Ю.Н.Тюрин, А.А.Макаров

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ НА КОМПЬЮТЕРЕ

Под редакцией В. Э. Фигурнова

МОСКВА 1998
ИНФРА • М

ББК **517.8, 32.973**
Т98
УДК **519.2, 681.3**

Тюрин Ю.Н., Макаров А.А.

Т98 Статистический анализ данных на компьютере/Под ред. В.Э.Фигурнова — М.: ИНФРА–М, 1998. — 528 с., ил.
ISBN 5-86225-662-8

Книга является учебным пособием по анализу данных и статистике, рассчитанным на прикладных специалистов, менеджеров и студентов. Излагаются основные сведения, необходимые на практике для анализа данных, на наглядных примерах рассматриваются основные постановки задач, а затем эти же примеры решаются с использованием популярных статистических пакетов STADIA, STATGRAPHICS, SPSS и Эвриста. В приложении дается обзор других программных средств для анализа данных. Большое внимание в книге уделено средствам анализа временных рядов и других методам, часто используемым в прикладных задачах.

ББК 517.8, 32.973

ISBN 5-86225-662-8

Тюрин Юрий Николаевич
Макаров Алексей Алексеевич

Статистический анализ данных на компьютере

Оригинал-макет подготовлен В.Э.Фигурновым
с помощью TEX (emTEX 3.1415 и dvips 5.54)

Подписано в печать 21.10.97. Формат $60\times90^1/_{16}$. Гарнитура «Антиква».
Печать офсетная. Усл. п. л. 33. Тираж 11 000 экз. Заказ 783.
ЛР № 070824 от 21.01.1993 г.

Издательство «ИНФРА-М»
127247, Москва, Дмитровское шоссе, 107

Тел. (095)485-71-77, (095)485-76-18 — отдел сбыта,
(095)485-70-63, (095)485-74-00 — заключение договоров.
Факс (095)485-53-18. E-mail — contract@infram.msk.ru

Мелкооптовый склад — Скорняжный пер., д. 7, корп. 1,
между ст. м. «Сухаревская» и «Красные ворота»,
за маг. «Военная книга», тел. (095)208-32-59

Отпечатано с готовых диапозитивов
в ОАО «Ярославский полиграфкомбинат»
150049, г. Ярославль, ул. Свободы, 97.

Об этой книге

На протяжении более десятка лет авторы читали курсы лекций и проводили практические занятия по анализу данных и прикладной статистике в МГУ им. М.В.Ломоносова, в других учебных заведениях России и иных стран. Материалы наших лекций и опыт общения со слушателями и легли в основу этой книги. Нашими слушателями были самые разные люди: психологи и экономисты, бизнесмены и инженеры, медики, социологи и т.д. Мы стремились как можно более просто и понятно передать им дух статистической науки, научить их самым необходимым и употребительным методам обработки данных, т.е. способам извлечения из них обоснованных выводов. Мы также стремились показать, как это делается с помощью персональных компьютеров и пакетов прикладных программ.

В 1994 г. авторы написали книгу «Анализ данных на компьютере», которая была выпущена издательством «ИНФРА-М». Эта книга быстро разошлась среди читателей, а во многих вузах была положена в основу преподавания курсов статистики. Однако за время, прошедшее с момента издания этой книги, заметно расширилось использование статистических методов анализа данных на практике, появилось много новых статистических компьютерных программ, а существующие программы значительно обновились. Поэтому, по многочисленным просьбам читателей, в книгу были включены главы, посвященные анализу временных рядов и описаны более новые версии статистических компьютерных программ, в том числе версии для Windows.

Материал, помещенный в книгу, значительно увеличился, и мы даже решили дать книге новое название «Статистический анализ данных на компьютере». Это название, на наш взгляд, более точно отражает содержание книги.

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.

Мы надеемся, что эта книга будет полезна всем, кто хочет освоить методы статистического анализа данных и применять их в своей деятельности.

Предисловие редактора

Моторы реактивного самолета взревели 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 незыблемость 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 потребуется 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 нашего века немало способствовало появление компьютеров, а начиная с 80-х годов — 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 общеупотребительным 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, а также гл. 11—14).

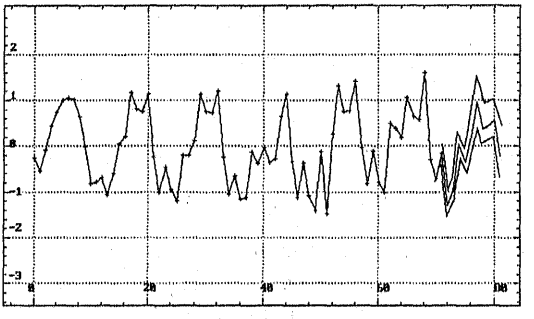


Рис. 1. График изменения объема транспортных перевозок и его прогноз

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 самими 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).

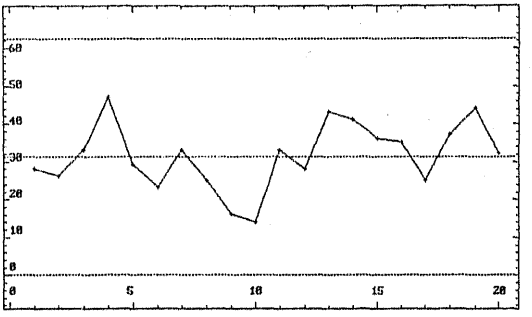


Рис. 2. Контрольная карта изменения показателя качества с зоной допустимых пределов изменения

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. 15) этих данных. Во многих случаях имеющиеся объекты удается 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 пиво, но близкое по совокупности свойств к очень

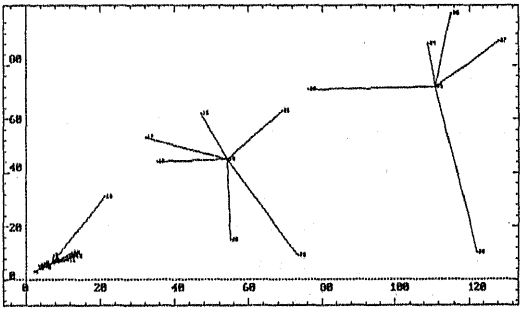


Рис. 3. С помощью кластерного анализа имеющаяся совокупность объектов разбита на три группы со схожими свойствами

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. Такой инструмент мог бы быть создан, если бы решаемые этими пользователями задачи были исключительно специфичны и не встречались ни в какой другой области. Как правило, это не так, и все нужные этим пользователям задачи могут быть решены с помощью универсальных пакетов (подобно тому, как практически для всех пользователей нужные им средства подготовки документов обеспечиваются универсальными редакторами документов типа Word или WordPerfect).

Разумеется, нет правил без исключений. Например, в Word или WordPerfect трудно на надлежащем уровне подготавливать документы с большим количеством формул (типа этой книги), и невозможно печатать ноты, поэтому в таких случаях используются специальные средства. Точно так же существуют и области человеческой деятельности, для которых 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 совершенно другая. В средней школе методы статистического анализа данных (хотя многие из них очень просты и весьма полезны) не упоминаются вовсе, а в высшей школе, даже в тех вузах и университетах, программы которых были просто перегружены математикой, методам анализа данных отводилось очень небольшое место. При этом обычно предметом изучения являются не столько эти методы, сколько формальные конструкции теории множеств, теории меры, функционального анализа и теории вероятностей, которые, может быть, нужны для строгих доказательств, но абсолютно не способствуют освоению и бесполезны при применении этих методов. А в гуманитарных и медицинских вузах курсы анализа данных чаще всего просто отсутствовали. В результате даже самые простейшие методы статистического анализа данных почти для всех отечественных руководителей и менеджеров остаются terra incognita[1]. Для исправления положения (что абсолютно необходимо для конкуренции с западным бизнесом), по-видимому, потребуется значительное время.

Таким образом, российские специалисты и менеджеры, исследователи и студенты, желающие применять методы анализа данных, нахо-

[1] Причины столь бедственного положения разнообразны, но одна из них вполне понятна. В стране, где важнейшие статистические данные, касающиеся экономики и сельского хозяйства, медицины и демографии, экологии и социологии тщательно скрывались с помощью режимов секретности, «форм доступа» и т.д. (даже от занимающихся этими вопросами специалистов), а зачастую и фальсифицировались, было вполне естественно сделать как можно менее распространенными и методы анализа данных.

дятся в гораздо более затруднительном положении по сравнению со своими западными коллегами. Им приходится изучать многие аспекты прикладной статистики самостоятельно, при этом часто по книгам, рассчитанным не на прикладных специалистов, а на профессиональных математиков (просто потому, что в основном именно такие книги имеются в наличии). Кроме того, при работе с западным статистическим пакетом им приходится читать объемистую и не всегда понятно написанную программную документацию на английском языке. Получить консультацию при затруднениях негде — разве что у своих же коллег. Ясно, что преодолеть эти препятствия далеко не всем под силу.

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 [68], [8], [14], [30], [85].

2. Использовать (если нет очень веских причин поступать иначе) отечественные, а не западные статистические пакеты — они, как правило, гораздо проще в использовании, снабжены понятной документацией и средствами интерпретации результатов. Особенно стоит порекомендовать пакеты STADIA (универсальный статистический пакет) и ЭВРИСТА (специализированный пакет для анализа временных рядов и регрессионного анализа).

3. Для статистических пакетов с хорошей документацией — читать эту документацию. Очень часто она фактически является популярным учебником, наглядно и неформально объясняющим применение средств анализа данных, в том числе и самых мощных многомерных методов. Особенно в этой связи можно рекомендовать документации пакетов STADIA, ЭВРИСТА и SPSS.

4. Практически применять в ходе изучения анализа данных статистические пакеты. 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 с первой по четвертую, а также частично пятая и десятая главы. 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, а также основы проверки статистических гипотез и оценивания параметров. Все изложение ведется не в строго формальном математическом ключе (который привлекателен только для математиков), а на общепонятийном уровне, с привлечением многочисленных примеров.

Вторая часть книги (главы 5—15) 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, анализ временных рядов, 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.

Примерно треть книги (ее составляют последние параграфы каждой главы и три приложения) посвящена современным статистическим пакетам и их использованию на персональных компьютерах. В этой части, во-первых, показано, как рассмотренные в книге задачи можно решать с помощью компьютера. В большей части книги для этого используются популярные в России статистические пакеты: отечественный — STADIA и американский — STATGRAPHICS. А решение задач анализа временных рядов показывается с помощью отечественного пакета Эвриста и американского пакета SPSS. Мы полагаем, что эти примеры будут полезны всем читателям, в том числе и пользователям других стати-

стических пакетов. Ведь входные данные и результаты статистической обработки, как правило, мало зависят от конкретного пакета, поскольку определяются общепринятыми традициями.

В приложении 1 и 3 дан обзор состояния и основных характеристик наиболее известных отечественных и зарубежных статистических пакетов и сведения о фирмах, их распространяющих.

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 применимости, 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 [83], Р.Готсданкера [26], М.Холлендера и Д.А.Вулфа [91] и др., а также из практической работы авторов.

Одной из особенностей этих примеров является сравнительно малый объем исходных данных. Это сделано не только из соображений облегчения демонстрации расчетов вручную. Другая причина состоит в том, что для большинства прикладных исследований, особенно в гуманитарных областях, характерны именно небольшие объемы данных (исключение здесь составляют, пожалуй, только демография и отдельные области медицинской статистики). А поскольку на подобных объемах выборок практически невозможна эффективная проверка гипотез об их распределении, а процедуры отбраковки грубых наблюдений бесполезны или малоэффективны, мы рассматривали в первую очередь непараметрические статистические методы, т.е. методы, свободные от предположений о распределении и потому более универсальные.

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 и 4. 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. Знакомство с работой типичных статистических процедур на персональном компьютере полезно начинать с приложений 1 и 2, где

описываются общая архитектура пакетов, их интерфейсы, возможности работы с данными и пр.

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 знакомства с частями 3 и 7 книги В.Э.Фигурнова [81].

Обозначения. При записи чисел мы придерживались американской системы записи, т.е. целая часть от дробной отделяется не запятой, а точкой (скажем, два с половиной — это 2.5, а не 2,5). Дело в том, что именно такая форма записи чисел принята в статистических пакетах, а кроме того, при этом списки чисел, которые нам иногда приходится использовать в книге, становятся проще для восприятия.

Числами в квадратных скобках обозначаются книги или статьи из списка литературы (так, [3] — ссылка на третью книгу в списке литературы).

Сведения об авторах

Тюрин Юрий Николаевич — д.ф.-м.н., профессор кафедры теории вероятностей механико-математического факультета МГУ им. М.В.Ломоносова. Много лет читал курсы теоретической и прикладной статистики на различных факультетах МГУ, в других учебных заведениях России и иных стран. Область научных интересов — многомерный непараметрический анализ, методы анализа нечисловой информации.

Макаров Алексей Алексеевич — к.ф.-м.н., ведущий научный сотрудник НИИ Механики МГУ. Область научных интересов — непараметрические методы анализа данных, статистические пакеты, прикладные задачи в бизнесе, маркетинге, экономике и т.д.

Предложения и замечания по данной книге просьба посылать по адресу: 117899, Москва, Воробьевы горы, МГУ, УНИР ректората МГУ, Макарову А.А.

Благодарности

В первую очередь, мы отдаем дань признательности и восхищения нашим учителям А.И.Колмогорову и Б.В.Гнеденко. Вместе с другими выдающимися учеными они были основателями школы математической статистики в СССР и учителями многих исследователей, внесших вклад в развитие отечественной статистики. Всех этих исследователей назвать невозможно, так что мы отметим лишь некоторых. Становлению прикладной статистики много способствовали регулярные семинары и школы, организатором которых был С.А.Айвазян (тоже ученик А.П.Колмогорова). Выдающуюся роль в развитии в СССР планирования эксперимента сыграл В.В.Налимов (который в свое время был сотрудником А.Н.Колмогорова). Кроме того, в распространении статистических знаний важную роль сыграли многочисленные полезные книги, как отечественные, так и переводные, выпущенные издательствами «Финансы и статистика», «Мир» и др. Всем им мы приносим свою глубокую благодарность.

Мы глубоко признательны Д.С.Шмерлингу, явившимся инициатором и вдохновителем создания этой книги, за постоянное внимание и поддержку, которую он оказывал авторам.

Мы благодарны нашему редактору В.Э.Фигурнову, который внес в текст много улучшений. Он также провел литературное и техническое редактирование, выполнил компьютерную верстку. Мы благодарны нашим рецензентам В.Н.Тутубалину и М.В.Болдину, а также многим другим нашим коллегам, внесшим ряд полезных замечаний и поправок.

Мы выражаем глубокую признательность фирмам НПО «Информатика и компьютеры», «ИнфоСтрой», «Центр Статистических Исследований МГУ», «Статистические системы и сервис», «СтатДиалог», «ТАНДЕМ» за предоставленные для ознакомления последние версии пакетов: STADIA, STATGRAPHICS, ЭВРИСТА, SPSS, SYSTAT, МЕЗОЗАВР, Статистик-Консультант.

Глава 1

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 таблице 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 игральных 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 год. Данные предоставлены А.И.Манелля, которому авторы выражают глубокую признательность.

Таблица 1.2

Урожайность зерновых культур в СССР с 1945 по 1989 гг.
(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

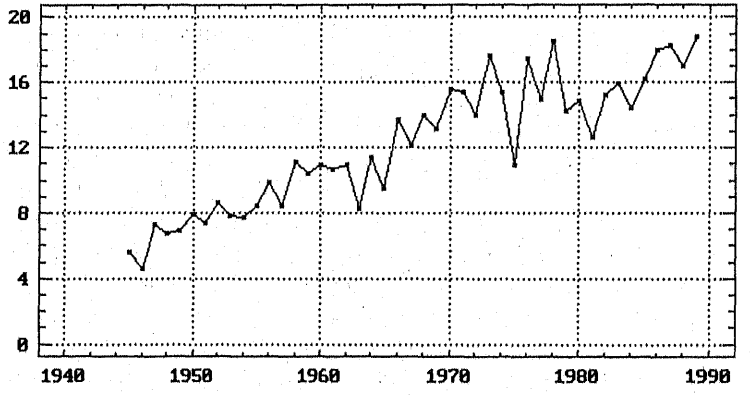


Рис. 1.1. Урожайность всех зерновых культур в СССР с 1945 по 1989 гг. (ц/га)

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: $0 \leqslant P(A) \leqslant 1$.

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:

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

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

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

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

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 $\overline{A}$. В то же время $A + \overline{A}$ — 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 $\overline{A}$, состоит в том, что в результате бросания кости выпадет 4, 5 или 6 очков.

Свойства вероятности. Теперь свойства вероятности перечислить просто:

1. $0 \leqslant P(A) \leqslant 1$ 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*

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

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:

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

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, $\sum_{\omega \in B} P(\omega \mid B)$ равна $P(B \mid B)$. Но при наступления B событие B является достоверным, поэтому согласно свойству 3 вероятностей $P(B \mid B)$ равно 1. Чтобы это условие выполнялось, естественно положить для $\omega \in B$:

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

Теперь мы можем определить условную вероятность для любого события A.

Определение. *Условная вероятность события A при условии B есть*

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

Из этого определения легко вывести, что:

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

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:

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

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

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

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

$$P(A \mid B) = P(A).$$

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 некому 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 главе 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 то 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 выше, 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, бросаемых независимо, вероятность получить в сумме 2 очка равна $1/6 \times 1/6 = 1/36$, получить 3 очка — равна $2/36$ и так далее. Распределение вероятностей суммы выпавших очков определяется следующей таблицей 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 в начале 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* $F(x) = P(\xi \leqslant x)$.

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

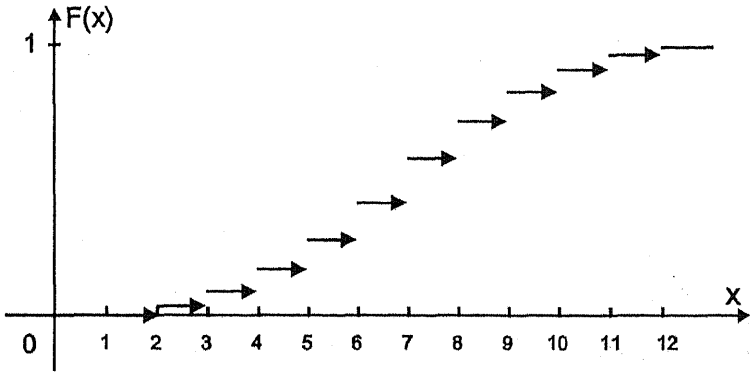


Рис. 1.2. График функции распределения суммы очков, выпавших на двух игральных костях

каждого отдельного значения случайной величины равна нулю. На этом 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 эквивалентных 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 любого[1] действи-

[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.

тельного t

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

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:

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

(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 $0 \leqslant p \leqslant 1$). В этом случае функция распределения имеет вид:

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

График этой функции изображен на рис. 1.3.

Рассмотрим функцию распределения случайной величины более общего вида. Пусть случайная величина ξ принимает конечное число значений $a_1, \ldots, a_n$, причем $P(\xi = a_k) = p_k \geqslant 0$, $\left(\sum_{k=1}^{n} p_k = 1\right)$. График функции этого дискретного распределения изображен на рис. 1.4. (Для удобства предположим, что возможные значения занумерованы в порядке возрастания.)

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

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

Практически, разумеется, используют наиболее регулярную и простую из возможных функций плотности.

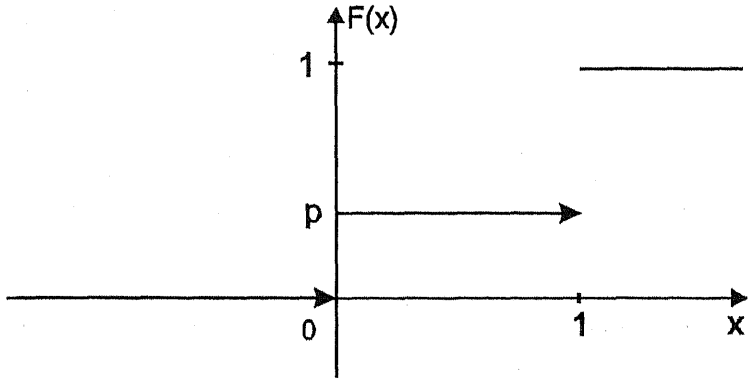


Рис. 1.3. График функции распределения, сосредоточенного в двух точках.

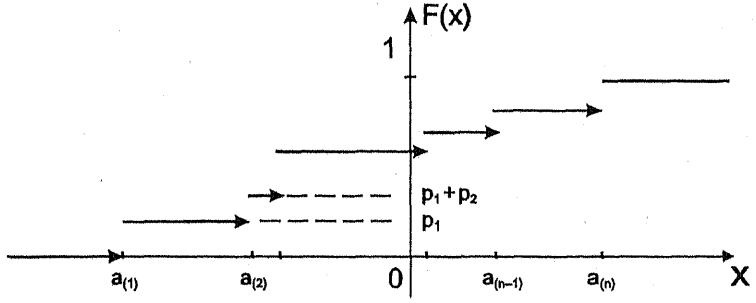


Рис. 1.4. График функции дискретного распределения

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

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

График этой функции приведен на рис. 1.5.

В приведенных примерах можно заметить, что $F(x) \to 0$ при $x \to -\infty$ и $F(x) \to 1$, при $x \to +\infty$, 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.

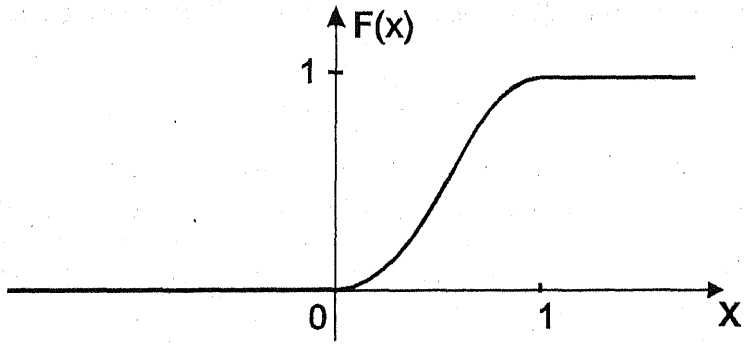


Рис. 1.5. Пример непрерывной функции распределения

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 $x_1, x_2, \ldots$, имеющих вероятности $p_1, p_2, \ldots$*

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

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, то есть, чтобы этот ряд сходился 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 $1^1, 2^2, \ldots, n^n, \ldots$ 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 ввиду 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.

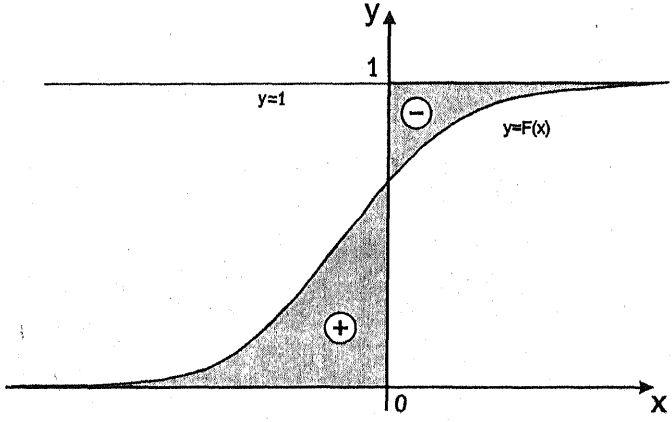


Рис. 1.6. Геометрическая интерпретация математического ожидания

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*

$$D\xi = M(\xi - M\xi)^2, \qquad \textit{или} \qquad D\xi = M\xi^2 - (M\xi)^2.$$

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

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

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 $x_1, x_2, \ldots$, имеющих вероятности $p_1, p_2, \ldots$, k-ым моментом $M\xi^k$ называется величина $M\xi^k = \sum_i x_i^k p_i$, а k-ым центральным моментом называется величина $\sum_i (x_i - M\xi)^k p_i$. Для непрерывной случайной величины с плотностью $p(x)$, k-ым моментом называется величина $\int_{-\infty}^{\infty} x^k p(x)\,dx$, а k-ым 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, то есть 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:

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

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 наблюденные 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 $0.1, 0.2, \ldots, 0.9$. 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 неединственное 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:

$$M\xi\eta = M\xi\, M\eta,$$
$$D(\xi + \eta) = D\xi + D\eta,$$

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* $\operatorname{cov}(\xi, \eta)$ *случайных величин* ξ *и* η *называют*

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

если указанное математическое ожидание существует.

Легко видеть, что верна и другая формула:

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

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

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

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* $\operatorname{corr}(\xi, \eta)$*, либо* $\rho(\xi, \eta)$*, либо просто* ρ*) называют*

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

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. $|\rho(\xi,\eta)| \leqslant 1$
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 $D(\eta - t\xi) = t^2 D\xi - 2t\operatorname{cov}(\xi,\eta) + D\eta$, т.е. она представляется квадратным трехчленом от t. Этот квадратный трехчлен неотрицателен, поскольку дисперсия всегда неотрицательна. Поэтому его дискриминант $[\operatorname{cov}(\xi,\eta)]^2 - D\xi\, D\eta \leqslant 0$, а это и означает, что $|\rho(\xi,\eta)| \leqslant 1$ (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, то есть была бы, как говорят, *репрезентативной*. Как этого добиться? Если генеральная совокупность нам мало известна или совсем неизвестна, не удается 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 году.

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.Гэллап 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 [41].

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 $x_1, x_2, \ldots, x_n$. Тогда $x_1, x_2, \ldots, x_n$ будут независимыми одинаково распределенными случайными величинами, то есть *выборкой*. Зная величины $x_1, x_2, \ldots, x_n$, 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 $x_1, x_2, \ldots, x_n$, называется функция $F_n(x)$, равная доле таких значений x_i, что $x_i \leqslant x$, $i = 1, \ldots, n$.*

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

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

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

Так, $x_{(1)}$ обозначает наименьшее из чисел $x_1, \ldots, x_n$, $x_{(2)}$ — наименьшее из оставшихся после удаления $x_{(1)}$ и т.д. В частности, $x_{(n)}$ обозначает наибольшее из $x_1, \ldots, x_n$. 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 $x_{(1)}, \ldots, x_{(n)}$, причем величина каждого скачка равна $1/n$ (рис. 1.7).

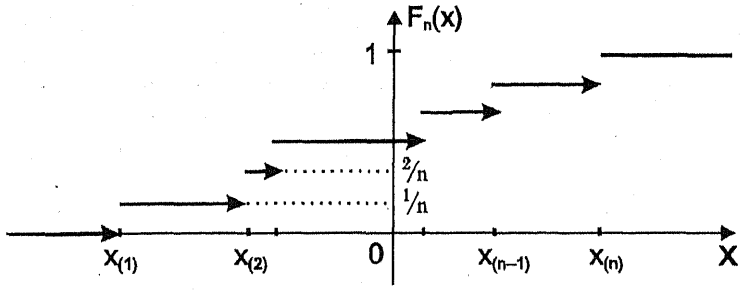


Рис. 1.7. Общий вид эмпирической функции распределения

Видно, что график эмпирической функции распределения напоминает график дискретного распределения вероятностей. Это не случайно: эмпирическую функцию выборки $x_1, \ldots, x_n$ можно рассматривать как функцию распределения вероятностей, где каждому значению x_i, $i = 1, \ldots, n$, 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 $(\xi \leqslant x)$ в ряду из n независимых повторений. Поэтому $F_n(x) \to F(x)$ при $n \to \infty$.

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 $n \to \infty$ 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. приводят 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 $x_{(1)}, \ldots, x_{(n)}$, 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*

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

Аналогично,

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

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

Однако в статистике чаще в качестве выборочной дисперсии используют

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

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 $x_{(k)} \leqslant x < x_{(k+1)}$ $F(x) = 0.5$. 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 $\left\{\frac{i}{2k+1},\ i = 0, 1, \ldots, 2k+1\right\}$. 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: $F_n\left(x_{(k+1)}\right) = \frac{(k+1)}{2k+1} \to \frac{1}{2}$.

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 главах 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 $(x_1, y_1), \ldots, (x_n, y_n)$, или *выборочным коэффициентом корреляции* называют величину

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

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

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 из 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 обыкновенно 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 $x_1, \ldots, x_n$ называют *связкой*; количество таких одинаковых наблюдений в данной связке называют ее размером.) Средний ранг x_i в ранжировке наблюдений $x_1, \ldots, x_n$ 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 $x_1, x_2, \ldots, x_n$, 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 $x_1, x_2, \ldots, x_n$ являются выборкой, то есть 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 $x_1, x_2, \ldots, x_n$ 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 $x_1, x_2, \ldots, x_n$ используют другие методы — методы описательной статистики.

Определение. *Методами описательной статистики принято называть методы описания выборок $x_1, x_2, \ldots, x_n$ 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 член 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 характеристик.
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 так же 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, то есть 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 $x_1, \ldots, x_n$, 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 $t_1, t_2, \ldots$ — центры (середины) выбранных промежутков. Тогда вместо выборочного среднего $\overline{x}$ используют величину $\overline{t}$:

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

а в качестве выборочной дисперсии s^2

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

Приведем ниже еще несколько полезных приемов описательной статистики для работы с выборкой. В качестве примера рассмотрим данные из таблицы 1.1, в которой приведены результаты измерения диаметров 200 головок заклепок. Здесь случайная величина — диаметр изготовляемой 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 таблицы 1.1 приведена на рис. 1.8.

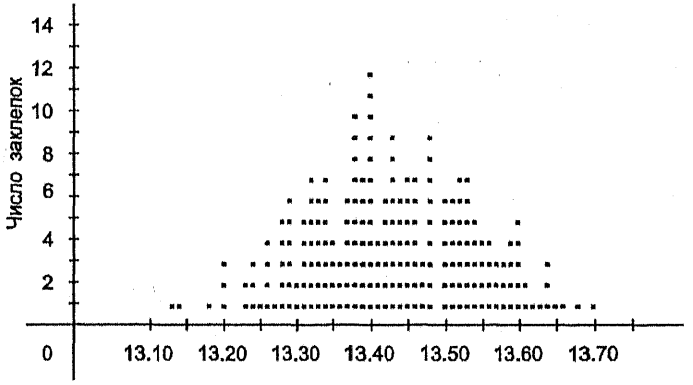


Рис. 1.8. Точечная диаграмма. Распределение диаметров 200 головок заклепок, выраженных в мм.

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 наблюденных результатов $x_1, \ldots, x_n$ на m интервалов, которые будем называть *интервалами группировки*. Длины интервалов обозначим через $\Delta_1, \ldots, \Delta_m$, а середины интервалов группировки — через $t_1, \ldots, t_m$.

Число наблюдений n_{ij} в j-м интервале группировки равно количеству x_i, $i = 1, \ldots, n$, 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-ый 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.

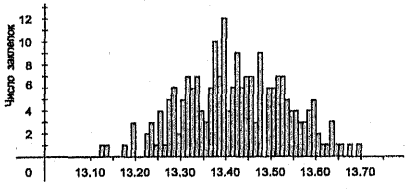


Рис. 1.9. Гистограмма. Длина интервала группировки равна 0.01 мм

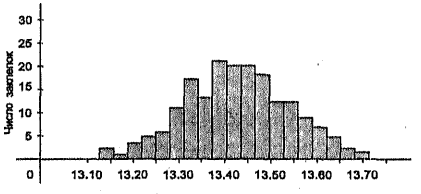


Рис. 1.10. Гистограмма. Длина интервала группировки равна 0.03 мм

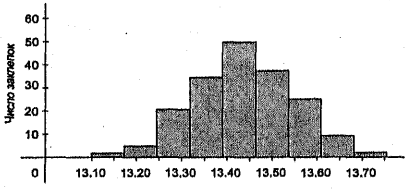


Рис. 1.11. Гистограмма. Длина интервала группировки равна 0.07 мм

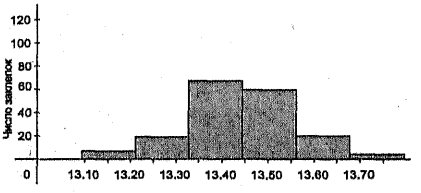


Рис. 1.12. Гистограмма. Длина интервала группировки равна 0.11 мм

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 Карлсберговской 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 пакетах STADIA и STATGRAPHICS

1.9.1. Пакет STADIA

В пакете STADIA довольно полно представлены методы описательной статистики, все они собраны воедино в разделе пакета «Параметрические тесты» меню Статистические методы (смотри описание структуры пакета в приложении 2). Проиллюстрируем их работу на рассмотренных

выше примерах. При этом будет рассмотрена версия пакета STADIA 6.0 для Windows. Решение задач в DOS-версии пакета STADIA 5.0 аналогично и отличается только видом окон ввода данных и параметров процедур.

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

Подготовка данных. Находясь в электронной таблице пакета, следует либо ввести данные таблицы 1.1 с клавиатуры, либо загрузить их из уже созданного файла. Пусть данные таблицы 1.1 находятся в текстовом (ASCII) файле DIAMZ.TXT в виде столбца с именем d. Для загрузки файла данных в пакет STADIA в пункте меню Файл выберите подпункт Ввести, как это показано на рис. 1.13.

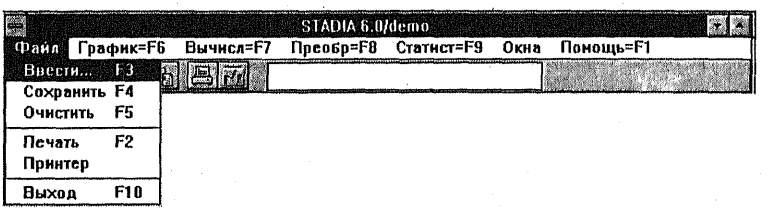


Рис. 1.13. Пакет STADIA. Вызов процедуры загрузки файла данных

В открывающемся при этом диалоговом окне Чтение файла (рис. 1.14) укажите необходимый тип файла данных, каталог и имя файла. Для выбора файла достаточно подвести указатель мыши к имени файла и дважды нажать левую кнопку мыши.

Результаты ввода (загрузки) данных в электронную таблицу представлены на рис. 1.16.

Выбор процедуры. После выбора пункта меню Статист или нажатия клавиши [F9] (см. рис. 1.13) программа выведет на экран меню Статистические методы (рис. 1.17).

С помощью мыши выберите в меню пункт 1 = Описательная статистика. На экране появится окно Анализ переменных (рис. 1.15). В нем можно выбрать одну или несколько переменных из электронной таблицы для дальнейшего анализа. Выделив переменную d в списке переменных, нажмите мышью на кнопку со стрелкой вправо. Выбранная переменная переместится в поле для анализа. Завершив выбор переменных, нажмите кнопку [Утвердить]

Результаты. На экране в окне Результаты появятся значения основных описательных статистик (см. верхний ряд чисел рис. 1.18) и запрос

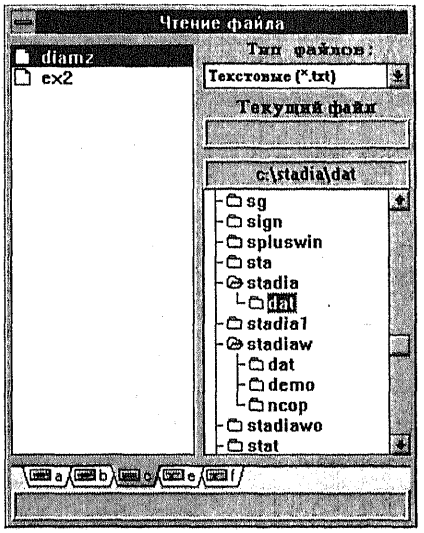


Рис. 1.14. Пакет STADIA
Окно чтения файла данных

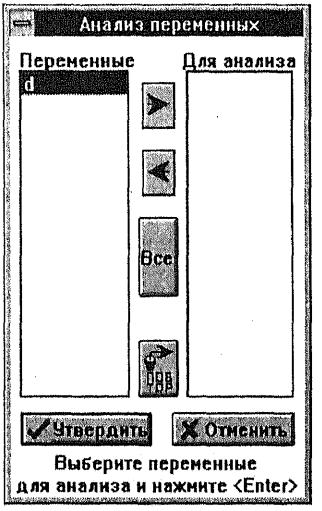


Рис. 1.15. Пакет STADIA. Окно выбора переменных для анализа

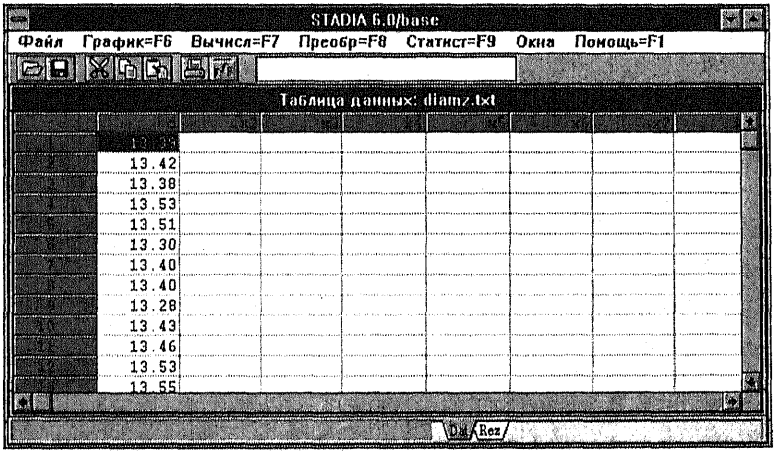


Рис. 1.16. Пакет STADIA. Электронная таблица с загруженными данными

системы Выдать дополнительную статистику. В ответ на запрос можно нажать [Да] (или [Yes]), и тогда программа выведет остальные описательные статистики (рис. 1.18). Заметим, что во встроенном справочнике программы имеются определения и сведения о назначениях всех этих описательных статистик. Для вывода данной информации на экран следует нажать [F1].

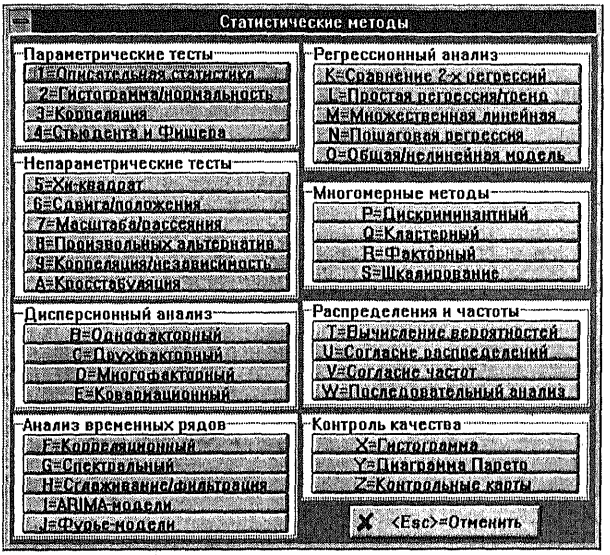


Рис. 1.17. Пакет STADIA. Меню статистических методов

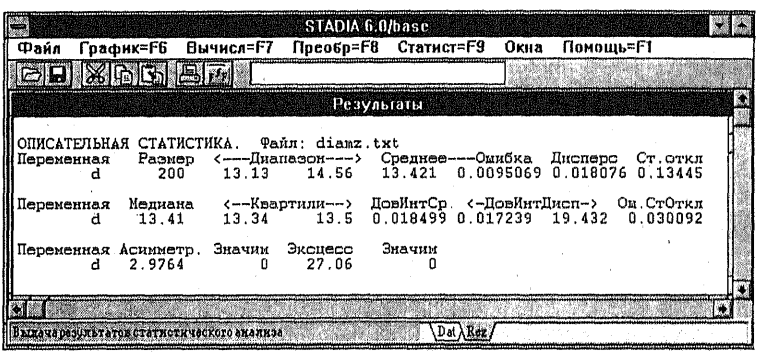


Рис. 1.18. Пакет STADIA. Окно результатов процедуры описательной статистики

Требуемые в примере максимальное и минимальное значение выборки находятся в графе <–Диапазон–>, верхняя и нижняя квартили — в графе <–Квартили–>. Названия остальных необходимых в примере характеристик присутствуют на экране в явном виде.

Комментарии. 1. Часть описательных статистик, вычисляемых этой процедурой, относится только к выборкам из нормального распределения. Это касается размера доверительного интервала для среднего и значений концов доверительного интервала для дисперсии.

2. Если в окне выбора переменных для анализа выбрано несколько переменных, то будут вычислены описательные статистики для каждой из этих переменных.

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.

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

Выбор процедуры. В меню статистических методов (рис. 1.17) следует выбрать процедуру 2=Гистограмма/нормальность, нажав на экране соответствующую кнопку мышью или нажав клавишу [2].

Заполнение полей ввода данных. На экране появится окно Анализ переменной (рис. 1.15), в котором следует выбрать переменную d для анализа. Далее последует запрос пакета о параметрах группировки данных (рис. 1.19). Введем число интервалов группировки равным 12, левую границу группировки данных — 13 и правую границу — 14.8, как это показано на рис. 1.19. Затем нажмите кнопку [Утвердить].

Рис. 1.19. Задание интервалов группировки

Результаты. На экране появятся результаты расчетов, включающие таблицу табуляции частот (рис. 1.20), значения статистик Колмогорова, омега-квадрат и хи-квадрат, а также заключение системы Гипотеза 1: Распределение отличается от нормального.

В первом столбце таблицы указан правый конец интервала группировки, во втором значения первого столбца трансформированы следующим образом: из каждого элемента первого столбца вычитается среднее значение выборки и полученная разность делится на стандартное отклонение выборки. Следующие четыре столбца содержат частоту, относительную частоту, накопленную частоту и относительную накопленную частоту соответственно.

После нажатия [Enter] появится запрос системы Вывести график?. При ответе [Да] (или [Yes]) программа выводит гистограмму и подобранную по выборке кривую плотности нормального распределения в специальное графическое окно. Полученные графики показаны на рис. 1.21 слева.

Комментарии. 1. Изучение таблицы табуляции частот показывает, что в выборке находится одно сильно выделяющееся наблюдение, которое, по-видимому, и оказало влияние на результат проверки нормальности. Влияние

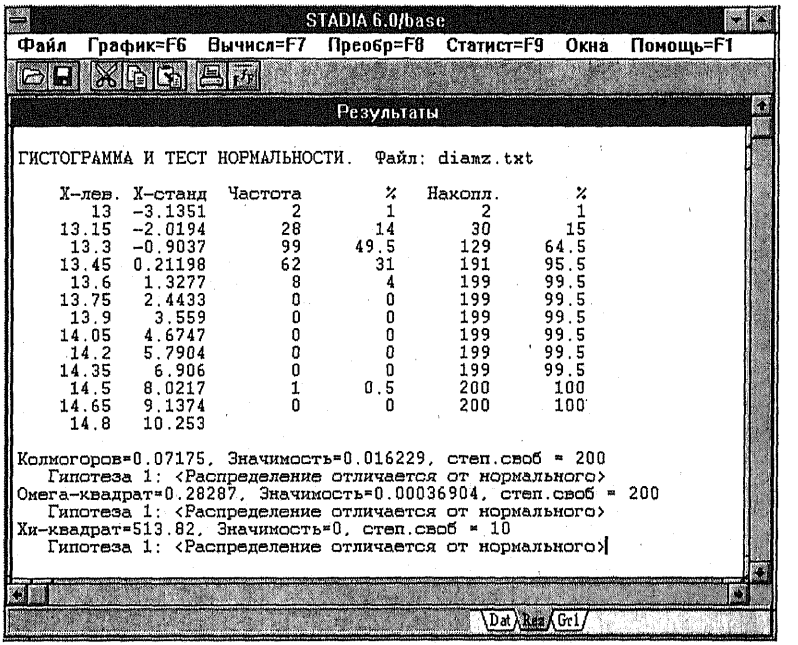


Рис. 1.20. Пакет STADIA. Экран результатов процедуры «Гистограмма и нормальность»

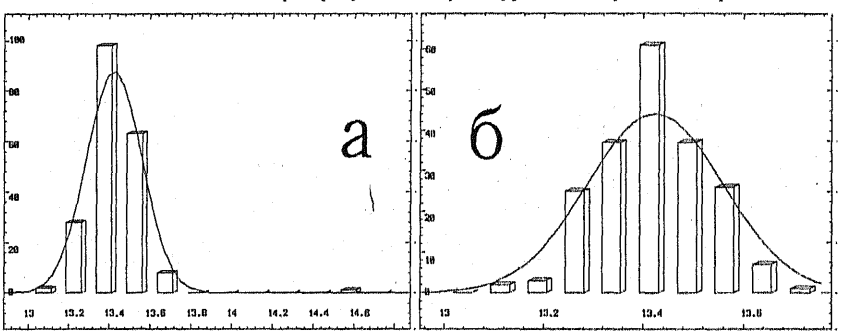


Рис. 1.21. Гистограмма с наложенным графиком нормальной кривой (а — исходные данные, б — без учета сильно выделяющегося наблюдения)

этого наблюдения на различные выборочные статистики будет рассмотрено в главах 5 и 10. В примере 1.3к мы проведем приведенные выше расчеты без учета сильно выделяющегося наблюдения.

2. Вопросы, связанные с проверкой нормальности в данной процедуре, обсуждаются в главах 5 и 10.

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 осуществляется так же, как в примере 1.1к для пакета STADIA.

Выбор процедуры. В блоке статистических методов нажатием клавиши [2] следует выбрать процедуру 2=Гистограмма и нормальность.

Заполнение полей ввода данных. На запрос системы Укажите число интервалов и диапазон гистограммы (Enter=вычисл) следует ввести скорректированные значения: 10, 13, 13.75, и затем нажать [Enter].

Результаты. На экране (рис. 1.22) появится результаты расчетов, включающие таблицу табуляции частот, значения статистик Колмогорова, омега-квадрат и хи-квадрат, а также заключение системы Гипотеза 0: Распределение не отличается от нормального. Выводимый для этого случая график показан на рис. 1.21 справа.

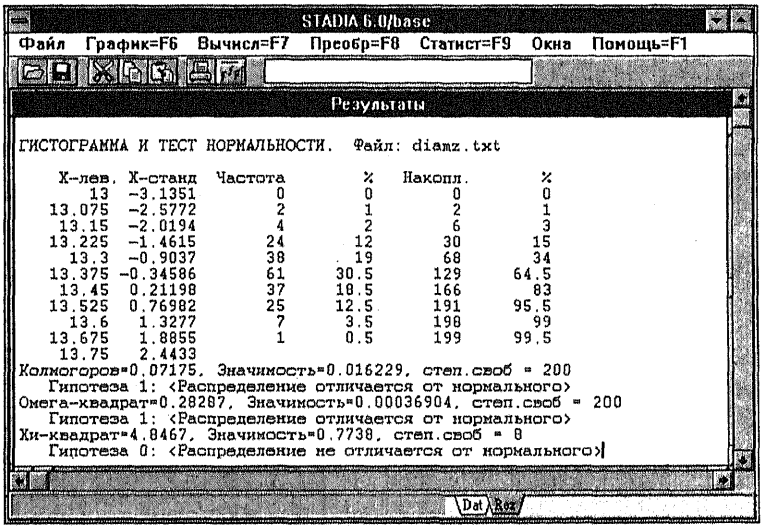


Рис. 1.22. Пакет STADIA. Экран результатов процедуры «Гистограмма и нормальность»

Прочие возможности. Из других графических методов описательной статистики в пакете STADIA представлен также матричный график, в котором значения каждой переменной, находящейся в текущий момент в блоке редактора данных, сгруппированы в отдельном столбце графика с указанием положения среднего значения и интервала стандартного отклонения.

