохастической финансовой математики. Т. 1. Факты. Модели. — М.: ФАЗИС, 1998. — 512 с.
120. *Ширяев А.Н.* Основы стохастической финансовой математики. Т. 2. Теория. — М.: ФАЗИС, 1998. — 544 с.
121. *Шураков В.В., Дайитбегов Д.М., Мизрохи С.В., Ясеновский С.В.* Автоматизированное рабочее место для статистической обработки данных. — М.: Финансы и статистика, 1990. — 190 с.
122. *Эддоус М., Стенсфилд Р.* Методы принятия решения / пер. с англ.; Под ред. член-корр. РАН И.И. Елисеевой. — М.: Аудит, ЮНИТИ, 1997. — 590 с.
123. *Яглом А.М.* Корреляционная теория стационарных случайных функций (с примерами из метеорологии). — Гидрометеиздат, 1981. — 280 с.
124. *Aczel A.D.* Complete business statistics. 3rd ed. — Richard D. Irwing, 1996. — 869 p.
125. *Biometrika Tables for Statisticians*, vol. 1/3rd ed, vol. 2 Pearson E.S. Hartley H.O., eds. — Cambridge: Cambridge Univ. Press, 1970, 1972. — XVI, 270 p., XVIII, 385 p.
126. *Vox G.E.P., Cox D.R.* An analysis of transformations, *J. Roy. Stat. Soc.*, 1964, B26. P. 211–243
127. *Chatfield C.* The Analysis of Time Series: an Introduction, 4th ed. — Chapman and Hall, 1989. — 242 p.
128. *Elliott A.C., Gray Y.L.* Directory of Statistical Microcomputer Software. — N.Y.: Basel, 1986.
129. *Everit B.* A Handbook of Statistical Analyses using S-PLUS. Chapman & Hall, 1994. — 143 p.
130. *Granger C.W.J., Newbold P.* Forecasting Economic Time Series, 2nd ed. — Academic Press, Inc., 1986. — 338 p.
131. *Hanke J.E., Reitsch A.G.* Business forecasting. 6th ed. — Prentice-Hall, Inc., 1998. — 581 p.
132. *Hartley H.O.* Testing of homogeneity of a set of variances. — 31. *Biometrika*, 1940, — P. 249–255.
133. *Lindley D.V., Scott W.F.* New Cambridge Elementary Statistical tables. — Cambridge University press, 1984. — 80 p.
134. *Mosteller F., Tukey J.W.* Data Analysis and Regression: A Second Course in Statistics. Reading, MA: Addison-Wesley, 1977.

135. *Neter J., Wasserman W., Whitmore G.A.* Applied Statistics, Allyn and Bacon, Inc., 1988. — 1006 p.
136. *Oden R.E.* Extended tables of the distribution of Fridman's S-statistic in two-way layout. — CS, 1977, B6, 29–48.
137. *Oden R.E., Owen D.B., Birnbaum Z.W., Fisher L.* Pocet Book of Statistical Tables. — N. Y. and Basel: Marsel Dekker, Inc., 1977. — X, 166 p.
138. *Sen P.K.* Nonparametric simultaneous inference for some MANOVA models/Handbook of Statistics. — V. 1. Holland, 1980.
139. Software Digest Rating Report. 1991. V. 8, № 5.
140. *Spector P.* An introduction to S and S-PLUS. Duxbury Press, 1994. — 286 p.
141. *Venables M.N., Ripley B.D.* Modern Applied Statistics with S-PLUS. Springer-Verlag, 1994. — 462 p.
142. *Wilcoxon F., Katti S.K., Wilcox Roberta A.* Critical values and probability levels for the Wilcoxon rank test. — In: Selected Tables in Mathematical Statistics, vol. 1/2-d ed. H.L.Harter, D.B.Owen, eds. — Providence, R. I.: Am. Math. Soc., 1973. — P. 171–235.
143. *Woodwant W.A, Elliott A.C., Gray Y.L., Mattlock D.C.* Directory of Statistical Microcomputer Software. — N.Y.: Basel, 1988.
144. *Zar J.H.* Significance testing of the Spearman rank correlation coefficient. — JASA, 1972, 67, 578–580.