1.9.2. Пакет STATGRAPHICS

В пакете STATGRAPHICS широко представлены численные и графические методы описательной статистики. Большинство из них сгруп-

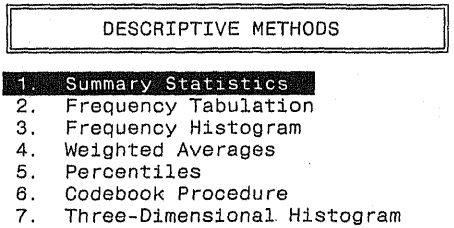


Рис. 1.23. Пакет STATGRAPHICS. Меню описательных методов статистики

пированы в пункте F. Descriptive Methods (описательные методы) головного меню пакета (рис. 1.23).

Рассмотрим меню рис. 1.23, кратко опишем назначения входящих в него процедур и разберем на примерах наиболее употребительные из них.

1. Summary Statistics (описание данных) — позволяет получить широкий набор числовых характеристик, включая среднее, дисперсию, стандартное отклонение, медиану, квартили, для одной или нескольких выборок. Работа этой процедуры разобрана в примере 1.1к.

2. Frequency Tabulation (табуляция частот) — выдает таблицу частот и накопленных частот после группирования данных в заданное число классов (интервалов группировки). В отдельных режимах эта процедура может выдавать диаграмму частот и эмпирическую функцию распределения. Ей посвящен пример 1.2к.

3. Frequency Histogram (гистограмма частот) — в графическом виде представляет таблицу частот после группирования данных в заданное число классов. Эта процедура подробно разбирается в примере 1.3к.

4. Weighted Averages (взвешенные средние) — вычисляет средние значения для сгруппированных данных или для данных, отношение к которым с точки зрения их точности или важности не равнозначно. Не разбирая этой процедуры подробно, мы ограничимся только замечаниями по вводу данных в процедуру для разных случаев.

5. Percentiles (процентили) — вычисляет выборочные процентили для указанных пользователем процентов, а так же находит процент наблюдений, лежащих левее указанного числа.

6. Codebook Procedure (описание группированных данных) — вычисляет те же описательные статистики, что и процедура 1. Summary Statistics, но только для указанных подмножеств вектора данных. Замечания по заполнению полей ввода этой процедуры будут даны ниже.

7. Three-Dimensional Histogram (трехмерная гистограмма) — строит трехмерную гистограмму частот для двумерных наблюдений, т.е. для двумерной выборки.

Ряд менее распространенных и привычных методов описательной статистики находятся в других разделах пакета. Так, в пункте I. Exploratory data anallysis (разведочный анализ) головного меню пакета можно получить доступ к описательным процедурам Box-and-Whisker Plot, Multiple Box-and-Whisker Plot и Notched Box-and-Whisker Plot. Прямой перевод термина «Box-and-Whisker Plot» на русский язык звучит как «ящик с усами».

Собственно, именно так и выглядят графики, которые строятся в этих процедурах. Описание этих процедур можно найти в [76]. Специфические описательные методы, возникающие в модельных задачах будут рассмотрены в следующих главах.

Рассмотрим несколько примеров.

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

Подготовка данных. В редакторе базы данных пакета (процедура File Operations) следует создать файл с именем DIAMZ, в этом файле в режиме редактирования создать числовую переменную с именем d и ввести в нее (для определенности, по столбцам) значения из табл. 1.1. Подробное описание процедуры File Operations смотри в п. П2.6

Выбор процедуры. В меню пункта F. Descriptive Methods клавишами управления курсором надо установить активное поле на 1. Summary Statistics и нажать F6. (Это стандартный способ порядка выбора любой процедуры пакета.) При этом осуществится переход к экрану ввода данных в процедуру (рис. 1.24).

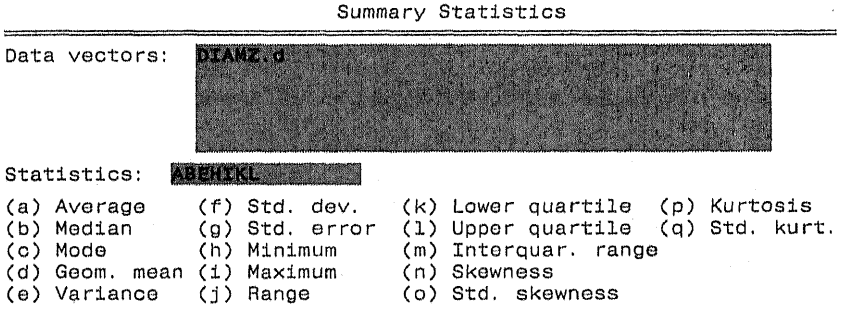


Рис. 1.24. Пакет STATGRAPHICS. Экран ввода данных в процедуру описательной статистики

Заполнение полей ввода данных. В любое место поля Data vectors надо ввести имя файла и переменной DIAMZ.d, содержащей данные примера. С помощью клавиши Tab следует перейти в поле Statistics. Используя клавиши управления курсором и клавишу Del, в этом поле надо оставить буквы, означающие требуемые характеристики (таблица соответствия приведена в нижней части экрана ввода). Названия некоторых статистических характеристик этой таблицы приведены нами в сокращении. Приведем их наименование в том виде, как они указываются в пакете, и их русский перевод:

Average — среднее значение
Median — медиана
Mode — мода
Geometric mean — геометрическое среднее
Variance — дисперсия
Std. deviation — стандартное отклонение
Std. error — стандартная ошибка
Minimum — минимум
Maximum — максимум
Range — размах
Lower quartile — нижняя квартиль
Upper quartile — верхняя квартиль
Interquar. range — межквартильный размах
Skewness — коэффициент асимметрии
Std. skewness — нормированный коэффициент асимметрии
Kurtosis — коэффициент эксцесса
Std. kurtosis — нормированный коэффициент эксцесса

Завершив выбор требуемых характеристик, надо нажать клавишу [F6].

Результаты. На экране появится таблица результатов вычислений (рис. 1.25).

Variable:	DIAMZ.d
Sample size	200
Average	13.4215
Median	13.41
Variance	0.0180761
Minimum	13.13
Maximum	14.56
Lower quartile	13.34
Upper quartile	13.5

Рис. 1.25. Пакет STATGRAPHICS. Экран выдачи результатов процедуры описательной статистики

После нажатия [Esc] или [F10] произойдет возврат в экран ввода данных процедуры, на котором появится всплывающее меню (рис. 1.26). С его помощью Вы можете, в частности, сохранить полученные результаты в базе данных пакета (пункт меню Save statistics).

Redisplay results	повторить вывод на экран
Save statistics	сохранить результаты
Save labels	сохранить метки

Рис. 1.26. Пакет STATGRAPHICS. Всплывающее меню дополнительных возможностей процедуры описательной статистики

Комментарии. 1. Если в поле ввода процедуры Data vectors указать только имя файла без имени переменной, то выборочные статистики будут вычислены для всех числовых переменных, находящихся в файле. Символьные переменные будут проигнорированы.

2. Числовые данные небольших объемов можно ввести в поле Data vectors непосредственно с клавиатуры. В этом случае они будут обрабатываться построчно, как разные переменные.

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.

Выбор процедуры. В меню пункта F. Descriptive Methods надо выбрать процедуру 2. Frequency Tabulation и нажать (F6). Появится экран ввода данных в процедуру (рис. 1.27)

```
                    Frequency Tabulation

Data:  DIAMZ.d
```

Рис. 1.27. Пакет STATGRAPHICS. Экран ввода данных в процедуру табуляции частот

Заполнение полей ввода данных. В поле Data надо ввести имя файла и имя переменной DIAMZ.d, содержащей данные примера. После нажатия клавиши (F6) появится запрос ввода параметров процедуры (рис. 1.28), заполненный пакетом по умолчанию.

```
                  Tabulation Input Panel
Primary Variable                          Secondary Variable

Type                Continuous            Type
Lower limit         13                    Lower limit
Upper limit         14.8                  Upper limit
No. of classes      12                    No. of classes

Length = 200                              Length =
Minimum = 13.13                           Minimum =
Maximum = 14.56                           Maximum =
```

Рис. 1.28. Пакет STATGRAPHICS. Экран ввода параметров процедуры табуляции частот

В столбце ввода параметров Primary Variable (первая переменная) в поле Type (тип данных), последовательно нажимая клавишу (Пробел), надо установить значение Continuous (непрерывный). Возможности других режимов заполнения этого поля смотри в комментариях. При необходимости следует скорректировать поля Lower limit (нижний предел) и Upper limit (верхний предел). Разделив величину диапазона группировки 1.8 на заданный шаг группировки 0.15, получим число интервалов группировки 12, которое и надо ввести в поле No. of classes (число классов группировки). Выбору параметров группировки помогает справочная информация о выборке, указывающая размер выборки (Length), ее

минимальный и максимальный элементы. Столбец ввода параметров, относящийся к Secondary Variable (второй переменной), в данной процедуре не используется. После заполнения полей ввода и нажатия клавиши F6 на экране появится всплывающее меню форм выдачи результатов процедуры (рис. 1.29). (Напротив каждой процедуры этого меню нами приведен перевод на русский язык.)

Display table	таблица
Plot frequency polygon	диаграмма частот
Plot rel. freq. polygon	диаграмма относительных частот
Plot cumulative freqs.	график накопленных частот
Plot cumulative rel. freqs.	график накопленных относительных частот

Рис. 1.29. Пакет STATGRAPHICS. Меню форм вывода результатов процедуры табуляции частот

Клавишами перемещения курсора установим активное поле на пункт Display table и нажмем клавишу F6.

Результаты. На экран будет выведена таблица, представленная на рис. 1.30.

Frequency Tabulation

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum.Rel. Frequen.
at or below	13.00			0	.00000	0	.0000
1	13.00	13.15	13.08	2	.01000	2	.0100
2	13.15	13.30	13.23	28	.14000	30	.1500
3	13.30	13.45	13.38	99	.49500	129	.6450
4	13.45	13.60	13.52	62	.31000	191	.9550
5	13.60	13.75	13.68	8	.04000	199	.9950
6	13.75	13.90	13.83	0	.00000	199	.9950
7	13.90	14.05	13.98	0	.00000	199	.9950
8	14.05	14.20	14.13	0	.00000	199	.9950
9	14.20	14.35	14.28	0	.00000	199	.9950
10	14.35	14.50	14.43	0	.00000	199	.9950
11	14.50	14.65	14.58	1	.00500	200	1.0000
12	14.65	14.80	14.73	0	.00000	200	1.0000
above	14.80			0	.00000	200	1.0000

Mean = 13.4215 Standard Deviation = 0.134448 Median = 13.41

Рис. 1.30. Пакет STATGRAPHICS. Экран вывода результатов процедуры табуляции частот в форме таблицы

Приведенная таблица включает в себя значения верхней (Lower Limit) и нижней (Upper Limit) границ интервала группировки, его середину (Midpoint), число (Frequency) и относительную частоту (Relative Frequency) попаданий в интервал группировки, а также их накопленные показатели.

Комментарии. Указание в поле ввода Type экрана ввода процедуры значения List означает отказ от режима группировки данных. В этом случае Вы сможете получить представление своих данных в виде точечной диаграммы (и

соответствующей ей таблицы) и посмотреть их эмпирическую функцию распределения. В противном случае в поле Type надо указать значение Continuous для данных действительного типа и значение Discret — для целочисленных данных.

Пример 1.3к. Для выборки диаметров головок заклепок построить гистограмму частот с шагом группировки 0.075 мм.

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

Выбор процедуры. В меню пункта F. Descriptive Methods следует выбрать процедуру 3. Frequency Histogram и нажать F6. При этом появится экран ввода данных, аналогичный приведенному на рис. 1.27.

Заполнение полей ввода данных. Ввод данных на первом этапе — такой же, как в процедуре 2. Frequency Tabulation. Он разобран в примере 1.2к. Затем появляется экран ввода параметров процедуры (рис. 1.31).

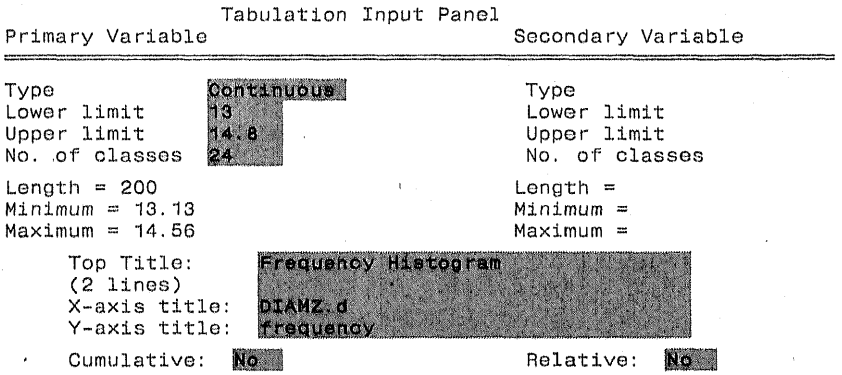


Рис. 1.31. Пакет STATGRAPHICS. Экран ввода параметров процедуры гистограммы частот

Заполнение полей Type, Lower limit, Upper limit, No. of classes аналогично процедуре 2. Frequency Tabulation (см. пример 1.2к). При требуемом шаге группировке 0.075 и выбранных верхнем и нижнем пределах число классов группировки равно 24. Поля Top Title: (2 lines) (заголовок графика), X-axis title (название оси X), Y-axis title (название оси Y) предназначены для оформления надписей на графике. Их заполнение не обязательно. Мы оставили этих полях значения, предложенные пакетом по умолчанию. Наконец, поля Cumulative (накопление) и Relative (относительный) задают типы графиков, выводимых на экран. В этих полях могут фигурировать значения Yes и No в произвольной комбинации. В зависимости от их комбинации на график будут выводиться данные одного из последних четырех столбцов таблицы, выдаваемой в результате работы процедуры 2. Frequency Tabulation (см. пример 1.2к). После заполнения необходимых

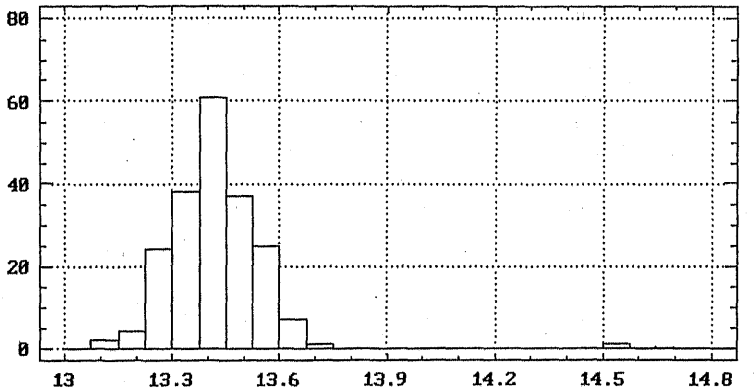


Рис. 1.32. Гистограмма. Длина интервала группировки равна 0.075 мм.

полей ввода параметров надо нажать клавишу F6, и на экране появится график, приведенный на рис. 1.32.

Комментарии. 1. При выводе графика на экран подсказка со значениями функциональных клавиш исчезает, однако функциональные клавиши работают в обычном режиме.

2. Для вывода на печать изображенного на экране графика надо нажать клавишу F4. Печать может быть произведена как на принтер, так и в файл.

3. Для изменения оформления графика можно воспользоваться клавишей F5 или режимом интерактивного редактирования графика. Часть возможностей этих режимов разобрана в п. П2.7.

Другие возможности. Кратко остановимся на особенностях ввода данных и работы других процедур блока DESCRIPTIVE METHODS.

Процедура 4. Weighted Averages (взвешенные средние). В случае сгруппированных данных в активное поле экрана Data vector or matrix (вектор или матрица данных) следует ввести середины интервалов группировки, а в активное поле Weight vector (вектор весов) — вектор, координатами которого являются числа попаданий наблюдений в соответствующие интервалы (или их частоты). Для несгруппированных данных в поле Weight vector необходимо указать числовой вектор, отражающий степень достоверности или важности каждого из элементов вектора данных. При этом учитываются не абсолютные значения вектора весов, а только их отношения друг к другу. Взвешенное среднее вычисляется здесь путем умножения каждого значения вектора данных на соответствующий вес и делением суммы этих чисел на сумму весов.

Если в поле Data vector or matrix введена матрица, то длина вектора весов должна быть равна числу строк матрицы. Каждый столбец матрицы будет обработан отдельно с этим вектором весов.

Процедура 5. Percentiles (процентили). Ввод данных в эту процедуру предусматривает заполнение активных полей Data vector и Percentages (проценты). В последнее вводятся значения процентов, для которых Вы хотите получить выборочные процентили. После нажатия клавиши F6 в графе Percentiles появятся ответы. Так, для получения значений выборочной медианы, нижней и верх-

ней квартили в поле Percentages необходимо ввести столбиком значения 50, 25 и 75, что соответствует уровням квантилей 0.5, 0.25 и 0.75. Предварительно нажав клавишу [F5] и выбрав в появившемся подменю пункт Switch input fields (переключение поля ввода), можно решить обратную задачу, то есть вычислить процент наблюдений, лежащих слева от введенного вами числа в активное поле Percentiles. Заметим, что вводимое при этом число не обязано быть элементом рассматриваемой выборки.

6. Codebook Procedure (описание группированных данных). Задание разбиения вектора данных на подмножества в этой процедуре осуществляется с помощью задания вектора классификации в активном поле Level codes (коды уровня). Его координаты могут быть как числовыми, так и символьными, а длина должна совпадать с длиной вектора данных. При этом для включения элементов вектора данных в одно подмножество необходимо в векторе классификации присвоить соответствующим координатам одно и то же значение.

Отметим, что подобная процедура очень удобна для работы с данными, хранение которых организовано по общим принципам баз данных и где возможные вектора классификации, как правило, является элементами самой базы данных.

Заполнение активного поля экрана Labels (метки) не обязательно. Работа с полем Statistics аналогична работе с этим полем в процедуре 1. Summary Statistics (см. пример 1.1к).

Активные поля Confidence level (доверительный уровень) и Point symbols (символы точек) определяют соответствующие параметры сопутствующих этой процедуре графиков, доступ к которым дает нажатие клавиши [F5] после того, как получены значения описательных статистик и осуществлен возврат к экрану ввода данных нажатием [Esc]. Один из сопутствующих этой процедуре графиков Plot standard errors (график стандартных ошибок) выводит значения среднего и стандартной ошибки для каждого подмножества, а также значения всех точек каждого подмножества, если в поле Point symbols (символы точек) указан ответ Yes. Другой график Plot confidence intervals (график доверительных интервалов), в отличие от предыдущего, вместо стандартных ошибок строит доверительные интервалы для средних с указанным Вами доверительным уровнем.

7. Three-Dimensional Histogram (трехмерная гистограмма). Экран ввода данных в эту процедуру требует заполнения двух активных полей Sample 1 (выборка 1) и Sample 2 (выборка 2). После нажатия клавиши [F6] появляется экран ввода параметров табуляции, работа с которым была описана при рассмотрении процедур 2. Frequency Tabulation и 3. Frequency Histogram. После коррекции параметров табуляции и нажатия клавиши [F6] на экране появится требуемый результат. Отметим, что графический редактор пакета предусматривает возможность пространственного вращения трехмерного графика (изменения точки обзора), что позволяет выбрать наиболее наглядный вид графика.

Глава 2

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 об 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 [16], [50], [60], [68] и [87].

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 $0, 1, \ldots, n$ с вероятностями:*

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

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:

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

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 $1, 2, \ldots$ отдельным испытаниям, то те же номера надо присвоить и соответствующим им случайным величинам $\xi : \xi_1, \xi_2, \ldots$ Тогда X можно представить в виде: $X = \xi_1 + \xi_2 + \cdots + \xi_n$, 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).

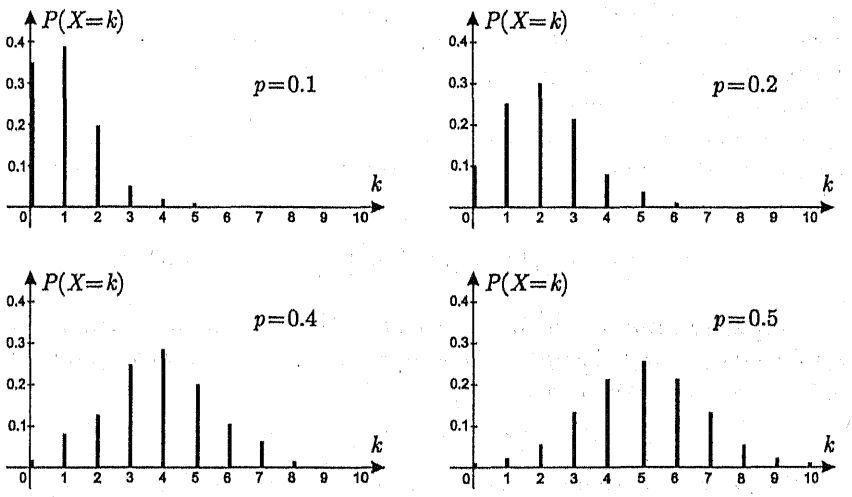


Рис. 2.1. Вид биномиального распределения для различных значений p при $n = 10$

Связь с другими распределениями. Биномиальное распределение тесно связано с многими другими распределениями. Ниже мы укажем наиболее часто используемые из этих связей. Описание других можно найти в [16], [87].

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 $0.1 \leqslant p \leqslant 0.9$. 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 [16] приведены таблицы $P(X = k \mid n, p)$ (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

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

Например, в [60], $P(X \leqslant k \mid n, p)$ 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, [16]).

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:

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

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 $\Delta \to 0$.
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 $0, 1, 2, \ldots$, имеет закон распределения Пуассона с параметром $\lambda > 0$, если*

$$P(\xi = k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad \text{для } k = 0, 1, 2, \ldots$$

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

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

Эти выражения несложно получить прямыми вычислениями. Имеем:

$$M\xi = \sum_{k=0}^{\infty} kP(\xi = k \mid \lambda) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{(k-1)}}{(k-1)!} e^{-\lambda}$$

$$= \lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} = \lambda.$$

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 $P(\xi = k \mid \lambda)$ для различных значений k и λ .

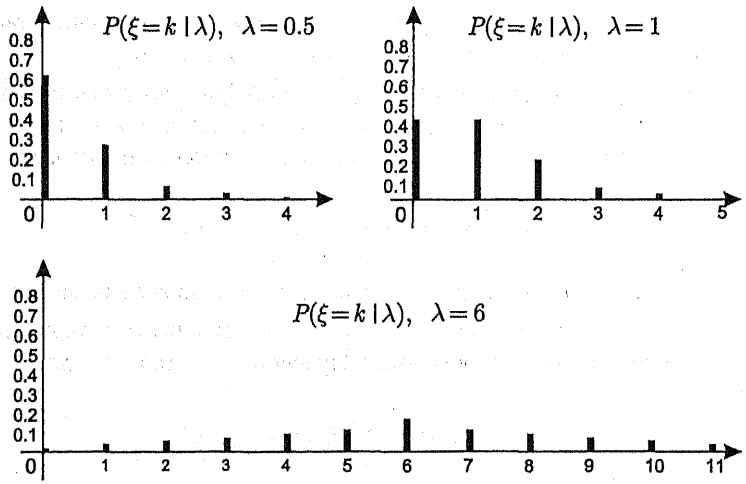


Рис. 2.2. Вид распределения Пуассона для различных значений k и λ

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:

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

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

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

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 $\lambda_1, \lambda_2, \ldots, \lambda_n$ соответственно, имеет также распределение Пуассона с параметром

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

Таблицы. Таблицы распределения Пуассона при различных значениях даны, например, в [16], [50], [60], а также в других сборниках таблиц и монографиях.

Дадим описание таблиц, приведенных в [16] для $P(\xi = k \mid \lambda)$. При этом значение λ изменяется от 0.1 (0.1) 15.0, а значение k изменяется с единичным шагом в таких пределах, где $P(\xi = k \mid \lambda) > 5 \cdot 10^{-7}$. Там

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

Более подробные таблицы распределения Пуассона даны в [50], где λ изменяется до 205. Отметим, что при больших значениях λ для вычисления $P(\xi = k \mid \lambda)$ можно использовать приближенную формулу

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

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

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

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

В [60] приведены таблицы $P(\xi \leqslant k \mid \lambda)$ 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 [42].

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*

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

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:

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

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

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

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.

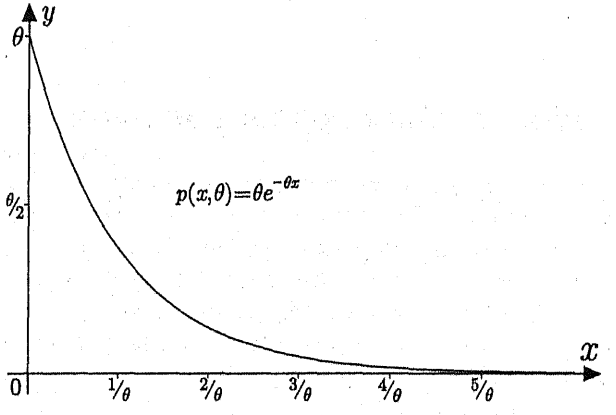


Рис. 2.3. Плотность показательного распределения с параметром θ

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

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

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)

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

для любых s, $t \geqslant 0$. 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 [87].

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:*

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

Смысл параметров нормального распределения наглядно показан на рис. 2.4.

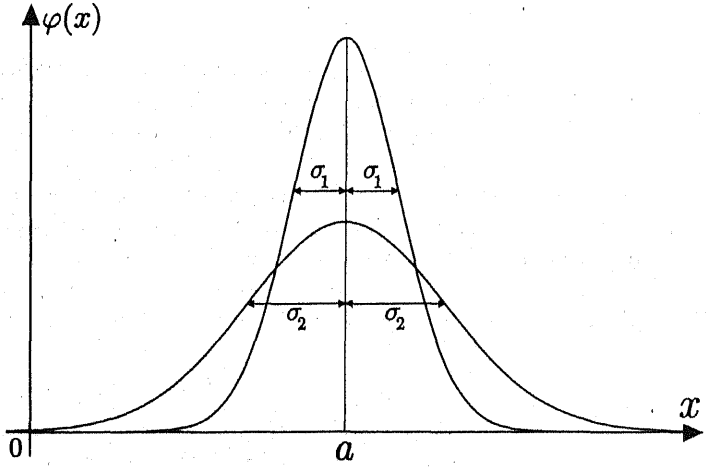


Рис. 2.4. Плотность нормального распределения со средним a и различными значениями дисперсии σ^2

Отметим, что $\varphi(x)$ стремится к нулю при $x \to -\infty$ и $x \to +\infty$. 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 $\sigma (\sigma > 0)$ 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

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

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 $x \geqslant 0$. 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 0.9933 — 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 0.9933 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

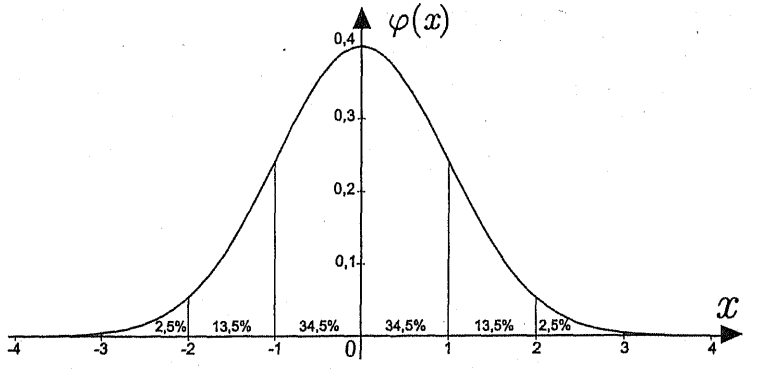


Рис. 2.5. Примерное распределение площадей под кривой функции плотности стандартного нормального распределения

таблицы разной степени подробности. Так, в [16] 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 $P(\xi \geqslant x) = 1 - \Phi(x)$. Например, в [91] дана таблица $P(\xi \geqslant x)$ 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 главе 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:

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

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:

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

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

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

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 $\xi_1, \xi_2, ..., \xi_n$ — независимы, и каждая из них имеет стандартное нормальное распределение $N(0,1)$. Говорят, что случайная величина χ^2_n, определенная как:

$$\chi^2_n = \xi_1^2 + \cdots + \xi_n^2,$$

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

Ясно, что χ^2_n (для любого $n \geqslant 1$) с вероятностью 1 принимает положительные значения. Функция плотности χ^2_n 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^2_n равны:

$$M\chi^2_n = n, \qquad D\chi^2_n = 2n.$$

Таблицы. Для случайной величины χ^2_n составлены разнообразные таблицы (см. [16], [50], [60]). Чаще всего они содержат значения p-квантилей случайных величин χ^2_n, $n = 1, 2, ..., m$ (если вероятность

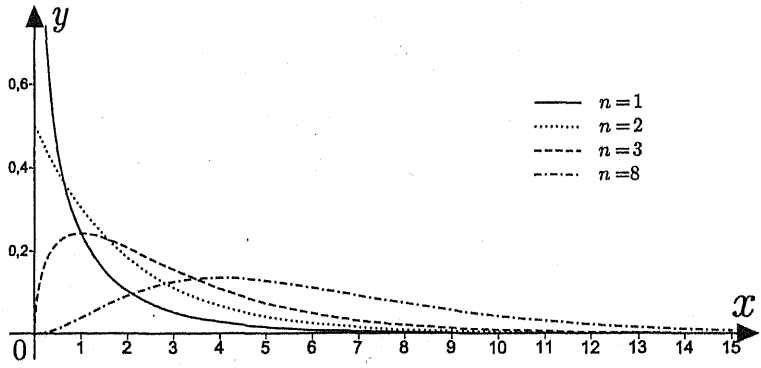


Рис. 2.6. Функции плотности распределения хи-квадрат с различным числом степеней свободы n

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 $\xi_0, \xi_1, \ldots, \xi_n$ — 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, $n = 1, 2, \ldots$ 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:

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

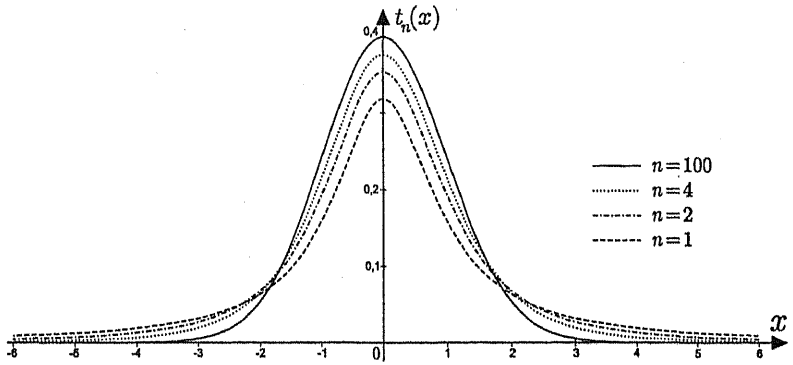


Рис. 2.7. Функции плотности распределения Стьюдента с различным числом степеней свободы n

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

2.6.3. F-распределение

Определение. Пусть $\eta_1, \ldots, \eta_m$; $\xi_1, \ldots, \xi_n$ (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

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

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:

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

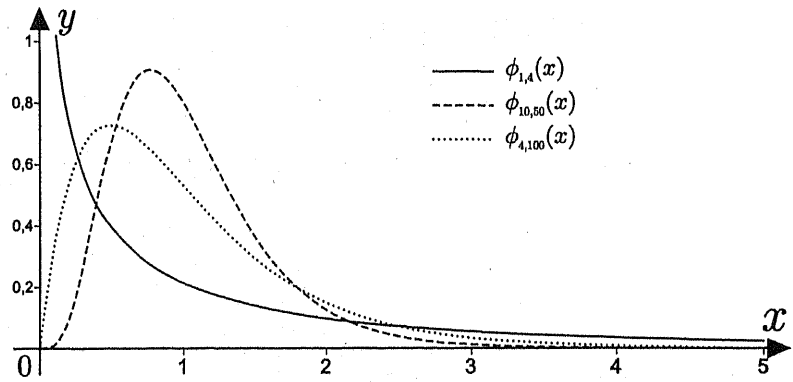


Рис. 2.8. Функции плотности F -распределения с различным числом степеней свободы

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 пакетах STADIA и STATGRAPHICS

Статистические пакеты могут предоставлять обширную справочную информацию по различным семействам вероятностных распределений, наглядно иллюстрируя их свойства и заменяя статистические таблицы.

2.7.1. Пакет STADIA

Пакет предоставляет возможность работать с пятью дискретными и восемью непрерывными распределениями вероятностей, приведенными ниже на рис. 2.9. Доступ к ним осуществляется из раздела **Распределения и частоты** меню блока статистических методов (см. рис. 1.17). Разберем несколько примеров.

Пример 2.1к. Построим графики плотности распределения вероятностей нормального распределения с параметрами $a = 0$, $\sigma^2 = 1$; $a = 0$, $\sigma^2 = 4$; $a = 2$, $\sigma^2 = 1$.

Выбор процедуры. В меню блока **Статистические методы** (рис. 1.17) щелкнем мышью кнопку **T = Вычисление вероятностей** или нажмем клавишу [T]. На экране появится меню выбора **Функция вероятности распределения** (рис. 2.9), в котором нужно выбрать пункт **6=нормальное**.

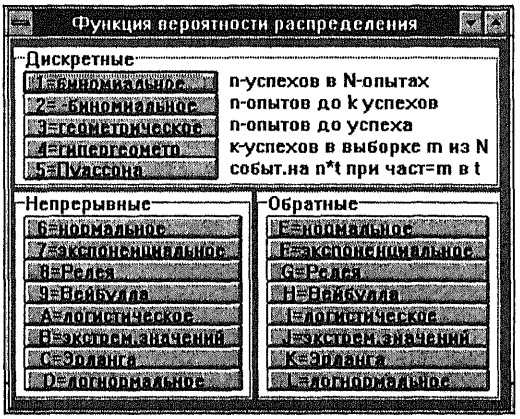


Рис. 2.9. Меню выбора функции распределения

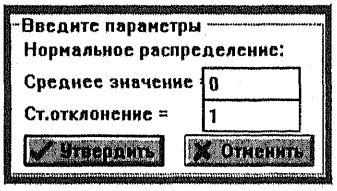


Рис. 2.10. Ввод параметров нормального распределения

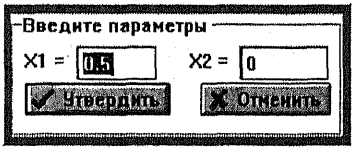


Рис. 2.11. Параметры расчета значения функции нормального распределения

Заполнение полей ввода данных. На экране появится окно ввода параметров нормального распределения (рис. 2.10). В поля Среднее значение и Ст. отклонение надо ввести параметры нужного распределения, например 0, 1. Система предложит вычислить значения функций нормального распределения $F(x1)$ и $F(x2)$, а также их разности $F(x1) - F(x2)$, в выбранных точках $x1$ и $x2$. Эти точки следует указать в окне ввода параметров (рис. 2.11). При вводе только одного значения x система выведет указанные величины для пары $-\infty$, x.

Результаты. При нажатии кнопки запроса Утвердить в окне ввода параметров на экран в графическое окно будет выдан график плотности и функции распределения для нормального распределения с заданными параметрами (рис. 2.12). Полученный график может быть сохранен в отдельном графическом окне и вызван затем на экран в течение всего сеанса работы с пакетом.

Комментарии. В пакете нет специальной опции для одновременного вывода графиков нескольких функций распределения. Для решения примера 2.1к следует последовательно применить разобранную процедуру для каждой группы параметров распределения. Сохраненные графические окна с этими

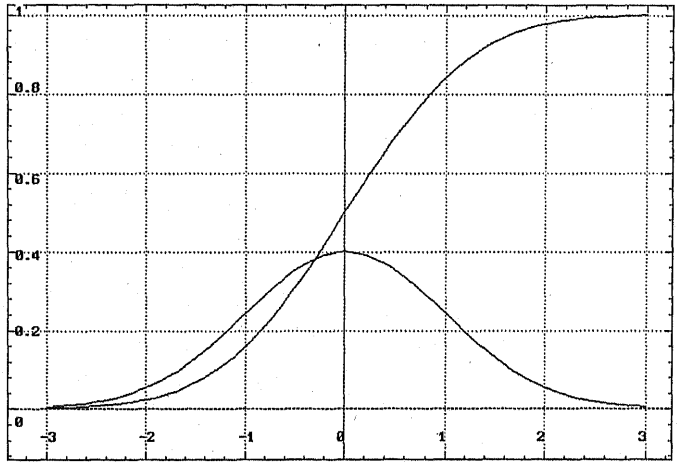


Рис. 2.12. График плотности и функции распределения нормального распределения

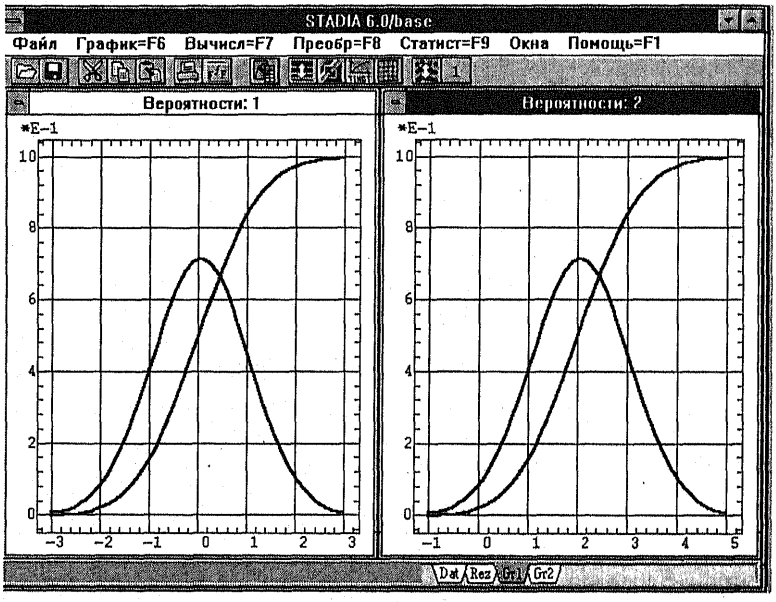


Рис. 2.13. Графические окна с графиками плотности и функции нормального распределения с параметрами (0, 1) и (2, 1)

графиками могут быть выведены затем одновременно, как это показано на рис. 2.13.

Пример 2.2к. Найдем p-квантили экспоненциального распределения со средним значением 4 для $p = 0.95; 0.975; 0.99$

Выбор процедуры. В меню блока Статистические методы (рис. 1.17) выберем пункт T = Вычисление вероятностей или нажмем его клавишу [T]. На экране появится меню выбора Функции вероятности распределения (рис. 2.9). Для получения квантилей в указанном меню следует выбрать требуемое распределение в группе обратных (правый столбец).

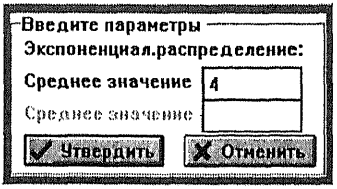


Рис. 2.13. Ввод параметров экспоненциального распределения

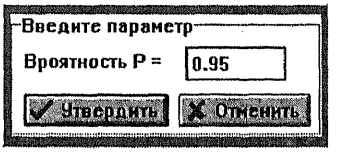


Рис. 2.14. Ввод параметров для расчета квантилей

Заполнение полей ввода данных. В окне ввода параметров экспоненциального распределения (рис. 2.13) укажите среднее значение 4 и нажмите кнопку запроса [Утвердить]. Система предложит ввести Вероятность P в окне ввода параметров (рис. 2.14).

Результаты. В запросе рис. 2.14 введите значение вероятности 0.95 и нажмите кнопку запроса [Утвердить]. В окне результатов (рис. 2.15). появится строка, указывающая соответствующее значение квантили X. Повторив эти действия для других значений вероятности, мы получим итоговый результат (рис. 2.15).

```
ВЫЧИСЛЕНИЕ ВЕРОЯТНОСТЕЙ.  Файл:
Распределение экспоненциальное: 4

Среднее=4, Дисперсия=16, Ст.отклонение=4
P=      0.95,    X=     11.983
P=     0.975,    X=     14.756
P=      0.99,    X=     18.421
```

Рис. 2.15. Результат вычисления квантилей экспоненциального распределения

Пример 2.3к. Создадим выборку размером 10 из равномерного распределения на отрезке [0, 5].

Выбор процедуры. Выберите в меню пакета пункт Преобр. или нажмите клавишу [F8]. В открывшемся меню преобразований выберите пункт 3=генератор чисел.

Заполнение полей ввода данных. На экране появится запрос режимов генератора (рис. 2.16). В поле Всего чисел укажем количество генерируемых чисел 10. В поля a= и b= в нижней части окна введем границы равномерного распределения 0 и 5. В меню типов генераторов выберем 3=равномерное.

Результаты. Сгенерированная выборка будет помещена в блок редактора данных системы (рис. 2.17).

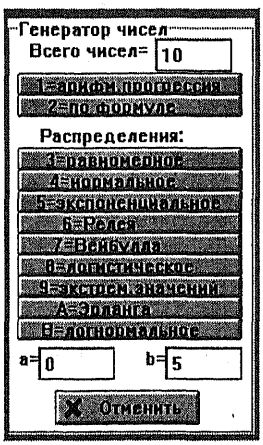


Рис. 2.16. Меню генератора чисел

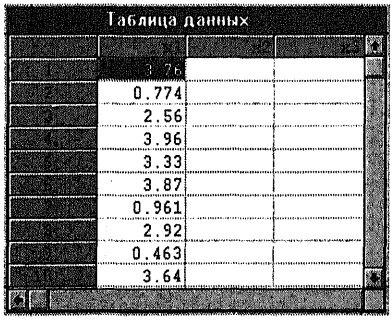


Рис. 2.17. Электронная таблица со сгенерированной выборкой

2.7.2. Пакет STATGRAPHICS

В пакете представлены 18 наиболее распространенных семейств распределения вероятностей (см. рис. 2.19). Среди них 6 дискретных: (Бернулли, биномиальное, дискретное равномерное, геометрическое, отрицательное биномиальное и Пуассоновское), и 12 непрерывных: бета, хи-квадрат, Эрланга, экспоненциальное, F-распределение, гамма, логнормальное, нормальное, t-распределение Стьюдента, треугольное, равномерное и Вейбулла).

Для доступа к процедурам, работающими с распределениями, в головном меню пакета необходимо выбрать пункт H. Distribution function (функции распределения). Меню этого пункта приведено на рис. 2.18.

```
DISTRIBUTION FUNCTIONS

1. Distribution Fitting
2. Distribution Plotting
3. Tail Area Probabilities
4. Critical Values
5. Random Number Generation
```

Рис. 2.18. Меню процедур, работающих с распределениями

Назначение процедур этого меню следующее:

1. Distribution Fitting (подбор распределения) — позволяет проверить с помощью двух критериев гипотезу о том, что введенная выборка принадлежит указанному Вами распределению. Теоретическая постановка этой проблемы и ее реализация на компьютере подробно разбираются в главе 10.

2. Distribution Plotting (графики распределений) — позволяет построить для перечисленных выше семейств плотности и функции распределения, а так же ряд других характеристик (см. ниже пример 2.1к).

3. Tail Area Probabilities (вероятности хвостов) — позволяет вычислить значение выбранной функции распределения в любой точке. Работа этой процедуры разобрана в примере 3.1к главы 3.

4. Critical Values (критические значения) — вычисляет по заданному значению функции распределения (вероятности) p квантили x_p этого распределения. Эта операция является обратной по отношению к процедуре Вероятности хвостов. Работа процедуры разобрана в примере 2.2к.

5. Random Number Generation (генератор случайных чисел) — порождает (имитирует) последовательность независимых одинаково распределенных случайных величин, подчиняющихся выбранному Вами закону распределения — одному из упомянутых восемнадцати. Этой процедуре посвящен пример 2.3к.

Рассмотрим, как приведенные выше примеры решаются с помощью пакета STATGRAPHICS.

Пример 2.1к. Построим графики нормальных плотностей с параметрами $(a = 0, \sigma^2 = 1)$; $(a = 0, \sigma^2 = 4)$; $(a = 2, \sigma^2 = 1)$.

Выбор процедуры. Из пункта H. Distribution function головного меню пакета выберем процедуру 2. Distribution Plotting. Экран ввода данных этой процедуры (с полями для ввода параметров нормального распределения) приведен на рис. 2.19.

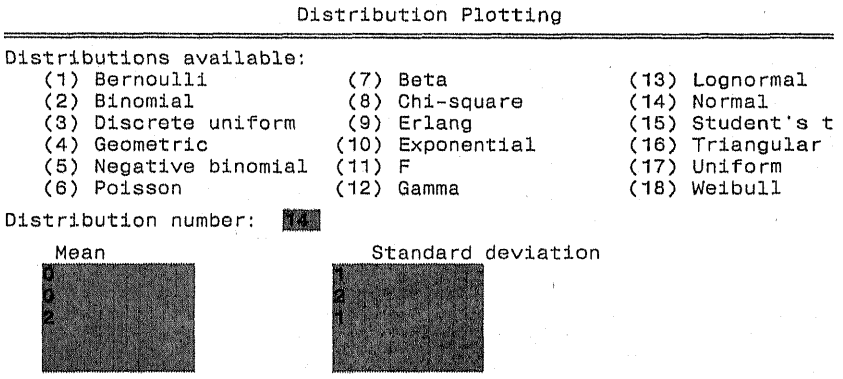


Рис. 2.19. Запрос параметров процедуры графика распределений

Заполнение полей ввода данных. В поле Distribution number (номер распределения) необходимо ввести номер требуемого распределения из приведенного на экране списка (нормальное распределение имеет номер 14) и затем нажать клавишу F6. На том же экране появятся поля ввода параметров выбранного распределения, в данном случае — Mean (среднее значение) и Standard deviation (стандартное отклонение, т.е. σ). В них надо указать данные примера 2.1к, как это показано на экране.

Закончив заполнение, надо нажать клавишу F6. На том же экране появится всплывающее окно выбора формы представления распределения (рис. 2.20, справа нами указан русский перевод).

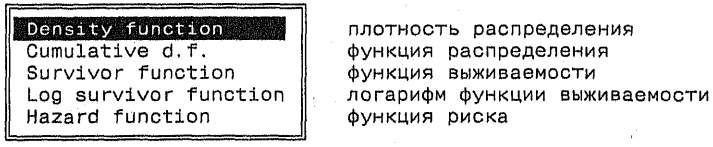


Рис. 2.20. Меню формы представления распределения

Первые две формы представления распределений из этого меню являются наиболее употребительными. Последние три — более специфичны и еще не рассматривались нами. Ограничимся формальным определением назначения последних процедур, не указывая области их применения. *Функция выживаемости* по определению равна единице минус функция распределения. Под *логарифмом функции выживаемости* понимается натуральный логарифм этой функции. *Функцией риска* называется частное от деления плотности распределения на функцию выживаемости.

Выберем в меню пункт 1. Density function и нажмем F6. На экране появится меню параметров графиков (рис. 2.21).

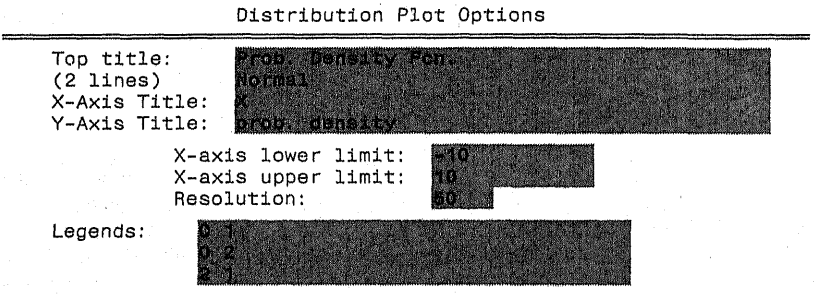


Рис. 2.21. Запрос параметров графиков распределений

Комментарии. Приведенное на рис. 2.21 заполнение полей ввода экрана было осуществлено самим пакетом, исходя из той информации, которая была введена на предыдущих шагах. Сделаем необходимые комментарии. Верхний блок ввода (поля Top title (заголовок графика), X-Axis Title (название оси X), X-Axis Title (название оси Y)) относится к оформлению заголовка графика и наименованию осей координат. Заполнение этих полей не обязательно.

В полях X-axis lower limit (нижняя граница по оси X) и X-axis upper limit (верхняя граница по оси X) необходимо указать, в каких пределах будут выведены на экран графики. Пакет вычисляет оптимальные пределы вывода. Их можно скорректировать, например, для просмотра части графика.

В поле Resolution (разрешение) указывается число, точек по которым будет происходить интерполяция графика. Чем больше это число, тем более точный график будет выводится на экран. Однако увеличение числа точек замедляет скорость вывода графика на экран.

Если на график выводится несколько кривых, то они обозначаются различными типами линий — непрерывной, пунктирной, точечной и др. Поля Legends (легенды) предназначены для указания связи между типами линий и параметрами кривых выводимых на график. Эта информация выводится в правом верхнем углу графика. В данном случае пакет поставил в эти поля параметры закона распределения вероятностей — среднее и стандартное отклонение. В эти поля могут вводиться и символьные записи.

Результаты. Заполнив поля запроса, нажмем клавишу F6. На экране появятся требуемые графики (рис. 2.22).

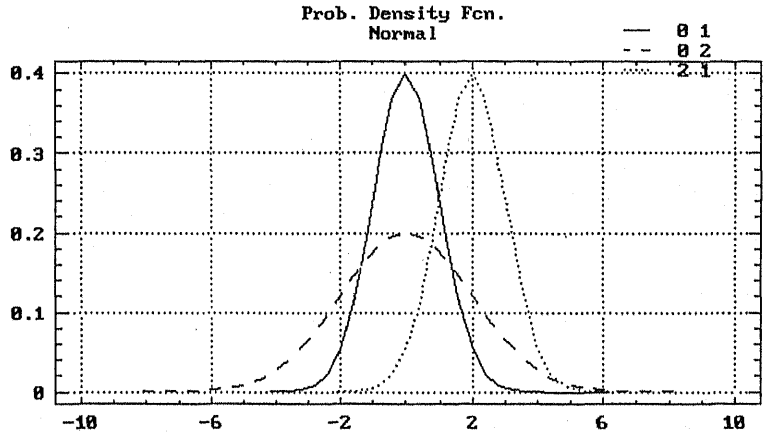


Рис. 2.22. Графики плотности нормального распределения с параметрами $a = 0$, $\sigma^2 = 1$; $a = 0$, $\sigma^2 = 4$; $a = 2$, $\sigma^2 = 1$

Комментарии. 1. В зависимости от выбранного распределения в процедуре будут изменяться поля ввода. Это относится прежде всего к полям ввода параметров распределения. Так, для распределений Стьюдента и хи-квадрат параметром является число степеней свободы (Degrees of freedom), для равномерного распределения — верхняя и нижняя граница интервала распределения (Lower limit) и (Upper limit), и т.д.

2. Содержание окна выбора формы представления распределения несколько меняется в зависимости от выбранного распределения. В нем указываются только те возможности, которые доступны для данного распределения.

3. Указанная процедура позволят выводить на экран до пяти однотипных кривых. Сложнее, но возможно, получить на одном графике функции распределения из разных семейств. Для этого следует воспользоваться процедурой наложения одного графика на другой (см. процедуру Splitscreen/Overlay Plotting (разделение листа и наложение графиков) пункта Report Writer and Graphics Replay (текстовой редактор и воспроизведение графиков) головного меню пакета.

Пример 2.2к. Найдем p -квантили экспоненциального распределения со средним значением 4 для $p = 0.95; 0.975; 0.99$

Выбор процедуры. В пункте H. Distribution function (рис. 2.18) головного меню пакета надо выбрать процедуру Critical Values (критические зна-

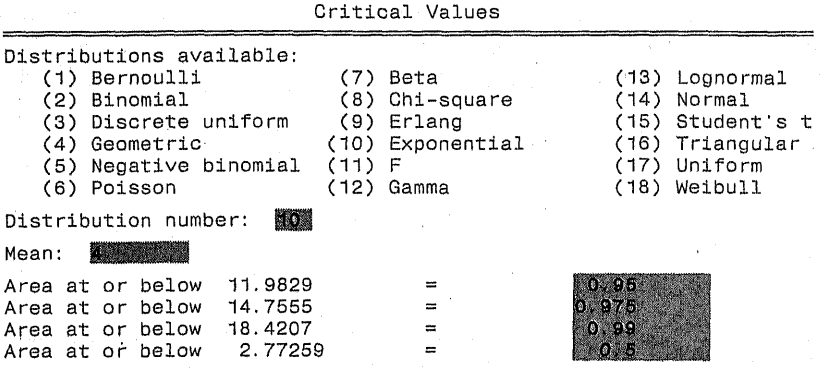


Рис. 2.23. Запрос параметров процедуры критических значений

чения). Приведем вид экрана этой процедуры с заполненными полями ввода данных и результатами проделанных расчетов (рис. 2.23).

Заполнение полей ввода данных. В поле Distribution number (номер распределения) надо ввести номер 10 экспоненциального распределения из приведенного на экране списка и нажать F6. На экране появятся поля ввода данных параметров распределения, в данном случае — поле Mean (среднее значение). Введя в него число 4, надо нажать F6. На том же экране появятся четыре строки ввода Area at or below = 0.5. Значение 0.5, поставленное пакетом по умолчанию, находится в активном поле ввода. Вместо него следует ввести необходимые значения p, как это показано на рис. 2.23.

Результаты. После нажатия F6 слева от каждого из введенных значений p появится значение квантили x_p.

Комментарии. 1. Эта процедура, наряду с процедурой 3. Tail Area Probabilities, заменяет традиционные таблицы математической статистики.

2. Порядок ввода данных разобранных выше процедур один и тот же на этапе выбора одного или нескольких конкретных распределений. Это относится ко всем процедурам пункта H. Distribution function.

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

Пример 2.3к. Создадим выборку размера 10 из равномерного распределения на отрезке [0, 5].

Выбор процедуры. В головном меню пакета выберем пункт H. Distribution function (рис. 2.18), а в нем — процедуру 5. Random Number Generation.

Заполнение полей ввода данных. Экран ввода данных этой процедуры на первом этапе выбора конкретного распределения совпадает с приведенными выше экранами для процедур 2. Distribution Plotting и 4. Critical Values. В поле Distribution number введем номер равномерного распределения — 17. В качестве параметров этого распределения выступают значения концов отрезка, на котором рассматривается распределение. Эти значения необходимо ввести в поля Lower limit (нижний предел) и Upper limit (верхний предел). Далее в поле Number of samples (объем выборки) введем значение 10. Поле Seed содержит некоторое число, используемое в качестве стартового значения для алгоритма моделирования. Это число должно быть целым, положительным и не превышающим по величине 2147483646. Начальное значение по умолчанию устанавливается самой системой во время запуска. Оно каждый раз новое, поскольку вычисляется исходя из показаний системного таймера. Это значение обновляется после каждого обращения к процедуре генерации случайных чисел.

Результаты. После заполнения всех полей ввода и нажатия клавиши F6 будут вычислены требуемые числа и сделан следующий запрос об имени файла и имени переменной в базе данных пакета, куда будет записаны полученные числа (рис. 2.24):

```
Enter variable in which to save the samples:
        File: UNIFORM      Variable: SAMPLES
```

Рис. 2.24. Запрос параметров сохранения результатов генерации выборки

В поля File и Variable введите необходимое Вам имя файла и имя переменной. После нажатия F6 сгенерированная выборка будет помещена в базу данных пакета.

Комментарии. Процедура также позволяет генерировать произвольные нормально распределенные векторы. Для этого надо заполнить поля vector of means (вектор средних) и variance-covariance matrix (матрица ковариаций). В последнее поле необходимо ввести матрицу ковариаций размером $k \times k$, где k — число генерируемых нормально распределенных векторов.

Глава 3

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 главе 1.

Если опустить детали и некоторые (хотя и важные) исключения, можно сказать, что весь статистический анализ основан на *идее 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 выборок.

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 главе 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 не разъясненным, 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, используясь 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 косвен-

ным. 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 века[1].

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 естественнонаучных 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 занимают

[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.

события практически невозможные. Причина этого проста: пригодных 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 $P(\cdot \mid H)$. Если H — простая гипотеза, то для всякого события A (A — множество в выборочном пространстве) его вероятность $P(A \mid H)$ определена однозначно. Если гипотеза H сложная (состоит из многих простых), то $P(A \mid H)$ 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. $P(A \mid H) < \varepsilon$. (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 $P(A \mid H) \leqslant \varepsilon$, 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 несоблюде-

нии 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 удается 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 предлагались 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 $\{y \mid y \geqslant a\}$, где a — некоторое число. При другом поведении статистики T множество A' может быть устроено по-другому, например $\{y \mid y \leqslant a\}$ или $\{y \mid y \leqslant a \text{ или } y \geqslant b\}$. Разумеется, следует выбирать множество A' так, чтобы $\max_{P \in H} P(A) \leqslant \varepsilon$, 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 ошибками первого и второго рода также неодинаково. При

построении статистических критериев мы фиксируем максимальную допустимую вероятность ошибки первого рода (то есть 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. [83]). 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 эксперимента 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 выдвижению 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 ощутить не состоянии. 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 естественнонаучной 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. [26]). 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, $i = 1, \ldots, 17$; 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 одинаковы? Логически эти вопросы тесно связаны: если мы отвечаем отрицательно на второй из них, мы тем самым признаем, что различия есть. После этого уже не-

трудно 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 $X_1, Y_1, \ldots, X_{17}, Y_{17}$ 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 $i = 1, \ldots, 17$, 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 $a_1, b_1, \ldots, a_{17}, b_{17}, \sigma^2$ нам неизвестны. При этих обозначениях выдвинутый вопрос о равном времени реакции на свет и на звук может быть сформулирован как статистическая гипотеза:

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

Если экспериментатор уверен, что группа испытуемых достаточно однородна, он может дополнительно предположить, что $a_1 = \ldots = a_{17}$ и $b_1 = \ldots = b_{17}$. Если обозначить общие значения параметров через a и b соответственно, то статистическую модель в этом случае можно сформулировать так: случайные величины $X_1, \ldots, X_{17}$ независимы и распределены по закону $N(a, \sigma^2)$; случайные величины $Y_1, \ldots, Y_{17}$ тоже независимы, не зависят от $X_1, \ldots, X_{17}$ 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 $MX_1 = \ldots = MX_{17}$ и $MY_1 = \ldots = MY_{17}$, 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 пункте 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 разобранное в главе 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 $P(A \mid H)$, отмечая, что эта вероятность вычислена при гипотезе H. Рассмотрим некоторые примеры событий и вычислим их вероятности. В табл. 3.2 приведены вероятности событий вида $\{S = k\}$ при $p = 1/3$.

Таблица 3.2

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

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 $\{S \geqslant k\}, k = 0, 1, 2, \ldots, 10$.

Таблица 3.3

k	0	1	2	3	4	5
$P(S \geqslant k \mid H)$	1.000	0.9827	0.8959	0.7009	0.4407	0.2131
k	6	7	8	9	10	
$P(S \geqslant k \mid H)$	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 $P(A \mid H)$ 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 таблице 3.3, когда они

Таблица 3.4

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

вычисляются при альтернативе $p = 0.9$. Соответствующие значения даны в таблице 3.4.

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 $S \geqslant 7$, имеет вероятность, равную 0.0197 (см. табл. 3.3). Из табл. 3.4 видно, что вероятность события $S \geqslant 7$, вычисленного при альтернативе, равна 0.9872, т.е. событие $S \geqslant 7$ 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 $\{S \geqslant 7\}$.

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 $A = \{S = 0\} \cup \{S \geqslant 9\}$ при альтернативе. Действительно, $P(A \mid H_3) = 0.7361$, т.е. это событие менее вероятно при альтернативе, чем выбранное выше.

Разобранный пример характеризует в некотором смысле идеальную ситуацию, когда удается 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 $p, 1/3 < p \leqslant 1$. 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 таблиц 3.2 и 3.4. Выше было установлено, что событие $A = \{S \geqslant 7\}$ при справедливости первой гипотезы маловероятно. В то же время, чем больше значение p, тем больше

вероятность этого события. Так, при $p = 0.9$ $P(S \geqslant 7 \mid H_3) = 0.9872$. 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 наблюденное 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 $\{S \geqslant 7\}$. 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 $\{S \geqslant 8\}$, 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 $\{S \geqslant C\}$, где C — некоторое критическое значение. Чем больше число C, тем менее вероятно при гипотезе H событие $\{S \geqslant C\}$. Тем больше поэтому уверенность, что H надо отвергнуть, если $S_{\text{набл.}} \geqslant C$. Наибольшей достижимой уверенности соответствует наименьший возможный уровень значимости, который в нашей задаче есть $P(S \geqslant S_{\text{набл.}} \mid H)$.

Наименьший уровень значимости полезно вычислять во всех случаях, так как он характеризует, насколько сильно наблюденное 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 $A = \{S \leqslant C\}$, а число C выбирается так, чтобы $P(S \leqslant C \mid H)$ 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 невоз-

можно 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 $P(S \leqslant y \mid H_0) + P(S \geqslant x \mid H_0)$ была малой. Гипотеза H_0 отвергается, если $S_{\text{набл.}} \leqslant y$, либо $S_{\text{набл.}} \geqslant x$. Уровень значимости этого критерия есть $P(S \leqslant y \mid H_0) + P(S \geqslant x \mid H_0)$.

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 $P(|Snp_0| \geqslant z \mid H_0)$ оказывается малой. Гипотезу $H_0 : p = p_0$ отвергают, если $|S_{\text{набл.}} - np_0| \geqslant z$. 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 $P(|S - np_0| \geqslant |S_{\text{набл.}} - np_0| \mid H)$.

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 $x_1, \ldots, x_N$. 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, $\operatorname{med} X = \theta$, то $P\{(X - \theta) > 0\} = P\{(X - \theta) < 0\} = 1/2$. В выборке $x_i - \theta$, $i = 1, \ldots, N$, подсчитаем число положительных разностей и обозначим его через S. Для формализации этого алгоритма удобно ввести функцию

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

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, то есть 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 бóльших 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 $F \leqslant G$ и $F \geqslant G$. Дадим определения этих понятий.

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

Смысл этого определения состоит в том, что при $F \geqslant G$ 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 $P(X < x) \geqslant P(Y < x)$.

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 главе 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 $x_1, \ldots, x_m$ (выборка x) и $y_1, \ldots, y_n$ (выборка y) объемов m и n. Обозначим закон распределения первой выборки через F, а второй — через G.

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

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

Гипотеза. Утверждение об однородности выборок $x_1, \ldots, x_m$ и $y_1, \ldots, y_n$, 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 $F \geqslant G$ (правосторонняя альтернатива, "перетекание" вероят-

ностей вправо) или альтернативы $F \leqslant G$ (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

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

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

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

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

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

Здесь критическая область примет вид

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

В таблицах (см. [60], [64], [91]) 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 $U_{\text{л.}}(\alpha, m, n)$ можно воспользоваться соотношением

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

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

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

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

$$\{U \leqslant U_{\text{л.}}(\alpha, m, n)\} \cup \{U \geqslant U_{\text{п.}}(\alpha, m, n)\}\,,$$

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

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

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 $U_{\text{л.}}(\frac{\alpha}{2}, m, n)$ и $U_{\text{п.}}(\frac{\alpha}{2}, m, n)$

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 [60], [64], [91] приводятся таблицы, по которым можно найти вероятность $P(U \geqslant k)$ 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 $F \geqslant G$), или левосторонних (если $F \leqslant G$). 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 ($F \leqslant G$) 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 наблюденное 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 $mn/2$ наблюденное значение U, т.е. $U_{\text{набл.}}$, тем сильнее мы сомневаемся в том, что H верна. Разумеется, U может значительно отклоняться от $M(U \mid H)$ 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 $F \geqslant G$ можно выразить количественно, вычислив $P(U \geqslant U_{\text{набл.}} \mid H)$. 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

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

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 пункте 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 имеющимся 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 повторений в выборках. На практике же такие повторения 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 $i, j (i = 1, \ldots, m;\ j = 1, \ldots, n)$. 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 $x_1, \ldots, x_m$ и $y_1, \ldots, y_n$. 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 $1, \ldots, m+n$. Верно и обратное: любая перестановка чисел $1, \ldots, m+n$ может оказаться ранговой последовательностью. Так что множество возможных ранговых последовательностей — это совокупность перестановок чисел $1, 2, \ldots, m+n$. Их общее число равно $(m+n)!$.

Зная распределения случайных величин $X_1, \ldots, X_m$ и $Y_1, \ldots, Y_n$, мы можем (по крайней мере, теоретически) вычислить вероятность того, что результат их ранжирования будет заданной перестановкой. Поэтому каждое распределение случайных величин $X_1, \ldots, X_m$ и $Y_1, \ldots, Y_n$ порождает некоторое распределение вероятностей на указанном множестве перестановок. Ясно, что если исходные данные однородны ($X_1, \ldots, X_m$ и $Y_1, \ldots, Y_n$ 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 $F \leqslant G$ и правосторонние альтернативы $F \geqslant G$. Для правосторонних альтернатив $P(x_i < y_j) > 0.5$, то есть 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 $1, 2, \ldots, m+n$. Если же отступление таково, что $P(x_i < y_j) < 0.5$, то ранги игреков чаще будут принимать значения из левой части ряда чисел $1, 2, \ldots, m+n$. Переход от рангов игреков 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 $x_1, \ldots, x_m$ и $y_1, \ldots, y_n$, объемов m и n. Обозначим закон распределения первой выборки через F, а второй — через G.

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

2. Законы распределения выборок F и G непрерывны.

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

Метод. 1. Рассмотрим ранги игреков в общей совокупности выборок x и y. Обозначим их через $S_1, \ldots, S_n$.

2. Вычислим величину

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

называемую **статистикой Уилкоксона**. Таблицы распределения статистики W (при гипотезе однородности) можно найти в [16], [60], [91] 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

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

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

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

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

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

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:

$$P(W \geqslant W_{\text{набл.}}),$$
$$P(W \leqslant W_{\text{набл.}}),$$
$$P\left(|W - n(m+n+1)/2| \geqslant |W_{\text{набл.}} - n(m+n+1)/2|\right).$$

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 $m \to \infty$ и $n \to \infty$. Перейдем от величины W к $W^* = (W - MW)/\sqrt{DW}$. Ниже будет показано, что $MW = n(m+n+1)/2$. Так же 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 $W^*_{\text{набл.}} \geqslant z_\alpha$;
- отвергнуть H на уровне α против альтернативы $P(x_i < y_j) < 0.5$, если $W^*_{\text{набл.}} \leqslant -z_\alpha$;
- отвергнуть H на уровне 2α против альтернативы $P(x_i < y_j) \neq 0.5$, если $|W^*_{\text{набл.}}| \geqslant z_\alpha$.

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 ей 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 игреков служат (в той или иной последовательности) числа $1, 2, \ldots, n$. Максимальное значение W возникает, когда этими рангами служат $m+1, m+2, \ldots, m+n$.

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

Распределение вероятностей. Статистика Уилкоксона была определена нами как сумма (упорядоченного) набора рангов игреков $S_1, \ldots, S_n$. Вероятность каждого такого упорядоченного набора при выдвинутой гипотезе H — одна и та же и равна $(C^m_{m+n})^{-1} = \frac{m!\,n!}{(m+n)!}$. 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 игреков. Оно равно $C^2_{3+2} = 10$. 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 $M(W \mid H) = n(m+n+1)/2$.

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, $M\,(W - n(m+n+1)/2\,|\,H) = \sum_{k=-nm/2}^{nm/2} kp_k = 0$. 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.)

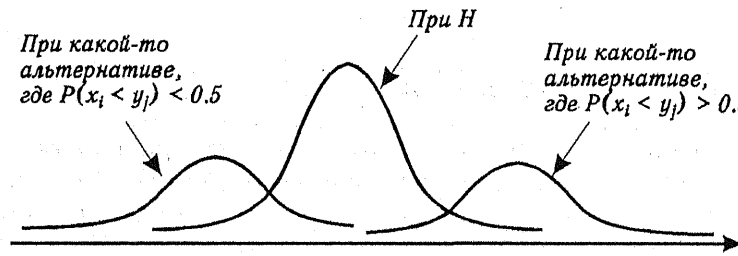


Рис. 3.1. Схематическое изображение распределений W

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 $z_1, z_2, \ldots, z_n$ 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, $t_2, \ldots, t_g$ — размеры наблюденных связок среди игреков, g — общее число связок среди игреков. 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 (попрежнему 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 $(x_1, y_1), \ldots, (x_n, y_n)$ объема n. Введем величины $z_i = y_i - x_i$, $i = 1, \ldots, n$.

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 $(x_1, y_1), \ldots, (x_n, y_n)$ можно записать в виде

$$H : P(x_i < y_i) = P(x_i > y_i) = 0.5 \qquad \text{для всех} \quad i = 1, \ldots, n.$$

Метод. 1. Перейдем от повторных парных наблюдений $(x_1, y_1), \ldots, (x_n, y_n)$ к величинам z_i, $i = 1, \ldots, n$, введенным выше.

2. К совокупности z_i, $i = 1, \ldots, n$ применим критерий знаков для проверки гипотезы о равенстве нулю медиан распределений величин z_i, $i = 1, \ldots, n$ (см. п. 3.4.2).

Приближение для больших совокупностей. Следует воспользоваться нормальной аппроксимацией биномиального распределения. Смотри пункт 2 раздела «Связь с другими распределениями» параграфа 2.1 главы 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 $z_i = \theta + e_i$, $i = 1, \ldots, n$, 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 $e_1, \ldots, e_n$. 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.

Осуществим переход от пар $(x_1, y_1), \ldots, (x_n, y_n)$ к величинам z_i, $i = 1, \ldots, n$ и запишем последние в виде: $z_i = \theta + e_i$, $i = 1, \ldots, n$.

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, $P(z_i > 0) = P(z_i < 0) = 1/2$, то есть 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 таблицы 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 $P(S \leqslant S_{\text{набл.}} \mid H)$. Для этого воспользуемся таблицами биномиального распределения при $p = 1/2$, $n = 16$ (см. [16], [60]). Учитывая, что в силу симметрии при $p = 1/2$ $P(S \leqslant S_{\text{набл.}} \mid H) = P(S \geqslant nS_{\text{набл.}} \mid H)$, получаем:

$$P(S \leqslant S_{\text{набл.}} \mid H) = P(S \geqslant 16 - 3 \mid H) = P(S \geqslant 13 \mid H) = 0.0106.$$

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 $\theta < 0$.

Таблица 3.5

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 $z_1, \ldots, z_n$ 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 $|z_1|, \ldots, |z_n|$. Пусть R_i обозначает ранг z_i в совместном упорядочении $|z_1|, \ldots, |z_n|$ от меньшего к большему.

2. Определим переменные ψ_i, $i = 1, \ldots, n$, где

$$\psi_i = \begin{cases} 1, & \text{если } z_i > 0; \\ 0, & \text{если } z_i < 0. \end{cases}$$

3. Вычислим наблюденное 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 $T_{\text{набл.}} \geqslant t(\alpha, n)$;
- принять H, если $T_{\text{набл.}} < t(\alpha, n)$,

где критическое значение $t(\alpha, n)$ удовлетворяет уравнению $P(T \geqslant t(\alpha, n) \mid H) = \alpha$. Таблицу критических значений можно найти в [91].

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 $T_{\text{набл.}} \leqslant \frac{n(n+1)}{2} - t(\alpha, n)$;
- 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 $T_{\text{набл.}} \geqslant t(\alpha, n)$ или $T_{\text{набл.}} \leqslant \frac{n(n+1)}{2} - t(\alpha, n)$;
- 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 $t(\alpha, n)$: $(P(T \geqslant t(\alpha, n)) = \alpha)$ 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 $n \to \infty$) распределение $N(0, 1)$. Приведем приближение нормальной теории для проверки H, для определенности, против правосторонней альтернативы: H отклоняется, если $T_{\text{набл.}} \geqslant z_\alpha$, 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 $|z_1|, \ldots, |z_n|$ 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 связок, $t_1,\ldots,t_g$ — их размеры. Определение связок смотри в разделе 3.5.2 при обсуждении статистики W^*.

3.7. Проверка статистических гипотез в пакетах STADIA и STATGRAPHICS

Ниже будет рассмотрено, как на компьютерах реализуются методы проверки гипотез в схеме испытаний Бернулли, как можно планировать подобного рода эксперименты, проверять гипотезы о равенстве медианы выборки заданному значению, а также применять критерии знаков и знаковых рангов Уилкоксона для парных сравнений.

3.7.1. Пакет STADIA

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.

Для решения этой задачи следует использовать процедуру вычисления значений функции заданного распределения вероятностей. Ее работа подробно рассмотрена в примере 2.1к. Здесь мы приведем диалог общения с ней для биномиального распределения.

Выбор процедуры. В меню блока Статистические методы (рис. 1.17) выберем пункт T = Вычисление вероятностей. В открывшемся при этом меню Функция вероятности распределения (рис. 2.9) выберем 1 = биномиальное или нажмем его клавишу [1].

Заполнение полей ввода данных. На экране появится окно ввода параметров распределения (рис. 3.2). Укажем в нем вероятность P успеха 0.333, число опытов N = 10, как показано на рис. 3.2, и нажмем кнопку [Утвердить]. Затем последует запрос о числе вычисляемых значений функции распределения (рис. 3.3). Здесь подразумевается задание всех или части возможных значений, которые принимает рассматриваемая случайная величина. По умолчанию система указывает в этом поле введенное число испытаний.

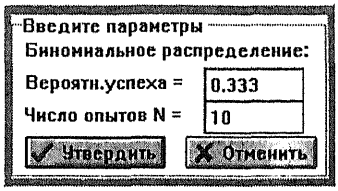


Рис. 3.2. Ввод параметров биномиального распределения

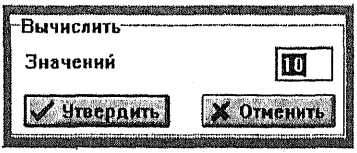


Рис. 3.3. Ввод числа вычисляемых значений распределения

Результаты. После нажатия кнопки Утвердить на экране в окне результатов появятся результаты работы процедуры (рис. 3.4). В графическое окно также будет выведен график вероятностей исходов и функции распределения.

```
ВЫЧИСЛЕНИЕ ВЕРОЯТНОСТЕЙ.  Файл: Распределение биномиальное: 0.333, 10
Среднее=3.33, Дисперсия=2.22, Ст.отклонение=1.49

 Функция распределения вероятностей
      r       P(=r)      P(X<=r)
      0      0.0174       0.0174
      1       0.087        0.104
      2       0.195          0.3
      3        0.26         0.56
      4       0.227        0.788
      5       0.136        0.924
      6      0.0567         0.98
      7      0.0162        0.997
      8     0.00303            1
      9    0.000336            1
```

Рис. 3.4. Результаты расчетов вероятностей дискретного распределения

Для получения минимального уровня значимости критерия для значения $S_{\text{набл.}} = 7$ против односторонних (правосторонних) альтернатив необходимо вычислить величину $1 - P(X \leqslant S_{\text{набл.}} - 1)$. В данном случае получаем $1 - 0.9805 = 0.0195$.

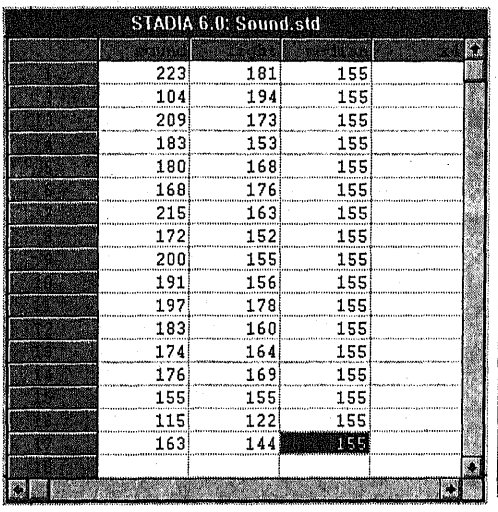
Пример 3.2к. В пакете STADIA отсутствует процедура, позволяющая планировать число испытаний Бернулли для проверки нулевой гипотезы при заданных ошибках первого и второго рода. Такая процедура есть в пакете STATGRAPHICS.

Пример 3.3к. В задаче о скорости реакции на звук и на свет проверим гипотезу о том, что медиана распределения скорости реакции на свет равна 155 мс.

Подготовка данных. В электронную таблицу пакета следует ввести данные таблицы 3.1 в соответствующие переменные, назовем их sound и light (см. рис. 3.5). Кроме того в переменную median надо ввести значения гипотетической медианы (это можно сделать с помощью блока преобразования данных пакета).

Замечание. Последнее требование связано с тем, что процедура критерия знаков реализована в пакете таким образом, что на ее вход всегда должны

подаваться две парные выборки. Поэтому для проверки гипотезы о равенстве медианы выборки заданной величине необходимо сформировать переменную такой же длины, что и у анализируемой выборки, все значения которой равны гипотетическому значению медианы.



223	181	155
104	194	155
209	173	155
183	153	155
180	168	155
168	176	155
215	163	155
172	152	155
200	155	155
191	156	155
197	178	155
183	160	155
174	164	155
176	169	155
155	155	155
115	122	155
163	144	155

Рис. 3.5. Данные для проверки гипотез о скорости реакций на свет и звук

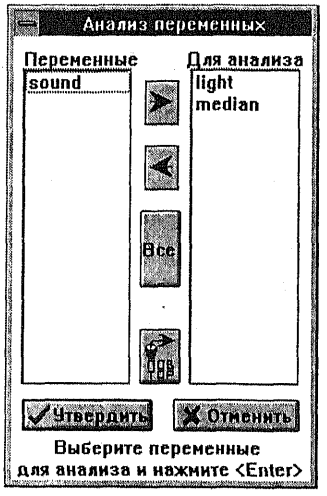


Рис. 3.6. Выбор переменных

Выбор процедуры. В меню блока Статистические методы в разделе Непараметрические тесты выберем пункт 6 = Сдвига/положения (для этого можно нажать клавишу 6).

Заполнение полей ввода данных. В появившемся окне Анализ переменных (рис. 3.6) выделим с помощью мыши переменные light и median. Нажав кнопку со стрелкой вправо, перенесем их из поля Переменные в поле Для анализа и затем нажмем кнопку Утвердить.

```
КРИТЕРИИ СДВИГА (ПОЛОЖЕНИЯ).  Файл: sound.std

                              Переменные: light, median
Вилкоксон=357, Z=-2.2552, Значимость=0.012, степ.своб = 17,17
   Гипотеза 1: <Есть различия между медианами выборок>
Ван дер Варден=4.797, Z=1.784, Значимость=0.0372, степ.своб = 17,17
   Гипотеза 1: <Есть различия между медианами выборок>

Для парных данных:
Вилкоксон=116, Z=3.1812, Значимость=0.0007, степ.своб = 2,15
   Гипотеза 1: <Есть различия между медианами выборок>
Знаков=11, Z=2.3240, Значимость=0.0101, степ.своб = 2,15
   Гипотеза 1: <Есть различия между медианами выборок>
```

Рис. 3.7. Результаты проверки гипотезы о положении выборки

Результаты. На экране появятся результаты расчетов для следующих непараметрических критериев положения: Уилкоксона, Ван дер Вардена (см. [21]) и знаков (рис. 3.7). Для критерия знаков выводится

значение статистики знаков S, нормальная аппроксимация этой статистики (Z-статистика) и минимальный уровень значимости нулевой гипотезы об отсутствии различий в сдвиге двух выборок или о равенстве медианы заданному значению против односторонних альтернатив. Указываемый минимальный уровень значимости всегда рассчитывается на основе нормальной аппроксимации. Следовательно, для выборок малого объема лучше воспользоваться соответствующими статистическими таблицами для величины S (см. [16], [91]). Кроме того на экран выводится «число степеней свободы», под которым в данном контексте понимается число пар несовпадающих элементов. Пары совпадающих элементов исключаются из обработки.

Обсуждение других критериев этой процедуры дано в примере 3.4к.

Пример 3.4к. С помощью критерия знаков и критерия рангов Уилкоксона проверить гипотезу об совпадении распределений времени реакции на звук и на свет в предположении, что оба распределения непрерывны.

Подготовка данных. Смотри пример 3.3к.

Выбор процедуры. Аналогичен примеру 3.3к.

Заполнение полей ввода данных. В окне Анализ данных (рис. 3.6) выделим с помощью мыши переменные sound и light в поле Переменные. Нажав кнопку со стрелкой вправо, перенесем в поле Для анализа. Затем нажмем кнопку [Утвердить].

Результаты. На экране появятся результаты расчетов для непараметрических критериев положения: Уилкоксона, Ван дер Вардена (см. [21]) и знаков (рис. 3.8).

```
КРИТЕРИИ СДВИГА (ПОЛОЖЕНИЯ).  Файл: sound.std
                          Переменные: sound, light
Вилкоксон=363.5, Z=-2.2747, Значимость=0.0114, степ.своб = 17,17
   Гипотеза 1: <Есть различия между медианами выборок>
Ван дер Варден=15.081, Z=5.5971, Значимость=0, степ.своб = 17,17
   Гипотеза 1: <Есть различия между медианами выборок>

Для парных данных:
Вилкоксон=128.5, Z=3.1294, Значимость=0.0008, степ.своб = 2,16
   Гипотеза 1: <Есть различия между медианами выборок>
Знаков=13, Z=2.75, Значимость=0.0029, степ.своб = 2,16
   Гипотеза 1: <Есть различия между медианами выборок>
```

Рис. 3.8. Результат проверки гипотезы о смещении выборок

Экран выдачи результатов критерия Уилкоксона включает в себя значение статистики W, нормальную аппроксимацию для этой статистики (Z-статистику) и минимальный уровень значимости нулевой гипотезы об отсутствии различий в сдвиге двух выборок против односторонних альтернатив. Указываемый минимальный уровень значимости всегда рассчитывается на основе нормальной аппроксимации.

Для парных данных процедура выдает значения статистик знаков и рангов Уилкоксона, соответствующие им Z-статистики и их минимальные уровни значимости против односторонних альтернатив. Полученные результаты говорят о том, что нулевая гипотеза должна быть отвергнута.

3.7.2. Пакет STATGRAPHICS

Схема испытаний Бернулли. В пакете STATGRAPHICS есть несколько возможностей осуществить проверку нулевой гипотезы о значении вероятности успеха p в схеме испытаний Бернулли. Во-первых, можно воспользоваться процедурами пункта Distribution function (функции распределения) головного меню пакета. Их описание дано в главе 2. Другую возможность проверки гипотезы в схеме испытаний Бернулли против различных альтернатив дает процедура Sample size – Binomial Proportion (размер выборки — биномиальные доли) пункта Sampling (Выборки) головного меню пакета. Разберем их работу на примерах.

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.

Подготовка данных. Согласно нулевой гипотезе вероятность «успеха» $p = 0.3333$. Число испытаний $n = 10$. Значение числа «успехов» $S_{\text{набл.}} = 7$. Эти значения будут вводиться с клавиатуры как параметры процедуры. Мы хотим найти $P(S \geqslant 7 \,|\, n = 10, p = 0.3333) = 1 - P(S \leqslant 6 \,|\, n = 10, p = 1/3)$.

Выбор процедуры. В пункте Distribution function (его описание дано в главе 2) головного меню пакета следует выбрать процедуру Tail Area Probabilities (вероятности хвостов). Приведем вид экрана этой процедуры с заполненными полями ввода данных и результатами проделанных расчетов.

Заполнение полей ввода данных. В поле Distribution number (номер распределения) необходимо ввести номер распределения из приведенного на экране списка (биномиальное распределение имеет номер 2), и нажать клавишу F6. На том же экране появятся поля ввода данных параметров выбранного распределения. В данном случае — Number of trials (число испытаний) и Event probability (вероятность «успеха»). После их заполнения надо нажать клавишу F6, и на том же экране появятся четыре строки ввода Area at or below (функции распределения в точке). По умолчанию эти все эти поля заполнены значением 5. В одно из

```
                    Tail Area Probabilities

Distributions available:
  (1) Bernoulli           (7) Beta              (13) Lognormal
  (2) Binomial            (8) Chi-square        (14) Normal
  (3) Discrete uniform    (9) Erlang            (15) Student's t
  (4) Geometric          (10) Exponential       (16) Triangular
  (5) Negative binomial  (11) F                 (17) Uniform
  (6) Poisson            (12) Gamma             (18) Weibull

Distribution number:  2

Number of trials:  10

Event probability:  0.3333

Area at or below  6          =  0.98035
Area at or below  5          =  0.923471
Area at or below  5          =  0.923471
Area at or below  5          =  0.923471
```

Рис. 3.9. Запрос параметров и результаты процедуры вычисления квантилей

этих полей следует ввести значение $S_{\text{набл.}} - 1$. В примере это значение введено в первое поле, а остальные поля оставлены без изменения.

Результаты. После нажатия F6 против каждого из введенных чисел появится значение функции распределения. Минимальный уровень значимости для статистики $S_{\text{набл.}}$ равен единице минус полученное значение функции распределения. В данном случае $1 - 0.98035 = 0.01965$.

Комментарии. 1. Приведенный порядок действий используется для вычисления функции распределения при $p \leqslant 0.5$. При $p > 0.5$ следует воспользоваться соотношением $P(S = k \mid n, p) = P(S = n - k \mid n, 1 - p)$.

2. Если Вы хотите использовать критерий с фиксированным уровнем значимости p против односторонних альтернатив, то для получения критических значений необходимо воспользоваться процедурой Critical Values (критические значения). Ее описание дано в главе 2.

3. Использовать описанные выше процедуры для проверки гипотез против двусторонних альтернатив можно, но неудобно.

На втором примере разберем работу процедуры Sample size – Binomial Proportion (размер выборки — биномиальные доли).

Пример 3.2к. Определим необходимое число испытаний в тройном тесте и соответствующее критическое значение для проверки гипотезы о неразличимости стимула (т.е. о том, что вероятность его обнаружения $p = 0.3333$) против той альтернативы, что вероятность обнаружения стимула равна 0.9 с вероятностью ошибки первого рода $\alpha = 0.02$ и вероятностью ошибки второго рода $\beta = 0.1$.

Подготовка данных. Вероятность «успеха» при нулевой гипотезе есть $p = 0.3333$, вероятность «успеха» p при альтернативе равна 0.9, уровень значимости $\alpha = 0.02$, вероятность ошибки второго рода $\beta = 0.1$. Эти значения мы введем с клавиатуры как параметры процедуры.

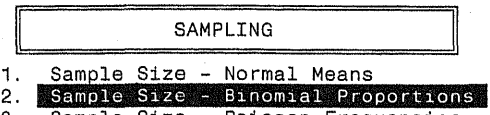


Рис. 3.10. Меню выбора процедуры из группы Sampling (выборки)

Выбор процедуры. В головном меню пакета выберем пункт Sampling (выборки), а в появившемся меню (рис. 3.10) — процедуру Sample Size – Binomial Proportions.

Приведем вид экрана этой процедуры с заполненными полями ввода данных и результатами проделанных расчетов (рис. 3.11).

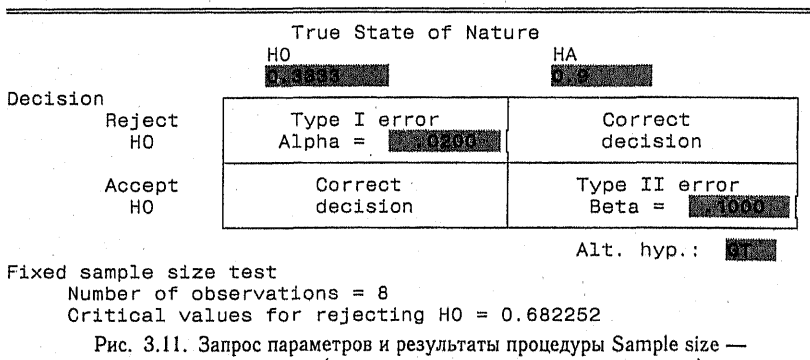


Рис. 3.11. Запрос параметров и результаты процедуры Sample size — Binomial Proportion (размер выборки — биномиальные доли)

Заполнение полей ввода данных. В поле H0 экрана необходимо ввести значение параметра биномиального распределения $p = 0.3333$, который определяет нулевую гипотезу. В поле HA надо ввести значение параметра биномиального распределения, задающего простую альтернативу (в нашем примере — 0.9). В поле ошибки первого рода Alpha = вставим вероятность отвергнуть нулевую гипотезу в том случае, если она верна, то есть 0.02. В поле ошибки второго рода Beta = введем вероятность принять нулевую гипотезу в том случае, если на самом деле верна альтернатива. Наконец, в поле Alt. hyp. надо указать тип альтернативной гипотезы. В этом поле могут стоять значения GT, LT, NE для правосторонних, левосторонних и двусторонних альтернатив соответственно. В нашем случае мы имеем дело с правосторонней альтернативой.

Результаты. Заполнив все активные поля экрана, нажмем клавишу F6. На экране появится сообщение о необходимом числе повторений опыта Number of observations и величине критического значения Critical value for rejection H0 для достижения поставленных целей. Для значений параметров разбираемого примера получаем необходимое число

испытаний $n = 8$, и критическое значение равное 0.682 при выбранном параметре $\beta = 0.1$. Под критическим значением здесь понимается доля успехов от общего числа испытаний. Значение доли успехов при пяти успехах в восьми испытаниях равно 0.625, а при шести успехах соответственно 0.75. Сравнивая эти значения с критическим значением, выданным процедурой, заключаем, что мы должны отвергнуть нулевую гипотезу, если число успехов в 8 испытаниях будет больше или равно 6 при заданных значениях ошибок первого и второго рода.

Комментарии. 1. Описанную процедуру целесообразно применять до начала проведения эксперимента. Она строит фиксированные и последовательные планы эксперимента для оценки параметров биномиального распределения. В фиксированных планах число экспериментов выбирается заранее, а в последовательных оно определяется с учетом результатов уже проведенных экспериментов.

2. Процедуру можно использовать для обычной проверки гипотез в схеме испытаний Бернулли, когда объем выборки задан заранее. Порядок действий при этом иной, чем описанный выше. Выбрав саму процедуру и войдя в экран ее параметров необходимо сразу нажать клавишу F5 и получить на экране подменю, в котором следует выбрать пункт Enter sample size (введение размера выборки). При этом осуществляется возврат в модифицированный экран ввода параметров процедуры. Поле ошибки второго рода становится пассивным, то есть его уже не надо заполнять. Кроме того, появляется приглашение к вводу объема выборки n. В данном случае для заданной ошибки первого рода и объема выборки будут вычислены критическое значение и величина ошибки второго рода. Введя в указанные поля данные примера 3.1к, получим следующие результаты: критическое значение равняется 0.639 и ошибка второго рода — 0.04. То есть следует отвергнуть гипотезу при числе успехов больше или равном семи, что соответствует выведенным ранее результатам.

3. Процедура расчета использует нормальную аппроксимацию биномиального распределения и ее результаты для малого числа испытаний (как это было в нашем случае) должны рассматриваться как приблизительные.

Два следующих примера посвящены проверке гипотез с помощью непараметрических критериев знаков и рангов. В случае одной выборки критерий знаков используется для проверки гипотезы о равенстве медианы заданному значению (пример 3.3к). Для парных данных критерий знаков и критерии рангов Уилкоксона используются при проверке гипотезы об отсутствии эффекта обработки (пример 3.4к).

Пример 3.3к. 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 базы данных пакета (процедура 2. File Operations пункта A. Data Management головного меню пакета) в файле SPEEDR создадим 2 целочисленных переменных с именами sound и light,

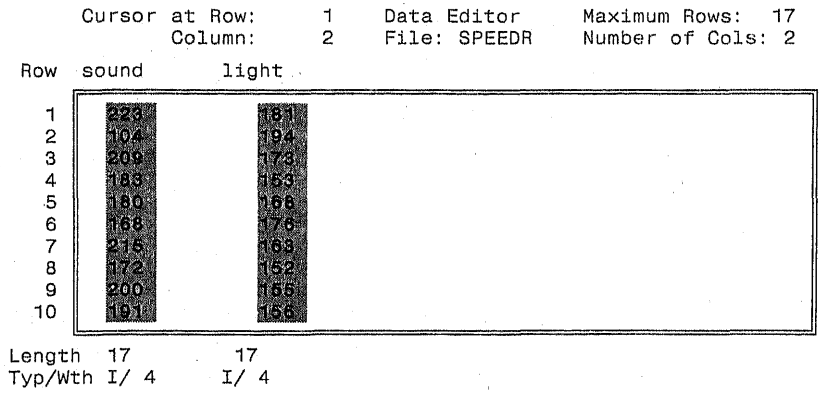


Рис. 3.12. Данные для проверки гипотез о времени реакции на свет и звук

каждая из которых содержит результаты испытуемых в соответствующей группе (см. табл. 3.1). Вид экрана редактора базы данных с частью введенных данных приведен на рис. 3.12.