Краткий путеводитель по списку литературы

Для удобства читателей мы помещаем краткие пояснения к списку литературы.

- Справочники по прикладной статистике — [87], [81].
- Стандартные учебники теории вероятностей и статистики — (строгий, аксиоматический подход) — [30], [24], [118], [57].
- Учебники и пособия по статистике, рассчитанные на прикладных специалистов — [16], [85], [39], [87], [109], [10], [3], [4], [2], [68], [58], [1], [99], [100].
- Вероятностные распределения — [19], [50], [65], [87], [111].
- Таблицы распределений — [19], [65], [77], [79], [88], [92], [93], [125].
- Случайный выбор — [54], [38].
- Непараметрические методы статистики — [115], [83], [113], [28], [18], [78].
- Компьютерные алгоритмы статистики — [76], [10], [94], [89].
- Анализ данных на компьютере — [44], [61], [72], [80], [35], [23], [141], [140], [129], [121], [84], [11], [21], [131], [99], [100].
- Разведочный анализ данных — [98].
- Анализ данных с пропусками — [66].
- Регрессионный анализ — [41], [36], [27], [116], [134], [68], [40].
- Дисперсионный анализ — [41], [43], [39], [114], [2], [35].
- Планирование эксперимента — [39], [114], [43].
- Таблицы сопряженности — [7], [106], [49].
- Меры связи признаков — [53], [83], [102].
- Анализ временных рядов — [14], [17], [25], [37], [52], [87], [51], [123], [11], [130], [127], [6], [9], [23], [119], [120].
- Многомерные методы — [5], [87], [103], [47], [51], [2], [1], [42].

Факторный анализ — [10], [103], [110], [51], [2], [1].

Дискриминантный анализ — [10], [103], [2], [1].

Кластерный анализ — [47], [87], [103], [2], [1].

Многомерное шкалирование — [59], [87], [89].

Методы контроля качества — [73], [90].

Замечания. Из книг, приведенных в списке литературы, выделим двухтомный «Справочник по прикладной статистике» [87], написанный в основном английскими учеными. Он отражает добротный уровень английской статистической науки (и некоторые ее особенности). Этот справочник содержит постановки основных задач анализа данных и сведения о методах их решения.

Кроме того, мы хотели бы отметить трехтомник «Прикладная статистика» С.А. Айвазяна и его соавторов В.М. Бухштабера, И.С. Енюкова, Л.Д. Мешалкина [3], [4], [2]. Это справочное издание вобрало в себя многолетний опыт работы как его авторов, так и всей школы прикладной статистики в СССР. Издание отражает широкий круг статистических приложений, причем в единстве основных проблем прикладной статистики: построения статистической модели, развития математической теории, проведения численных расчетов. Библиография трехтомника содержит множество ссылок на книги и статьи на русском языке.

Из книг, носящих обзорный, справочный характер в конкретных областях прикладной статистики, обратим внимание на следующие: в области вероятностных распределений — [19], [111]; в области регрессионного и дисперсионного анализа — [41]; в области непараметрических методов — [115]. Те, кого интересуют вопросы разработки компьютерных алгоритмов статистики, могут найти полезную информацию в [76], [94]. Более строгое аксиоматическое изложение основ теории вероятностей и статистики содержится в [30], [24], [118], [57], [16].