Выбор процедуры. В головном меню пакета выберем пункт R. Nonparametric Methods. Меню этого пункта представлено на рис. 3.13. Выберем в нем процедуру 3. Tests for Location (критерии положения).

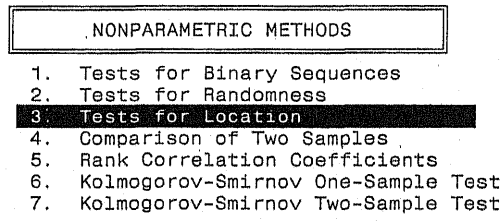


Рис. 3.13. Меню непараметрических методов

Заполнение полей ввода данных. Экран ввода данных выбранной процедуры (рис. 3.14) содержит активные поля: Data (данные), Hypothesized median (гипотетическая медиана) и Test based on (тест основан на). Заполним их данными примера, как показано на рис. 3.14. При этом в поле Test based on следует указать значение Signs, означающее выбор критерия знаков.

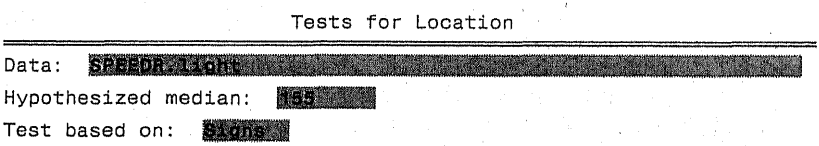


Рис. 3.14. Запрос параметров процедуры проверки гипотезы о положении выборки

Результаты. После нажатия F6 на том же экране появятся результаты расчетов для критерия знаков. Они приведены отдельно на рис. 3.15 и включают в себя значение выборочной медианы (Sample median), число наблюдений, превосходящих гипотетическую медиану (Number of values above hypothesized median), число наблюдений, меньших, чем гипотетическая медиана (Number of values below hypothesized median), ожидаемое число наблюдений, превосходящих гипотетическую медиану (Expected number), значение нормальной аппроксимации Z для статистики критерия знаков (Large sample test statistic Z) и минимальный уровень значимости критерия против двусторонних альтернатив (Two-tailed probability of equaling or exceeding Z). Строка комментария (NOTE) указывает общее число анализируемых наблюдений и число выборочных значений, совпавших с гипотетической медианой и игнорированных при расчетах.

```
Sample median = 163
Number of values above hypothesized median = 11
Number of values below hypothesized median = 4
Expected number = 7.5
Large sample test statistic Z = 1.54919
Two-tailed probability of equaling or exceeding Z = 0.121335

NOTE: 17 observations. 2 values equal to hypothesized median ignored.
```

Рис. 3.15. Результаты расчетов для одновыборочного критерия знаков

Полученный минимальный уровень значимости критерия против двусторонних альтернатив достаточно велик и не позволяет однозначно отвергнуть выдвинутую гипотезу.

Комментарии. 1. Значение минимального уровня значимости, выдаваемое процедурой, является приблизительным, так как базируется на использовании нормальной аппроксимации для статистики критерия знаков. Для столь малого числа наблюдений, как в разбираемом примере, точность нормальной аппроксимации недостаточна.

2. При указании значения Ranks (ранги) в поле Test based on экрана ввода данных процедуры (рис. 3.14) будет выполнен критерий рангов Уилкоксона [91].

Пример 3.4к. С помощью критерия знаков и критерия Манна–Уитни (Уилкоксона) проверим гипотезу о совпадении распределений времени реакции на звук и на свет в предположении, что оба распределения непрерывны.

Подготовка данных. См. пример 3.3к.

Выбор процедуры. В меню пункта R. Nonparametric Methods (рис. 3.13) выберем процедуру 4. Comparison of Two Samples (сравнение двух выборок).

Заполнение полей ввода данных. На рис. 3.16 приведен экран ввода данных процедуры (первые три поля) с результатами расчетов для критерия знаков.

```
                    Comparison of Two Samples

Sample 1:  SPEEDR.sound
Sample 2:  SPEEDR.light
Test based on:  Signs

Number of positive differences = 13
Number of negative differences = 3
Expected number = 8
Large sample test statistic Z = 2.25
Two-tailed probability of equaling or exceeding Z = 0.0244488

NOTE:  17 total pairs.  1 tied pairs ignored.
```

Рис. 3.16. Запрос параметров и результаты сравнения положения двух выборок с помощью критерия знаков

В поля Sample 1 и Sample 2 надо ввести переменные со значениями первой и второй выборки. В случае использования критерия знаков длины выборок должны совпадать. Для выполнения критерия знаков в поле Test based on следует указать значение Signs. Для выполнения критерия Манна–Уитни (Уилкоксона) в этой графе надо указать Pairs. В последнем случае анализируемые выборки могут быть произвольной длины.

Результаты. На рис. 3.16 приведены результаты расчетов для критерия знаков. Они включают в себя вычисление числа положительных (Number of positive differences) и отрицательных (Number of negative differences) разностей, ожидаемое число положительных разностей при нулевой гипотезе (Expected number), нормальную аппроксимацию Z-статистики критерия знаков для больших выборок (Large sample test statistic Z), минимальный уровень значимости критерия против двусторонних альтернатив (Two-tailed probability of equaling or exceeding Z), вычисленный с помощью нормальной аппроксимации. Комментарий к процедуре (NOTE) указывает число сравниваемых пар и число игнорированных нулевых разностей.

Сравним полученные результаты с рассчитанными ранее вручную. Заметим, что при вычислении разностей в пакете значение первой выборки вычитается из значения второй. Поэтому полученное процедурой число положительных разностей есть не что иное, как число отрицательных разностей, полученных при расчетах вручную.

Результаты вычислений для критерия рангов Уилкоксона приведены на рис. 3.17. Они включают число положительных разностей и их средний ранг (Number of positive differences = 13 with average rank = 8.88462), аналогичные значения для отрицательных разностей, нормальную аппроксимацию Z статистики критерия рангов Уилкоксона для больших выборок (Large sample test statistic Z) и минимальный уровень значимости критерия против двусторонних альтернатив (Two-tailed probability of equaling or exceeding Z), рассчитанный с помощью нормальной аппроксимации. В

```
Test based on: Ranks

Number of positive differences = 13 with average rank = 8.88462
Number of negative differences = 3 with average rank = 6.83333
Large sample test statistic Z = 2.43031
Two-tailed probability of equaling or exceeding Z = 0.0150858

NOTE:  17 total pairs.  1 tied pairs ignored.
```

Рис. 3.17. Результаты сравнения положения двух выборок с помощью критерия рангов Уилкоксона

комментарии (NOTE) указывается общее число пар наблюдений, а также число пар с совпадающими значениями.

Полученные с помощью разных критериев минимальные уровни значимости для проверки нулевой гипотезы о совпадении распределений близки между собой и достаточно малы, что позволяет скорее отвергнуть гипотезу, чем принять ее.

Комментарии. 1. При использовании этих процедур для малых выборок их результаты (минимальные уровни значимости) должны рассматриваться как приблизительные.

2. При малых объемах выборок для получения минимального уровня значимости критерия знаков против односторонних альтернатив следует подставить полученные значения числа ненулевых разностей и положительных разностей в процедуру Tail Area Probabilities для биномиального распределения с вероятностью успеха $p = 0.5$ (см. пример 3.1к) в качестве числа испытаний и критического значения.

Глава 4

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 (то есть 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, то есть 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 [84] и [92]. А изложение классических результатов теории оценивания можно найти в [14], [49] 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 главе 5 будет более подробно рассмотрено оценивание параметров для нормально распределенных выборок. А в главах 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 главе 1, иногда эту вероятность можно установить из соображе-

ний 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 [13]). 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 $n \to \infty$, то $S/n \to p$ 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 $a_n \to a$ при $n \to \infty$, 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 непременное выполнение соотношения $\left|\frac{S}{n} - p\right| < \varepsilon$. Поэтому для случайных последовательностей используется другое понятие предела:

$$P\left(\left|\frac{S}{n} - p\right| < \varepsilon\right) \to 1$$

(для любого $\varepsilon > 0$) при $n \to \infty$. Когда требуется отличать это понятие предела от того, которое используется в математическом анализе, говорят: «последовательность случайных величин сходится по вероятности».

Итак, событие $\left|\frac{S}{n} - p\right| < \varepsilon$ 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, $i = 1, \ldots, n$, 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 $X_1 + \cdots + X_n$, где случайные величины $X_1, \ldots, X_n$ независимы и одинаково распределены, причем для любого i: $MX_i = p$. Тогда мы можем переформулировать утверждение теоремы Бернулли в виде:

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - MX\right| < \varepsilon\right) \to 1 \quad \text{при } n \to \infty.$$

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 $X_1, \ldots, X_n$ — 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 $n \to \infty$*

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - a\right| < \varepsilon\right) \to 1.$$

В статистике среднее арифметическое величин $X_1, \ldots, X_n$ обозначают $\overline{X}$. Так что кратко теорему Чебышева можно записать так: $\overline{X} \to a$.

Доказательство. Для доказательства теоремы нам потребуется неравенство Чебышева: пусть ξ — неотрицательная случайная величина, имеющая математическое ожидание. В таком случае для любого $\varepsilon > 0$

$$P(\xi \geqslant \varepsilon) \leqslant \frac{M\xi}{\varepsilon}.$$

Проведем доказательство этого неравенства для случая, когда непрерывная случайная величина имеет плотность распределения вероятностей $f(x)$:

$$P(\xi \geqslant \varepsilon) = \int_{\varepsilon}^{\infty} f(x)\,dx \leqslant \int_{\varepsilon}^{\infty} \frac{x}{\varepsilon} f(x)\,dx \leqslant \frac{1}{\varepsilon}\int_{\varepsilon}^{\infty} x f(x)\,dx = \frac{M\xi}{\varepsilon},$$

что и требовалось доказать.

Применив это неравенство к неотрицательной случайной величине $\xi = (X - MX)^2$, получим, что

$$P(|X - MX| \geqslant \varepsilon) = P((X - MX)^2 \geqslant \varepsilon^2) \leqslant \frac{1}{\varepsilon^2} M(X - MX)^2 = \frac{1}{\varepsilon^2} DX,$$

то есть для любой случайной величины, имеющей математическое ожидание и дисперсию, $P(|X - MX| > \varepsilon) < DX/\varepsilon^2$.

Применим это утверждение к $\overline{X}$. Легко видеть, что $M\overline{X} = a$, $D\overline{X} = \sigma^2/n$, где $\sigma^2 = DX_i$. По неравенству Чебышева

$$P(|\overline{X} - a| \geqslant \varepsilon) \leqslant \frac{D\overline{X}}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \to 0.$$

Поэтому вероятность противоположного события $\{|\overline{X} - a| < \varepsilon\}$ стремится к 1, что и требовалось доказать.

Продолжение примера. Вернемся к измерению времени на «Бигле». Показание каждого прибора x_i, $i = 1, \ldots, n$ — 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: $MX_i = a$, $i = 1, \ldots, n$. 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 $D\overline{X}$ в n раз меньше DX_i. Поэтому «среднее время» $\overline{X}$ 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, $i = 1, \ldots, n$, связанные с отдельными испытаниями. Пусть $X_i = 1$, если испытание i закончилось «успехом», и $X_i = 0$ — в противном случае. Ясно, что $S = X_1 + \cdots + X_n$, а случайные величины $X_1, \ldots, X_n$ независимы и одинаково распределены. Легко видеть, что $MX_i = p$, $DX_i = p(1-p)$, а $\frac{S}{n} = \overline{X}$. По теореме Чебышева $P(|S/n - p| < \varepsilon) \to 1$ при $n \to \infty$, 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 $n \to \infty$ $\theta_n \to \theta$ 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 $n \to \infty$) 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 будем 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): $\varphi(x) = \frac{1}{\sqrt{2\pi b}} \exp(\frac{(x-a)^2}{2b})$.

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, $\theta = (\theta_1, \ldots, \theta_r)$. Обычно параметризацию семейства распределений вводят так, чтобы между значениями параметров и элементами семейства устанавливалось взаимно-однозначное соответствие, т.е. чтобы разным наборам $\theta = (\theta_1, \ldots, \theta_r)$ и $\theta' = (\theta'_1, \ldots, \theta'_r)$ 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 $x_1, \ldots, x_n$ — 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 $x_1, \ldots, x_n$, скажем, $t(x_1, \ldots, x_n)$. При этом надо, чтобы

$$t(x_1, \ldots, x_n) \simeq \theta. \tag{4.1}$$

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 сколько-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: $\overline{x} \simeq a$. Выведем аналогичный результат для дисперсии распределения.

Оценка дисперсии распределения. Пусть $x_1, \ldots, x_n$ — 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

$$(x_1 - \overline{x})^2, \quad (x_2 - \overline{x})^2, \ldots, \quad (x_n - \overline{x})^2.$$

Остается применить усреднение вместо внешнего символа M. Получаем приближенное выражение для дисперсии: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Докажем закон больших чисел для дисперсии. Нам надо показать, что при $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \to D\xi. \tag{4.2}$$

Для этого прежде преобразуем $\sum_{i=1}^{n}(x_i - \overline{x})^2$ следующим образом:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}[(x_i - a)^2 - (\overline{x} - a)]^2 =$$

$$\sum_{i=1}^{n}(x_i - a)^2 - 2(\overline{x} - a)\sum_{i=1}^{n}(x_i - a) + n(\overline{x} - a)^2 = \sum_{i=1}^{n}(x_i - a)^2 - n(\overline{x} - a)^2.$$

Поэтому левая часть соотношения (4.2) равна

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 - (\overline{x} - a)^2. \tag{4.3}$$

Так как $\overline{x} \to a$, второй член выражения (4.3) стремится при $n \to \infty$ к нулю. Первый же член выражения (4.3) при $n \to \infty$ сходится к $M(\xi - a)^2$, т.е. к $D\xi$, что и доказывает утверждение (4.2).

Выражение $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ можно назвать выборочной дисперсией (иногда говорят — *дисперсия выборки*). Однако чаще вместо него используют

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

Понятно, что уменьшение n на 1 в знаменателе левой части (4.2) не сказывается на предельном поведении этого выражения и $s^2 \to D\xi$ при $n \to \infty$. В то же самое время s^2 обладает тем свойством, что

$$Ms^2 = D\xi \quad \text{при любом } n, \tag{4.4}$$

что считается достоинством. Говорят, что s^2 является *несмещенной оценкой* $D\xi$.

Для доказательства (4.4) надо обратиться к (4.3) и учесть, что $M(\overline{x} - a)^2 = D\overline{x}$, так как $M\overline{x} = a$. Как отмечалось ранее, $D\overline{x} = \frac{1}{n}D\xi$, поэтому $M\sum_{i=1}^{n}(x_i - \overline{x})^2 = nD\xi - D\xi = (n-1)D\xi$. 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 (то есть 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 $X_1, \ldots, X_n$ — 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 $\sigma^2 + a^2$. Приравнивая 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

$$a^* = \overline{x}, \qquad \sigma^{2*} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

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 $x_1, \ldots, x_n$, 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:

$$a^* = Q_n(0.5) = \mathrm{med}(x_1, \ldots, x_n), \quad \sigma^* = \frac{1}{2\Phi^{-1}(0.75)}[Q_n(0.75) - Q_n(0.25)].$$

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 $\theta_n \to \theta$ по вероятности, когда $n \to \infty$;
- 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 $MW(\theta_n, \theta)$ 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 зависит 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:

$$P_\theta\left\{\hat{\theta} \in A(\theta, \alpha)\right\} \geqslant \alpha \qquad \text{для любого } \theta.$$

Тогда *доверительной областью* (в одномерном случае — *доверительным интервалом*) с уровнем доверия α для неизвестного нам истинного значения θ, построенной по наблюденному в опыте значению оценки $\hat{\theta}$, является множество

$$\left\{\theta \,|\, \hat{\theta} \in A(\theta, \alpha)\right\}.$$

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 главе 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 $X_1, \ldots, X_n$ — выборка из распределения, плотность которого в точке x зависит от неизвестного параметра θ. Обозначим плотность отдельного наблюдения X_i $(i = 1, \ldots, n)$ через $p(x, \theta)$. Поскольку случайные величины $X_1, \ldots, X_n$ независимы, плотность вероятностей вектора $(X_1, \ldots, X_n)$ равна

$$p(x_1, \theta)\, p(x_2, \theta) \ldots p(x_n, \theta), \tag{4.5}$$

где θ — неизвестное нам истинное значение параметра.

Метод наибольшего правдоподобия состоит в следующем. Подставим в (4.5) вместо переменных $(x_1, \ldots, x_n)$ элементы выборки, т.е. реализации случайных величин $X_1, \ldots, X_n$, 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:*

$$p(X_1, \theta)\, p(X_2, \theta) \ldots p(X_n, \theta). \tag{4.6}$$

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:

$$p(X_1,\theta)\,p(X_2,\theta)\ldots p(X_n,\theta)\to\max_{\theta}.$$

Ясно, что такой выбор $\hat\theta$ происходит в зависимости от значений $X_1,\ldots,X_n$, поэтому $\hat\theta$ является функцией от $X_1,\ldots,X_n$, 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

$$\Big(\frac{1}{\sqrt{2\pi b}}\Big)^n\exp\Big\{-\frac{1}{2b}\sum_{i=1}^n(x_i-a)^2\Big\}. \tag{4.7}$$

Надо выбрать параметры a,b так, чтобы выражение (4.7) было максимальным (при заданных значениях $x_1,\ldots,x_n$). 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 $a=\overline{x}$. Следовательно, оценка наибольшего правдоподобия $\hat a$ для a равна $\overline{x}$.

Для того, чтобы найти оценку наибольшего правдоподобия параметра b, вычислим

$$\max_b\left[(2\pi b)^{-\frac12}\exp\left\{-\frac{1}{2b}\sum_{i=1}^n(x_i-\overline{x})^2\right\}\right].$$

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 $\hat b=\frac1n\sum_{i=1}^n(x_i-\overline{x})^2$.

Итак, оценка наибольшего правдоподобия для (a,b) равна

$$\Big(\overline{x},\quad\frac1n\sum_{i=1}^n(x_i-\overline{x})^2\Big).$$

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 $p(X_1,\theta)\ldots p(X_n,\theta)$ и

$$\frac1n\ln[p(X_1,\theta)\ldots p(X_n,\theta)] \tag{4.8}$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\ln p(X_i,\theta) \tag{4.9}$$

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 $\theta=\theta^0$: $\int p(x,\theta^0)\ln p(x,\theta^0)\,dx \geqslant \int p(x,\theta^0)\ln p(x,\theta)\,dx$ для любого θ, и более того, для *любой* функции плотности $q(x)$:

$$\int p(x,\theta^0)\ln p(x,\theta^0)\,dx \geqslant \int p(x,\theta^0)\ln q(x)\,dx. \tag{4.10}$$

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 $p_1,\ldots,p_n$ и $q_1,\ldots,q_n$ — два набора положительных величин, причем $\sum_{i=1}^n p_i=1$, $\sum_{i=1}^n q_i=1$. Тогда

$$\sum_{i=1}^{n} p_i\ln q_i \leqslant \sum_{i=1}^{n} p_i\ln p_i\,.$$

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, [44]). 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 — вольно выражаясь, они используют всю информацию, ничего не оставляя в качества запаса прочности.

4.7. Оценивание параметров вероятностных распределений в пакетах STADIA и STATGRAPHICS

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 почти все пакеты выводят значения наиболее распространенных оценок параметров стандартных вероятностных распределений (см. пп. 2.6.1, 2.6.2). В примере 4.1 мы рассмотрим, как эти возможности реализованы в пакетах STADIA и STATGRAPHICS.

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 4.2.

Задача оценивания параметров нормального распределения в статистических пакетах рассматривается отдельно в главе 5.

4.7.1. Пакет STADIA

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. Решение первой части этой задачи осуществляет в пакете процедура 3=Генератор чисел меню Преобразования. Ее работа была подробно рассмотрена в примере 2.3к.

Для экспоненциального распределения в пакете используется следующая параметризация плотности распределения: $p(x, b) = \frac{1}{b}e^{-x/b}$, где $x \geqslant 0$, а параметр b является средним значением распределения (см. п. 2.3).

После вызова меню Преобразования (функциональная клавиша F8) и выбора пункта 3=Генератор чисел, в открывшемся запросе (рис. 2.16) укажите размер выборки 100 в поле Всего чисел и величину среднего значения 3 в поле a=. (Содержание поля b= в данном случае несущественно.) Затем укажите тип распределения — экспоненциальное. Результат генерации помещается в первый свободный столбец электронной таблицы (на рис. 4.1 это переменная x1).

Выбор процедуры. В меню Статистические методы (рис. 1.17) в разделе Распределения и частоты выберите пункт U = Согласие распределений.

Заполнение полей ввода данных. Программа запросит тип вероятностного распределения выборки (рис. 4.2).

Результаты. На рис. 4.3 приведены результаты работы процедуры. При этом в строке

Распределение экспоненциальное: 2.996, 3.0272

первое число 2.996 является требуемой оценкой, а второе — стандартной ошибкой среднего значения. Процедура также вычисляет статистики

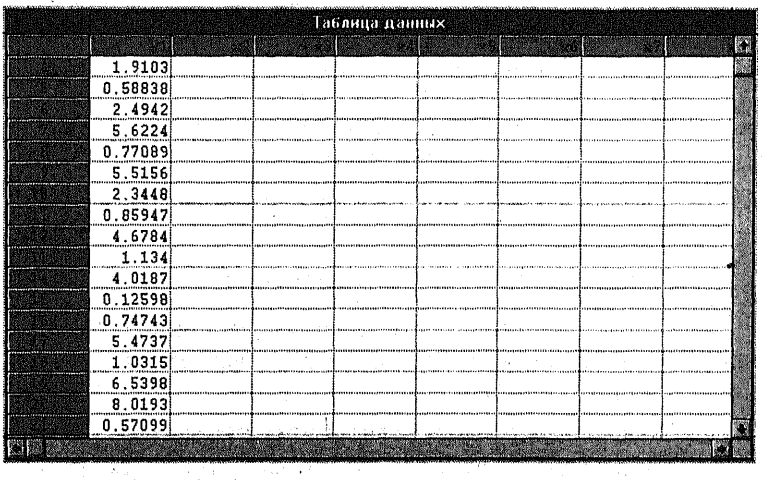


Рис. 4.1. Пакет STADIA. Электронная таблица со сгенерированной выборкой

Рис. 4.2. Пакет STADIA. Меню выбора вида распределения

```
СОГЛАСИЕ РАСПРЕДЕЛЕНИЙ. Файл:
Распределение экспоненциальное: 2.996, 3.0272

Колмогоров=0.056018, Значимость=0.6, степ.своб = 100
   Гипотеза 0: <Распределение не отличается от теоретического>
Омега-квадрат=0.063088, Значимость=0.52224, степ.своб = 100
   Гипотеза 0: <Распределение не отличается от теоретического>
```

Рис. 4.3. Пакет STADIA. Результат оценки параметров распределения и проверки согласия

критериев согласия Колмогорова и омега-квадрат (они будут рассмотрены в главе 10) и строит графики плотности и функции подобранного распределения вероятностей.

Комментарии. 1. Полученная оценка является оценкой максимального правдоподобия.

2. В документации пакета указаны формулы, используемые для получения оценок для распределений из списка рис. 4.2. По этим формулам можно судить о теоретических свойствах вычисляемых оценок.

3. Для решения примера можно также воспользоваться процедурой **Описательная статистика**, поскольку параметр экспоненциального распределения является его средним значением.

Следующий пример иллюстрирует возможности вспомогательных вычислительных процедур для построения оценок параметров распределений. В связи с этим его характер отчасти искусственен.

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 базы данных пакета (рис. 4.4).

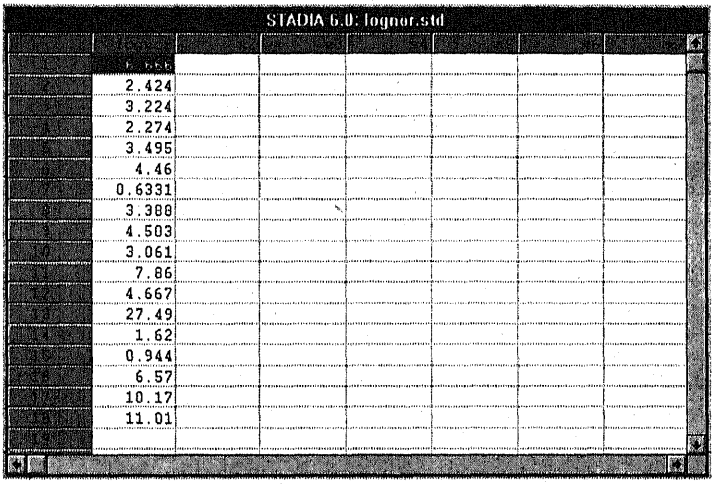


Рис. 4.4. Пакет STADIA. Электронная таблица с выборкой из логнормального распределения

Оценка максимального правдоподобия параметра μ задается выражением

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \ln(x_i).$$

Для ее непосредственного вычисления воспользуемся процедурами преобразования данных пакета.

Выбор процедуры. В меню Преобразования выберите пункт 1 = стандартная функция. На экране появится всплывающее окно с перечнем стандартных функций (рис. 4.5).

Рис. 4.5. Меню стандартных функций для преобразования данных

Заполнение полей ввода данных. В окне рис. 4.5 выберем пункт 2 = LN (нажав (2) или воспользовавшись мышью). После этого в переменной lognor будет находиться массив натуральных логарифмов элементов первоначального массива. Так как искомая оценка является средним арифметическим полученных значений, для ее вычисления можно использовать процедуру Описательная статистика из меню Статистические методы. Работа этой процедуры подробно описана в примере 1.1к.

Результаты. На рис. 4.6 приведены результаты работы процедуры Описательная статистика. Искомая оценка параметра μ находится в графе Среднее и равна 1.3805. Требуемую оценку в пакете можно было бы получить и сразу, воспользовавшись процедурой U=согласие распределений (пример 4.1).

```
ОПИСАТЕЛЬНАЯ СТАТИСТИКА. Файл:
Переменная   Размер  <---Диапазон--->  Среднее---Ошибка  Дисперс  Ст.откл
   lognor       18 -0.45713   3.3138   1.3805  0.21126  0.80338  0.89632

Переменная  Медиана  <---Квартили--->  ДовИнтСр. <-ДовИнтДисп->  ОшСтОткл
   lognor   1.3732  0.86945   1.9382   0.44046   0.4524    1.806  0.36983

Переменная Асимметр. Значим  Эксцесс   Значим
   lognor -0.045086  0.4633   3.1527   0.2717
```

Рис. 4.6. Результы работы процедуры описательной статистики

Комментарии. 1. Запись результата вычисления значений стандартной функции на место первоначальной переменной не всегда удобна, так как ведет к потере первоначального массива данных. Прежде чем выполнять подобные процедуры, можно создать копию первоначальной переменой в электронной таблице, используя буфер обмена.

2. Указанное выше неудобство отчасти компенсируются простотой проведения преобразований.

3. Для устойчивого оценивания в блоке преобразования данных пакета имеется процедура С=пропущ. значения, которая выделяет в матрице данных возможные «выбросы». Под «выбросом» в пакете понимается величина, отклоняющаяся от среднего значения переменной более чем на два стандартных отклонения. Каждый из выбросов может быть заменен либо «пропущенным значением» (при этом не происходит нарушения общей структуры матрицы данных и она может быть использована целиком в различных методах анализа модельных задач),

либо средним значением по переменной, либо, если данные имеют соответствующую структуру, — регрессионными оценками.

4. В целом пакет STADIA предоставляет достаточно простых возможностей для самостоятельного построения различных оценок.

4.7.2. Пакет STATGRAPHICS

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. Решение первой части задачи осуществляет процедура 5. Random Number Generation пункта H. Distribution function (рис. 2.16) головного меню пакета. Ее работа разобрана в примере 2.3к. В пакете используется следующая параметризация плотности экспоненциального распределения: $p(x, b) = \frac{1}{b}e^{-x/b}$, где $x \geqslant 0$, а параметр b является средним значением распределения (см. пункт 2.3). Поэтому процедура генерации запрашивает значение Mean в качестве параметра распределения. Пусть результат работы процедуры помещен в переменную EXPO.var.

Выбор процедуры. В головном меню пакета выберем пункт H. Distribution function, а в этом пункте — процедуру 1. Distribution Fitting (подбор распределения, см. рис. 2.16).

Заполнение полей ввода данных. Экран ввода данных процедуры (с результатами расчетов для экспоненциального распределения) приведен на рис. 4.7. В поле Data vector необходимо ввести анализируемую выборку, а в поле Distribution number — номер требуемого распределения из списка на экране.

```
                    Distribution Fitting
═══════════════════════════════════════════════════════════════
Data vector:  EXPO.var
Distributions available:
   (1) Bernoulli          (7) Beta              (13) Lognormal
   (2) Binomial           (8) Chi-square        (14) Normal
   (3) Discrete uniform   (9) Erlang            (15) Student's t
   (4) Geometric         (10) Exponential       (16) Triangular
   (5) Negative binomial (11) F                 (17) Uniform
   (6) Poisson           (12) Gamma             (18) Weibull
Distribution number:  10
Mean:  3.08867
```

Рис. 4.7. Подбор параметров распределения по выборке

Результаты. После заполнения полей ввода следует нажать клавишу F6, и на том же экране в поле Mean появится значение оценки максимального правдоподобия параметра b.

Комментарии. 1. Основное назначение данной процедуры — проверка критериев согласия выборочных данных с указанным распределением. Эта часть ее работы обсуждается в главе 10.

2. В документации к пакету указана параметризация распределений из списка рис. 4.7, но отсутствуют используемые формулы для оценок и указания на тип оценок. В связи с этим, в качестве иллюстрации разнообразия подходов к оцениванию, приведем конкретные формулы и тип оценок для некоторых из указанных выше распределений.

Распределение Пуассона имеет стандартную параметризацию, при которой

$$P(\xi = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, \ldots, \quad \lambda > 0.$$

В качестве оценки параметра λ по выборке $x_1, x_2, \ldots, x_n$ 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:

$$f(x, a, b) = \frac{1}{b - a}, \qquad (a \leqslant x \leqslant b),$$

где параметры a и b задают левую и правую границу распределения. В качестве оценок параметров a и b в пакете используются оценки максимального правдоподобия:

$$\hat{a} = x_{(1)}, \qquad \hat{b} = x_{(n)},$$

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:

$$a^* = \frac{1}{n-1}(nx_{(1)} - x_{(n)}), \qquad b^* = \frac{1}{n-1}(nx_{(n)} - x_{(1)}).$$

Для распределения хи-квадрат (8) Chi-square с параметризацией плотности распределения $f(x, \nu)$

$$f(x, \nu) = \frac{x^{(\nu-2)/2} e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}, \qquad x \geqslant 0$$

в качестве оценки параметра ν используется оценка, построенная по методу моментов на базе первого выборочного момента

$$\nu^* = \overline{x}.$$

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 EX.lognor базы 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. Укажем один из возможных алгоритмов получения искомой оценки с помощью прямых вычислений. Для этого будет использован командный режим работы в пакете.

Из любой процедуры пакета (меню, подменю, экрана ввода параметров и т.п.) можно перейти в командный режим работы, нажав клавишу [F8]. При этом появится всплывающее окно (рис. 4.8) с запросом Enter command (введите команду). В ответ на этот запрос надо ввести EXEC — имя процедуры, позволяющей проводить различные арифметические и функциональные преобразования данных. На рис. 4.8 приведен вид окна командного режима, вызванного из редактора базы данных пакета.

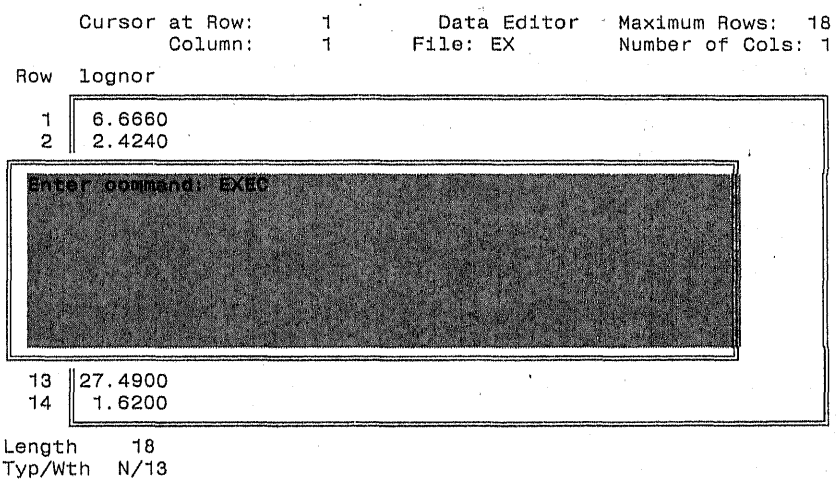


Рис. 4.8. Окно командного режима пакета STATGRAPHICS

Заполнение полей ввода данных. После ввода имени процедуры EXEC надо нажать клавишу [F6]. На экране появится окно процедуры EXEC (рис. 4.9), в которое надо ввести выражение

SUM (LOG EX.lognor)/18

и нажать [F6]. Указанное выражение приведет к следующим вычислениям. Сначала будет выполнен оператор (LOG EX.lognor), который сформирует переменную, элементы которой являются натуральными логарифмами от элементов переменной EX.lognor. Затем каждый элемент полученной переменной будет разделен на 18 (число элементов выборки). Наконец оператор SUM вычислит сумму элементов переменной, вычисленной с помощью предыдущих преобразований.

Результаты. На рис. 4.9 приведен экран с результатами требуемых вычислений.

```
: SUM (LOG EX.lognor)/18
1.38097
:
```

Рис. 4.9. Результаты вычислений оценки по заданной формуле

Далее можно продолжить вычисления или нажать клавишу [Esc], что приведет к возврату в то состояние пакета, из которого был вызван командный режим работы.

Комментарии. 1. Широкий круг стандартных математических функций и преобразований загружается непосредственно при запуске пакета. К ним относятся, кроме приведенных выше, квадратный корень (SQRT), экспонента (EXP) и др. Дополнительно пользователь может загрузить функции: синус (SIN), косинус (COS), тангенс (TAN), арксинус (ASIN), арккосинус (ACOS), арктангенс (ATAN), возведение в степень десяти (EXP10), десятичный логарифм (LOG10), гамма-функцию (GAMMA) и др. Каждая из этих функций может преобразовывать поэлементно одномерный массив. Для их загрузки необходимо выполнить процедуру Mathematical functions (математические функции) пункта V. Supplementary Operations (дополнительные операции) головного меню пакета.

2. Непосредственное вычисление оценки можно также осуществить с помощью операции J. Update процедуры 2. File Operation пункта A. Data Management головного меню пакета.

3. Параметризация плотности логнормального распределения осуществляется через среднее значение этого распределения, поэтому процедура 1. Distribution Fitting (подбор распределения), разобранная в примере 4.1к, не пригодна для получения оценки параметра μ.

Глава 5

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 главе 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. [17], [23], [94]) 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 (то есть 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 этого 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 не устойчивы к грубым (ошибочным) наблюдениям или выбросам. Поэтому при их использовании следует соблюдать определенную осторожность и внимательно изучать другие сопутствующие описательные характеристики выборок (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: $F(x) = \Phi\left((x-a)/\sigma\right)$, 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 $x_1, \ldots, x_n$ 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 $x_{(1)}, \ldots, x_{(n)}$. 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 $(x_{(i)}, \Phi^{-1}(\frac{2i-1}{2n}))$ 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 таблице 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 $\Phi^{-1}(1/34)$, $\Phi^{-1}(3/34), \ldots \Phi^{-1}(33/34)$. Заметим, что $\Phi^{-1}((2k-1)/2n) = -\Phi^{-1}((2n-2k+1)/2n)$. Отсюда имеем:

$$\Phi^{-1}(17/34) = \Phi^{-1}(1/2) = 0,$$
$$\Phi^{-1}(19/34) = -\Phi^{-1}(15/34) = 0.1479, \quad \Phi^{-1}(21/34) = -\Phi^{-1}(13/34) = 0.2993,$$
$$\Phi^{-1}(23/34) = -\Phi^{-1}(11/34) = 0.4578, \quad \Phi^{-1}(25/34) = -\Phi^{-1}(9/34) = 0.6289,$$
$$\Phi^{-1}(27/34) = -\Phi^{-1}(7/34) = 0.8208, \quad \Phi^{-1}(29/34) = -\Phi^{-1}(5/34) = 1.0494,$$
$$\Phi^{-1}(31/34) = -\Phi^{-1}(3/34) = 1.3517, \quad \Phi^{-1}(33/34) = -\Phi^{-1}(1/34) = 1.8895.$$

На рис. 5.1 приведены значения $F_n(x)$ в плоскости (x, z). Глазомерный метод позволяет нам судить, насколько правдоподобна гипотеза о нормальности распределения выборки. Однако четкого критерия отклонения гипотезы он не дает.

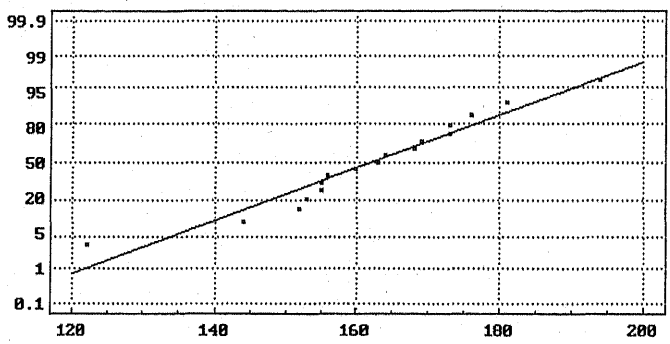


Рис. 5.1. Значения скачков эмпирической функции распределения $F_n(x)$ на плоскости (x, z) (вдоль оси ординат приведены значения $\Phi(z)$ в процентах)

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 $x_1, \ldots, x_n$ — 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 $x_1, \ldots, x_n$. Например, в качестве оценки параметра a часто используют медиану выборки $x_1, \ldots, x_n$ 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 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ можно рассматривать величину $\left[\frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|\right]^2$ 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 $\overline{x}$ и s^2, начав с $\overline{x}$.

Свойства выборочного среднего. Мы уже знаем, что по закону больших чисел (см. гл. 4) выборочное среднее $\overline{x}$ стремится к a с увеличением объема выборки n, т.е. $\overline{x}$ приблизительно равно a при больших объемах выборки. Нас будет интересовать, насколько точным является это приближенное равенство. Близость $\overline{x}$ к a подразумевает существование некоторого малого числа ε, такого, что

$$|\overline{x} - a| < \varepsilon. \tag{5.1}$$

Так как $\overline{x}$ является случайной величиной, $|\overline{x} - a|$ 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 вероятность выполнения неравенства (5.1) надо найти распределение оценки $\overline{x}$.

Из свойств нормального распределения, приведенных в п. 5.1, легко следует, что $\overline{x}$ также имеет нормальное распределение. При этом

$$M\overline{x} = a, \qquad D\overline{x} = D\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2} D\sum_{i=1}^{n} x_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Чтобы найти вероятность выполнения неравенства (5.1), рассмотрим величину $\eta = \sqrt{n}\,(\overline{x} - a)/\sigma$. 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.

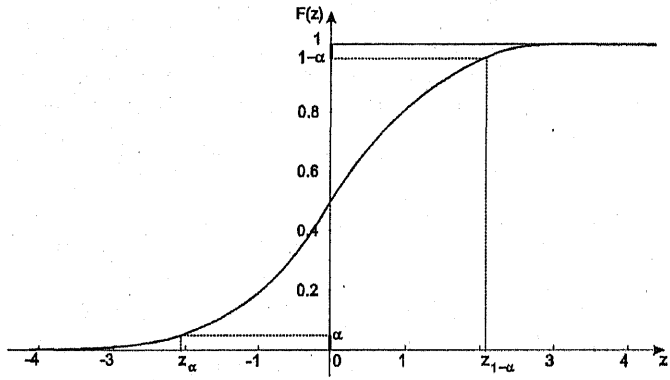


Рис. 5.2. Квантили стандартного нормального распределения

Заменяя η выражением $\sqrt{n}\,\frac{(\overline{x}-a)}{\sigma}$, находим, что

$$P\left(\left|\sqrt{n}\,\frac{(\overline{x}-a)}{\sigma}\right| < z_{1-\alpha}\right) = 1 - 2\alpha$$

или

$$P\left(|\overline{x}-a|<\frac{\sigma}{\sqrt{n}}z_{1-\alpha}\right)=1-2\alpha.$$

Это означает, что с вероятностью $1-2\alpha$ точность приближения $\overline{x}$ 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

$$|\overline{x}-a|<\frac{\sigma}{\sqrt{n}}z_{1-\alpha}$$

представляет собой интервал $\left(\overline{x}-\frac{\sigma}{\sqrt{n}}z_{1-\alpha},\ \overline{x}+\frac{\sigma}{\sqrt{n}}z_{1-\alpha}\right)$ с центром $\overline{x}$ 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 ($z_{1-\alpha}\to\infty$ при $\alpha\to 0$). 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 $\eta=\sqrt{n}(\overline{x}-a)/\sigma$ 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

$$t = \sqrt{n}\,\frac{(\overline{x} - a)}{s}.$$

Ее часто называют стьюдентовской дробью, или стьюдентовским 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 $\overline{x} = a + \sigma\overline{\xi}$, а

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\sigma^2\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2. \tag{5.2}$$

Поэтому

$$t = \sqrt{n}\,\frac{(\overline{x} - a)}{s} = \frac{\sqrt{n}\,\sigma\xi}{\frac{\sigma}{\sqrt{n-1}}\sqrt{\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2}} = \sqrt{n}\,\frac{\overline{\xi}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2}}.$$

Видно, что t является функцией от стандартно распределенных величин $\xi_1, \ldots, \xi_n$ 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

$$P\left(\sqrt{n}\left|\frac{\overline{x} - a}{s}\right| < t_{1-\alpha}\right) = 1 - 2\alpha,$$

или

$$P\left(|\,\overline{x} - a\,| < \frac{s}{\sqrt{n}}\,t_{1-\alpha}\right) = 1 - 2\alpha.$$

Как и в случае с известной дисперсией σ^2, последние соотношения характеризуют точность приближения $\overline{x}$ 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 $\left(\overline{x}-\frac{s}{\sqrt{n}}t_{1-\alpha},\ \overline{x}+\frac{s}{\sqrt{n}}t_{1-\alpha}\right)$ с центром $\overline{x}$ 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

$$s^2 = \frac{\sigma^2}{n-1}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2.$$

Заметим, что каждая случайная величина $\xi_i - \overline{\xi}$ 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, $i = 1, \ldots, n$, 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 $\xi_1 - \overline{\xi}, \xi_2 - \overline{\xi}, \ldots, \xi_n - \overline{\xi}$ не являются независимыми. Действительно, в каждое выражение

$$\xi_j - \overline{\xi} = \xi_j - \frac{1}{n}(\xi_1 + \cdots + \xi_i + \cdots + \xi_n)$$

входят остальные случайные величины.

Но все же оказывается, что сумму $\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2$ можно представить в виде суммы независимых квадратов $\sum_{i=1}^{n-1}\eta_i^2$, где η_i $(i = 1, \ldots, n-1)$ — независимые случайные величины с распределением $N(0,1)$. Таким образом получается, что величина $\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2$ 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^2_α и $\chi^2_{1-\alpha}$ так, что

$$P(\chi^2_\alpha < \chi^2 < \chi^2_{1-\alpha}) = 1 - 2\alpha.$$

(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:

$$\frac{s^2(n-1)}{\sigma^2} = \sum_{i=1}^{n}(\xi_i - \overline{\xi})^2.$$

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 $1-2\alpha$

$$P\left\{\chi^2_\alpha < \frac{s^2(n-1)}{\sigma^2} < \chi^2_{1-\alpha}\right\} = 1-2\alpha,$$

или

$$P\left\{\frac{1}{n-1}\chi^2_\alpha < \frac{s^2}{\sigma^2} < \frac{1}{n-1}\chi^2_{1-\alpha}\right\} = 1-2\alpha.$$

Этому утверждению часто придают другую форму, тождественно преобразовав неравенство в скобках:

$$P\left\{s^2\frac{(n-1)}{\chi^2_{1-\alpha}} < \sigma^2 < s^2\frac{(n-1)}{\chi^2_\alpha}\right\} = 1-2\alpha.$$

Таким образом, доверительный интервал для дисперсии имеет вид:

$$\left(s^2\frac{(n-1)}{\chi^2_{1-\alpha}},\ s^2\frac{(n-1)}{\chi^2_\alpha}\right).$$

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 $x_1, \ldots, x_n$ — выборка из нормального распределения $N(a, \sigma^2)$ с некими 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:

$$\eta = \sqrt{n}\frac{(\overline{x} - a_0)}{\sigma}.$$

(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

$$\sqrt{n}\left|\frac{(\overline{x} - a_0)}{\sigma}\right| < z_{1-\alpha/2};$$

- 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

$$t = \sqrt{n}\frac{\overline{x} - a_0}{s}.$$

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

$$\sqrt{n}\left|\frac{(\overline{x} - a_0)}{s}\right| < t_{1-\alpha/2},$$

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 $x_1, \ldots, x_n$; $y_1, \ldots, y_m$ — 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 $y_1, \ldots, y_m$ в выборку $y_1 + \Delta, \ldots, y_m + \Delta$.

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 $x_1, \ldots, x_n$ и $y_1, \ldots, y_m$ перейдем к выборочным средним $\overline{x}$ и $\overline{y}$. Согласно свойствам нормального распределения и выдвинутой гипотезе, величины $\overline{x}$ и $\overline{y}$ 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 $\overline{x}$, $\overline{y}$ 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) $\frac{\overline{x}-\overline{y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}$. Статистика имеет стандартное нормальное распределение, так как является линейной комбинацией независимых нормальных величин. Гипотеза H принимается на уровне значимости α, если

$$\left| \frac{\overline{x} - \overline{y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \right| < z_{1-\alpha/2};$$

в противном случае гипотеза отвергается в пользу альтернативы $a_1 \neq a_2$.

б) $\frac{\overline{x}-\overline{y}}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}}$. 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

$$\frac{\overline{x}-\overline{y}}{s\sqrt{\frac{1}{n}+\frac{1}{m}}},$$

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

$$\frac{\overline{x}-\overline{y}}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{m}}}.$$

В этой ситуации указать точное распределение введенной статистики затруднительно. Известно, однако, что это распределение близко к распределению Стьюдента с числом степеней свободы, равным

$$\frac{\left(s_1^2/n + s_2^2/m\right)^2}{\frac{\left(s_1^2/n\right)^2}{n-1} + \frac{\left(s_2^2/m\right)^2}{m-1}}.$$

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 v (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 [16].

Критерий Фишера. Кратко остановимся на вопросе проверки гипотезы о равенстве дисперсий двух нормальных выборок.

Рассмотрим отношение оценок дисперсий первой и второй выборки 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

$$F_{\alpha/2,\,n-1,\,m-1} \leqslant F \leqslant F_{1-\alpha/2,\,n-1,\,m-1};$$

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 $\frac{\sigma_1^2}{\sigma_2^2}$. Если единица (значение отношения σ_1^2/σ_2^2 при гипотезе) принадлежит доверительному интервалу для $\frac{\sigma_1^2}{\sigma_2^2}$, то гипотеза принимается. В противном случае она отвергается.

5.4.3. Парные данные

В пункте 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, $i = 1, \ldots, n$, 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 пункте 3.6.

Данные — совокупность пар случайных величин (x_1, y_1), ..., (x_n, y_n), где n — объем совокупности (число пар). Обозначим $z_i = y_i - x_i$, $i = 1, \ldots, n$.

Допущения (частично повторяют допущения пункта 3.6, а частично их усиливают).

1. Все z_i, $i = 1, \ldots, n$ — взаимно независимы.

2. Предположим, что z_i можно представить в виде:

$$z_i = \theta + e_i,$$

где $e_1, \ldots, e_n$ — 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 $e_1, \ldots, e_n$ распределены по нормальному закону $N(0, \sigma^2)$, где дисперсия σ^2 обычно неизвестна. Это предположение дополняет и усиливает перечень свойств случайных величин $e_1, \ldots, e_n$, принятый в пункте 3.6.2.

Приняв эти допущения, мы свели задачу о парных данных к задаче об одной нормальной выборке, уже рассмотренной в пункте 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 пункте 5.4.1 об одной нормальной выборке, так что нет нужды повторяться.

Пример (продолжение примера из пункта 3.6). Проиллюстрируем описанный метод на примере сравнения времени реакции на звук и на свет, рассмотренном в пункте 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.)

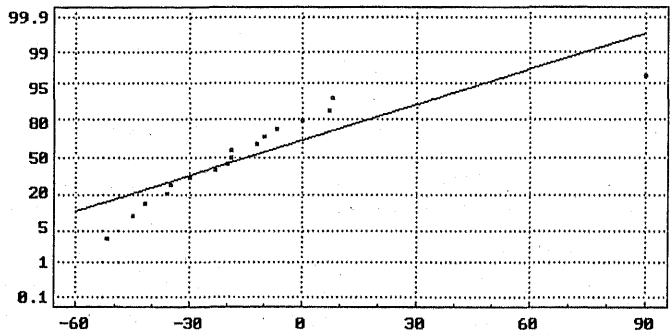


Рис. 5.3. Эмпирическая функция распределения на нормальной вероятностной бумаге

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: $\overline{z} = -14.4$, $s^2 = 1033.7.6$, $s = 32.15$.

Для проверки гипотезы $H : \theta = 0$ составляем отношение Стьюдента:

$$t = \frac{\overline{z} - \theta}{s}\sqrt{n} = -\frac{14.4}{32.15}\sqrt{17} = -1.846.$$

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, то есть 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 $\overline{z}$ и s изменяется следующим образом:

$$\overline{z} \simeq -22, \qquad s^2 \simeq 331.6, \qquad s \simeq 18.2, \qquad t \simeq -\frac{22}{18.2}\sqrt{16} \simeq -4.835.$$

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 [16], [60], [67]. 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 пакетах STADIA и STATGRAPHICS

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.

5.5.1. Пакет STADIA

В пакете STADIA (в отличие от STATGRAPHICS, см. п. 5.5.2) отсутствует возможность глазомерной проверки нормальности с помощью графика на нормальной вероятностной бумаге. Проверку нормальности в пакете осуществляет другая процедура (Гистограмма/нормальность), сравнивающая гистограмму частот с графиком подобранной плотности распределения. Работа этой процедуры будет рассмотрена в главе 10 при обсуждении критерия согласия хи-квадрат.

Пример 5.2к. Построим 95% доверительные интервалы для среднего значения и дисперсии по выборке диаметров головок заклепок (табл. 1.1) и проверим гипотезу о равенстве среднего значения выборки заданной величине 13.4.

Решение этой задачи в пакете осуществляет процедура 1=Описательная статистика из меню Статистические методы. Ее работа была подробно рассмотрена в примере 1.1к. Экран выдачи результатов этой процедуры для данных диаметров головок заклепок приведен на рис. 1.18.

Для получения левого конца доверительного интервала для среднего следует вычесть из полученной оценки для среднего 13.421 величину Доверит. ср., то есть 0.018499. Для получения правого конца доверительного интервала для среднего следует прибавить к среднему указанную выше величину.

В пакете отсутствует процедура, в явном виде реализующая критерий Стьюдента для проверки гипотезы о равенстве среднего значения нормально распределенной выборки заданному числу. Для решения этой задачи при уровне значимости $\alpha = 0.05$ против двусторонних альтернатив следует посмотреть, попадает ли гипотетическое значение 13.4 в полученный доверительный интервал для среднего. В данном случае гипотетическое значение лежит правее нижней границы 95% доверительного интервала, которая равна 13.4015. Поэтому гипотезу $H : a = 13.4$ следует отвергнуть на указанном уровне значимости 0.05.

Пример 5.3к. 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.

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								

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. Поместим наблюдения из таблиц 5.1 и 5.2 в переменные girl и boy электронной таблицы пакета (см. рис. 5.4).

Выбор процедуры. В меню Статистические методы (рис. 1.17) выберем пункт 4 = Стьюдента и Фишера.

Заполнение полей ввода данных. На экране появится окно Анализ переменных. С помощью мыши выделим в левом поле этого окна имена переменных girl и boy. Нажав кнопку со стрелкой вправо, перенесем их в правое поле и нажмем кнопку запроса [Утвердить].

Результаты. На рис. 5.5 приведены значения статистик Фишера и Стьюдента для проверки гипотез о равенстве дисперсий и средних значений двух нормальных выборок. Указаны их минимальные уровни значимости и числа степеней свободы соответствующих распределений.

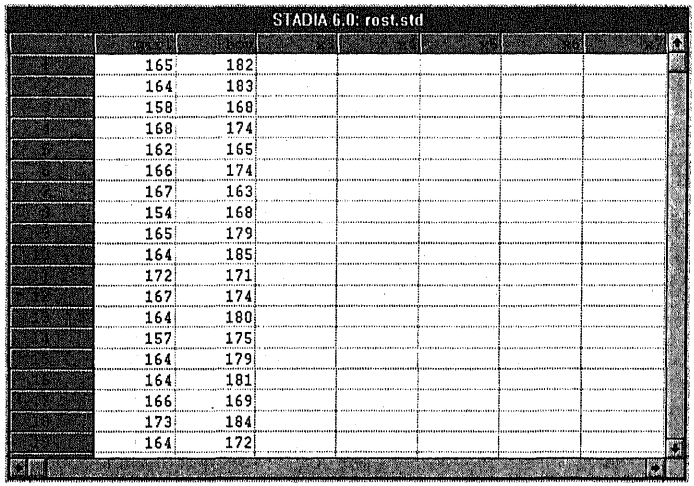


Рис. 5.4. Электронная таблица с данными из табл. 5.1 и 5.2

В случае совпадения объемов анализируемых данных выводится значение статистики Стьюдента для парных наблюдений (см. п. 5.4.3). Для каждого из перечисленных критериев выводится заключение системы о принятии или отвержении гипотезы на уровне значимости $\alpha = 0.05$ против двусторонних альтернатив.

На основании полученных результатов можно заключить, что имеется различие между средними значениями анализируемых выборок. Имеющиеся данные не противоречат гипотезе о равенстве дисперсий выборок.

```
КРИТЕРИЙ ФИШЕРА И СТЬЮДЕНТА.  Файл:

                              Переменные: girl, boy
Статистика Фишера=0.64116, Значимость=0.1039, степ.своб = 19,52
   Гипотеза 0: <Нет различий между выборочными дисперсиями>

Статистика Стьюдента=6.6794, Значимость=0, степ.своб = 71
   Гипотеза 1: <Есть различия между выборочными средними>
```

Рис. 5.5. Результаты проверки различия между средними и дисперсиями выборок

Комментарии. Процедура указывает минимальные уровни значимости критериев против двусторонних альтернатив. Чтобы получить значения минимальных уровней значимости для критерия Стьюдента против односторонних альтернатив, следует разделить значение минимального уровня значимости против двусторонних альтернатив пополам.

5.5.2. Пакет STATGRAPHICS

Часть основных процедур для анализа нормальных выборок собрана в разделе G. Estimation and Testing (оценивание и тестирование) головного

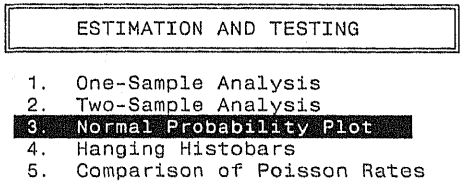


Рис. 5.6. Меню процедур оценивания и тестирования

меню пакета. Меню этого пункта приведено на рис. 5.6. Укажем назначение входящих в него процедур.

1. One-Sample Analysis (анализ одной выборки) — осуществляет построение доверительных интервалов для среднего и дисперсии нормальных выборок, а также проверку гипотез о возможных значениях этих параметров. Работа данной процедуры разобрана в примере 5.2к.

2. Two-Sample Analysis (анализ двух выборок) — вычисляет доверительные интервалы для разности средних значений и отношения дисперсий нормальных выборок, проверяет гипотезу о равенстве средних этих выборок. Данной процедуре посвящен пример 5.3к.

3. Normal Probability Plot (график на нормальной вероятностной бумаге) — строит график эмпирической функции распределения на нормальной вероятностной бумаге. Процедура используется в примере 5.1к.

4. Hanging Histobars («висячие гистобары») — сравнивает гистограмму частот с подобранным графиком плотности нормального распределения.

5. Comparison of Poisson Rates (сравнение интенсивности пуассоновских потоков событий) — сравнивает числа событий, наблюденных в двух независимых периодах времени, и проверяет гипотезу о равенстве интенсивностей наступления рассматриваемых событий в указанные периоды времени. Эта процедура не имеет отношения к анализу нормальных выборок.

Пример 5.1к. С помощью графика на нормальной вероятностной бумаге проверим нормальность распределения данных для выборки диаметров головок заклепок (табл. 1.1).

Подготовка данных. Поместим данные табл. 1.1 в переменную DIAMZ.d базы данных пакета (см. пример 1.1к).

Выбор процедуры. В головном меню пакета выберем пункт G. Estimation and Testing, а в этом пункте (рис. 5.6) — процедуру 3. Normal Probability Plot.

Заполнение полей ввода данных сводится к указанию числового вектора в поле ввода Data vector (вектор данных), как это показано на рис. 5.7.

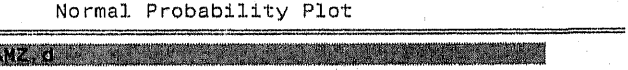


Рис. 5.7. Запрос данных процедуры вывода графика распределения на нормальной вероятностной бумаге

Результаты. После заполнения поля ввода надо нажать клавишу F6, и на экране появятся точки скачков эмпирической функции распределения, построенные на нормальной вероятностной бумаге, а также линия простой линейной регрессии (см. главу 8) для этих данных (рис. 5.8а).

Обсуждение. По замыслу разработчиков процедуры, подгонка с помощью простой линейной регрессии должна помогать выявлять отклонение построенных точек от прямой. На самом деле, эта прямая иногда может и мешать. Во всяком случае, это не та прямая линия, которую проводим мы с вами, чтобы приблизить ею график ступенчатой ломаной. Глазомерный анализ эмпирической функции распределения на нормальной вероятностной бумаге для рассматриваемых данных показывает, что они довольно хорошо ложатся на прямую — за исключением одной точки. Эта точка соответствует максимальному значению выборки, которое, скорее всего, является грубой ошибкой, «выбросом». Происхождение этого наблюдения может быть самым различным: сбой в работе измерительной аппаратуры, ошибка при записи или переписывании выборочных значений, попадание в партию заклепок заклепки другого диаметра и т.п. Присутствие этого наблюдения в выборке будет тем или иным образом сказываться на результатах расчетов. В следующих примерах, а также в примерах главы 10 мы будем отмечать и исследовать влияние таких «исключительных наблюдений» (выбросов) на статистические выводы. Например, мы сравним результаты, получаемые в присутствии выбросов и после их удаления (иначе говоря, после «цензуры», или «цензурирования» исходных данных). Для разбираемого примера это означает удаление из выборки ее максимального элемента, далеко отстоящего от остальных. Именно это наблюдение отклонило подобранную линию регрессии на рис. 5.8а. На рис. 5.8б приведены результаты работы процедуры для цензурированных данных.

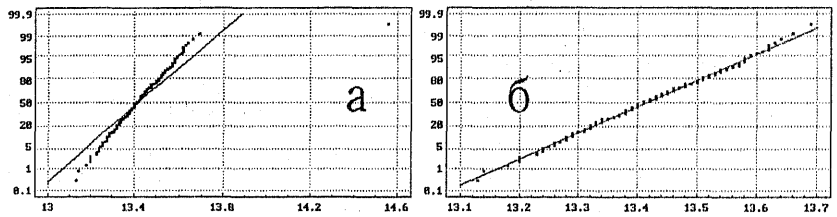


Рис. 5.8. График скачков эмпирической функции распределения на нормальной вероятностной бумаге: а — исходные данные, б — цензурированные данные

Пример 5.2к. 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. Смотри примеры 1.1к. и 5.1к.

Выбор процедуры. В головном меню пакета выберем пункт G. Estimation and Testing, а в нем (рис. 5.6) — процедуру 1. One-Sample Analysis.

Заполнение полей ввода данных. Экран ввода данных и порядок его заполнения такой же как в процедуре 3. Normal Probability Plot (см. пример 5.1к.)

Результаты. Введя вектор данных, надо нажать клавишу (F6). На рис. 5.9 приведен экран результатов, выдаваемых процедурой.

```
                One-Sample Analysis Results
═══════════════════════════════════════════════════════════
                                        DIAMZ.d
Sample Statistics:  Number of Obs.      200
                    Average             13.4215
                    Variance            0.0180761
                    Std. Deviation      0.134448
                    Median              13.41

Confidence Interval for Mean:           95    Percent
                    Sample 1            13.4027 13.4403      199 D.F.

Confidence Interval for Variance:       95    Percent
                    Sample 1            0.0149905 0.022229   199 D.F.

Hypothesis Test for H0: Mean= 13.4  Computed t statistic=2.26152
                       vs Alt: NE   Sig. Level=0.0248083
                     at Alpha= 0.05 so reject H0.
```

Рис. 5.9. Результаты анализа выборки из табл. 1.1

Поля ввода указанного экрана

Confidence Interval for Mean: 95 Percent
Confidence Interval for Variance: 95 Percent

позволяют задать уровень доверия для построения доверительных интервалов для среднего значения и дисперсии. Под этими полями процедура выдает границы построенных доверительных интервалов.

Экран выдачи результатов также включает:

Number of Obs. — число наблюдений в анализируемой выборке;
Average — среднее значение;
Variance — дисперсию;
Std. Deviation — стандартное отклонение;
Median — медиану.

Для проверки гипотезы о равенстве среднего значения выборки заданному числу необходимо ввести это число в поле ввода Hypothesis Test for H0: Mean=, а также указать тип альтернатив в поле vs Alt (NE — двусторонние альтернативы, LT — левосторонние, GT — правосторонние). В поле at Alpha= указывается уровень значимости критерия.

В результате проверки процедура выдает значение t-статистики Стьюдента (Computed t statistic=2.26152), ее минимальный уровень значимости Sig. Level=0.0248083 и заключение об отвержении (so reject H0, как в данном случае) или о принятии гипотезы.

```
                    One-Sample Analysis Results

                                          DIAMZ.d1
Sample Statistics:  Number of Obs.        199
                    Average               13.4158
                    Variance              0.0115882
                    Std. Deviation        0.107648
                    Median                13.41

Confidence Interval for Mean:                 95    Percent
                    Sample 1            13.4007 13.4308     198 D.F.

Confidence Interval for Variance:             95    Percent
                    Sample 1         9.60571E-3 0.0142583 198 D.F.

Hypothesis Test for H0: Mean= 13.4   Computed t statistic=2.06774
                        vs Alt:      NE    Sig. Level=0.0399645
                     at Alpha =      0.05  so reject H0.
```

Рис. 5.10. Результаты анализа подвергнутой «цензурированию» выборки из табл. 1.1

Аналогичным образом мы можем провести анализ для «цензурированных» данных (см. пример 5.1к). Результаты вычислений показаны на рис. 5.10.

Сравнение полученных результатов показывает, что влияние «выброса» на среднее значение и длину его доверительного интервала незначительно. Отчасти это связано с относительно большим объемом выборки. Выборочная дисперсия более чувствительна к присутствию выбросов.

Пример 5.3к. Проведем анализ однородности двух нормальных выборок для данных о росте девушек и юношей. (Описание этих данных дано в пункте 5.4.1.) Проверим гипотезу о равенстве их средних значений и дисперсий.

Подготовка данных. Поместим данные столбцов табл. 5.1 и 5.2 в переменные LENGTH.girl и LENGTH.boy соответственно. Экран редактора базы данных пакета с частью введенных данных показан на рис. 5.11.

Выбор процедуры. В головном меню пакета выберем пункт G. Estimation and Testing (рис. 5.6), а в нем — процедуру 2. Two-Sample Analysis.

Заполнение полей ввода данных. Экран ввода данных предполагает указание двух анализируемых переменных (рис. 5.12).

```
                        Two-Sample Analysis

Sample 1:  LENGTH.girl
Sample 2:  LENGTH.boy
```

Рис. 5.12. Запрос данных процедуры сравнения двух выборок

Результаты. Введя вектора данных, следует нажать клавишу F6. На рис. 5.13 приведен экран результатов вычислений процедуры.

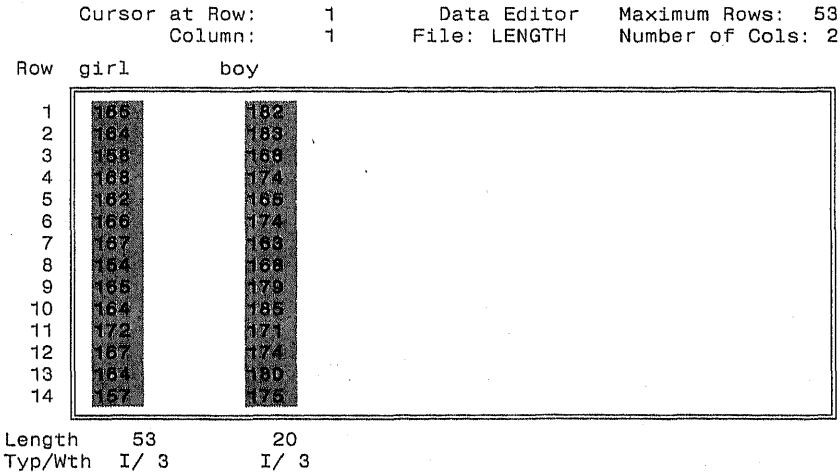


Рис. 5.11. Экран редактора данных с данными из примера 5.3к

Two-Sample Analysis Results

		LENGTH.girl	LENGTH.boy	Pooled
Sample Statistics:	Number of Obs.	53	20	73
	Average	164.226	175	167.178
	Variance	26.7939	41.7895	30.8068
	Std. Deviation	5.17628	6.46448	5.55039
	Median	164	174	165

Difference between Means = -10.7736

Conf. Interval For Diff. in Means:	95 Percent			
(Equal Vars.)	Sample 1 - Sample 2	-13.6786	-7.86861	71 D.F.
(Unequal Vars.)	Sample 1 - Sample 2	-14.0706	-7.4766	28.7 D.F.

Ratio of Variances = 0.641164

Conf. Interval for Ratio of Variances:	95 Percent				
	Sample 1 - Sample 2	0.279953	1.28335	52	19 D.F.

Hypothesis Test for H0: Diff = 0 Computed t statistic = -7.39654
vs Alt: LT Sig. Level = 1.02464E-7
at Alpha = 0.05 so reject H0.

Рис. 5.13. Результаты сравнения двух выборок

Информацию экрана выдачи результатов процедуры можно разбить на четыре блока. В первом (Sample Statistics) представлены основные выборочные характеристики (число наблюдений, среднее значение, дисперсия, стандартное отклонение и медиана) каждой из переменных и, в столбце Pooled, объединенной совокупности.

Во втором блоке даны характеристики доверительных интервалов для разности средних значений совокупности. Требуемый уровень доверия (в процентах) задается в строке

Conf. Interval For Diff. in Means: 95 Percent .

При этом в строке

(Equal Vars.) Sample 1 – Sample 2 –13.6786 –7.86861 71 D.F.

приводятся границы доверительного интервала для разности средних в предположении о равенстве неизвестных дисперсий выборок. Величина 71 D.F. означает число степеней свободы распределения Стьюдента, процентная точка которого используется для построения доверительного интервала. В строке

(Unequal Vars.) Sample 1 – Sample 2 –14.0706 –7.4766 28.7 D.F.

границы доверительного интервала для разности средних даны в предположении несовпадения неизвестных дисперсий. Величина 28.7 D.F. означает число степеней свободы распределения Стьюдента, процентная точка которого используется для приближенного построения доверительного интервала.

Обратим внимание, что оба 95% доверительных интервала не включают значение 0, то есть гипотеза о равенстве средних значений при этом уровне значимости может быть отвергнута.

Третий блок экрана вывода результатов процедуры рассматривает частное от деления (Ratio of Variances) дисперсии первой и второй выборки. Здесь приводятся границы доверительного интервала для частного с заданной степенью доверия (в процентах, в данном случае 95%). Заметим, что полученный доверительный интервал (0.279953, 1.28335) включает значение 1, то есть гипотеза о равенстве дисперсий на выбранном уровне значимости не может быть отвергнута.

Последний блок экрана вывода результатов позволяет проводить проверку гипотез о равенстве разности средних значений выборок заданному числу против различных альтернатив. Работа с этим блоком осуществляется так же, как в процедуре 1. One-Sample Analysis и описана в примере 5.2к. Для данных нашего примера видно, что гипотеза о равенстве средних значений выборок против левосторонних альтернатив на уровне значимости $\alpha = 0.05$ должна быть отвергнута.

Глава 6

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}
	⋮	⋮		⋮
	$x_{n_1 1}$	$x_{n_2 2}$	...	$x_{n_k k}$

Здесь $n_1, \ldots, n_k$ — объемы выборок, $N = n_1 + n_2 + \ldots + n_k$ — 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 главе 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. пп. 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 вли-

яния двух или нескольких способов обработки между собой, то есть 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 на сколько она больше или меньше для разных способов обработки почвы.

Но прежде чем судить о количественном влиянии фактора на измеряемый признак, полезно спросить себя, есть ли такое влияние вообще. Нельзя ли объяснить расхождения наблюденных 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, опирающееся 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 неучтенного нами фактора. Например, при изучении влияния

способов обработки почвы на урожайность таким фактором может быть тип почвы. В главе 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 не значимы, 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. Мы покажем, как это делается, в п. 6.7.2. А при применении методов, основанных на предположении о нормальности распределения данных, может быть проведено исследование нормальности остатков (то есть 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. 6.7.2.

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. главу 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 наблюденное 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 (пп 6.2.1 и 6.2.2). Построение непараметрических оценок эффектов обработки изложено в пункте 6.4. Параметрические методы (дисперсионный однофакторный анализ) описаны в пп. 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

$$R_j = \sum_{i=1}^{n_j} r_{ij} \quad \text{и} \quad R_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} r_{ij},$$

где $R_{\cdot j}$ — это средний ранг, рассчитанный по столбцу. Если между столбцами нет систематических различий, средние ранги $R_{\cdot j}$, $j = 1, \ldots, k$ не должны значительно отличаться от среднего ранга, рассчитанного по всей совокупности $||r_{ij}||$. Ясно, что последний равен $(N+1)/2$. Поэтому величины

$$\left(R_{\cdot 1} - \frac{N+1}{2}\right)^2, \ldots, \left(R_{\cdot k} - \frac{N+1}{2}\right)^2$$

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

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(R_{\cdot k} - \frac{N+1}{2}\right)^2. \tag{6.1}$$

Эта величина называется *статистикой Краскела–Уоллеса*. Множитель $12/[N(N+1)]$ нужен для стабилизации ее распределения при большом числе наблюдений (см. ниже). Другая форма для вычисления H:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1). \tag{6.2}$$

Таблицы и асимптотика. Небольшие таблицы распределения статистики H при гипотезе H_0 можно найти в сборниках статистических таблиц. При больших объемах $n_1, \ldots, n_k$, 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 [50]). Так что при использовании этого приближения мы отвергаем H_0 (на уровне значимости α), если $H_{\text{набл.}} > \chi^2_{1-\alpha}$, где $\chi^2_{1-\alpha}$ — 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':

$$H' = \frac{H}{1 - \left(\sum_{j=1}^{g} T_j / [N^3 - N]\right)}, \tag{6.3}$$

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 [91].

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. Табл. 6.1 в этом случае имеет два столбца. Фактически здесь речь идет о проверке однородности двух выборок. Напомним, что в главе 3 для решения этой задачи была предложена статистика Манна–Уитни. А именно: пусть $x_1, \ldots, x_m$ и $y_1, \ldots, y_n$ — 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 $1 \leqslant u < v \leqslant k$, 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

$$J = \sum_{1 \leqslant u < v \leqslant k} U_{u,v}.$$

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 небольшим k распределение статистики J табулировано (см., например, [91]). Для больших выборок в отношении J действует нормальная аппроксимация: $J \overset{ac}{\sim} N(MJ, DJ)$, где MJ и DJ равны:

$$MJ = \frac{1}{4}(N^2 - \sum_{j=1}^{k} n_j^2), \quad DJ = \frac{1}{72}\left[N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)\right].$$

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 [50]).

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 [26].

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–Уоллиса 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 пп. 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 $R_{\cdot j} = R_j/n_j$ 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
$R_1 = 28.5$	$R_2 = 45$	$R_3 = 62.5$	$R_4 = 80.5$	$R_5 = 117$	$R_6 = 131.5$
$R_{\cdot 1} = 5.7$	$R_{\cdot 2} = 9$	$R_{\cdot 3} = 12.5$	$R_{\cdot 4} = 16.1$	$R_{\cdot 5} = 23.4$	$R_{\cdot 6} = 26.3$

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 $n_j = 5$, $j = 1, \ldots, 6$. 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 оправдано 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:

$$
\begin{array}{llll}
U_{12} = 17 & U_{23} = 17 & U_{34} = 16.5 & U_{45} = 22 \\
U_{13} = 18.5 & U_{24} = 20.5 & U_{35} = 23,5 & U_{46} = 23,5 \\
U_{14} = 24 & U_{25} = 24.5 & U_{36} = 25 & U_{56} = 18 \\
U_{15} = 25 & U_{26} = 25 & & \\
U_{16} = 25 & & &
\end{array}
$$

Отсюда

$$
J = \sum_{\substack{u=1,\ldots,6 \\ v=1,\ldots,6 \\ u<v}} U_{u,v} = 325.
$$

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:

$$x_{ij} = a_j + e_{ij}, \qquad j = 1, \ldots, k, \quad i = 1, \ldots, n, \tag{6.4}$$

где $a_1, a_2, \ldots, a_k$ — 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 $\tau_1 + \tau_2 + \cdots + \tau_k = 0$. Тогда $x_{ij} = a_j + e_{ij}$ можно записать в виде:

$$x_{ij} = \mu + \tau_j + e_{ij}, \qquad j = 1, \ldots, k, \quad i = 1, \ldots, n.$$

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 $\tau_1, \ldots, \tau_k$. Гипотеза об однородности данных означает равенства $a_1 = a_2 = \cdots = a_k$, то есть $\tau_1 = \tau_2 = \cdots = \tau_k = 0$. Альтернатива об упорядоченности эффектов обработки превращается в $\tau_1 \leqslant \tau_2 \leqslant \cdots \leqslant \tau_k$, а различие между эффектами i-ой и j-ой 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}:

$$z_{ij} = \operatorname{med}(x_{ui} - x_{vj},\ u = 1, \ldots, n_i,\ v = 1, \ldots, n_j).$$

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

$$\overline{\Delta_i} = \frac{\sum_{u=1}^{k} n_u z_{iu}}{N}, \qquad i = 1, \ldots, k,$$

где $z_{ii} = 0$, $i = 1, \ldots, k$. $\overline{\Delta_i}$ отражает сдвиг выборки i относительно всех остальных выборок, усредненный с весами $n_1, \ldots, n_k$.

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:

$$z_{11} = 0,\ z_{12} = -2,\ z_{13} = -4,\ z_{14} = -6,\ z_{15} = -10,\ z_{16} = -13,$$
$$z_{41} = 6,\ z_{42} = 4,\ z_{43} = 2,\ z_{44} = 0,\ z_{45} = -4,\ z_{46} = -8.$$

Таким образом,

$$\overline{\Delta_1} = \frac{5}{30} - (z_{11} + z_{12} + z_{13} + z_{14} + z_{15} + z_{16}) = -5\frac{5}{6},$$

$$\overline{\Delta_4} = \frac{5}{30} - (z_{41} + z_{42} + z_{43} + z_{44} + z_{45} + z_{46}) = 0, \quad W_{14} = \overline{\Delta_1} - \overline{\Delta_4} = -5\frac{5}{6}.$$

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 $c_1, \ldots, c_k$ — 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 $j = 1, \ldots, k$, $i = 1, \ldots, k$. 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 [91].

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 $e_{ij} \sim N(0, \sigma^2)$, то есть 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 $H_0 : a_1 = \cdots = a_k$. 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

$$x_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}, \qquad j = 1, \ldots, k,$$

и далее вычислим $\sum_{i=1}^{n_j} (x_{ij} - x_{\cdot j})^2$. 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 $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - x_{\cdot j})^2$ имеет распределение $\sigma^2\chi^2$ с $N - k$ степенями свободы. Отсюда получаем первую (основную) оценку σ^2:

$$\sigma^{2*} = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - x_{\cdot j})^2. \tag{6.5}$$

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 $x_{\cdot j}$. Согласно свойствам нормального распределения,

$$x_{\cdot j} \sim N(a_j, \sigma^2/n_j). \tag{6.6}$$

Кроме того, $x_{\cdot j}$ и $\sum_{i=1}^{n_j}(x_{ij} - x_{\cdot j})^2$ статистически независимы. Найдем центр совокупности (6.6) с учетом «весов» средних значений n_j, т.е. найдем, при каком z достигается минимум выражения

$$\sum_{j=1}^{k}(x_{\cdot k} - z)^2 n_j \to \min_z. \tag{6.7}$$

С помощью стандартных средств математического анализа легко видеть, что минимум (6.7) достигается при $z = \overline{x}$, где

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j}x_{ij}. \tag{6.8}$$

Заметим, что при выполнении гипотезы H_0 значение выражения (6.7) при $z = \overline{x}$ 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:

$$\sigma^{2**} = \frac{1}{k-1}\sum_{j=1}^{k} n_j(x_{\cdot j} - \overline{x})^2. \tag{6.9}$$

Поскольку, как было отмечено, случайные величины $x_{\cdot j}$ 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 $\frac{1}{k-1}\sum_{j=1}^{k} n_j(a_j - \overline{a})^2$.

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,

$$F = \frac{\frac{1}{k-1}\sum_{j=1}^{k} n_j (x_{\cdot j} - \overline{x})^2}{\frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - x_{\cdot j})^2}, \tag{6.10}$$

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 $P(F \geqslant F_{\text{набл.}})$, т.е. вероятность получить за счет действия случайности значение статистики F большее или равное $F_{\text{набл.}}$. Гипотезу H_0 следует отвергнуть, если вероятность $P(F \geqslant F_{\text{набл.}})$ – мала. К сожалению, мы не располагаем столь подробными таблицами F-распределения, в них приводятся только процентные точки. Поэтому вместо вычисления $P(F \geqslant F_{\text{набл.}})$ 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 $a_1, \ldots, a_k$. Ранее мы уже видели, что их оценками могут служить внутригрупповые средние $x_{\cdot j}$, которые имеют распределения $N(a_j, \sigma^2/n_j)$ и статистически независимы от оценки дисперсии σ^{2*} (6.5). Поэтому отношение

$$t = \frac{x_{\cdot j} - a_j}{\sigma^*}\sqrt{n_j} \tag{6.11}$$

подчиняется распределению Стьюдента с $N-k$ степенями свободы. С помощью (6.11) можно указать доверительный интервал для a_j с

произвольным коэффициентом доверия $1-2\alpha$:

$$P\left\{\left|\sqrt{n_j}\,\frac{x_{\cdot j}-a_j}{\sigma^*}\right|<t_{1-\alpha}\right\}=1-2\alpha.$$

Здесь $t_{1-\alpha}$ — квантиль уровня $(1-\alpha)$ соответствующего распределению Стьюдента. Отсюда получаем доверительный вывод об a_j (с коэффициентом доверия $1-2\alpha$):

$$|\,x_{\cdot j}-a_j\,|<\frac{\sigma^*}{\sqrt{n_j}}t_{1-\alpha}. \tag{6.12}$$

Доверительные интервалы для контрастов. Можно указать доверительный интервал также и для любой линейной комбинации $\theta=\sum_{j=1}^{k}c_ja_j$, где $c_1,\ldots,c_k$ — произвольные коэффициенты. В частности, нередко приходится обращаться к сравнениям групп попарно, т.е. к разностям a_j-a_l, $(j,l=1,\ldots,k)$. В любом случае стьюдентово отношение (с $N-k$ степенями свободы) имеет вид

$$t=\frac{\theta^*-\theta}{\sigma^*\sqrt{\sum_{j=1}^{k}c_j^2/n_j}}, \tag{6.13}$$

где $\theta^*=\sum_{j=1}^{k}c_jx_{\cdot j}$. С помощью (6.13) доверительные суждения о различных θ получаем аналогично сказанному ранее:

$$|\theta^*-\theta|<\sigma^*\sqrt{\sum_{j=1}^{k}c_j^2/n_j}\,t_{1-\alpha}. \tag{6.14}$$

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 $x_{\cdot j}$ следует, что случайная величина $\sum_{j=1}^{k}n_j(x_{\cdot j}-a_j)^2$ имеет вид $\sigma^2\chi^2(k)$, при этом она не зависит от σ^{2*}. Поэтому величина

$$F=\frac{\frac{1}{k}\sum_{j=1}^{k}n_j(x_{\cdot j}-a_j)^2}{\sigma^{2*}}$$

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

$$P\left\{\sum_{j=1}^{k} n_j(a_j - x_{\cdot j})^2 < k\sigma^{2*}F_{1-\alpha}\right\} = 1-\alpha. \tag{6.15}$$

Множество точек $a = (a_1, \ldots, a_k)$ k-мерного пространства, удовлетворяющих (6.15), образует эллипсоид с центром $a^* = (x_{\cdot 1}, \ldots, x_{\cdot k})$. Проведем к нему необходимое нам число пар параллельных касательных плоскостей. Уравнение каждой пары таких плоскостей имеет вид:

$$\sum_{j=1}^{k} c_j(a_j - x_{\cdot j}) = \pm d. \tag{6.16}$$

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 $(x_{\cdot 1}, \ldots, x_{\cdot k})$ и σ^{2*}. Истинное значение $a = (a_1, \ldots, a_k)$, 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

$$\sum_{j=1}^{k} c_j x_{\cdot j} - d < \sum_{j=1}^{k} c_j a_j < \sum_{j=1}^{k} c_j x_{\cdot j} + d$$

из выделенного выбора плоскостей (6.16). Если мы рассмотрим вообще все плоскости, многогранник превратится в эллипсоид. Остается для каждого $c = (c_1, \ldots, c_k)$ определить соответствующее d (6.16). Такое $d > 0$ есть максимальное значение выражение $\sum_{j=1}^{k} c_j(a_j - x_{\cdot j})$ при условии, что точка $a = (a_1, \ldots, a_k)$ лежит на поверхности эллипсоида, т.е. удовлетворяет соотношению $\sum_{j}^{n_j}(a_j - x_{\cdot j})^2 = k\sigma^{2*}F_{1-\alpha}$. Расчет дает

$$d(c_1, \ldots, c_k) = k\sigma^{2*}F_{1-\alpha}\sum_{j=1}^{k} c_j^2/n_j.$$

Вывод. Для любой совокупности векторов $(c_1, \ldots, c_k)$ вероятность одновременного выполнения всех неравенств

$$\left|\sum_{j=1}^{k} c_j(a_j - x_{\cdot j})\right| < \sqrt{k\sigma^{2*}F_{1-\alpha}}\sqrt{\sum_{j=1}^{k} c_j^2/n_j} \tag{6.17}$$

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).

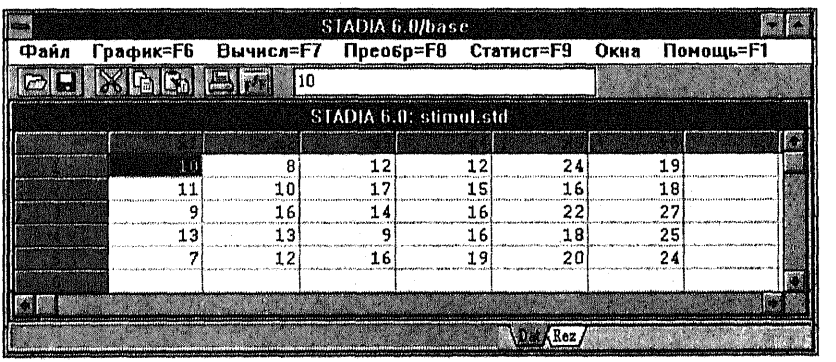
6.7. Однофакторный анализ в пакетах STADIA и STATGRAPHICS

6.7.1. Пакет STADIA

В пакете представлены следующие методы однофакторного анализа: непараметрические критерии Краскела–Уоллиса и Джонкхиера, а также методы дисперсионного анализа. Обращение к ним осуществляется из раздела Дисперсионный анализ меню Статистические методы. Проиллюстрируем использование этих методов на примерах.

Пример 6.1к. Проверим гипотезу об отсутствии эффектов обработки с помощью критерия Краскела–Уоллиса для данных о влиянии стимулирования на производительность труда (таблица 6.3).

Подготовка данных. В электронной таблице пакета введем данные первого столбца таблицы 6.3 в переменную x1, второго — в переменную x2 и так далее, как это показано на рис. 6.1.



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	

Рис. 6.1. Электронная таблица с данными для однофакторного анализа

Замечание. Процедуры однофакторного анализа пакета STADIA требуют, чтобы данные, отвечающие различным способам обработки (уровням фактора)

находились в отдельных переменных. При этом в файле данных недопустимо наличие посторонних переменных. Отсюда следует, что если мы хотим провести факторный анализ только для части способов обработки или объединить несколько способов обработки в один, следует завести новый файл данных и осуществить в нем требуемые преобразования.

Выбор процедуры. В меню Статистические методы (рис. 1.17) выберите пункт B = Однофакторный. В появившемся на экране запросе нажмите кнопку 2=Крускал-Уоллиса (можно также нажать клавишу [2]).

Результаты. Программа выдаст в окно результатов значения статистики Краскела–Уоллиса, минимального уровня значимости и числа степеней свободы распределения хи-квадрат, которое используется в качестве асимптотического приближения распределения статистики Краскела–Уоллиса. Сравнение минимального уровня значимости статистики с фиксированным уровнем значимости 0.05 позволяет системе сделать заключение «Есть влияние фактора на отклик» (рис. 6.2).

Для выполнения критерия Джонкхиера на появившийся запрос системы Значения 1-го фактора упорядочены? следует нажать кнопку [Да] (или [Yes]). На экране появятся значения статистики Джонкхиера, ее минимальный уровень значимости и заключение системы «Есть влияние фактора на отклик».

```
1-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: stimul

Краскал-Уоллис=21.219, Значимость=0.0007, степ.своб = 5
   Гипотеза 1: <Есть влияние фактора на отклик>
Джонкхиер=327, Значимость=0, степ.своб = 6, 30
   Гипотеза 1: <Есть влияние фактора на отклик>
```

Рис. 6.2. Результаты однофакторного непараметрического анализа

Комментарии. 1. Процедуры непараметрического однофакторного анализа в пакете допускают также ввод в виде таблицы рангов данных (при ранжировании по всей совокупности). То есть можно было использовать для ввода и данные таблицы 6.4.

2. В пакете STADIA (как и в STATGRAPHICS) отсутствует процедура оценки 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 меню Статистические методы (рис. 1.17) выберите пункт B = Однофакторный. В появившемся на экране запросе нажмите кнопку 1=параметрический (можно также нажать клавишу [1]).

Результаты. Программа выведет в окно результатов результаты анализа. Сначала выводятся *базовая таблица дисперсионного анализа*

и значения оценок параметров модели (рис. 6.3). Назначение базовой таблицы дисперсионного анализа — дать ответ на вопрос о наличии значимого влияния уровней факторов на исследуемый отклик, или, другими словами, о присутствии эффектов обработки.

```
1-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл:stimul
          параметрический

Источник   Сум.квадр  Ст.своб Ср.квадр Сила влияния
Факт.1        590.8        5   118.16  0.14957
Остат.        224.4       24     9.35
Общая.        815.2       29    28.11

F(фактор1)=12.637, Значимость=0, степ.своб = 5,24
   Гипотеза 1: <Есть влияние фактора на отклик>

          Параметры модели:
Среднее = 15.6, доверит.инт.=6.4652
Эффект1 = -5.6, доверит.инт.=9.4646
Эффект2 = -3.8, доверит.инт.=9.4646
Эффект3 = -2, доверит.инт.=9.4646
Эффект4 = 0, доверит.инт.=9.4646
Эффект5 = 4.4, доверит.инт.=9.4646
Эффект6 = 7, доверит.инт.=9.4646
```

Рис. 6.3. Базовая таблица дисперсионного анализа и оценки параметров модели

Таблица дисперсионного анализа. Дадим определения величин, приведенных в таблице дисперсионного анализа. Сначала рассмотрим столбец **Сум.квадр.** В строке **Общая** указана общая сумма квадратов разностей наблюдений и их среднего значения:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\overline{x})^2.$$

В строке **Факт.1** приведен вклад в общую сумму квадратов, обусловленный различиями в уровнях фактора a_j. Часто эту величину называют *суммой квадратов между группами*:

$$\sum_{j=1}^{k} n_j(x_{\cdot j}-\overline{x})^2, \tag{6.18}$$

где $x_{\cdot j}=\frac{1}{n_j}\sum_{i=1}^{n_j}x_{ij}$, а $\overline{x}$ определяется выражением (6.8).

В строке **Остат.** указан вклад в общую сумму квадратов, вызванный случайной изменчивостью данных внутри групп. Его часто называют *суммой квадратов внутри групп*:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-x_{\cdot j})^2. \tag{6.19}$$

Легко видеть, что сумма величин первой и второй строк столбца **Сумма квадр.** таблицы дисперсионного анализа (рис. 6.3) 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 указывано во втором столбце Ст.своб. таблицы). Наконец, в третьем столбце таблицы Ср.квадр. находятся частные от деления величин первого столбца на соответствующие величины второго 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 снизу от таблицы дисперсионного анализа, как раз и используется для проверки нулевой гипотезы. Справа от F-отношения указывается минимальный уровень значимости указанной F-статистики (здесь он практически равен нулю), и числа степеней свободы соответствующего F-распределения. Как обычно, если значимость F-статистики близка к нулю, есть основание отвергнуть нулевую гипотезу. Система сравнивает уровень значимости F-статистики с 0.05, и на основе этого сравнения выводит на экран заключение «Есть влияние фактора на отклик».

В четвертой строке таблицы рис. 6.3 выводится *сила влияния фактора (по Снедекору)*, т.е. величина $h_x^2 = \left(s_x^2 - s_e^2\right) / \left(s_x^2 + (n-1)s_e^2\right)$, где s_x^2 — средние квадраты между группами, s_e^2 — средние квадраты внутри групп, а величина n равна числу наблюдений в группе, если в каждой группе одинаковое число наблюдений. Если число наблюдений для каждого уровня фактора различно, в качестве n в этой формуле используют величину $n = \frac{1}{k-1}\left(N - (\sum_{j=1}^{k} n_j^2)/N\right)$, где, как обычно, N — общее число наблюдений, k — число уровней фактора, n_j — число наблюдений на уровне j фактора. Величина силы влияния показывает, какую долю вариации данных определяет модель. Для данных примера мы получили, что доля стимулирования составляет 14.9% в производительности.

Оценки параметров модели. Раздел выдачи результатов Параметры модели (рис. 6.3) включает оценку общего среднего значения и оценки отклонений от среднего для каждого уровня фактора в строках Эффект1, Эффект2 и т.д. Для каждого из этих отклонений указан размах доверительного интервала.

Парные сравнения. Вслед за описанными выше таблицами в окне результатов располагается заголовок Парные сравнения Шеффе, после которого для всех возможных пар факторов приводятся оценки разностей

влияния этих факторов, размахи доверительных интервалов и уровни значимости для гипотезы об отсутствии различий влияния этих двух факторов (рис. 6.4).

Парные сравнения Шеффе

Переменные	Разность	Интервал	Значим	Гипотеза Н1
1-2	1.8	6.9891	0.9683	
1-3	3.6	6.9891	0.6356	
1-4	5.6	6.9891	0.178	
1-5	10	6.9891	0.0022	Да
1-6	12.6	6.9891	0.0002	Да
2-3	1.8	6.9891	0.9683	
2-4	3.8	6.9891	0.5806	
2-5	8.2	6.9891	0.0143	Да
2-6	10.8	6.9891	0.001	Да
3-4	2	6.9891	0.9518	
3-5	6.4	6.9891	0.0884	
3-6	9	6.9891	0.0061	Да
4-5	4.4	6.9891	0.4204	
4-6	7	6.9891	0.0495	Да
5-6	2.6	6.9891	0.8698	

Рис. 6.4. Результаты дисперсионного анализа. Парные сравнения Шеффе

В нашем случае полученные результаты показывают, что только уровни фактора 5 и 6 значимо отличны от остальных. Поэтому целесообразно провести объединение различных уровней фактора в две группы и сравнить их между собой. Для этого пакет предлагает выделить соответствующие группы (см. рис. 6.5).

Рис. 6.5. Выбор двух групп факторов для сравнения

Выделим мышью в поле запроса Исходные переменные имена переменных x1, x2, x3, x4 и перенесем их в поле Группа 1, нажав верхнюю кнопку со стрелкой вправо. Аналогично перенесем переменные x5 и x6 в поле Группа 2. Таким образом мы сформировали две новых переменных, соответствующих различным группам уровней фактора. После нажатия кнопки Утвердить на экран будут выданы результаты сравнения двух групп уровней фактора (рис. 6.6).