Оглавление

Предисловие авторов	3
Предисловие редактора	6
Как читать эту книгу	13
Глава 1. Основные понятия прикладной статистики	15
1.1. Случайная изменчивость	15
1.2. События и их вероятности	19
1.3. Измерения вероятности	23
1.4. Случайные величины. Функции распределения	24
1.5. Числовые характеристики распределения вероятностей	30
1.6. Независимые и зависимые случайные величины	34
1.7. Случайный выбор	36
1.8. Выборки и их описание	38
1.8.1. Что такое выборка	38
1.8.2. Выборочные характеристики	39
1.8.3. Ранги и ранжирование	42
1.8.4. Методы описательной статистики	44
1.8.5. Наглядные методы описательной статистики	46
1.9. Методы описательной статистики в пакете SPSS	49
Глава 2. Важные законы распределения вероятностей	59
2.1. Биномиальное распределение	60
2.2. Распределение Пуассона	63
2.3. Показательное распределение	66
2.4. Нормальное распределение	68
2.5. Двумерное нормальное распределение	71
2.6. Распределения, связанные с нормальным	73
2.6.1. Распределение хи-квадрат	74
2.6.2. Распределение Стьюдента	75
2.6.3. F-распределение	76
2.7. Законы распределения вероятностей в пакете SPSS	77

Глава 3. Основы проверки статистических гипотез	82
3.1. Статистические модели	82
3.2. Проверка статистических гипотез (общие положения)	85
3.3. Примеры статистических моделей и гипотез	91
3.4. Проверка статистических гипотез (прикладные задачи)	96
3.4.1. Схема испытаний Бернулли	96
3.4.2. Критерий знаков для одной выборки	100
3.5. Проверка гипотез в двухвыборочных задачах	101
3.5.1. Критерий Манна–Уитни	103
3.5.2. Критерий Уилкоксона	107
3.6. Парные наблюдения	113
3.6.1. Критерий знаков для анализа парных повторных наблюдений	114
3.6.2. Анализ повторных парных наблюдений с помощью знаковых рангов (критерий знаковых ранговых сумм Уилкоксона)	116
3.7. Проверка статистических гипотез в пакете SPSS	118
Глава 4. Начала теории оценивания	125
4.1. Введение	125
4.2. Закон больших чисел	126
4.3. Статистические параметры	131
4.3.1. Параметры распределения	131
4.3.2. Параметры модели	132
4.4. Оценивание параметров распределения по выборке	133
4.5. Свойства оценок. Доверительное оценивание	136
4.6. Метод наибольшего правдоподобия	138
4.7. Оценивание параметров вероятностных распределений в пакете SPSS	141
Глава 5. Анализ одной и двух нормальных выборок	147
5.1. Об исследовании нормальных выборок	147
5.2. Глазомерный метод проверки нормальности	149
5.3. Оценки параметров нормального распределения и их свойства	151
5.4. Проверка гипотез, связанных с параметрами нормального распределения	156
5.4.1. Одна выборка	156
5.4.2. Две выборки	158
5.4.3. Парные данные	160

5.5. Анализ нормальных выборок в пакете SPSS.....	163
Глава 6. Однофакторный анализ	170
6.1. Постановка задачи	170
6.2. Непараметрические критерии проверки однородности.....	174
6.2.1. Критерий Краскела–Уоллиса (произвольные альтернативы)	175
6.2.2. Критерий Джонкхиера (альтернативы с упорядочением)	176
6.3. Практический пример.....	177
6.4. Оценивание эффектов обработки (непараметрический подход).....	180
6.5. Дисперсионный анализ	183
6.6. Оценивание эффектов обработки в нормальной модели.....	185
6.6.1. Доверительные интервалы	185
6.6.2. Метод Шеффе множественных сравнений	186
6.7. Однофакторный анализ в пакете SPSS	188
Глава 7. Двухфакторный анализ.....	194
7.1. Связь задач двухфакторного и однофакторного анализа	194
7.2. Таблица двухфакторного анализа	195
7.3. Аддитивная модель данных двухфакторного эксперимента при независимом действии факторов	196
7.4. Непараметрические критерии проверки гипотезы об отсутствии эффектов обработки.....	197
7.4.1. Критерий Фридмана (произвольные альтернативы).....	197
7.4.2. Критерий Пейджа (альтернативы с упорядочением)	199
7.5. Практический пример.....	200
7.6. Двухфакторный дисперсионный анализ.....	202
7.7. Двухфакторный анализ в пакете SPSS.....	205
Глава 8. Линейный регрессионный анализ.....	208
8.1. Модель линейного регрессионного анализа.....	208
8.2. О стратегии, методах и проблемах регрессионного анализа.....	210
8.3. Простая линейная регрессия	213
8.4. О проверке предпосылок в задаче регрессионного анализа.....	217
8.5. Непараметрическая линейная регрессия.....	219
8.6. Практический пример.....	225
8.7. Регрессионный анализ в пакете SPSS	230