Переменные	Разность	Интервал	Значим	Гипотеза Н1
x1,x2,x3,x4-x5,x6:	0.55	4.6884	0.0001	Да

Рис. 6.6. Результаты сравнения двух групп уровней фактора

Углубленный анализ. В пакете STADIA (как и в STATGRAPHICS) отсутствуют возможности непосредственной проверки правомерности применения методов дисперсионного анализа. Поэтому мы рекомендуем пользователям по крайней мере сравнивать результаты однофакторного дисперсионного анализа и критерия Краскела–Уоллиса.

6.7.2. Пакет STATGRAPHICS

В пакете довольно широко отражены стандартные методы факторного анализа. Доступ к части из них осуществляется из пункта J. Analysis of variance (анализ вариаций) раздела ANOVA AND REGRESSION ANALYSIS (АНОВА и РЕГРЕССИОННЫЙ АНАЛИЗ) головного меню. (Сокращение ANOVA происходит от выражения «Analysis of variance». В отечественной литературе вместо термина «анализ вариации» чаще используется термин «дисперсионный анализ».)

Меню пункта J. Analysis of variance приведено на рис. 6.7. Опишем кратко назначение входящих в него процедур.

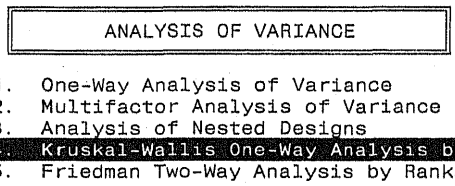


Рис. 6.7. Меню процедур дисперсионного анализа

1. One-Way Analysis of Variance (однофакторный дисперсионный анализ) — исследует эффект действия одного качественного фактора на одну переменную отклика, в предположении, что данные относится к нормальному семейству распределений с одной и той же дисперсией. Работа этой процедуры разобрана ниже в примере 6.2к.

2. Multifactor Analysis of Variance (многофакторный дисперсионный анализ) — исследует эффект действия одного или нескольких качественных факторов на одну переменную отклика, в предположении, что данные относятся к нормальному семейству распределений с одной и той же дисперсией. Данная процедура может быть использована для ковариационого анализа с одной или несколькими ковариатами. Основные элементы многофакторного дисперсионного анализа рассмотрены в главе 7 на примере двухфакторного анализа. Частичный разбор работы этой процедуры дан в примере 7.2к.

3. Analysis of Nested Designs (анализ групповых планов) — исследует эффект действия одного или нескольких качественных факторов на одну переменную отклика, если данные полностью сгруппированные или иерархические, и число наблюдений равно всем комбинациям уровней факторов. Содержание и работа этой процедуры в книге не рассматриваются.

4. Kruskal-Wallis One-Way Analysis by Ranks (ранговый однофакторный анализ Краскела–Уоллиса) — исследует эффект действия одного фактора классифи-

кации для сбалансированного или несбалансированного однофакторного плана. Эта процедура подробно разбирается ниже в примере 6.1к.

5. Friedman Two-Way Analysis by Ranks (ранговый двухфакторный анализ Фридмана) — исследует эффект действия двух факторов классификации для сбалансированного двухфакторного плана. Описание критерия Фридмана дано в главе 7. Работе процедуры посвящен пример 7.1к.

Многие другие типы планов эксперимента в пакете можно исследовать с помощью процедуры множественной регрессии (глава 8).

Разберем работу 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 (таблица 6.3).

Подготовка данных. В редакторе базы данных пакета (процедура 2. File Operations пункта A. Data Management головного меню пакета в файле STIMUL создать 6 целочисленных переменных с именами gr1, gr2, gr3, gr4, gr5, gr6, каждая из которых содержит результаты испытуемых в соответствующей группе. Вид экрана редактора базы данных с введенными данными приведен на рис. 6.8.

```
        Cursor at Row:    1     Data Editor      Maximum Rows:   5
              Column:     1     File: STIMUL     Number of Cols: 6

Row   gr1      gr2      gr3      gr4      gr5      gr6
  1    10        8       12       12       24       19
  2    11       10       17       15       16       18
  3     9       16       14       16       22       27
  4    13       13        9       16       18       25
  5     7       12       16       19       20       24
  6
  7
  8
  9
 10
Length    5        5        5        5        5        5
Typ/Wth I/ 2     I/ 2     I/ 2     I/ 2     I/ 2     I/ 2
```

Рис. 6.8. Экран редактора данных с данными для однофакторного анализа

Ниже будет показано, что такая форма ввода в виде таблицы не является обязательной.

Выбор процедуры. В меню пункта J. Analysis of variance (рис. 6.7) выберем пункт 4. Kruskal-Wallis One-Way Analysis by Ranks (ранговый однофакторный анализ Краскела–Уоллиса) и нажмем Enter.

Заполнение полей ввода данных. Экран ввода данных в эту процедуру (рис. 6.9) содержит активные поля: Data (данные), Level codes (коды уровня), Labels (метки).

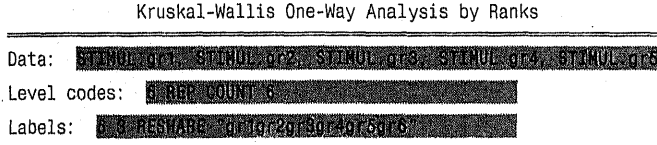


Рис. 6.9. Запрос данных процедуры однофакторного анализа Краскела–Уоллиса

Обратим внимание на особенности заполнения этих полей. В поле **Data** должен быть введен вектор данных, объединяющий все имеющиеся значения переменной отклика. При этом порядок расположения наблюдений в этом векторе, вообще говоря, не существен. В нашем случае данные, отвечающие различным способам обработки, были записаны в отдельные переменные. Объединить их в один вектор можно, введя через запятую имена переменных, в которых они хранятся (см. рис. 6.9), например:

```
STIMUL.gr1, STIMUL.gr2, STIMUL.gr3, STIMUL.gr4, STIMUL.gr5, STIMUL.gr6,
```

или просто:

```
gr1, gr2, gr3, gr4, gr5, gr6,
```

если в других файлах базы данных нет переменных с подобными именами.

Поле **Level codes** предназначено для указания, к какому способу обработки (уровню фактора) относится то или иное наблюдение в векторе данных, введенном в поле **Data**. В него необходимо ввести числовой или символьный вектор той же размерности, что и вектор данных. Этот вектор должен быть устроен следующим образом. Следует пометить элементы вектора данных, относящихся к одному и тому же способу обработки одним и тем же числом (или символом) и заменить этими числами (символами) соответствующие элементы вектора данных. Полученный вектор надо записать в поле **Level codes**. Например, пометим наблюдения, относящиеся к каждой группе испытуемых, номером их группы. Тогда вектор кодов уровня будет иметь вид:

```
1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 5 5 5 5 6 6 6 6 6.
```

Сокращенно с помощью операторов пакета этот вектор можно записать следующим выражением:

```
5 REP COUNT 6,
```

которое устроено так: сначала оператор **COUNT 6** сформирует вектор: **1 2 3 4 5 6**, а затем оператор **5 REP** повторит каждый из элементов вектора 5 раз. Таким образом будет достигнут тот же результат, что был приведен выше.

Заполнение поля **Labels** необязательно, так как значения этого поля используются лишь для оформления таблицы итоговых результатов. В

случае, если это поле оставлено пустым, пакет сам порождает метки из вектора кодов уровня. При заполнении этого поля в него следует вводить символьную матрицу так, чтобы каждая ее строка соответствовала одному и только одному значению кода уровня (способа обработки).

В нашем примере, если оставить поле Labels незаполненным, то в качестве меток уровней будут использованы числа 1, 2, 3, 4, 5, 6. Для демонстрации заполнения этого поля введем в него выражение, использующее оператор RESHAPE пакета.

```
6 3 RESHAPE 'gr1gr2gr3gr4gr5gr6',
```

которое расшифровывается так: оператор RESHAPE превратит символьный вектор, стоящий в кавычках в правой части выражения, в символьную матрицу размера 6 × 3, имеющую вид:

```
gr1
gr2
gr3
gr4
gr5
gr6
```

Каждая из строк этой матрицы будет использована при выдаче результатов для обозначения соответствующего способа обработки.

Результаты. После заполнения полей ввода и нажатия клавиши F6, на экране появляются результаты обработки (рис. 6.10).

Kruskal-Wallis analysis of STIMUL.gr1, STIMUL.gr2, STIMUL.gr3,

Level	Sample Size	Average Rank
gr1	5	5.70000
gr2	5	9.00000
gr3	5	12.5000
gr4	5	16.1000
gr5	5	23.4000
gr6	5	26.3000

Test statistic = 21.219 Significance level = 7.36387E-4

Рис. 6.10. Результаты выполнения процедуры однофакторного анализа Краскела–Уоллиса

В колонке Level (уровень) стоят метки соответствующих способов обработки, взятые из поля ввода Labels. В колонке Sample size (размер выборки) — число наблюдений для каждого способа обработки. В колонке Average Rank (средний ранг) — соответствующая величина для каждой группы. Сравните полученные значения с аналогичными, приведенными в таблице 6.4. Под таблицей приведены значения для асимптотической аппроксимации скорректированной для случая совпадающих наблюдений статистики Краскела–Уоллиса (Test statistic) и минимального уровня

значимости этой статистики (Significance level). Полученные результаты совпадают с полученными ранее вручную.

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.

3. В пакете отсутствует процедура, реализующая критерий Джонкхиера. Однако наличие в пакете критерия Манна-Уитни для сравнения двух выборок (см. главу 3) позволяет провести вычисления для получения составляющих частей $U_{u,v}$ статистики J. Но для большого числа способов обработки этот путь весьма утомителен.

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. Смотри пример 6.1к.

Выбор процедуры. В меню пункта J. Analysis of variance (рис. 6.7) выберем пункт 1. One-Way Analysis of Variance.

Заполнение полей ввода данных. Экран ввода данных и параметров этой процедуры (рис. 6.11) отличается от аналогичного экрана процедуры критерия Краскела–Уоллиса (рис. 6.9) наличием двух дополнительных полей ввода Range test (множественные сравнения) и Confidence level (уровень доверия), связанных с определением методов построения доверительных интервалов для эффектов обработки или их разностей.

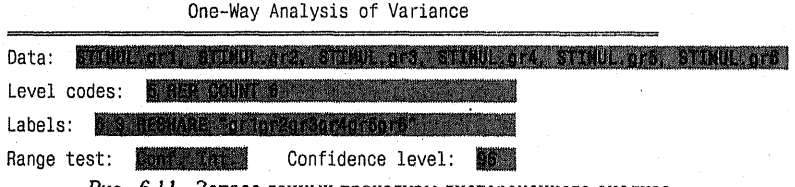


Рис. 6.11. Запрос данных процедуры дисперсионного анализа

Заполнение первых трех полей ввода (рис. 6.11) аналогично примеру 6.1к и разобрано выше. В поле Range test указать выражение Conf. Int. (доверительные интервалы для эффектов обработки). Назначение других значений в этом поле смотри в комментариях. В поле Confidence level указать значение 95, означающее выбор 95% уровня доверия. После заполнения всех необходимых полей ввода надо нажать клавишу F6.

Результаты. На экране ввода данных и параметров процедуры появится базовая **таблица дисперсионного анализа** (рис. 6.12). Ее назначение — дать ответ на вопрос о наличии значимого влияния уровней факторов на исследуемый отклик или, другими словами, о присутствии эффектов обработки. Процедуры оценивания эффектов обработки и анализа адекватности модели будут разобраны ниже.

Analysis of variance

Source of variation	Sum of Squares	d.f.	Mean square	F-ratio	Sig. level
Between groups	590.80000	5	118.16000	12.637	.0000
Within groups	224.40000	24	9.35000		
Total (corrected)	815.20000	29			

0 missing value(s) have been excluded.

Рис. 6.12. Базовая таблица дисперсионного анализа

Эта таблица аналогична таблице, выводимой пакетом STADIA (см. рис. 6.3). Дадим перевод терминов, фигурирующих в этой таблице:

Source of Variation — источник вариации;
Between groups — между группами;
Within groups — внутри групп;
Total (correct) — итого (скорректированное значение);
Sum of Squares — сумма квадратов;
d.f. — степени свободы;
Mean square — средние квадраты;
F-ratio — F-отношение;
Sig.level — уровень значимости;
0 Missing value(s) have been excluded — 0 пропущенных значений было исключено.

Теперь объясним значение величин, содержащихся в таблице рис. 6.12. В строке Between groups выводятся характеристики, связанные с действием анализируемого фактора: сумма квадратов между группами (6.18), соответствующее число степеней свободы (d.f.) и частное этих величин, т.е. оценка σ^{2**} (6.9). В строке Within groups выводятся сумма квадратов внутри групп (6.19), соответствующее число степеней свободы, и частное этих величин, т.е. оценка σ^{2*} (6.5). В строке Total (corrected) выводится сумма квадратов отклонений наблюдений от их среднего значения и число степеней свободы распределения этой величины при выполнении нулевой гипотезы. В столбце F-ratio выводится значение F-статистики (6.10), а в столбце Sig.level — ее уровень значимости. Как обычно, если эта величина близка к нулю, есть основание отвергнуть нулевую гипотезу.

Приведенное под таблицей сообщение 0 missing value(s) have been excluded свидетельствует о том, что алгоритмы однофакторного анализа пакета способны обрабатывать неполные таблицы данных, с которыми часто приходится сталкиваться на практике.

Для данных нашего примера из приведенной таблицы дисперсионного анализа (рис. 6.12) 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 оценок эффектов обработки и построение для них доверительных интервалов.

Углубленный анализ. Для продолжения анализа следует нажать клавишу [F5]. При этом на экране в накладывающемся окне появится меню процедур углубленного анализа (рис. 6.13). (Напротив каждой процедуры этого меню нами указан русский перевод.)

Means table	таблица средних
Means plot	график средних
Multiple boxplot	множественный график
Notched boxplot	график с нишами
Residual plots	график остатков
Variance check	тесты дисперсий
Multiple range tests	множественные сравнения
Save residuals	сохранить остатки
Save intervals	сохранить интервалы

Рис. 6.13. Меню процедур углубленного анализа для дисперсионного анализа

Далее можно в любом порядке выполнить все или некоторые из процедур этого меню. Они в различной форме дают информацию по двум группам вопросов: оценивание эффектов обработки и анализ адекватности применения метода дисперсионного анализа.

Получить оценки эффектов обработки и построить для них доверительные интервалы позволяет процедура Means table. Результаты ее работы приведены на рис. 6.14.

Дадим перевод терминов и определения величин, фигурирующих в полученной таблице.

Level — уровень фактора, способ обработки

Count — число наблюдений на данном уровне

Average — среднее значение на данном уровне

Stnd. Error (internal) — стандартная ошибка (внутренняя)

Stnd. Error (pooled s) — стандартная ошибка (объединенная)

95 Percent Confidence intervals for mean — 95% доверительный интервал для среднего значения

Table of means for STIMUL.gr1, STIMUL.gr2, STIMUL.gr3, STIMUL.gr4, STIMUL.gr5,

Level	Count	Average	Stnd. Error (internal)	Stnd. Error (pooled s)	95 Percent Confidence intervals for mean	
gr1	5	10.000000	1.0000000	1.3674794	7.176989	12.823011
gr2	5	11.800000	1.3564660	1.3674794	8.976989	14.623011
gr3	5	13.600000	1.4352700	1.3674794	10.776989	16.423011
gr4	5	15.600000	1.1224972	1.3674794	12.776989	18.423011
gr5	5	20.000000	1.4142136	1.3674794	17.176989	22.823011
gr6	5	22.600000	1.7492856	1.3674794	19.776989	25.423011
Total	30	15.600000	.5582711	.5582711	14.447510	16.752490

Рис. 6.14. Результаты вычисления эффектов обработки

Под внутренней стандартной ошибкой (Stnd. Error (internal)) в приведенной таблице понимается величина

$$\sqrt{\frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_{ij} - x_{\cdot j})^2}, \qquad (6.20)$$

являющаяся оценкой стандартного отклонения по группе наблюдений, в то время как под объединенной стандартной ошибкой (Stnd. Error (pooled s)) понимается квадратный корень σ^* из оценки (6.5). Еще раз заметим, что эти две оценки могут применяться вне зависимости от того, верна нулевая гипотеза или нет.

Как отмечалось в пункте 6.6.1, оценками эффектов обработки могут служить внутригрупповые средние $x_{\cdot j}$, находящиеся в третьем столбце таблицы рис. 6.14. Границы доверительных интервалов этих оценок приведены в двух последних столбцах указанной таблицы. Так как при заполнении экрана ввода параметров процедуры (рис. 6.11) в поле Range test было указано выражение Conf. Int., то для получения верхней и нижней границы доверительных интервалов использовалось выражение (6.15).

Кратко изложим назначение остальных процедур меню углубленного анализа (рис. 6.11).

Процедура Means plot реализует графическое представление данных таблицы, выдаваемой процедурой Means table (рис. 6.15)

Две следующих графических процедуры меню Multiple boxplot и Notched boxplot позволяют для каждого уровня обработки получить в компактной форме информацию о медианах, верхних и нижних квартилях, размахе и ряде других параметров. Эти процедуры дают полезные сведения о симметрии закона распределения данных на каждом уровне обработки, о возможных грубых выбросах в наблюдениях, а так же наглядно сравнивают положения основных масс данных для разных уровней фактора. Подробное описание этих описательных методов дано в [76].

Процедура Residual plots предлагает построение графика остатков в одной из трех возможных форм: в зависимости от уровня фактора, в зависимости

от предсказанных значений или просто в зависимости от индекса наблюдения в векторе ввода данных. Каждая из этих форм подчеркивает свой аспект в возможных причинах нарушения однородности распределения остатков.

С точки зрения обоснованности применения в целом метода дисперсионного анализа весьма важной является процедура Variance check. Она включает в себя результаты трех статистических критериев — Кокрена, Бартлетта и Хартли, — для сравнения разбросов наблюдений на разных уровнях фактора. Для данных примера эта процедура выдает на экран следующие результаты (рис. 6.15):

```
Tests for Homogeneity of Variances

Cochran's C test: 0.272727    P = 0.925796
Bartlett's test: 1.06547    P = 0.925721
Hartley's test: 3.06
```

Рис. 6.15. Результаты критериев проверки однородности дисперсии

При этом по первым двум критериям (Кокрена и Бартлетта), кроме значений статистик этих критериев приведены также значения минимальных уровней значимости в виде выражений P = 0.925796 и P = 0.925721. Эти значения говорят о том, что у нас нет оснований отвергнуть нулевую гипотезу этих критериев о том, что данные на разных уровнях обработки имеют одну и ту же дисперсию. Однако заметим, что критерии Кокрена и Бартлетта весьма чувствительны к отклонению от предположения, что нормированные оценки дисперсии для каждого уровня обработки подчиняются распределению хи-квадрат (с соответствующим числом степеней свободы). Это требование соблюдается, если данные для каждого уровня обработки принадлежат нормальному семейству распределений. Для других семейств распределений это требования чаще всего не соблюдается. Таким образом, в интерпретации результатов этих критериев нужна определенная осторожность. Подробную информацию о критериях Кокрена и Бартлетта можно найти в [16], [68]. Критерий Хартли описан в [102].

Процедура Multiple range tests выдает результаты анализа множественных сравнений для средних в виде, указанном на рис. 6.16:

```
Multiple range analysis for gr1,gr2,gr3,gr4,gr5,gr6 by 5 REP COUNT 6

Method: 95 Percent Confidence Intervals
Level    Count    Average      Homogeneous Groups
gr1      5        10.000000    *
gr2      5        11.800000    *
gr3      5        13.600000    *
gr4      5        15.600000    **
gr5      5        20.000000     **
gr6      5        22.600000      *
```

Рис. 6.16. Результаты анализа множественных сравнений для средних

В столбце Homogeneous Groups (однородные группы) вертикальными столбцами звездочек выделены возможные однородные группы уровней обработки. В нашем случае таких групп три. В первую из них попали способы обработки с первого по четвертый (крайний левый столбец звездочек), во вторую — четвертый и пятый способы обработки (средний столбец звездочек), и в третью группу — пятый и шестой способы обработки (крайний правый столбец звез-

дочек). Таким образом отвергнув нулевую гипотезу об отсутствии эффектов обработки в целом по всем данным, мы можем ставить вопрос об объединении в однородные группы некоторых из уровней фактора.

Перечисленные выше процедуры довольно слабо затрагивали вопрос о правомерности применения дисперсионного анализа к анализируемым данным. Для более детального рассмотрения этого вопроса пакет предлагает пользователю сохранить полученные остатки (процедура Save residuals) для дальнейшего анализа в других разделах пакета. В частности к ним далее могут быть применены критерий хи-квадрат и критерий Колмогорова-Смирнова для проверки согласия с нормальным распределением. Также можно воспользоваться глазомерным методом проверки нормальности с помощью графика остатков на нормальной вероятностной бумаге или критериями асимметрии и эксцесса. Все эти вопросы подробно рассмотрены в главах 5 и 10. Для нашего примера трудно всерьез говорить о возможности достоверной проверки нормальности выборки объема 30. Отметим только, что определенные сомнения в нормальности распределения остатков вызывает уже гистограмма остатков и их график на нормальной вероятностей бумаге.

Последняя процедура из меню рис. 6.13 Save intervals полезна как для дальнейшего составления отчета, так и для передачи этих данных в некоторые графические процедуры пакета.

Замечание. В поле Range test экрана ввода данных процедуры однофакторного анализа (рис. 6.11), кроме значения Conf.Int могут фигурировать значения Sheffe, LSD, Tukey. Они задают различные способы построения доверительных интервалов. Метод Шеффе (значение параметра Sheffe) кратко описан выше. Сведения о других методах можно найти в [76].

Глава 7

Двухфакторный анализ

7.1. Связь задач двухфакторного и однофакторного анализа

Продолжая тему исследования зависимостей, начатую в главе 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 удается 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 [32].

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 уравновешивают 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, [30], [90]).

7.2. Таблица двухфакторного анализа

Рассмотрим, как изменяется таблица однофакторного анализа, приведенная в пункте 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*, то есть 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

Таблица 7.1

	Обработки			
Блоки	1	2	...	k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
⋮	⋮	⋮		⋮
n	x_{n1}	x_{n2}	...	x_{nk}

и A сложилась 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, то есть 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 таблицы 7.1 могут содержаться несколько наблюдений). Более подробно об этом можно прочесть в [30], [90].

7.3. Аддитивная модель данных двухфакторного эксперимента при независимом действии факторов

Для описания данных таблицы 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:

$$x_{ij} = b_i + t_j + e_{ij}; \quad i = 1, \ldots, n; \quad j = 1, \ldots, k. \tag{7.1}$$

При этом числа $b_1, \ldots, b_n$ являются результатом влияния на отклик мешающего фактора B, действие которого разбивает все данные на блоки. Поэтому величины $b_1, \ldots, b_n$ называют *эффектами блоков*. Числа $t_1, \ldots, t_k$ 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 разобраны 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 $t_1, \ldots, t_k$, то есть эффектами обработки. Некоторые ослабления можно сделать и в условии независимости e_{ij} (см. например, [89]). 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:

$$x_{ij} = \mu + \beta_i + \tau_j + e_{ij}; \quad i = 1, \ldots, n; \quad j = 1, \ldots, k. \qquad (7.2)$$

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: $H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0$. 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}, $i = 1, \ldots, n$, $j = 1, \ldots, k$ 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, то есть рассматривается каждая отдельная строка таблицы 7.1 и при фиксированном индексе i осуществляется ранжирование величин x_{ij} при $j = 1, \ldots, k$. 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 $1, 2, \ldots, k$. 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 $H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0$ каждая строка рангов $r_{i1}, r_{i2}, \ldots, r_{ik}$ будет представлять случайную перестановку чисел от 1 до k, причем все $k!$ перестановок равновероятны. Введем величину: $r_{\cdot j} = \frac{1}{n}\left(\sum_{i=1}^{n} r_{ij}\right)$, являющуюся средним значением рангов по столбцу j. При гипотезе H_0 в силу равновероятности всех перестановок рангов в каждой строке значение $r_{\cdot j}$ для каждого j не должно сильно отличаться от величины $r_{\cdot\cdot} = (k+1)/2$, 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:

$$S = \frac{12n}{k(k+1)} \sum_{j=1}^{k} (r_{\cdot j} - r_{\cdot\cdot})^2. \tag{7.3}$$

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:

$$S = \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(\sum_{i=1}^{n} r_{ij} \right)^2 \right] - 3n(k+1). \tag{7.4}$$

Как отмечалось выше, при справедливости гипотезы H_0 величины $(r_{\cdot j} - r_{\cdot\cdot})^2$ 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 $H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0$ 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 $S \geqslant S(\alpha, k, n)$. Критическое значение $S(\alpha, k, n)$ находят как решение уравнения $P\{S \geqslant S(\alpha, k, n)\} = \alpha$, 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 [25] и [91]. При больших n для выбора критических значений приходится пользоваться аппроксимацией. Она основана на том факте, что при справедливости гипотезы H_0 и $n \to \infty$ статистика Фридмана S асимптотически распределена как хи-квадрат с $(k-1)$ степенями свободы (сведения о более точной аппроксимации можно найти в [50]). 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 [91].

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 ($H_0 : \tau_1 = \tau_2 = \cdots = \tau_k$) против альтернатив с упорядочением: $\tau_1 \leqslant \tau_2 \leqslant \cdots \leqslant \tau_k$, 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:

$$L = \sum_{j=1}^{k} j r_j = r_1 + 2r_2 + \cdots + k r_k. \tag{7.5}$$

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 $L \geqslant l(\alpha, k, n)$,

где функция $l(\alpha, k, n)$ удовлетворяет уравнению $P\{L \geqslant l(\alpha, k, n)\} = \alpha$.

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 [91]. В случае больших значений k и n для нахождения процентных точек следует использовать асимптотическое распределение статистики L. Рассмотрим величину L^*:

$$L^* = \frac{L - nk(k+1)^2/4}{[n(k^3-k)^2/144(k-1)]^{1/2}}. \tag{7.6}$$

При справедливости H_0 статистика L^* имеет при $n \to \infty$ асимптотическое распределение $N(0,1)$ (сведения о более точной аппроксимации можно найти в [50]). 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 таблице 7.2 приведены данные из [91]. Они являются результатом исследования зависимости частоты самопроизвольного дрожания

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

мышц рук (тремора) от тяжести специального браслета, одеваемого на запястье.

Каждое табличное значение — среднее из 5 экспериментальных измерений частоты тремора у испытуемого. Каждая обработка соответствует весу браслета, измеренного в фунтах. Перейдем от таблицы 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
$r_{\cdot j}$	5	4	2.8333	2	1.1667

В двух последних строках таблицы 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 таблицы 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 таблицы 7.3 в обратном порядке. То есть номер $j = 1$ будет соответствовать пятому столбцу таблицы 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 [91] 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 $L \geqslant l(0.001, 5, 6)$ то, следовательно, гипотеза H_0 должна быть отвергнута в пользу альтернативы $\tau_1 \leqslant \tau_2 \leqslant \cdots \leqslant \tau_k$ 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 $H_0 : \tau_1 = \tau_2 = \cdots = \tau_k$ 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 дисперсионного

анализа в пункте 6.5 главы 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 достигает минимума выражение:

$$\sum_{i,j}(x_{ij} - \mu - \beta_i - \tau_j)^2 \tag{7.7}$$

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 $\hat{\tau}_j$:

$$\hat{\mu} = x_{..} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}x_{ij} \quad \hat{\beta}_i = x_{i.} - x_{..} = \frac{1}{k}\sum_{j=1}^{k}x_{ij} - x_{..} \tag{7.8}$$

$$\hat{\tau}_j = x_{.j} - x_{..} = \frac{1}{n}\sum_{i=1}^{n}x_{ij} - x_{..}$$

Полученные оценки параметров модели имеют следующие распределения:

$$\hat{\mu} \sim N(\mu, \frac{\sigma^2}{nk}); \quad \hat{\beta}_i \sim N(\beta_i, \frac{\sigma^2(n-1)}{nk}); \quad \hat{\tau}_j \sim N(\tau_j, \frac{\sigma^2(k-1)}{nk}).$$

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:

$$\sigma^{2*} = \frac{1}{(n-1)(k-1)}\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - x_{i.} - x_{.j} + x_{..})^2 \tag{7.9}$$

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 $x_{\cdot 1}, \ldots, x_{\cdot k}$, 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: $n \sum_{j=1}^{k} (x_{\cdot j} - x_{\cdot\cdot})^2$, имеющую распределение $\sigma^2 \chi^2$ с $(k-1)$ степенями свободы. При этом сама оценка σ^{2**} есть:

$$\sigma^{2**} = \frac{n}{k-1} \sum_{j=1}^{k} (x_{\cdot j} - x_{\cdot\cdot})^2 . \tag{7.10}$$

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 $\tau_1, \tau_2, \ldots, \tau_k$.

Критерий для проверки гипотезы $H_0 : \tau_1 = \tau_2 = \cdots = \tau_k$. Составляя, так же как в гл. 6, F-отношение двух оценок дисперсий, получаем:

$$F = \frac{\frac{n}{k-1} \sum_{j=1}^{k} (x_{\cdot j} - x_{\cdot\cdot})^2}{\frac{1}{(n-1)(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - x_{i\cdot} - x_{\cdot j} + x_{\cdot\cdot})^2} .$$

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 $F \geqslant F_{1-\alpha}$;
- 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 $x_{\cdot\cdot}$ представима в виде:

$$\sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - x_{\cdot\cdot})^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - x_{i\cdot} - x_{\cdot j} + x_{\cdot\cdot})^2 +$$

$$\sum_{i=1}^{n} \sum_{j=1}^{k} (x_{i\cdot} - x_{\cdot\cdot})^2 + \sum_{i=1}^{n} \sum_{j=1}^{k} (x_{\cdot j} - x_{\cdot\cdot})^2 =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij}-x_{i\cdot}-x_{\cdot j}+x_{\cdot\cdot})^2+k\sum_{i=1}^{n}(x_{i\cdot}-x_{\cdot\cdot})^2+n\sum_{j=1}^{k}(x_{\cdot j}-x_{\cdot\cdot})^2.$$

Отсюда и идет название «дисперсионный анализ», то есть 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 $\tau_1,\ldots,\tau_n$ довольно устойчивы даже при нарушении основных предположений о нормальном распределении и о равенстве дисперсий.

7.7. Двухфакторный анализ в пакетах STADIA и STATGRAPHICS

7.7.1. Пакет STADIA

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, одеваемого на запястье (табл. 7.2).

Подготовка данных. В электронной таблице пакета введем данные первого столбца таблицы 7.2 в переменную x1, второго — в переменную x2 и т.д., как это показано на рис. 7.1.

Таблица данных

	3.11	2.85	2.62	2.63	2.58		
	3.47	3.43	3.15	2.83	2.7		
	3.35	3.14	3.02	2.71	2.78		
	3.1	2.86	2.58	2.49	2.36		
	3.41	3.32	3.08	2.96	2.67		
	3.07	3.06	2.85	2.5	2.43		

Рис. 7.1. Электронная таблица с данными для двухфакторного анализа

Процедуры непараметрического двухфакторного анализа пакета STADIA однозначно требуют, чтобы данные, отвечающие различным способам обработки (уровням фактора), находились в отдельных переменных. Число наблюдений в каждой переменной должно быть одина-

ковым, наблюдения, соответствующие одному блоку, должны стоять в одной строке. Как и в процедурах однофакторного анализа, наличие посторонних переменных в файле данных недопустимо.

Выбор процедуры. В меню Статистические методы (рис. 1.17) щелкнем мышью кнопку С = 2-факторный (можно также нажать клавишу [С]). Программа выведет запрос (рис. 7.2), в котором надо выбрать нужный метод двухфакторного анализа. Следует щелкнуть мышью кнопку 6 = Фридмана или нажать клавишу [6].

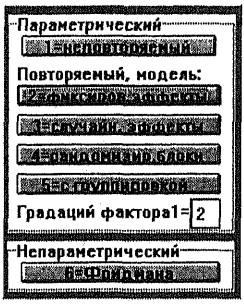


Рис. 7.2. Пакет STADIA. Запрос выбора метода двухфакторного анализа

Результаты. Программа выведет в окно результатов (рис. 7.3) значение статистики Фридмана, ее уровень значимости, вычисленный с помощью асимптотического распределения хи-квадрат, и число степеней свободы этого распределения. Сравнивая полученный уровень значимости с фиксированным (равным 0.05) система выдает сообщение о наличии или отсутствии влияния фактора на отклик.

Далее следует запрос системы «Значения 1-го фактора упорядочены?». При положительном ответе на этот вопрос (кнопка [Да] или [Yes]) программа выдает значение статистики Пейджа и нормальную аппроксимацию ее уровня значимости.

```
2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл:tremor

Фридман=22.533, Значимость=0.0001, степ.своб = 4
   Гипотеза 1: <Есть влияние фактора на отклик>
Пейдж=212, Значимость=0, степ.своб = 5,6
   Гипотеза 1: <Есть влияние фактора на отклик>
```

Рис. 7.3. Результаты применения критериев Фридмана и Пейджа

Пример 7.2к. Проведем двухфакторный дисперсионный анализ для данных примера 7.1к: проверим нулевую гипотезу об отсутствии эффектов обработки, оценим значения этих эффектов и построим для них 95% доверительные интервалы.

Подготовка данных. Мы будем использовать в качестве исходных те же данные, что и в примере 7.1к выше.

Выбор процедуры. В меню Статистические методы выберем пункт C = 2-факторный (клавиша [C]). В появившемся запросе выбора метода двухфакторного анализа (рис. 7.2) следует щелкнуть мышью кнопку 1 = неповторяемый (можно также нажать клавишу [1]).

Под повторяемым планом эксперимента в программе подразумевается план, содержащий повторные измерения при каждом сочетании значений двух исследуемых факторов. А неповторяемым планом программа называет план эксперимента, не содержащий таких повторных измерений. В нашем случае мы имеем только одно числовое значение для каждой комбинации факторов, поэтому должны выбрать неповторяемый план эксперимента.

Результаты. Экран вывода результатов этой процедуры содержит базовую таблицу дисперсионного анализа (рис. 7.4) (ее описание смотри в примере 6.2к для пакетов STADIA и STATGRAPHICS, рис. 6.3, 6.12). В нашем примере значения F-статистик для каждого из факторов F(фактор1) и F(фактор2) и их уровни значимости показывают, что имеется влияние каждого из факторов на отклик.

```
2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: tremor.std
          Факторный план: неповторяемый

Источник    Сум.квадр  Ст.своб  Ср.квадр Сила влияния
  Факт.1       1.8017        4   0.45044     0.9835
  Факт.2      0.90726        5   0.18145     0.9364
  Остат.      0.16449       20 0.0082247
  Общая.       2.8735       29  0.099086

F(фактор1)=54.767, Значимость=0, степ.своб = 4,20
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор2)=22.062, Значимость=0, степ.своб = 5, 20
   Гипотеза 1: <Есть влияние фактора на отклик>
```

Рис. 7.4. Результаты процедуры двухфакторного дисперсионного анализа

Вслед за таблицей дисперсионного анализа в окне результатов находятся оценки влияния для уровней каждого из двух факторов (рис. 7.5). Здесь приводятся оценка общего среднего μ (строка Среднее), оценки величин β_i (в строках Эффект1-1—Эффект1-5) и τ_j (в строках Эффект2-1—Эффект2-6). Оценки вычисляются по формулам (7.8).

```
          Параметры модели:
Среднее = 2.9003, доверит.инт.=0.043188
Эффект1-1 = 0.33467, доверит.инт.=0.11685
Эффект1-2 = 0.20967, доверит.инт.=0.11685
Эффект1-3 = -0.017, доверит.инт.=0.11685
Эффект1-4 = -0.21367, доверит.инт.=0.11685
Эффект1-5 = -0.31367, доверит.инт.=0.11685
Эффект2-1 = -0.16233, доверит.инт.=0.11696
Эффект2-2 = 0.21567, доверит.инт.=0.11696
Эффект2-3 = 0.099667, доверит.инт.=0.11696
Эффект2-4 = -0.22233, доверит.инт.=0.11696
Эффект2-5 = 0.18767, доверит.инт.=0.11696
Эффект2-6 = -0.11833, доверит.инт.=0.11696
```

Рис. 7.5. Оценки параметров модели в процедуре двухфакторного дисперсионного анализа

Замечания. 1. При использовании повторяемого и неповторяемого планов эксперимента исходные данные для программы STADIA надо готовить поразному. Исходные данные эксперимента без повторных измерений должны представлять собой матрицу размером $n \times k$, в которой столбцы отвечают различным способам обработки (k уровням первого фактора), строки отвечают различным блокам (n уровням второго фактора), а каждый элемент есть отклик, измеренный при соответствующем сочетании уровней исследуемых факторов. Этим требованиям в точности соответствуют рассмотренные выше данные (рис. 7.1). Для экспериментов с повторными наблюдениями в матрице данных должно быть nk переменных, в каждой из которых записаны повторные измерения для некоторого сочетания значений факторов (для сочетания факторов (i, j) измерения должны содержаться в столбце с номером $i + (k-1)j$). Значение k — число уровней первого фактора, — указывается в запросе выбора метода двухфакторного анализа.

2. Внимательный читатель может вспомнить, что в п. 7.5 говорилось о том, что на самом деле для каждого сочетания уровней факторов проводилось 5 наблюдений частоты тремора. Однако нам известны только результаты усреднения этих повторных наблюдений, а исходная информация нам не доступна. Поэтому при анализе мы должны считать наблюдениями эти известные нам средние значения повторных измерений тремора. Для каждого сочетания уровня факторов мы имеем одно такое усредненное наблюдение, поэтому на запрос программы о плане эксперимента следует указать неповторяемый план. Можно сказать, что в этой задаче на этапе сбора данных был осуществлен переход от повторяемого плана эксперимента к неповторяемому.

7.7.2. Пакет STATGRAPHICS

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, одеваемого на запястье (табл. 7.2).

Подготовка данных. В редакторе базы данных пакета (процедура 2. File Operations пункта A. Data Management головного меню пакета) в файле TREMOR следует создать 5 переменных с именами var1, var2, ..., var5 в формате с двумя десятичными знаками после точки (в поле Type описания типа переменной надо указать значение 2), каждая из которых содержит данные соответствующего столбца табл. 7.2. Вид экрана редактора базы данных с введенными наблюдениями приведен на рис. 7.6. Важным моментом ввода данных для разбираемой процедуры является то, что наблюдения, относящиеся к одному и тому же блоку (уровню второго фактора), должны иметь один и тот же порядковый номер в своей переменной, то есть образовывать строку в матрице данных.

Выбор процедуры. В головном меню пакета надо выбрать пункт J. Analysis of variance, а в меню указанного пункта (рис. 6.7) — проце-

Cursor at Row: 1 Data Editor Maximum Rows: 6
Column: 1 File: TREMOR Number of Cols: 5

Row	var1	var2	var3	var4	var5
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
7	.	.	.	.	.
8	.	.	.	.	.
9	.	.	.	.	.
10	.	.	.	.	.
Length	6	6	6	6	6
Typ/Wth	2/ 8	2/ 7	2/ 8	2/ 8	2/ 8

Рис. 7.6. Экран редактора данных с данными для двухфакторного анализа

дуру 5. Friedman Two-Way Analysis by Ranks (ранговый двухфакторный анализ Фридмана).

Заполнение полей ввода данных. Экран ввода данных процедуры приведен на рис. 7.7.

Friedman Two-Way Analysis by Ranks

Data: TREMOR.var1, TREMOR.var2, TREMOR.var3, TREMOR.var4, TREMOR.var5

Level codes: 6 REP COUNT 5

Labels:

Рис. 7.7. Запрос параметров процедуры двухфакторного анализа Фридмана

Заполнение полей ввода производится так же как в процедурах рангового однофакторного анализа Краскела-Уоллиса и однофакторного анализа (примеры 6.1к и 6.2к в п. 6.7.2).

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.*

Результаты. После заполнения полей ввода и нажатия клавиши F6 на экран выводятся результаты обработки данных в виде, указанном на рис. 7.8.

Форма представления результатов вычислений этой процедуры полностью совпадает с процедурой рангового однофакторного анализа Краскела-Уоллиса (рис. 6.10), разобранной в примере 6.1к. Для определения уровня значимости статистики Фридмана в пакете используется аппроксимация распределения этой статистики с помощью распределения хи-квадрат с числом степеней свободы $(k-1)$.

Friedman analysis of TREMOR.var1, TREMOR.var2, TREMOR.var3, TREMOR.var4, TREMOR

Level	Sample Size	Average Rank
1	6	5.00000
2	6	4.00000
3	6	2.83333
4	6	2.00000
5	6	1.16667

Test statistic = 22.5333 Significance level = 1.56919E-4

Рис. 7.8. Результаты процедуры двухфакторного анализа Фридмана

Cursor at Row: 1 Data Editor Maximum Rows: 6
Column: 1 File: TREMOR Number of Cols: 5

Row	n	var1	var2	var3	var4	var5
1	1	3.01	2.85	2.62	2.63	2.58
2	2	3.47	3.43	3.15	2.83	2.70
3	3	3.35	3.14	3.02	2.71	2.78
4	4	3.10	2.86	2.58	2.49	2.36
5	5	3.41	3.32	3.08	2.96	2.67
6	6	3.07	3.06	2.85	2.50	2.43
7		.	.	.	.	.
8		.	.	.	.	.
9		.	.	.	.	.
10		.	.	.	.	.
Length	6	6	6	6	6	6
Typ/Wth		I/1	2/ 8	2/ 7	2/ 8	2/ 8	2/ 8

Рис. 7.9. Экран редактора данных с данными для двухфакторного анализа (введена переменная n)

Комментарии. 1. В пакете отсутствует процедура, реализующая критерий Пейджа, однако значение статистики этого критерия, как было показано в пункте 7.5, легко получается из таблицы 7.3 или из аналогичной ей таблицы рис. 7.8.

2. Для получения непараметрических оценок эффектов обработки в двухфакторном анализе можно использовать процедуру **4. Median Polish of Two-Way Table** (декомпозиция двухфакторной таблицы). Ее вызов осуществляется из пункта **I. Exploratory Data Analysis** (разведочный анализ данных) головного меню пакета. Описание работы этой процедуры можно найти в [76].

Пример 7.2к. Проведем двухфакторный дисперсионный анализ для данных примера 7.1к: проверим нулевую гипотезу об отсутствии эффектов обработки, оценим значения этих эффектов и построим для них 95% доверительные интервалы.

Подготовка данных. Смотри пример 7.1к. Дополнительно в файле **TREMOR** создадим целочисленную переменную с именем **n** и введем в нее значения 1, 2, 3, 4, 5, 6. Эта переменная будет использована при заполнении экрана ввода параметров процедуры для указания уровней одного из факторов. Результат ввода данных представлен на рис. 7.9.

Выбор процедуры. В головном меню пакета выберем пункт **J. Analysis of variance**, в меню указанного пункта (рис. 6.7) — процедуру **2. Multifactor Analysis of Variance** (многофакторный дисперсионный анализ).

Заполнение полей ввода данных. Экран ввода данных процедуры приведен на рис. 7.10. Ввод данных в эту процедуру во многом схож с вводом в процедуру **One-Way Analysis of Variance** (однофакторный дисперсионный анализ), разобранную в п. 6.7. Он сводится к следующему. В поле **Data** (данные) надо ввести в виде числового вектора в произвольном порядке данные двухфакторной (многофакторной) таблицы дисперсионного анализа. Поле **Covariates** (ковариаты) оставить незаполненным. Назначение этого поля смотри в комментариях.

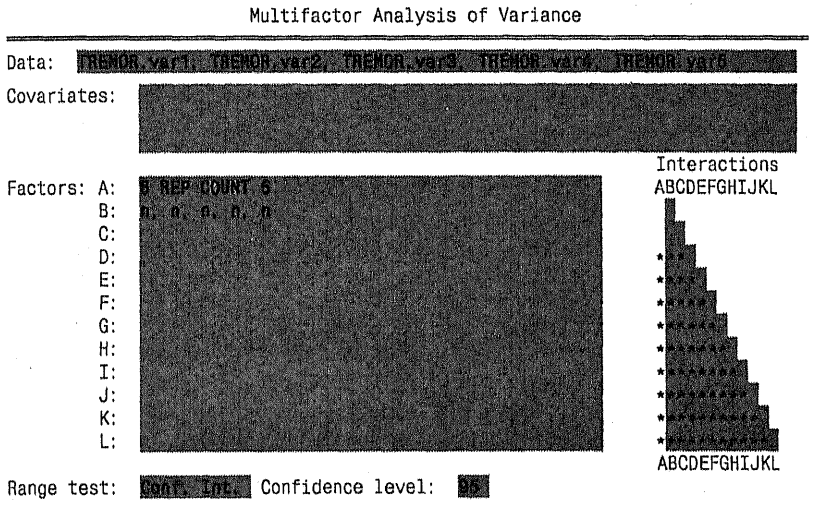


Рис. 7.10. Запрос данных процедуры многофакторного дисперсионного анализа

Заполнение поля **Factors** (факторы) аналогично заполнению этого поля в процедуре однофакторного дисперсионного анализа. При этом, как отмечалось выше, коды уровней позволяют достаточно гибко осуществлять ввод данных многофакторной таблицы. Для определенности, пусть фактор A соответствует способу обработки, а фактор B — наблюдаемому испытуемому. Обратим внимание на то, что фактор A, соответствующий весу браслета, является, вообще говоря, количественным. Однако в том случае, когда нас не интересует конкретный вес браслета или характер влияния величины веса не предсказуем, мы можем интерпретировать его как качественный. Второй фактор в этом примере — конкретный испытуемый — является по определению качественным.

В поле Factors необходимо для каждого из факторов указать его коды уровня. Учитывая порядок ввода в поле Data, коды уровней фактора A должны иметь следующую структуру:

aaaaaabbbbbbccccccddddddeeeeee

где a, b, c, d, e могут быть числами, символами или строками символьной матрицы. Это выражение показывает, что первые шесть наблюдений в векторе поля Data относятся к первому способу обработки (уровню фактора A, вторые шесть наблюдений — ко второму способу обработки и т.д. В частности, его можно задать выражением:

111111222222333333444444555555

которое кратко записывается с помощью операторов пакета в следующем виде

6 REP COUNT 5

в первой строке A поля Factors (описание этого выражения дано в примере 6.1к). Коды уровней фактора B должны иметь структуру:

abcdefabcdefabcdefabcdefabcdef

где a, b, c, d могут быть числами, символами или строками символьной матрицы. Это выражение показывает, что первое наблюдение в векторе поля Data относятся к первому блоку (уровню фактора B), второе наблюдение — ко второму блоку и т.д. В частности, его можно задать выражением:

123456123456123456123456123456

для краткой записи которого использована специально созданная нами переменная n (см. рис. 7.9).

Порядок заполнения полей Range test (множественные сравнения) и Confidence level (уровень доверия) совпадает с порядком заполнения этих полей для процедуры однофакторного анализа. Оставим в этих полях значения, предлагаемые пакетом по умолчанию, а именно: Conf. Int. и 95.

Поле Interactions (взаимодействия) позволяет дополнительно учитывать в модели (7.2) взаимодействие факторов. При рассмотрении моделей без взаимодействия факторов в этом поле следует удалить все звездочки (*) в строках рассматриваемых факторов. В строках, где коды уровней факторов не заданы, звездочки можно не удалять.

Результаты. После заполнения полей ввода и нажатия клавиши F6 на экран выводится базовая таблица дисперсионного анализа (рис. 7.11).

Описание столбцов этой таблицы было дано при разборе процедуры однофакторного анализа (пример 6.2к). В первой строке таблицы MAIN

Analysis of Variance for TREMOR.var1, TREMOR.var2, TREMOR.var3, TREMOR.var4,

Source of variation	Sum of Squares	d.f.	Mean square	F-ratio	Sig. level
MAIN EFFECTS	2.7090033	9	.3010004	36.597	.0000
6 REP COUNT 5	1.8017467	4	.4504367	54.767	.0000
n, n, n, n, n	.9072567	5	.1814513	22.062	.0000
RESIDUAL	.1644933	20	.0082247		
TOTAL (CORR.)	2.8734967	29			

0 missing values have been excluded.

Рис. 7.11. Результаты выполнения двухфакторного дисперсионного анализа

EFFECTS (главные эффекты) указана величина вариации данных, приходящаяся на рассматриваемую двухфакторную модель. Две следующие строки таблицы **6 REP COUNT 5** и **n, n, n, n, n** показывают, как вариация данных первой строки распределяется между факторами A и B. Строка **RESIDUAL** (остатки) содержит вариацию данных, не объясненную рассматриваемой моделью. Строка **TOTAL (CORR.)** (общая вариация, скорректированная на среднее значение), как и в однофакторном анализе, содержит общую вариацию данных.

В качестве **F-ratio** (F-отношения) рассматривается частное от деления **Mean square** (средних квадратов), обусловленных факторами, на средние квадраты остатков. Эти статистики стандартным образом используются для проверки гипотезы об отсутствии влияния каждого из факторов. В последнем столбце таблицы приведены значения минимальных уровней значимости соответствующих F-отношений.

Как следует из приведенной таблицы, мы должны отвергнуть гипотезу об отсутствии влияния способов обработки (фактор A). Минимальный уровень значимости F-отношения для проверки гипотезы об отсутствии влияния фактора B (третья строка таблицы) показывает, что эта гипотеза также должна быть отвергнута. Другими словами, испытуемых нельзя считать однородной группой и описывать данные с помощью однофакторной модели.

Для получения оценок эффектов обработки и их доверительных интервалов следует нажать Esc. Произойдет возврат к экрану ввода параметров процедуры (рис. 7.10), на котором появится всплывающее меню углубленного анализа (см. рис. 6.13). Назначение процедур этого меню указано при разборе однофакторного анализа в примере 6.2к в п. 6.7.2.

Необходимые в примере оценки выдает процедура **Means table**, разобранная выше. Результаты ее работы приведены на рис. 7.12.

Форма выдачи результатов этой процедуры совпадает с аналогичной процедурой однофакторного анализа (рис. 6.14) и описана в приме-

Table of means for TREMOR.var1, TREMOR.var2, TREMOR.var3, TREMOR.var4, TREMOR.v

Level	Count	Average	Stnd. Error (internal)	Stnd. Error (pooled s)	95 Percent Confidence for mean	
6 REP COUNT 5						
1	6	3.2350000	.0806536	.0370240	3.1577506	3.3122494
2	6	3.1100000	.0966092	.0370240	3.0327506	3.1872494
3	6	2.8833333	.0984773	.0370240	2.8060839	2.9605828
4	6	2.6866667	.0758361	.0370240	2.6094172	2.7639161
5	6	2.5866667	.0666166	.0370240	2.5094172	2.6639161
n, n, n, n, n						
1	5	2.7380000	.0827889	.0405578	2.6533775	2.8226225
2	5	3.1160000	.1549064	.0405578	3.0313775	3.2006225
3	5	3.0000000	.1172604	.0405578	2.9153775	3.0846225
4	5	2.6780000	.1336563	.0405578	2.5933775	2.7626225
5	5	3.0880000	.1320379	.0405578	3.0033775	3.1726225
6	5	2.7820000	.1356982	.0405578	2.6973775	2.8666225
Total	30	2.9003333	.0165576	.0165576	2.8657863	2.9348803

Рис. 7.12. Результаты вычисления эффектов обработок

ре 6.2к в п. 6.7.2. Для получения собственно оценок эффектов обработки, при записи модели в форме (7.2) следует из средних значений для каждого уровня обработки (пять первых строк таблицы) вычесть общее среднее (последняя строка таблицы).

Комментарии. 1. Процедура позволяет наряду с качественными факторами учесть влияние на отклик до трех количественных переменных, которые часто рассматриваются как мешающие. Для ввода этих переменных следует использовать поле ввода **Covariates**. Их длина должна совпадать с длиной вектора, введенного в поле **Data**.

2. При рассмотрении более сложных моделей двухфакторного анализа, учитывающих взаимовлияние факторов и (или) ковариаты, в базовой таблице дисперсионного анализа (рис. 7.11) появятся дополнительные строки, отражающие вариацию данных, которая приходится на эти эффекты.

Глава 8

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: $x_1, \ldots, x_p$ и $y_1, \ldots, y_q$. Для краткости мы объединим $x_1, \ldots, x_p$ в многомерную переменную $\mathbf{x}$, а $y_1, \ldots, y_q$ — 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 $\mathbf{x} = (x_1, \ldots, x_p)$ — *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 $\mathbf{x}$, то есть 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

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon, \tag{8.1}$$

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:

$$\mathbf{y}_i = f(\mathbf{x}_i) + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{8.2}$$

Предположения об ошибках. Разделение $\mathbf{y}_i$ 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 $\varepsilon_1, \ldots, \varepsilon_n$ статистически независимы и одинаково распределены.

В последние десятилетия активно развиваются методы, позволяющие находить решение задачи при изменении и ослаблении этих предположений (см., например, [20]).

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:

$$\mathbf{y}_i = f(\mathbf{x}_i, \theta) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{8.3}$$

и восстановление зависимости между $\mathbf{x}$ и $\mathbf{y}$ оказывается эквивалентным указанию значения θ (точнее, ее *оценки* $\hat{\theta}$) по исходным данным $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, \ldots, n$. 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, то есть $f(\mathbf{x}, \theta) = A(\mathbf{x})\theta$, 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 $f(\mathbf{x}, \theta)$ не линейна по θ, можно познакомиться в [31], или более подробно в [28].

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 $x_1, \ldots, x_n$) 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 $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, \ldots, n$ от этого не меняется.

8.2. О стратегии, 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 $x_1, x_2, \ldots, x_n$ — значения x и $y_1, y_2, \ldots, y_n$ — 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

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. 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 $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$ провести расчет (оценку) θ, то есть входящих в f неизвестных коэффициентов (параметров). Тем самым мы полностью определили подобранную регрессионную функцию:

$$y = f(x, \hat{\theta}), \qquad \text{где } \hat{\theta} \text{ — оценка } \theta.$$

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:

$$r_i = y_i - \hat{y}_i = y_i - f(x_i, \hat{\theta}), \qquad i = 1, \ldots, n$$

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 (пп. 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 [31]. 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 искажать 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 пункте 8.7.

Множественная регрессия. В более общем случае задача регрессионного анализа предполагает установление линейной зависимости между группой независимых переменных $x_1, x_2, \ldots, x_k$ (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 [28], [31]. 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 $x_1, x_2, \ldots, x_k$. 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 $x_1, x_2, \ldots, x_k$ действует на отклик не только независимо от других, но и порождает совместное воздействие. Для учета этого в модель, кроме переменных $x_1, x_2, \ldots, x_k$ 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 оправдано, 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:

$$y_i = A + bx_i + \varepsilon_i \qquad i = 1, \ldots, n. \tag{8.4}$$

Здесь $x_1, \ldots, x_n$ — заданные числа (значения фактора); $y_1, \ldots, y_n$ — наблюденные значения отклика; $\varepsilon_1, \ldots, \varepsilon_n$ — 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 $\mathbf{y}_i - f(\mathbf{x}_i, \theta)$ (в случае простой линейной регрессии — величины $y_i - A - bx_i$) должны в совокупности быть близки к нулю. Меру 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:

$$\sum_{i=1}^{n} [y_i - A - bx_i]^2 \to \min_{A,\,b}$$

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 $\varepsilon_1, \ldots, \varepsilon_n$, 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

$$y_i = a + b(x_i - \overline{x}) + \varepsilon_i \qquad i = 1, \ldots, n. \tag{8.5}$$

где $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $a = A + b\overline{x}$. Этот переход означает перенос начала отсчета на оси абсцисс в точку $\overline{x}$, которая служит центром совокупности (выборки) $x_1, \ldots, x_n$.

Для нахождения оценок по методу наименьших квадратов нам надо выяснить, при каких (a, b) достигается минимум выражения

$$\sum_{i=1}^{n} [y_i - a - b(x_i - \overline{x})]^2 . \tag{8.6}$$

Приравнивая нулю частные производные по a и b выражения (8.6), получим систему уравнений относительно неизвестных a и b :

$$\begin{cases} \sum_{i=1}^{n}[y_i - a - b(x_i - \overline{x})] = 0 \\ \sum_{i=1}^{n}(x_i - \overline{x})[y_i - a - b(x_i - \overline{x})] = 0 \end{cases}$$

Ее решение $(\hat{a}, \hat{b})$ легко найти:

$$\hat{a} = \overline{y} \qquad (\text{где } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i), \tag{8.7}$$

$$\hat{b} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})\,(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{8.8}$$

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 $D\hat{b} = \sigma^2/\sum_{i=1}^{n}(x_i - \overline{x})^2$;
3 $\operatorname{cov}(\hat{a}, \hat{b}) = 0$;
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 $M\hat{b} = b$.

$$M\hat{b} = M\frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\,M\,(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

поскольку величины $x_1, \ldots, x_n$ и $\overline{x}$ не случайны и содержащие только их выражения можно вынести из-под знака математического ожидания. Далее, поскольку $M\varepsilon_i = 0$ и $M\overline{\varepsilon} = 0$, то

$$M(y_i - \overline{y}) = My_i - M\overline{y} = a + b(x_i - \overline{x}) - a = b(x_i - \overline{x}).$$

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 $y_1, \ldots, y_n$ 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 $\hat{b}$:

$$\hat{a} \sim N\left(a, \frac{\sigma^2}{n}\right), \qquad \hat{b} \sim N\left(b, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right). \tag{8.9}$$

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

$$\sum_{i=1}^{n}[y_i - \hat{a} - \hat{b}(x_i - \overline{x})]^2. \tag{8.10}$$

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:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}[y_i - \hat{a} - \hat{b}(x_i - \overline{x})]^2. \tag{8.11}$$

Поскольку s^2 не зависит от $\hat{a}$ и $\hat{b}$, отношения

$$\sqrt{n}\,\frac{\hat{a} - a}{s} \quad \text{и} \quad \frac{\hat{b} - b}{s}\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{8.12}$$

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 [31].

Замечание. Стоит обратить внимание на сходство результата (8.11) с тем, что мы уже встречали, имея дело с нормальной выборкой. Пусть сейчас $y_1, \ldots, y_n$ — выборка из $N(a, \sigma^2)$. Оценками a, σ^2 служат, соответственно, $\hat{a} = \overline{y}$ и $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$. При этом a и s^2 независимы как случайные величины, и $\sum_{i=1}^{n}(y_i - \overline{y})^2$ 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 единичным 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 $\varepsilon_1,\ldots,\varepsilon_n$ 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 $\hat{y}_i=\hat{a}-\hat{b}(x_i-\overline{x})$, 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 $y_i-\hat{y}_i$, $i=1,\ldots,n$, где

$$\hat{y}_i=\hat{a}+\hat{b}(x_i-\overline{x})\,. \tag{8.13}$$

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 $x = x_i$, $i = 1, \ldots, n$ проводится m независимых измерений отклика. Их результаты при данном i удобно обозначить через $y_{i1}, \ldots, y_{im}$. При этом y_{ij} как случайные величины независимы при всех $j = 1, \ldots, m$, $i = 1, \ldots, n$. (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 $y_{i1}, \ldots, y_{im}$ перейдем к

$$y_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} y_{ij}\,, \qquad s_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_{ij} - y_{i\cdot})^2\,. \tag{8.14}$$

Мы уже вспоминали, что величины $(m-1)s_i^2$, $i = 1, \ldots, n$ распределены как $\sigma^2\chi^2(m-1)$ и стохастически независимы от $y_{i\cdot}$. Объединяя, мы получим, что

$$(m-1)\sum_{i=1}^{n} s_i^2 = \sigma^2\chi^2[n(m-1)]\,. \tag{8.15}$$

Как мы видели в п. 8.3, другую оценку дисперсии ошибок дает остаточная сумма квадратов

$$\sum_{i=1}^{n}[y_{i\cdot} - \hat{a} - \hat{b}(x_i - \overline{x})]^2 = \frac{\sigma^2}{m}\chi^2(n-2)\,. \tag{8.16}$$

Мы использовали формулу (8.10). Роль y_i в ней играет теперь $y_{i\cdot}$, причем $Dy_{i\cdot} = \sigma^2/m$. 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):

$$F = \frac{\frac{m}{n-2}\sum_{i=1}^{n}[y_{i\cdot} - \hat{a} - \hat{b}(x_i - \overline{x})]^2}{\frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - y_{i\cdot})^2} \tag{8.17}$$

с числом степеней свободы $(n-2,\ n(m-1))$.

Гипотеза о линейности должна быть отвергнута, если наблюденное 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 [31].

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

$$y_i = A + bx_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{8.18}$$

где $\varepsilon_1, \ldots, \varepsilon_n$ — 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 удается, так как изменение всех y_i на одну и ту же постоянную величину не изменяет рангов $y_1, \ldots, y_n$. 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

$$x_1 < x_2 < \cdots < x_n .$$

При такой нумерации легче следить за поведением остатков.

Если из наблюденных величин y_i вычесть истинные значения bx_i, то остатки $y_i - bx_i = A + \varepsilon_i$, $i = 1, \ldots, n$ 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 $y_i - \beta x_i$, $i = 1, \ldots, n$ 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

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

Наименьшей зависимости остатков $y_i - \beta x_i$ от x_i $(i = 1, \ldots, n)$ соответствует значение $r = 0$. По отношению к β это дает уравнение

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \beta \sum_{i=1}^{n}(x_i - \overline{x})^2$$

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

$$\begin{gathered} y_1 - \beta x_1, y_2 - \beta x_2, \ldots, y_n - \beta x_n \\ x_1, x_2, \ldots, x_n \end{gathered} \tag{8.19}$$

составим коэффициенты ранговой корреляции: ρ Спирмена и τ Кендэлла. 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,

$$\rho = \frac{\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)\left(R_i - \frac{n+1}{2}\right)}{\sqrt{\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)^2}\sqrt{\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)^2}}, \tag{8.20}$$

где R_i — ранг величины $y_i - \beta x_i$. Поскольку R_i принимает значения от 1 до n, найдем: $\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)^2 = \sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)^2 = \frac{n(n^2-1)}{12}$. Преобразовав числитель выражения (8.20), окончательно запишем ρ в виде:

$$\rho = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(i - R_i)^2. \tag{8.21}$$

Коэффициент корреляции τ Кендэлла определяется как

$$\tau = \frac{2(P-Q)}{n(n-1)} = \frac{2K}{n(n-1)}, \tag{8.22}$$

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 Кендэлла*. Ее можно записать в следующем виде, учитывая что $x_1 < \cdots < x_n$:

$$K = \sum_{1 \leqslant i < j \leqslant n} \operatorname{sign}(y_j - \beta x_j - y_i + \beta x_i) = \sum_{1 \leqslant i < j \leqslant n} \operatorname{sign}(R_j - R_i).$$

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

$$\tau(\beta) = 0, \tag{8.23}$$

или

$$\rho(\beta) = 0. \tag{8.24}$$

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 ($\beta \to -\infty$) 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:

$$y_i - \beta x_i = y_j - \beta x_j \tag{8.25}$$

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

$$\beta_{ij} = \frac{y_j - y_i}{x_j - x_i} \qquad \text{где } 1 \leqslant i < j \leqslant n, \tag{8.26}$$

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:

$$\hat{\beta} = \operatorname{med}\left\{ \frac{y_j - y_i}{x_j - x_i}, \quad \text{все } 1 \leqslant i < j \leqslant n \mid x_i \neq x_j \right\}. \tag{8.27}$$

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,

$$P\{|\tau| \leqslant \tau_\alpha\} = 1-2\alpha \quad \text{и} \quad P\{|\rho| \leqslant \rho_\alpha\} = 1-2\alpha. \tag{8.28}$$

(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 $P\{|\tau| \leqslant \tau_\alpha\} \geqslant 1-2\alpha$ (для ρ_α — аналогично).

Доверительные интервалы для b с коэффициентом доверия не меньше $1-2\alpha$ имеют вид:

$$\{\beta: \quad |\tau(\beta)| \leqslant \tau_\alpha\} \quad \text{или} \quad \{\beta: \quad |\rho(\beta)| \leqslant \rho_\alpha\}, \tag{8.29}$$

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.

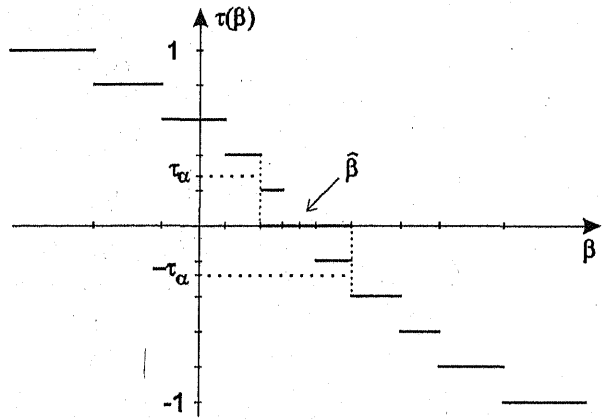


Рис. 8.1. Доверительный интервал для коэффициента корреляции $\tau(\beta)$ при $n=5$

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 [89].

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:

$$\{\beta: \quad |K(\beta)| \leqslant K_\alpha\},$$

где K_α есть решение уравнения $P\{|K| \leqslant K_\alpha\} = 1-2\alpha$. При этом вероятность P рассматривается в случае справедливости выдвинутой гипотезы о независимости двух рядов чисел (8.19). В [91] приведена таблица вероятностей хвостов распределения статистики K для $n = 4(1)40$. Чтобы воспользоваться этими таблицами, заметим, что $K_\alpha + 2$ удовлетворяет соотношению $P(K \geqslant K_\alpha + 2) = \alpha/2$.

Затем совокупность чисел $\frac{y_j - y_i}{x_j - x_i}$, $1 \leqslant i < j \leqslant n$, надо расположить в порядке возрастания. Мы предположим сейчас, что среди чисел $x_1, \ldots, x_n$ нет совпадающих. Обозначим элементы этой упорядоченной совокупности через $S^{(1)} \leqslant S^{(2)} \leqslant \cdots \leqslant S^{(N)}$, $N = \frac{n(n-1)}{2}$. 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 связок среди иксов) , а через t_l — число совпадающих элементов в группе с

номером $l : l = 1, \ldots, g$. Тогда значение K_α, получаемое при использовании нормальной аппроксимации для распределения коэффициента ранговой корреляции τ при гипотезе независимости, имеет вид:

$$K_\alpha \sim \frac{\sqrt{n(n-1)(2n+5) - \sum_{l=1}^{g} t_l(t_l-1)(2t_l+5)}}{18}\, u_{1-\alpha/2}. \qquad (8.30)$$

Этот результат был получен П.Сеном [105]. Соответствующие значения M_1 и M_2 равны

$$M_1 = \left[\frac{N-K_\alpha}{2}\right], \qquad M_2 = \left[\frac{N+K_\alpha}{2}\right] + 1. \qquad (8.31)$$

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 $i=1,\ldots..6$		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 $y_{i\cdot}$		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 таблице 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 $y_{i\cdot}$ и стандартные отклонения s_i^2. Графическое изображение этих данных дано на рис. 8.2.

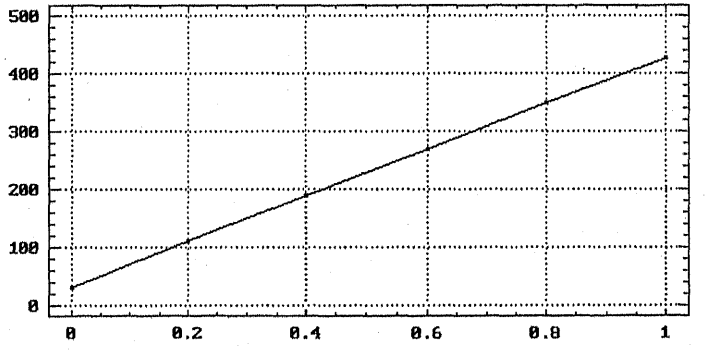


Рис. 8.2. Графическая зависимость $y_{i\cdot}$ от x_i.

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

$$y = A + bx + \varepsilon\,, \tag{8.32}$$

то есть 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):

$$y_{ij} = a + b(x_i - \overline{x}) + \varepsilon_{ij} \quad i = 1, \dots, 6, \quad j = 1, \dots, 6\,.$$

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:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}[y_{ij} - a - b(x_i - \overline{x})]. \tag{8.33}$$

Приравнивая к нулю производные по переменным a и b в выражении (8.33) получаем:

$$\begin{aligned} &\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - a - b(x_i - \overline{x})) = 0\,, \\ &\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i - \overline{x})(y_{ij} - a - b(x_i - \overline{x})) = 0\,. \end{aligned} \tag{8.34}$$

Проводя суммирование в уравнениях (8.34) по индексу j, и деление каждого из уравнений на компоненту m, имеем:

$$\begin{aligned} &\sum_{i=1}^{n}(y_{i\cdot} - a - b(x_i - \overline{x})) = 0, \\ &\sum_{i=1}^{n}(x_i - \overline{x})(y_{i\cdot} - a - b(x_i - \overline{x})) = 0, \end{aligned} \tag{8.35}$$

где $y_{i\cdot}$ определено в (8.14).

Полученная система уравнений отличается от системы, рассмотренной в п. 8.3, заменой y_i на $y_{i\cdot}$. 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 $y_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$. Оценки параметров a и b, являющиеся решением системы (8.35), согласно (8.7), (8.8) суть

$$\hat{a} = \overline{y}, \quad \text{где } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_{i\cdot} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\,, \tag{8.36}$$

$$\hat{b} = \frac{\sum_{i=1}^{n}(y_{i\cdot} - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\,. \tag{8.37}$$

Подставляя в (8.36) и (8.37) соответствующие значения из таблицы 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:

$$\hat{a} \sim N\left(a, \frac{\sigma_1^2}{n}\right), \quad \hat{b} \sim N\left(b, \frac{\sigma_1^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right),$$

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:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_{i\cdot} - \hat{a} - \hat{b}(x_i - \overline{x}))^2\,.$$

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:

$$\hat{a} - \frac{s}{\sqrt{n}}\,t_{1-\alpha/2} < a < \hat{a} + \frac{s}{\sqrt{n}}\,t_{1-\alpha/2}\,,$$

$$\hat{b} - \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\,t_{1-\alpha/2} < b < b + \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\,t_{1-\alpha/2}\,.$$

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. [16]) находим $t_{1-\alpha/2} \simeq 2.79$. Отсюда 95% доверительные интервалы для a и b равны:

$$227.8 < a < 229.6, \qquad 394.5 < b < 399.9. \tag{8.38}$$

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):

$$F = \frac{\frac{m}{n-2}\sum_{i=1}^{n}[y_{i\cdot} - \hat{a} - \hat{b}(x_i - \overline{x})]^2}{\frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - y_{i\cdot})^2} = \frac{ms^2}{\frac{1}{n}\sum_{i=1}^{n} s_i^2}.$$

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 [16] 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, $x_2, \ldots$, x_2, $x_3, \ldots, x_3, \ldots, x_6, \ldots, x_6$, где каждое значение x_i повторяется 6 раз. Таким образом, объем массива x равен $N = 36$. Рассмотрим массив:

$$\beta_{ij} = \left\{ \frac{y_j - y_i}{x_j - x_i},\quad \text{все } 1 \leqslant i < j \leqslant N, \text{ для которых } x_i \neq x_j \right\}.$$

Объем массива β_{ij} в нашем случае, с учетом повторений значений в массиве x, равен: $C_{36}^2 - 6\,C_6^2 = 540$. Согласно (8.27), новой оценкой коэффициента b будет являться величина $\tilde{\beta}$, равная:

$$\tilde{\beta} = \underset{\substack{i,j=1,\ldots,n \\ x_i \neq x_j}}{\operatorname{med}} \left\{ \frac{y_j - y_i}{x_j - x_i} \right\} = \frac{\beta^{(270)} + \beta^{(271)}}{2}.$$

Здесь $\beta^{(k)}$ обозначает k-ый 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 $l = 1, \ldots, g$. Выбирая значение $\alpha = 0.05$ по таблице [16] 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

$$395.8 < b < 399.4. \qquad (8.39)$$

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 пакетах STATGRAPHICS и STADIA

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 [31]. 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 [28], [65], [95]. В целом отметим, что комплектация статистических пакетов регрессионными программами сильно варьируется. В пакетах STATGRAPHICS и STADIA представлен довольно традиционный набор регрессионных методов, позволяющий решать весьма широкий круг задач, однако считать его полным не в коем случае не следует. В обоих пакетах полностью отсутствуют непараметрические методы регрессионного анализа.

8.7.1. Пакет STATGRAPHICS

Методы регрессионного анализа в пакете представлены в пункте K. Regression analysis головного меню. Процедуры этого пункта приведены на рис. 8.3. Опишем кратко их назначение.

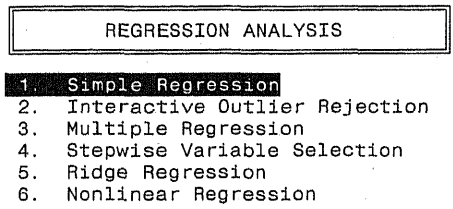


Рис. 8.3. Меню процедур регрессионного анализа

1. Simple Regression (простая регрессия) — выполняет простую регрессию для одной независимой переменной методом наименьших квадратов. Оценивает параметры для линейной и ряда нелинейных (сводящихся к линейной) моделей. Работа этой процедуры разобрана в примере 8.1к.

2. Interactive Outlier Rejection (интерактивное отбрасывание) — позволяет выборочно исключать резко выделяющиеся наблюдения из графика простой линейной регрессии. Этой процедуре посвящен пример 8.2к.

3. Multiple Regression (множественная регрессия) — выполняет обычную регрессию для нескольких независимых переменных методом наименьших квадратов. Частичный разбор работы этой процедуры дан в примере 8.3к.

4. Stepwise Variable Selection (шаговая регрессия) — выполняет прямую (включение) или обратную (исключение независимых переменных) пошаговую множественную регрессию. Описание этой процедуры дано, например в [31].

5. Ridge Regression (ридж-регрессия или гребневая регрессия) — выполняет гребневую регрессию на стандартизированных переменных и строит график результатов. Содержание и работа этой процедуры в книге не рассматриваются.

6. Nonlinear Regression (нелинейная регрессия) — вычисляет оценки наименьших квадратов для параметров в заданной пользователем нелинейной регрессионной модели (см. [10], [28], [31]).

Разберем на примерах работу наиболее употребительных из этих процедур.

Пример 8.1к. С помощью метода наименьших квадратов вычислим оценки параметров в модели простой линейной регрессии для данных калибровочного эксперимента (табл. 8.1). Построим 95% доверительный интервал для среднего значения отклика.

Подготовка данных. В редакторе базы данных пакета (процедура 2. File Operations пункта A. Data Management головного меню пакета) в файле TENZOV создадим переменные nagr и otkl. Вид экрана редактора базы данных с введенными переменными приведен на рис. 8.4.

Для переменной nagr удобно выбрать тип с одним десятичным знаком после точки (значение 1 в поле Type), а для переменной otkl — с тремя десятичными знаками после точки (значение 3 в поле Type). В первую переменную запишем в порядке возрастания значения эталонной силы x_i из второй строки табл. 8.1. Как было показано выше, в случае равного числа наблюдений для каждого из значений независимой переменной x_i, в качестве зависимой переменной можно рассматривать средние значения соответствующих откликов. Поэтому в переменную otkl запишем значения из предпоследней строки табл. 8.1.

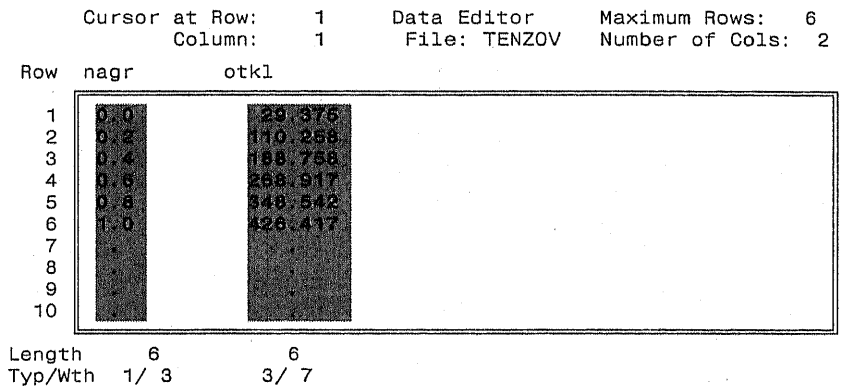


Рис. 8.4. Экран редактора с данными для примера 8.1к

Выбор процедуры. В меню пункта K. Regression analysis (рис. 8.3) выберем процедуру 1. Simple Regression.

Заполнение полей ввода данных. Экран ввода данных выбранной процедуры (рис. 8.5) содержит активные поля: Dependent variable (зависимые переменные), Independent variable (независимые переменные), Model (модель), Confidence limits (доверительные границы для среднего значения отклика), Prediction limits (доверительные границы для прогноза), Point labels (метки точек).

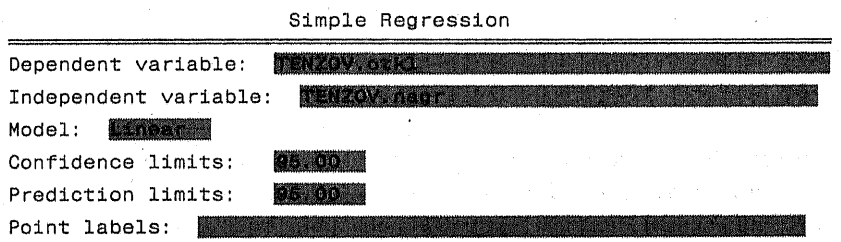


Рис. 8.5. Запрос параметров простой регрессии

Введем в поле **Dependent variable** переменную **TENZOV.otkl**, а в поле **Independent variable** — **TENZOV.nagr**.

В поле **Model** укажем значение **Linear**. Описание назначения других возможных значений в этом поле смотри в комментариях.

Поле **Confidence limits** предполагает ввод в процентах уровня доверия для границ доверительного интервала для среднего отклика при заданном значении x. Укажем в нем значение 95, как это требуется в примере.

Поле **Prediction limits** предполагает ввод в процентах уровня доверия для прогноза новых наблюдений. Оставим в этом поле значение 95, выставленное пакетом по умолчанию.

Заполнение поля **Point labels** не обязательно. Вводимая в это поле переменная должна содержать значения меток точек, если Вы хотите провести интерактивную разметку точек на графике регрессии.

Результаты. После заполнения всех необходимых полей ввода следует нажать клавишу F6. На экране появятся результаты обработки (рис. 8.6) в виде двух таблиц.

Regression Analysis - Linear model: Y = A+bX

Dependent variable: TENZOV.otkl Independent variable: TENZOV.nagr

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	30.124	0.581373	51.8152	.00000
Slope	397.174	0.960106	413.678	.00000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	Prob. Level
Model	110423.27	1	110423.27	171129.3	.00000
Error	2.5810491	4	.6452623		
Total (Corr.)	110425.85	5			

Correlation Coefficient = 0.999988 R-squared = 100.00 percent
Stnd. Error of Est. = 0.803282

Рис. 8.6. Результаты регрессионного анализа для примера 8.1к

В верхней таблице рис. 8.6 приведены оценки параметров простой линейной модели $Y = A + bX$ и их статистические характеристики. Строка **Intercept** (свободный член) относится к параметру A, а строка **Slope** (наклон) — к параметру b. Столбец **Estimate** содержит оценки этих параметров (заметим, что в пакете простая линейная модель рассматривается в форме (8.4), а не в форме (8.5), и оценка $\hat{A}$ вычисляется по формуле $\hat{A} = \overline{y} - \hat{b}\overline{x}$). Столбец **Standard Error** (стандартная ошибка) содержит значения стандартных ошибок указанных коэффициентов. Два последних столбца таблицы **T Value** и **Prob. Level** содержат значения стью-

дентовых отношений (см. 8.12) для полученных оценок и их минимальные уровни значимости для проверки гипотезы о равенстве значений коэффициентов нулю. Полученные уровни значимости говорят, что обе оценки значимо отличаются от нуля.

Таблица Analysis of Variance рис. 8.6 является базовой таблицей анализа вариации и служит для оценки адекватности предлагаемой модели данных. Описание этой таблицы дано в примерах 6.2к и 7.2к.

В случае регрессионного анализа общая вариация отклика относительного его среднего значения распадается на вариацию, обусловленную моделью, и остаточную вариацию, приписываемую случайным ошибкам. Формальный вид этого разложения для простой линейной модели следующий:

$$\sum_{i=1}^{N}(y_i-\overline{y})^2=\sum_{i=1}^{N}(\hat{A}+\hat{b}x_i-\overline{y})^2+\sum_{i=1}^{N}(y_i-\hat{A}-\hat{b}x_i)^2\,,$$

где величина в левой части называется общей вариацией или суммой квадратов относительно среднего (Total (Corr.) Sum of Squares), первое слагаемое в правой части — суммой квадратов, обусловленной регрессией или моделью (Model Sum of Squares), второе слагаемое — суммой квадратов относительно модели регрессии или суммой квадратов ошибок (Error Sum of Squares).

Для проверки гипотезы о равенстве коэффициента b нулю (в общем случае — об адекватности предлагаемой модели) в пакете используется F-Ratio (F-отношение). Оно вычисляется как частное от деления средних квадратов относительно модели на средние квадраты ошибок. Если $b=0$, это отношение имеет F-распределение с числом степеней свободы 1 и $N-2$. Когда же b отлично от нуля, эта величина имеет тенденцию к возрастанию с ростом значения b. Полученный в разбираемом примере минимальный уровень значимости F-отношения свидетельствует о том, что линейная зависимость между x и y значима, т.е. значение b отлично от нуля.

Еще одним показателем качества подобранной модели традиционно считается квадрат выборочного коэффициента корреляции Пирсона R (Correlation Coefficient). Нетрудно показать, что R^2 (R-squared) является отношением суммы квадратов, обусловленных регрессией, к общей сумме квадратов откликов, скорректированной на среднее. Другими словами, R^2 показывает долю общего разброса y относительно среднего $\overline{y}$, объясняемую регрессией. Величину R^2 также часто именуют *коэффициентом детерминации* и измеряют не в долях единицы, а в процентах. Чем ближе значение R^2 к ста процентам, тем лучше подобранная модель описывает данные эксперимента.

Наконец, последняя характеристика экрана выдачи результатов — Stnd. Error of Est. (стандартная ошибка оценки). Эта величина по определению равна

$$\sqrt{\frac{1}{N-p-1}\sum_{i=1}^{N}\left(y_i-\hat{A}-\hat{b}x_i\right)^2},$$

где p — число степеней свободы суммы квадратов, обусловленных регрессией. В данном случае $p=1$. Другими словами, стандартная ошибка оценки есть квадратный корень из средней суммы квадратов ошибок. Согласно формулам (8.12), она может использоваться для построения доверительных интервалов для истинных параметров модели.

Углубленный анализ. Для построения доверительных границ для среднего значения отклика следует обратиться к процедурам дополнительных методов анализа простой линейной регрессии. При нажатии Enter в экране вывода результатов (рис. 8.6) происходит возврат к экрану ввода параметров процедуры (рис. 8.5), на котором появляется всплывающее меню дополнительных методов (рис. 8.7)

Plot fitted line	график подобранной линии
Plot residuals	график остатков
Save residuals	сохранение остатков
Save predictions	сохранение предсказаний

Рис. 8.7. Меню дополнительных методов регрессионного анализа

Процедура Plot fitted line выводит на экран график линии регрессии, доверительные границы для среднего отклика в виде пары пунктирных линий, ближайших к линии регрессии. Область между этими линиями обычно называется *доверительной трубкой*. Пунктирные линии, более удаленные от линии регрессии, очерчивают доверительные границы (доверительную трубку) для прогноза значений новых наблюдений. На рис. 8.8 приведена часть этого графика в диапазоне изменений значения x от 0.2 до 0.3. Кроме того, в правом нижнем углу экрана с указанными графиками пользователь может задавать значения x или y, для которых он хочет получить соответствующие значения y или x из уравнения регрессии. Так, введя значение x, равное 0.25 и нажав Enter, из уравнения регрессии получаем значение y, равное 129.4176. Для ввода значения y перед ним необходимо набрать запятую.

Процедура Plot residuals полезна для представления о том, насколько подобранная модель соответствует исходным данным и насколько выполняются условия применения метода наименьших квадратов. При правильно подобранной модели и справедливости предположения о нор-

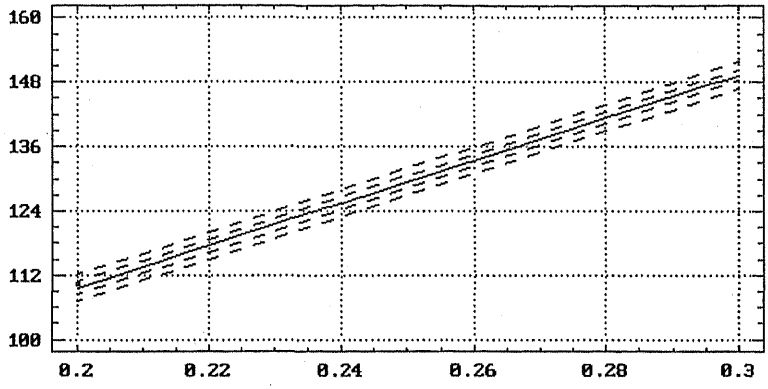


Рис. 8.8. График регрессии и доверительные трубки для среднего отклика и прогнозов новых наблюдений

мальном распределении остатков, вычисленные остатки также имеют нормальное распределение, хотя и являются статистически зависимыми. Как показывает опыт, при графическом анализе остатков последним обстоятельством можно пренебречь. Подробное описание процедур анализа остатков дано, например, в [31]. Для более детального анализа остатков пакет предусматривает возможность сохранения остатков (процедура Save residuals) в переменной, заданной пользователем. Некоторые методы проверки распределения остатков на нормальность в пакете приведены в гл. 5 и 10.

Процедура Save predictions записывает в переменную, задаваемую пользователем, вектор оцененных значений $\hat{y}_i = \hat{A} + \hat{b}x_i$. Эта процедура особенно полезна для мультипликативной и экспоненциальной моделей, в которых вычисление $\hat{y}_i$ требует определенных затрат.

Комментарии. 1. Длины векторов данных, вводимых в поля Dependent variable и Independent variable, всегда должны совпадать. Это приводит к дублированию значений x_i в независимой переменной, если в соответствующей точке имеются повторные наблюдения y_{ij} и их число варьируется в зависимости от индекса i.

2. В поле Model могут фигурировать следующие значения (в правом столбце указаны соответствующие им функции):

Linear (линейная)	$y = a + bx$
Multiplicative (мультипликативная)	$y = ax^b$
Exponential (экспоненциальная)	$y = e^{ax+b}$
Reciprocal (обратная)	$1/y = a + bx$

Заметим, что вторая и третья модели сводятся к простой линейной модели путем логарифмирования вектора y, а последняя — преобразованием $1/y$ компонент вектора y. Однако в последних трех моделях стандартные предположения о характере распределения ошибок делаются уже относительно преобразованных величин. В связи с этим вторую модель часто называют моделью с мультипликативной ошибкой. Если же мы хотим иметь во второй модели аддитив-

ную ошибку, то для нахождения оценок следует применять методы нелинейной регрессии. Это замечание относится также к третьей и четвертой моделям.

3. При вводе в поля Confidence limits или Prediction limits значения 0 соответствующие доверительные границы не строятся.

4. Процедура Plot fitted line по умолчанию строит график регрессии во всем диапазоне изменения значений x. Для изменения этого диапазона следует после построения графика регрессии использовать вызов процедур настройки графического вывода (клавиша F5).

Следующий пример наглядно демонстрирует недостатки использования метода наименьших квадратов для получения регрессионных оценок в случае, когда происходит нарушение исходных предпосылок модели. Для этого особенно удобна процедура 2. Interactive Outlier Rejection (интерактивное отбрасывание). В ней также решается задача простой линейной регрессии, однако пользователь получает возможность прямо на графике регрессии указать точки, которые он считает необходимым удалить из расчетов, и получить на том же графике новую линию регрессии. Мы продемонстрируем работу этой процедуры на примере П.Хьюбера [92], являющегося одним из основателей теории робастного (устойчивого) оценивания.

Пример 8.2к. Методом наименьших квадратов вычислим оценки параметров простой линейной регрессии, когда вектор независимых переменных x есть $(-4, -3, -2, -1, 0, 10)$, а в вектор зависимых переменных y — $(2.48, 0.73, -0.04, -1.44, -2.32, 0)$. Оценим адекватность подобранной модели. Проведем повторные расчеты, исключив из данных резко выделяющееся наблюдение.

Подготовка данных. Учитывая небольшой объем данных, их ввод будет осуществлен непосредственно с клавиатуры при заполнении полей экрана ввода параметров процедуры (см. ниже).

Выбор процедуры. В меню пункта K. Regression analysis (рис. 8.3) выберем процедуру 2. Interactive Outlier Rejection.

Заполнение полей ввода данных. Порядок ввода данных этой процедуры — такой же, как в процедуре 1. Simple Regression, разобранной выше (рис. 8.5). Отсутствует только поле выбора модели, так как в процедуре предполагается модель простой линейной регрессии.

В поля Dependent variable и Independent variable надо осуществить ввод с клавиатуры значений векторов y и x. Значения компонент этих векторов при вводе следует разделять пробелами (рис. 8.9).

Результаты. Экран вывода результатов процедуры приведен на рис. 8.10. Он содержит график подобранной регрессии, доверительные

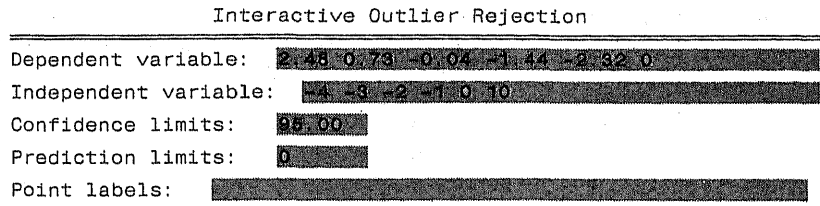


Рис. 8.9. Запрос параметров процедуры интерактивного отбрасывания

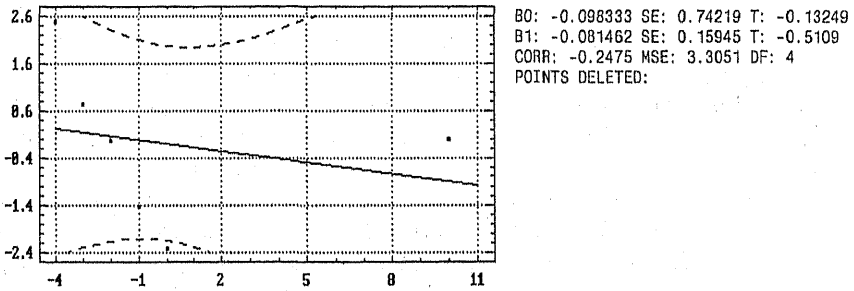


Рис. 8.10. Интерактивная регрессия в пакете STATGRAPHICS.

границы для среднего значения отклика и прогноза, а также точки исходных значений.

В левом нижнем углу экрана (на рис. 8.10 — справа) приведены оценки коэффициентов A и b (обозначенных через B0 и B1), их стандартные ошибки SE и значения t-статистик Стьюдента для проверки значимого отличия коэффициентов от нуля. Значение коэффициента множественной корреляции CORR и средняя сумма квадратов ошибок MSE с указанным числом степеней свободы в определенной мере характеризуют качество подобранной модели.

Обратим внимание на то, что значения t-статистик Стьюдента для каждого из коэффициентов не позволяет отвергнуть гипотезу о равенстве их нулю. Значение коэффициента корреляции показывает, что линейная модель довольно слабо объясняет вариацию данных. Обращает на себя внимание последняя точка с координатами $(10, 0)$, которая, как будет видно ниже, и является причиной полученных неудовлетворительных результатов. Проведем повторные расчеты без учета этой точки. Для этого подведем к этой точке с помощью клавиш управления курсором графический курсор, обозначенный на рисунке знаком (+), и нажмем клавишу [E]. При этом данная точка изменит свой цвет, что означает, что она удалена из обработки. Так можно удалить произвольное количество точек. После этого, нажав [F6], получаем на экране на старом графике новую линию регрессии и ее характеристики (рис. 8.11).

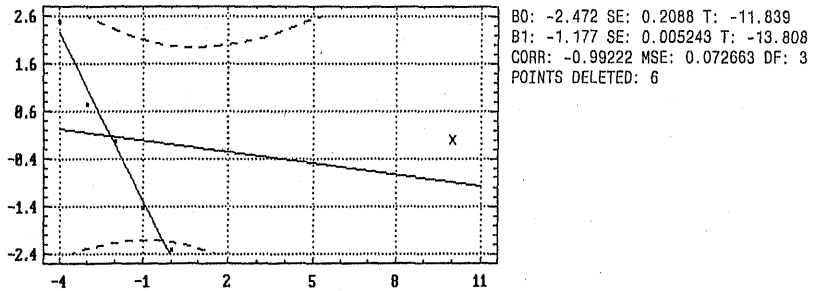


Рис. 8.11. Интерактивная регрессия в пакете STATGRAPHICS. Крестиком помечена удаленная из обработки точка

Рис. 8.11 показывает, как сильно грубые наблюдения влияют на расчеты методом наименьших квадратов в регрессионных задачах.