Глава 9. Независимость признаков	240
9.1. О шкалах измерений	240
9.2. Инструменты и стратегия исследования связи признаков	243
9.3. Связь номинальных признаков (таблицы сопряженности)	244
9.4. Связь признаков, измеренных в шкале порядков	253
9.5. Связь признаков в количественных шкалах	257
9.5.1. Коэффициент корреляции	257
9.5.2. Нормальная корреляция	260
9.6. Замечания о связи признаков, измеренных в разных шкалах	263
9.7. Анализ таблиц сопряженности и коэффициенты корреляции в пакете SPSS	263
Глава 10. Критерии согласия	271
10.1. Введение	271
10.2. Критерии согласия Колмогорова и омега-квадрат в случае простой гипотезы	272
10.3. Практический пример (закон Менделя)	276
10.4. Критерий согласия хи-квадрат К. Пирсона для простой гипотезы	278
10.5. Критерии согласия для сложной гипотезы	280
10.6. Критерий согласия хи-квадрат Фишера для сложной гипотезы	283
10.7. Другие критерии согласия. Критерий согласия для пуассоновского распределения	286
10.8. Критерии согласия в пакете SPSS	290
Глава 11. Выборочные обследования	296
11.1. Введение	296
11.2. Выборки. Простой случайный выбор	296
11.3. Точность выборочной оценки	299
11.4. Выборки. Сложные планы	306
11.5. Основные выводы	312
Глава 12. Многомерный анализ и другие статистические методы	315
12.1. Введение	315
12.2. Многомерный статистический анализ	315

12.3. Факторный анализ	317
12.4. Дискриминантный анализ	318
12.5. Кластерный анализ	318
12.6. Многомерное шкалирование	319
12.7. Методы контроля качества	320
12.8. Использование статистических пакетов	321

Приложения. Таблицы математической статистики 323

П1. Верхние процентные точки стандартного нормального распределения	325
П2. Верхние процентные точки распределения Стьюдента	326
П3. Верхние процентные точки распределения хи-квадрат	328
П4. Верхние процентные точки F-распределения	331
П5. Верхние процентные точки биномиального распределения вероятностей	332
П6. Верхние критические значения для статистики Уилкоксона	342
П7. Верхние критические значения статистики Краскела—Уоллиса для различных планов эксперимента	344
П8. Верхние критические значения для статистики Фридмана	349
П9. Верхние критические значения для коэффициента ранговой корреляции Кендэла	351
П10. Верхние критические значения для коэффициента ранговой корреляции Спирмена	352

Литература 355

Юрий Николаевич Тюрин
Алексей Алексеевич Макаров

Анализ данных на компьютере

Учебное пособие

Подписано в печать 18.01.2016. Формат 60×90/16. Печать офсетная.
Гарнитура «Антиква». Усл. печ. л. 23. Уч.-изд. л. 23,5.
Бумага офсетная. Тираж 2000 экз. Заказ .

Отпечатано с электронных носителей издательства
в ООО «Тверской полиграфический комбинат».
170024, г. Тверь, пр-т Ленина, 5.
Тел.: (4822) 44-42-15, (495) 748-04-67. Тел./факс: (4822) 55-42-15.

Книги издательства МЦНМО можно приобрести в магазине
«Математическая книга», Москва, Большой Власьевский пер., д. 11.
Тел. (495) 745-80-31. E-mail: biblio@mccme.ru