Более внимательное изучение графика данных может навести на мысль, что простые линейные модели просто не годятся для этих данных и следует воспользоваться параболической моделью. Результаты обработки для параболы приведены ниже в примере 8.3к. Как пишет П.Хьюбер [92]: «Совершенно очевидно, что для того, чтобы делать умозаключения о достоинствах или недостатках этих подгонок, приведенных данных недостаточно». Если судить по остаточной ошибке, то следует склониться к параболической модели, которой соответствует самое низкое значение остаточной ошибки. Однако учитывая происхождение данных (см. комментарий 1), подходящей будет подгонка простой линейной моделью с исключением из обработки шестой точки.

Комментарии. 1. Данные примера являются искусственными. Они были получены следующим образом: к шести точкам, лежащим на прямой $y = -2 - x$ были добавлены случайные ошибки: с первой по пятую — нормальные ошибки (с нулевым средним и стандартным отклонением 0.6), а к шестой точке — большая ошибка 12.

2. Для того, чтобы обратно включить удаленную точку в обработку, подведите к ней курсор и нажмите клавишу (I).

Следующий пример проиллюстрирует начало работы с процедурой 3. Multiple Regression. Детальный разбор проблем, возникающих в задачах множественной регрессии, дан в [31].

Пример 8.3к. Методом наименьших квадратов вычислим оценки параметров параболической модели регрессии $y = A + Bx + Cx^2$ для данных примера 8.2к.

Подготовка данных. В редакторе базы данных пакета стандартным образом в файле HUBER создадим переменные x и y с данными примера 8.2к.

Выбор процедуры. В меню пункта K. Regression analysis выберем процедуру 3. Multiple Regression.

Заполнение полей ввода данных. Экран ввода данных указанной процедуры приведен на рис. 8.12.

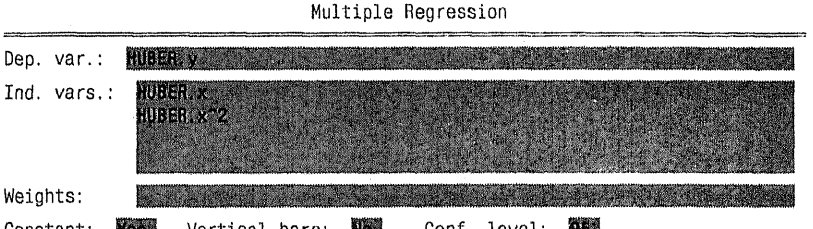


Рис. 8.12. Запрос параметров процедуры множественной регрессии

Поле Dep. var. предназначено для ввода зависимой переменной. В поле Ind. vars. (независимые переменные) вводятся независимые переменные. В рассматриваемом примере это HUBER.x и HUBER.x^2. Последняя переменная состоит из квадратов значений переменной HUBER.x, то есть равна следующему вектору

16, 9, 4, 1, 0, 100

Активное поле Weights (веса) следует заполнять только в случае использования взвешенного метода наименьших квадратов. (В него вносится вектор весов той же длины, что и вектор зависимой переменной.)

Для включения в модель регрессии константы в поле Constant надо указать значение Yes.

Поле Vertical bars позволяет задать дополнительное оформление различных графиков, связанных с процедурой регрессионного анализа вертикальными линиями.

Задание доверительного уровня (в процентах) при построении доверительных интервалов для коэффициентов регрессии осуществляется в поле Conf. level.

Результаты. После заполнения полей ввода следует нажать клавишу F6, и на экране появятся результаты расчетов для оценок коэффициентов модели, их стандартных ошибок, значений t-статистик и их уровней значимости (рис. 8.13).

Подробное описание всех статистик, указанных на экране, можно найти в [31].

Углубленный анализ. После нажатия Esc произойдет возврат к экрану ввода данных и параметров процедуры (рис. 8.12), на котором появится меню углубленного анализа (рис. 8.14).

Model fitting results for: HUBER.y

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	-2.266531	0.154469	-14.6731	0.0007
HUBER.x	-0.774258	0.044868	-17.2564	0.0004
HUBER.x^2	0.100071	0.005892	16.9828	0.0004

R-SQ. (ADJ.) = 0.9839 SE= 0.212992 MAE= 0.109383 DurbWat= 3.229
Previously: 0.0000 0.000000 0.000000 0.000
6 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

Рис. 8.13. Результаты вычислений процедуры множественной регрессии

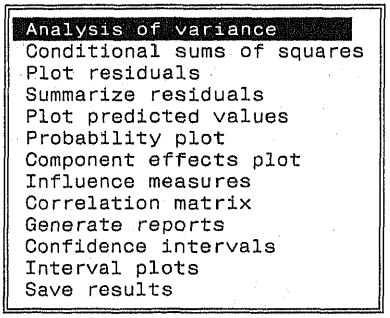


Рис. 8.14. Меню процедур углубленного анализа для множественной регрессии

Не разбирая работы указанных на рис. 8.14 процедур, кратко укажем их назначение (согласно документации пакета).

Analysis of variance (анализ вариации) — проводит дисперсионный анализ модели регрессии.

Conditional sums of squares (условные суммы квадратов) — вычисляет ряд дополнительных статистики дисперсионного анализа. Оценивает вклад каждой из переменных модели в общую сумму квадратов, обусловленную регрессией.

Plot residuals (график остатков) — строит график остатков в зависимости от указанной пользователем переменной.

Summarize residuals (анализ остатков) — вычисляет основные описательные статистики для остатков (среднее, дисперсию, стандартную ошибку, коэффициенты асимметрии и эксцесса и статистику Дурбина–Уотсона).

Plot predicted values (график прогнозов) — строит график предсказанных значений в зависимости от наблюдаемых. Эта процедура полезна для выявления случаев, в которых дисперсия зависимых переменных не постоянна.

Probability plot (график нормальной вероятности) — строит график эмпирической функции распределения остатков на нормальной вероятностной бумаге.

Component effects plot (график эффектов компонент) — графически иллюстрирует вклад каждой компоненты модели в общую вариацию зависимой переменной.

Influence measures (меры влияния) — указывает число наблюдений, влияние которых на полученные оценки модели выше определенного уровня.

Correlation matrix (матрица корреляции) — выдает матрицу коэффициентов корреляции Пирсона для оцененных коэффициентов модели.

Generate reports (генерация отчета) — позволяет получить (частично или полностью) информацию о наблюдаемых и подобранных по модели значениях, остатках, стандартных ошибках для прогноза, доверительных границах для прогноза.

Confidence intervals (доверительные интервалы) — вычисляет границы доверительных интервалов для оценок коэффициентов регрессии на уроне значимости, указанном при заполнении экрана ввода данных и параметров процедуры.

Interval plots (график доверительных интервалов) — строит график доверительных интервалов для прогнозов относительно указанной пользователем переменной.

Save results (сохранение результатов) — позволяет сохранить всю или часть полученной информации в файле базы данных пакета.

Перечень этих процедур показывает, что построение множественной регрессионной модели может включать в себя много различных аспектов и требует высокой квалификации исследователя. Обсуждение вопросов, связанных с множественной регрессией, довольно подробно дано в [31].

Комментарии. Задание сложных моделей множественной регрессии требует привлечения ряда специфических операторов преобразования независимых переменных. Их назначение и работа подробно описана в документации пакета.

8.7.2. Пакет STADIA

В пакете широко представлены различные методы регрессионного анализа, включая простую, множественную, пошаговую, нелинейную регрессию и др. (см. меню **Статистические методы** на рис. 1.17). Следует сразу обратить внимание на не совсем традиционную классификацию регрессионных моделей в пакете.

Для общего обозначения моделей данных, обрабатываемых методами регрессионного анализа, в справочнике пакета используется термин **Экспериментальные зависимости**. Последние делятся в пакете на однопараметрические и многопараметрические, линейные и нелинейные по параметрам. При этом под однопараметрической зависимостью понимается произвольная функция $y = f(x)$, где x — простая действительная переменная. Это определение может привести к путанице, так как число параметров в подобной зависимости может быть любое. В частности, все полиномиальные модели при этом попадают в процедуру «Простой регрессии». Скорее, эти зависимости следовало бы назвать одномерными или однофакторными. Более подробно анализ списка данных моделей (рис. 8.17) дан в комментариях к примеру 8.1к.

Пример 8.1к. Методом наименьших квадратов вычислим оценки параметров в модели простой линейной регрессии для данных калибровочного эксперимента (табл. 8.1). Построим 95% доверительную трубку для среднего значения отклика.

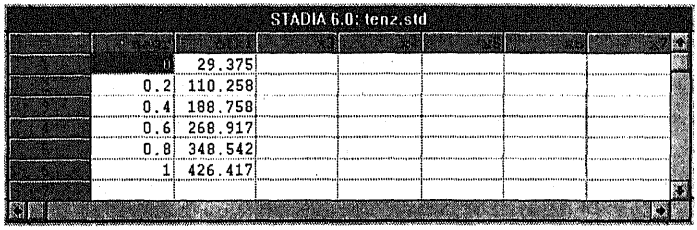


Рис. 8.15. Данные для примера 8.1к

Подготовка данных. Введем в электронную таблицу пакета данные таблицы 8.1 в переменные nagr и otkl (см. рис. 8.15).

Выбор процедуры. В меню Статистические методы (рис. 1.17) в разделе Регрессионный анализ выберите пункт L = Простая регрессия/тренд.

Заполнение полей ввода данных. В появившемся на экране запросе Переменные регрессии (рис. 8.16) укажите в качестве Y-переменной переменную otkl, а в качестве X-переменной — переменную nagr. Для этого следует выделить с помощью мыши нужную переменную в поле Переменные и нажать соответствующую кнопку со стрелкой вправо. После нажатия кнопки запроса [Утвердить] программа выдаст меню моделей регрессии (рис. 8.17), отнесенных в пакете к однопараметрическим. Выберите в нем пункт 1=линейная или просто нажмите клавишу [1].

Рис. 8.16. Запрос выбора переменных регрессии

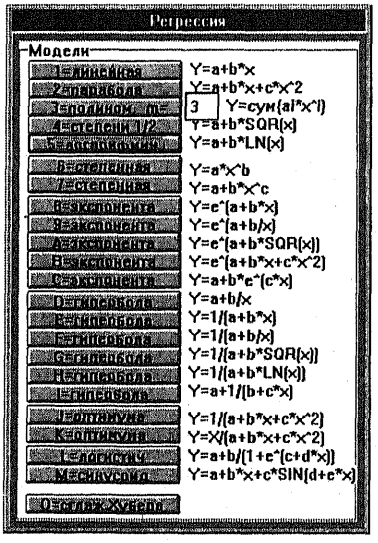


Рис. 8.17. Меню моделей однопараметрической регрессии

Результаты. Экран вывода результатов процедуры (рис. 8.18) содержит три блока информации. В первом из них представлены оценки коэффициентов модели, их стандартные ошибки и уровни значимости t-отношений для проверки гипотез об отличии соответствующих коэффициентов от нуля. Второй блок информации содержит базовую таблицу дисперсионного анализа (см. пример 6.2к), показывающую, как общая вариация отклика распределяется между вариацией, обусловленной введенной моделью, и вариацией остатков. Третий блок информации содержит абсолютную величину коэффициента множественной корреляции R, коэффициент детерминации R^2, несмещенную оценку коэффициента детерминации R^2прив, а также F-отношение и его уровень значимости для проверки гипотезы о соответствии выбранной модели наблюденным данным. Сравнивая полученный уровень значимости с пятипроцентным, процедура делает заключение об адекватности модели.

```
          Модель: линейная  Y = a0+a1*x
  Коэфф.         a0          a1
Значение     30.124      397.17
Ст.ошиб.    0.58137     0.96011
 Значим.     0.0001      0.0001

Источник  Сум.квадр. Степ.св  Средн.квадр.
Регресс. 1.1042E5         1     1.1042E5
Остаточн    2.581         4      0.64526
     Вся 1.1043E5         5

Множеств R      R^2    R^2прив  Ст.ошиб.         F     Значим
  0.99999  0.99998    0.99997  0.80328 1.17113E5          0
Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>

   nagr=0.75     Y=328
```

Рис. 8.18. Результаты расчетов процедуры простой линейной регрессии

Процедура также предлагает пользователю рассчитать с помощью подобранной модели значение зависимой переменной для указанного значения независимой переменной. Для этого в окне Интерполяция (рис. 8.19) следует указать требуемое значение переменной nagr, например 0.75 и нажать кнопку Утвердить. Рассчитанное значение отклика помещается в окно результатов (рис. 8.18).

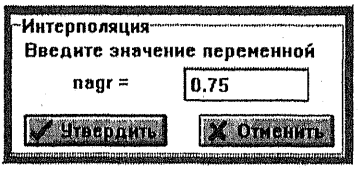


Рис. 8.19. Запрос вычисления значения зависимой переменной

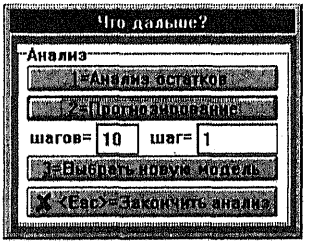


Рис. 8.20. Меню дополнительных возможностей процедуры регрессии

Далее процедура предлагает построить график экспериментальных точек и регрессионной кривой (рис. 8.21, левая часть). Построенный график при этом может быть сохранен в отдельном графическом окне. Содержание окна сохраняется в течение всего сеанса работы с программой и может быть просмотрено по требованию пользователя.

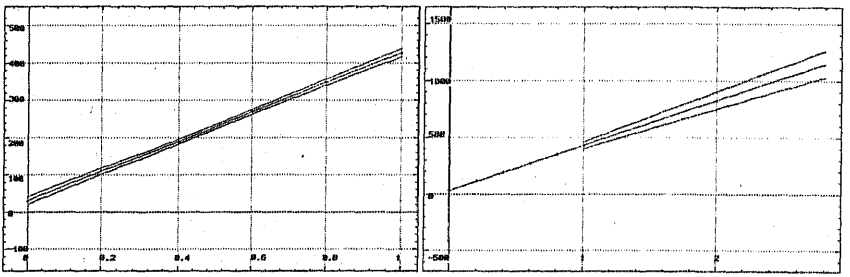


Рис. 8.21. Пакет STADIA. График экспериментальных точек и регрессионной кривой. Справа — график прогноза

Дополнительные возможности. Затем пользователю предлагается меню дополнительных возможностей процедуры (рис. 8.20).

Результаты расчетов процедуры 1 = **Анализ остатков** представлены на рис. 8.22. Кроме значений экспериментальных данных, они содержат подобранные значения модели, остатки и их стандартизированные значения, а также стандартные ошибки остатков и доверительные интервалы для них (в виде допустимого отклонения для 95% уровня доверия). Процедура также позволяет вывести график остатков и сохранить остатки в отдельной переменной базы данных пакета.

Хэксп	Уэксп	Урегр	остаток	Ст.остат	Ст.ошиб	Довер.инт
0	29.375	30.124	-0.74895	-1.0424	0.99159	2.7544
0.2	110.26	109.56	0.69916	0.97312	0.9142	2.5394
0.4	188.76	188.99	-0.23572	-0.32809	0.87294	2.4248
0.6	268.92	268.43	0.48839	0.67976	0.87294	2.4248
0.8	348.54	347.86	0.6785	0.94436	0.9142	2.5394
1	426.42	427.3	-0.88138	-1.2267	0.99159	2.7544

Рис. 8.22. Результаты анализа остатков

Процедура 2 = **Прогнозирование** позволяет получить прогноз вперед с заданным числом точек прогноза и шагом прогноза (см. рис. 8.23).

Хпрогн	Упрогн	Ст.ошиб	Довер.инт
1.1	467.02	1.0415	2.8929
1.2	506.73	1.0975	3.0485
1.3	546.45	1.1588	3.2187
1.4	586.17	1.2245	3.4014
1.5	625.89	1.2941	3.5945
1.6	665.6	1.3668	3.7966
1.7	705.32	1.4423	4.0063
1.8	745.04	1.5201	4.2224
1.9	784.76	1.5999	4.4439
2	824.47	1.6813	4.6702

Рис. 8.23. Результаты вычисления прогноза

На рис. 8.23 приведены результаты расчетов этой процедуры при числе точек прогноза 10 и шаге прогноза 0.1. Первое значение Хпрогн равно сумме максимального значения наблюденной независимой переменной и величины шага прогноза. В столбце Довер.инт фигурирует величина допустимого отклонения от прогноза при 95% уровне доверия. Процедура также строит график прогноза, он приведен на рис. 8.21 справа.

Комментарии. 1. Количества наблюдений в зависимой и независимой переменных должны быть одинаковыми.

2. Большинство функций, обрабатываемых процедурой однопараметрической регрессии (см. рис. 8.17), являются нелинейными относительно входящих в них параметров. В таких случаях для решения задачи регрессии возможны два подхода. Наиболее общий из них сводится к применению нелинейного метода наименьших квадратов для нахождения оценок неизвестных параметров в модели (8.3) с аддитивной ошибкой. Другой, частный метод, основан на преобразовании векторов зависимой и независимой переменных таким образом, чтобы преобразованная функциональная зависимость была линейной относительно параметров. Например, для функции $y = 1/(a + b\sqrt{x})$ преобразование вектора y вида $u_i = 1/y_i$ переводит ее в линейную относительно параметров функцию. Аналогичные преобразования допустимы для большинства функций, указанных в списке (кроме функций 7, C, I, L, M) . Для этих функций программа сначала осуществляет необходимые преобразования векторов независимой и зависимой переменных x и y, а затем применяет к преобразованной модели стандартный метод наименьших квадратов для нахождения оценок параметров. Но в связи с этим следует помнить, что требования аддитивности, одинаковой распределенности и нормальности случайной ошибки ε_i относятся к преобразованной модели, а не к первоначальной. Более того, за исключением нескольких частных случаев, сформулировать статистические требования к характеру случайной ошибки в исходной модели крайне трудно. Поэтому последующий анализ остатков для первоначальной модели не имеет смысла и необходимо исследовать остатки преобразованной модели. Мы специально обращаем на это внимание, так как процедура однопараметрической регрессии выводит график и предусматривает возможность сохранения остатков только для первоначальной модели.

Для зависимостей с номерами 7, C, I, L, M, не допускающих сведения к линейным относительно параметров функциям, программа использует нелинейный метод наименьших квадратов. Для этих зависимостей дальнейший анализ остатков можно проводить стандартными способами.

3. Чтобы пользователю было легче выбрать наиболее подходящую зависимость, во встроенном справочнике пакета STADIA дается краткая классификация различных функциональных зависимостей, представленных в процедуре простой регрессии, с точки зрения скорости изменения (поведения производных), максимумов, асимптот, периодичности и т.п.

4. Пакет STADIA позволяет использовать и другие модели однопараметрической регрессии. Выбрав пункт О=Общая/нелиннейная модель в меню выбора статистических методов, Вы можете задать вид зависимости формулой.

Пример 8.3к. Методом наименьших квадратов вычислим оценки параметров параболической модели регрессии $y = A + Bx + Cx^2$ для данных примера 8.2к из п. 8.6.1.

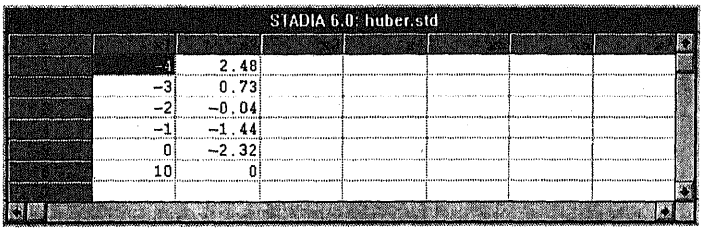


Рис. 8.24. Данные для примера 8.3к

Подготовка данных. На рис. 8.24 приведен экран редактора базы данных пакета с введенными данными примера.

Выбор процедуры. В меню Статистические методы (рис. 1.17) в разделе Регрессионный анализ выберите пункт L = Простая регрессия/тренд. Так же, как в примере 8.1к, укажем в ответ на запрос программы номера независимой и зависимой переменных. Из меню моделей регрессии (рис. 8.17) выберем нажатием кнопки [2] параболическую модель.

Результаты. На рис. 8.25 приведены результаты расчетов процедуры.

```
ПРОСТАЯ РЕГРЕССИЯ.  Файл:huber.std

                                Переменные: x1, x2

          Модель: парабола  Y = a0+a1*x+a2*x^2
  Коэфф.        a0          a1          a2
Значение   -2.2665     -0.77426     0.10007
Ст.ошиб.   0.15447     0.044868    0.058925
 Значим.    0.0006       0.0004      0.0004

 Источник  Сум.квадр.  Степ.св  Средн.квадр.
Регресс.      13.947         2       6.9734
Остаточн      0.1361         3     0.045366
     Вся      14.083         5

Множеств R      R^2    R^2прив   Ст.ошиб.        F     Значим
  0.99516   0.99034    0.98389    0.21299   153.72     0.0009
  Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>
```

Рис. 8.25. Результаты вычислений процедуры параболической регрессии

Дальнейший порядок работы может быть таким же, как в предыдущем примере.

Глава 9

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 17 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 не оправдано, 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, то есть 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 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 пунктах 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 Кендэлла 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 [72], 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 высокореактивных (+P) и низкореактивных (−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 на 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, $A_2, \ldots, A_r$, признак B подразделяется на s градаций B_1, $B_2, \ldots, B_s$. 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):

$$P(A_iB_j) = P(A_i)\,P(B_j) \tag{9.1}$$

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.:

$$P(B_j/A_i) = P(B_j) \tag{9.2}$$

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_iB_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:

$$\sum_{j=1}^{s} n_{ij} = n_{i1} + n_{i2} + \cdots + n_{is}. \tag{9.3}$$

Обозначим эту сумму через $n_{i\cdot}$. Аналогично, частота появления B_j равна

$$n_{\cdot j} = n_{1j} + n_{2j} + \cdots + n_{rj}\,. \tag{9.4}$$

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

$$p_{ij} = P(A_i B_j). \tag{9.5}$$

Теперь

$$P(A_i) = \sum_{j=1}^{s} p_{ij} = p_{i.}, \quad P(B_j) = \sum_{i=1}^{r} p_{ij} = p_{.j}. \tag{9.6}$$

Гипотеза о независимости признаков в принятых обозначениях записывается так:

$$p_{ij} = p_{i.} p_{.j} \tag{9.7}$$

для всех пар (i, j), $i = 1, \ldots, r$, $j = 1, \ldots, s$.

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 таблицы 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 $n \to \infty$:

$$\frac{n_{ij}}{n} \to p_{ij}; \qquad \frac{n_{i\cdot}}{n} \to p_{i\cdot}; \qquad \frac{n_{\cdot j}}{n} \to p_{\cdot j}, \tag{9.8}$$

а поэтому для независимых признаков: $n_{ij} \simeq n_{i\cdot} n_{\cdot j}/n$.

Определение. *Величины $n_{i\cdot} n_{\cdot j}/n$ 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 таблица 9.1 принимает вид:

Таблица 9.3

Предпочтение различных видов инструкций в группах высокореактивных ($+P$) и низкореактивных ($-P$) 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
	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:

$$X^2 = \sum \frac{(H-T)^2}{T}, \tag{9.9}$$

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_iB_j. Оно равно rs. Подсчитаем *число связей*. Во-первых, $\sum_{i,j} n_{ij} = n$ (одна связь). Во-вторых, определяя $n_{i\cdot}$ (и $n_{\cdot j}$), мы воспользовались соотношениями

$$\sum_{j=1}^{s} n_{ij} = n_{i\cdot} \qquad \text{и} \qquad \sum_{i=1}^{r} n_{ij} = n_{\cdot j}.$$

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 суще-

ствованию 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:

$$Y^2 = 2\sum H \ln \frac{H}{T}. \tag{9.10}$$

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: $n_{i\cdot}n_{\cdot j}/n \geqslant 3$. 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:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}, \tag{9.11}$$

$$Y^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} \ln\left(\frac{n_{ij}}{\frac{n_{i\cdot}n_{\cdot j}}{n}}\right). \tag{9.12}$$

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 (приближенно, 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:

$$\frac{X^2}{n} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(\frac{n_{ij}}{n} - \frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}} \simeq \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}. \qquad (9.13)$$

Если гипотеза H_0 неверна (и только тогда), правая часть (9.13) отлична от нуля. В этом случае X^2 стремится к бесконечности (при $n \to \infty$). 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 $P\{\chi^2 \geqslant X^2\} < 0.005$. 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

$$\lim_{n\to\infty} \frac{1}{n} X^2 = \sum_{i,j} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

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 $(n \to \infty)$ $Q \to (p_{11}p_{22} - p_{12}p_{21})/(p_{11}p_{22} + p_{12}p_{21})$.

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 другие. Подробную информацию по изложенным выше вопросам можно найти в [6], [40], [64], [78], [82].

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) $X_1, \ldots, X_n$ измерения определенного признака A для объектов $O(1), \ldots, O(n)$, сколько ранги $r_1, \ldots, r_n$ чисел $X_1, \ldots, X_n$. (Здесь r_i — ранг X_i среди чисел $X_1, \ldots, X_n$.)

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, $X_1, \ldots, X_n$ суть значения признака A для объектов $O(1), \ldots, O(n)$, а $Y_1, \ldots, Y_n$ — значения признака B для тех же объектов. Каждый объект $O(i)$, $i = 1, \ldots, n$, теперь характеризуется парой чисел (X_i, Y_i) — своими значениями признаков A и B. От чисел $Y_1, \ldots, Y_n$ (так же как ранее для признака A) переходим к их рангам $s_1, \ldots, s_n$. (Здесь s_i — ранг Y_i среди $Y_1, \ldots, Y_n$). Будем считать, что среди чисел $X_1, \ldots, X_n$ (и среди чисел $Y_1, \ldots, Y_n$) 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 $x_1, \ldots, x_n$, в определенной степени влияет на порядок, в котором следуют числа $y_1, \ldots, y_n$. Иными словами, последовательность рангов $r_1, \ldots, r_n$ в какой-то мере влияет на ранговую последовательность $s_1, \ldots, s_n$. Чем более тесно связаны эти признаки, тем в большей степени последовательность $r_1, \ldots, r_n$ предопределяет последовательность $s_1, \ldots, s_n$.

Если же признаки такой связи не проявляют, то порядок среди игреков *случаен* по отношению к порядку среди иксов. В этом случае все $n!$ перестановок чисел 1, 2, ..., n, которые могут выступать как ранги $s_1, \ldots, s_n$, оказываются равновозможными, т.е. равновероятными при любом порядке чисел $r_1, \ldots, r_n$. Это центральный момент обсуждения: при гипотезе H_0 и любом наборе $r_1, \ldots, r_n$ все возможные последовательности $s_1, \ldots, s_n$ 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 Кендэлла, соответственно. С этими ранговыми коэффициентами мы уже встречались в параграфе 8.4. Начнем с той меры, которую предложил Ч.Спирмен.

Коэффициент Спирмена. Близость двух рядов чисел $r_1, \ldots, r_n$ и $s_1, \ldots, s_n$ 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 $(r_1, \ldots, r_n)$ и $(s_1, \ldots, s_n)$, 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: $|\rho| \leqslant 1$. 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 $1, 2, \ldots, n$. Вторую последовательность обозначим, скажем, $z_1, \ldots, z_n$. При этом

$$S = \sum_{i=1}^{n} (r_i - s_i)^2 = \sum_{k=1}^{n} (k - z_k)^2.$$

Коэффициент Кендэлла. Другой коэффициент ранговой корреляции получил популярность после работ М.Кендэлла, в особенности после выхода его книги [40]. 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 Кендэллу сначала введем статистику Кендэлла K. Выберем в качестве первона-

чальной нумерации упорядочение объектов по признаку A и подсчитаем K, сопоставляя $(1,\ 2, \ldots, n)$ и $(z_1,\ z_2, \ldots, z_n)$. Оказывается, что K равно числу *инверсий* в ряду $(z_1, \ldots, z_n)$. Пусть, например, $n = 4$ и $(z_1, \ldots, z_4) = (4, 3, 1, 2)$. 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 Кендэллу):

$$\tau = 1 - \frac{4K}{n(n-1)}.$$

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 $(z_1, \ldots, z_n)$ 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 [25], [82], [89], [91].

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 главы 1 мы знаем, что для независимых случайных величин α, β их ковариация

$$\operatorname{cov}(\alpha, \beta) = M\alpha\beta - M\alpha\, M\beta$$

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 $\operatorname{cov}(\alpha, \beta)$, а безразмерную величину — коэффициент корреляции $\rho(\alpha, \beta)$:

$$\rho = \frac{\operatorname{cov}(\alpha, \beta)}{\sqrt{D\alpha}\sqrt{D\beta}}. \quad (9.15)$$

Свойства коэффициента корреляции мы уже описывали в главе 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 $(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n)$, которую получают, наудачу 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 главе 1, его

вычисляют, заменяя усреднения по генеральной совокупности (математические ожидания) усреднениями по выборке. Выборочные аналоги для дисперсий, согласно п. 1.8, суть:

$$s_\alpha^2 = \frac{1}{n}\sum_{i=1}^{n}(\alpha_i - \overline{\alpha})^2, \qquad s_\beta^2 = \frac{1}{n}\sum_{i=1}^{n}(\beta_i - \overline{\beta})^2.$$

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

$$r = \frac{\overline{\alpha\beta} - \overline{\alpha}\,\overline{\beta}}{s_\alpha s_\beta}. \tag{9.16}$$

Можно r выразить и по-другому, например:

$$r = \frac{\sum_{i=1}^{n}(\alpha_i - \overline{\alpha})(\beta_i - \overline{\beta})}{\sqrt{\sum_{i=1}^{n}(\alpha_i - \overline{\alpha})^2}\sqrt{\sum_{i=1}^{n}(\beta_i - \overline{\beta})^2}}. \tag{9.17}$$

В силу закона больших чисел $r \to \rho$ при неограниченном росте объема выборки, т.е. при $n \to \infty$. 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 [16]. 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 $n \geqslant 20$ 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 ею 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 [16] 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 [36], [82].

В общем случае, к сожалению, методы анализа связи признаков становятся гораздо сложнее, чем приведенные выше. Для упрощения исследования часто приходится одну из шкал измерений приходится 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 пакетах STADIA и STATGRAPHICS

9.7.1. Пакет STADIA

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. В редакторе базы данных пакета введем значения из табл. 9.1, как это показано на рис. 9.1. Пакет предполагает, что данные уже сведены в таблицу сопряженности, то есть находятся в матрице размера $m \times n$, где столбцы отвечают различным значениям первого признака, а строки — различным значениям второго признака. При этом каждый элемент матрицы указывает число объектов с данным сочетанием признаков. Другая возможная форма ввода данных, когда они не сведены в таблицу сопряженности, описана в комментариях.

Рис. 9.1. Пакет STADIA. Экран блока редактора данных с загруженной таблицей сопряженности

Выбор процедуры. В меню Статистические методы выберем пункт A=Кросстабуляция (см. рис. 1.17).

Результаты. Выдача результатов процедуры включает 6 таблиц, соответствующих по размеру матрице кросстабуляции, в которых приведены следующие данные:

1) наблюденные частоты признаков x_{ij};
2) процентные частоты признаков для рядов;
3) процентные частоты признаков для столбцов;
4) процентные частоты признаков для всей таблицы;
5) ожидаемые частоты признаков в случае их независимости E_{ij};
6) остаточные частоты $x_{ij} - E_{ij}$.

Далее выдаются значения статистики хи-квадрат, ее уровень значимости, число степеней свободы для проверки гипотезы о независимости признаков, а также ряд статистик, используемых для оценки различных аспектов понятия связи между двумя номинальными переменными. Указанные результаты приведены на рис. 9.2–9.3. Полученный уровень значимости (0.0019) статистики хи-квадрат позволяет отвергнуть гипотезу о независимости признаков. Кроме того, процедура предлагает графическое представление матрицы кросстабуляции.

В наши задачи не входит подробный разбор назначения всех представленных мер связи признаков (см. [6], [40], [64], [78]). Отметим лишь удобство и исчерпывающий характер работы этой процедуры.

Замечание. Если исходные данные не сведены в таблицу кросстабуляции, а представляют собой значения парных переменных для n объектов, то процедура сама выполняет предварительное кросстабулирование. Критерием запуска

```
 Файл:                                Переменных=2          Измерений=4
КРОССТАБУЛЯЦИЯ.   Файл:chizh
       Наблюденные частоты признаков:
       63        42|       105
       34        56|        90
 ------------------|
       97        98|       195

       Процентная встречаемость признаков по рядам:
       60        40
  37.778    62.222

       Процентная встречаемость признаков по столбцам:
  64.948    42.857
  35.052    57.143

       Общая процентная встречаемость признаков:
  32.308    21.538|     53.846%
  17.436    28.718|     46.154%
 ------------------|
  49.744%   50.266%
```

Рис. 9.2. Результаты вычисления частот признаков

```
     Ожидаемые частоты признаков:
  52.231    52.769
  44.769    45.231

     Остаточные частоты признаков (набл-ожид):
  10.769   -10.769
 -10.769    10.769

Хи-квадрат =9.5729, Значимость=0.0019, степ.своб = 1
   Гипотеза 1: <Есть связь между признаками>

Коэфф. Фи =0.22157

Коэфф.сопряж. Пирсона =0.21632

V-коэфф. Граммера =0.22157

Лямбда Гудмана и Крускала: симметр, ряд, столб =0.18717, 0.15556, 0.21649

Тау-b Кендала =0.22157

Тау-c Кендала =0.22091

Гамма Гудмана и Кендала =0.42373

d(x,y) Соммера=0.22222, 0.22091
```

Рис. 9.3. Результаты проверки гипотезы о независимости признаков

этого преобразования является наличие в матрице только двух переменных с числом значений больше 5. Поэтому во избежание коллизии таблицы кросстабуляции, имеющие для одного признака две градации, а для другого — больше, необходимо размещать в матрице данных горизонтально. Исходные парные переменные должны иметь целочисленные положительные значения, максимальное из которых не превосходит n, где n — число значений, в противном случае операция кросстабулирования будет прервана с ошибкой. Кросстабулирование можно произвести также и самостоятельно посредством одноименной операции в блоке «Преобразования данных».

Следующий пример посвящен задаче выявления связи признаков, измеренных в порядковых или количественных шкалах.

Пример 9.2к. С помощью коэффициентов корреляции Спирмена, Кендэлла и Пирсона выясним связь между скоростями реакции на звук и на свет по данным табл. 3.1.

Подготовка данных. Указанные данные уже рассматривались нами в примере 3.3к. (Экран редактора базы данных с частью введенных данных таблицы 3.1 приведен на рис. 3.5.) Как и прежде, будем считать, что они находятся в двух переменных sound и light файла SOUND.

Выбор процедуры. Для вычисления коэффициентов ранговой корреляции в меню Статистические методы следует выбрать пункт 9 = Корреляция (независимость), а для вычисления коэффициента корреляции Пирсона — пункт 3 = Корреляция.

Заполнение полей ввода данных. Порядок работы двух указанных выше процедур совпадает. В окне Анализ переменных (рис. 9.4) надо выбрать переменные для анализа. Для этого следует выделить мышью в поле Переменные переменные sound и light и, нажав кнопку со стрелкой вправо, перенести их в поле Для анализа. Затем надо нажать кнопку запроса [Утвердить].

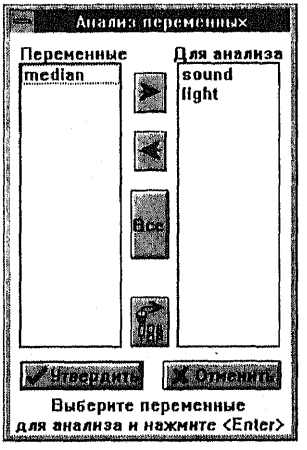


Рис. 9.4. Запрос выбора переменных для анализа

Результаты. На рис. 9.5 представлены результаты процедуры 9 = Корреляция (независимость).

```
НЕПАРАМЕТРИЧЕСКАЯ КОРРЕЛЯЦИЯ.  Файл: sound.std
                              Переменные: sound, light
Кендал=0.22059, Z=1.2358, Значимость=0.1082, степ.своб=17
   Гипотеза 0: <Нет корреляции между выборками>
Спирмен=0.27696,  Z=1.0844, Значимость=0.139, степ.своб=17
   Гипотеза 0: <Нет корреляции между выборками>
```

Рис. 9.5. Результаты непараметрической проверки независимости признаков

Они включает значения коэффициентов Кендэлла и Спирмена, значения их нормальной аппроксимации Z и уровни значимости, вычисленные с помощью указанной нормальной аппроксимации, для проверки

гипотезы о равенстве коэффициентов нулю против односторонних альтернатив. (Для получения уровня значимости критерия против двусторонних альтернатив следует увеличить полученный выше уровень значимости вдвое.) Запись степ.своб= трактуется как объем анализируемых переменных.

Сравнивая полученные уровни значимости с пятипроцентным, система выдает заключение о принятии или отвержении гипотезы о равенстве коэффициентов нулю.

На рис. 9.6 представлены результаты процедуры 3 = Корреляция.

```
ПАРАМЕТРИЧЕСКАЯ КОРРЕЛЯЦИЯ.  Файл: sound.std
                              Переменные: sound, light
Коэфф.корреляции=0.21324 T:=0.84531, Значимость=0.5841, степ.своб = 15
   Гипотеза 0: <Коэффициент корреляции не отличен от нуля>
```

Рис. 9.6. Результаты параметрической проверки независимости признаков

Процедура вычисляет значение коэффициента корреляции Пирсона, статистику Стьюдента, используемую для проверки гипотезы о равенстве нулю коэффициента корреляции и ее уровень значимости. Сравнивая полученный уровень значимости с пятипроцентным, система выдает заключение о принятии или отвержении гипотезы о равенстве коэффициентов нулю.

9.7.2. Пакет STATGRAPHICS

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. Объем данных этого примера невелик и будет введен в процедуру непосредственно с клавиатуры, как это показано на рис. 9.8.

Выбор процедуры. В головном меню пакета выберем пункт P. Categorical Data Analysis (анализ категорий), а в появившемся на экране меню (рис. 9.7) — процедуру 2. Contingency Tables (таблицы сопряженности). Назначение других пунктов этого меню смотри в комментариях.

Заполнение полей ввода данных. В поле ввода данных Table matrix (матрица данных) надо ввести данные из табл. 9.1, как это показано на рис. 9.8. Оператор пакета RESHAPE превращает вектор данных в его правой части в матрицу с числом строк и столбцов, указанных в его левой части.

```
CATEGORICAL DATA ANALYSIS

1.  Crosstabulation
2.  Contingency Tables
3.  Chi-Square Goodness-of-Fit Statistic
4.  Log-Linear Analysis
5.  Numeric Coding of Classification Factors
6.  Recoding Variables
```

Рис. 9.7. Меню процедур анализа таблиц сопряженности

```
Contingency Tables

Table matrix:  2 2 RESHAPE 63 42 34 56
```

Рис. 9.8. Ввод данных для анализа таблицы сопряженности

Результаты. После заполнения полей ввода и нажатия клавиши F6 на экран выводятся значение статистики хи-квадрат (Chi-square), ее число степеней свободы (D.F.) и уровень значимости (Significance), а также значения различных мер связи признаков (рис. 9.9). Описание этих статистик можно найти в [6], [40], [64], [78], [82].

Summary Statistics for Contingency Tables (Page 1)

Chi-square	D.F.	Significance
9.57290	1	1.97470E-3
8.70462	1	3.17404E-3 with Yates correction

Statistic	Symmetric	With rows dependent	With columns dependent
Lambda	0.18717	0.15556	0.21649
Uncertainty Coeff.	0.03580	0.03587	0.03572
Somer's D	0.22157	0.22091	0.22222

Рис. 9.9. Результаты проверки сопряженности признаков

После нажатия Enter будет продолжен вывод результатов работы процедуры (рис. 9.10).

Summary Statistics for Contingency Tables (Page 2)

Statistic	Value	Significance
Contingency Coeff.	0.21632	
Cramer's V	0.22157	
Conditional Gamma	0.42373	
Kendall's Tau B	0.22157	0.00203
Kendall's Tau C	0.22091	

Рис. 9.10. Результаты проверки сопряженности признаков (продолжение)

Полученный уровень значимости статистики хи-квадрат (менее 0.002) говорит, что гипотеза о независимости признаков должна быть отвергнута.

Комментарии. 1. Процедура 1. Crosstabulation (кросстабуляция) (рис. 9.11) позволяет сформировать и проанализировать таблицу сопряженности из данных, имеющих несколько (до девяти) факторов классификации. Разобранная процедура 2. Contingency Tables является частным случаем указанной выше процедуры, когда данные уже оформлены в виде таблицы сопряженности.

2. Более детальный анализ многофакторной таблицы с целью получения описания взаимосвязей факторов осуществляет процедура 4. Log-Linear Analysis (лог-линейный анализ).

Следующий пример посвящен задаче выявления связи признаков, измеренных в порядковых или количественных шкалах.

Пример 9.2к. С помощью коэффициентов корреляции Спирмена, Кендэлла 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 3.3к. Вид экрана редактора базы данных с частью введенных данных таблицы 3.1 приведен на рис. 3.5. Как и прежде будем считать, что данные находятся в двух переменных sound и light файла SPEEDR.

Выбор процедуры. Для вычисления ранговых коэффициентов корреляции Спирмена и Кендэлла выберем в головном меню пакета пункт R. Nonparametric Methods (непараметрические методы), а в меню указанного пункта (рис. 3.10) — процедуру 5. Rank Correlation Coefficients (коэффициенты ранговой корреляции).

Для вычисления коэффициента корреляции Пирсона следует выбрать в головном меню пакета пункт Q. Multivariate Methods (многомерные методы), а меню этого пункта — процедуру 1. Correlation Analysis (корреляционный анализ).

Заполнение полей ввода данных. Порядок ввода данных в обе указанные процедуры почти полностью совпадает и будет разобран на примере процедуры 5. Rank Correlation Coefficients. На рис. 9.11 приведен экран ввода данных, который предлагает ввести в поле Data vectors анализируемые вектора данных. Поле Missing values (пропущенные значения) предназначено для указания режима обработки пропущенных значений в векторах данных, в нем может фигурировать один из двух возможных режимов: Listwise или Pairwise. Их различия описаны в комментариях. При сравнении только двух векторов данных оба режима приводят к одному и тому же результату. В поле Procedure указывается, какой именно коэффициент ранговой корреляции будет вычисляться. (Для процедуры 1. Correlation Analysis это поле отсутствует.)

Rank Correlation Coefficients

Data vectors: SPEEDR sound
SPEEDR light

Missing values: Listwise

Procedure: Spearman

Рис. 9.11. Запрос параметров процедуры ранговой корреляции

Chisquare Test

Spearman Rank Correlations

	sound	light
sound	1.0000	.2761
	(17)	(17)
	1.0000	.2695
light	.2761	1.0000
	(17)	(17)
	.2695	1.0000

Coefficient (sample size) significance level

Рис. 9.12. Результаты вычисления коэффициента корреляции Спирмена

Результаты. На рис. 9.12–9.14 приведены результаты работы процедур при вычислении коэффициентов корреляции Спирмена, Кендэлла и Пирсона соответственно.

Kendall Rank Correlations

	sound	light
sound	1.0000	.2222
	(17)	(17)
	1.0000	.2073
light	.2222	1.0000
	(17)	(17)
	.2073	1.0000

Coefficient (sample size) significance level

Рис. 9.13. Результаты вычисления коэффициента корреляции Кендэлла

Sample Correlations

	sound	light
sound	1.0000	.2132
	(17)	(17)
	.0000	.4112
light	.2132	1.0000
	(17)	(17)
	.4112	.0000

Coefficient (sample size) significance level

Рис. 9.14. Результаты вычисления коэффициента корреляции Пирсона

Форма выдачи результатов этих процедур, как видно из рис. 9.12–9.14, одна и та же. Она включает в себя матрицу корреляции, элемента-

ми которой служат коэффициенты корреляции для различных сочетаний переменных. Первые строки (для каждой переменной) содержат значения коэффициентов корреляции, далее в скобках — число наблюдений, участвовавших в обработке, а в третьей строке — уровень значимости полученного коэффициента корреляции в случае справедливости гипотезы о равенстве нулю коэффициента корреляции. Уровень значимости указан против двусторонних альтернатив. Из определения коэффициента корреляции следует, что коэффициент корреляции переменной с самой собой всегда равен единице. Так как порядок переменных при вычислении коэффициента корреляции не существен, то выводимая матрица коэффициентов корреляции всегда будет симметричной. Поэтому в случае двух переменных информативным является только одно значение коэффициента корреляции из четырех, присутствующих в матрице.

Полученные значения коэффициентов корреляции (и их уровни значимости) не позволяют отвергнуть гипотезу о равенстве их нулю.

Заметим, что использование выданного процедурой уровня значимости коэффициента корреляции Пирсона для дальнейших статистических выводов в разбираемом примере неправомерно. Это связано с тем, что анализируемые данные не являются независимыми, а имеют структуру парных данных.

При завершении работы процедур пакет предлагает сохранить вычисленную матрицу корреляций в базе данных для дальнейшего использования ее в процедурах факторного анализа и метода главных компонент.

Комментарии. 1. Обе описанные процедуры позволяют вычислять коэффициенты корреляции для всех возможных пар переменных, указанных в поле Data vectors. Результатом подобных вычислений является корреляционная матрица.

2. При одновременном анализе более двух переменных представим их для удобства в виде матрицы, столбцами которой являются указанные переменные. Режим обработки пропущенных значений Listwise соответствует игнорированию всех строк в матрице данных, в которых есть хоть одно пропущенное значение. Режим обработки пропущенных значений Pairwise осуществляет попарное исключение данных с пропусками из тех переменных, корреляция которых вычисляется. Тем самым в обработке может участвовать больше данных, чем в первом случае.

Глава 10

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 удается 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 выборке[1].

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 (пп. 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 пп. 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 главе 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 $F_n(x) \to F(x)$ при $n \to \infty$. Для количественного выражения сходства функций F_n и F используют различные способы, о которых будет говориться ниже.

Статистика Колмогорова. Для выражения сходства функций можно использовать то или иное расстояние между этими функциями. Например, можно сравнить F_n и F в равномерной метрике, т.е. рассмотреть величину:

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F(x)|. \qquad (10.1)$$

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 $n \to \infty$, то $F_n(x) \to F(x)$ при всяком x. Поэтому естественно, что при этих условиях $D_n \to 0$. Если же гипотеза H неверна, то $F_n \to G$ и $G \neq F$, а потому $\sup_{-\infty<x<\infty} |F_n(x) - F(x)| \to \sup_x |G(x) - F(x)|$. 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 [16], 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 $D_n \to 0$ 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 $n \to \infty$ имеет предел, и дает его выражение:

$$\lim_{n\to\infty} P(\sqrt{n}\,D_n < z) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2z^2}. \qquad (10.2)$$

В сборниках таблиц можно найти значения функции (10.2) (см., например, [16], 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

$$D_n = \max_{1\leqslant k\leqslant n}\left[\frac{k}{n} - F(x_{(k)}),\ F(x_{(k)}) - \frac{k-1}{n}\right]. \qquad (10.3)$$

Здесь через $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ 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*:

$$\omega_n^2 = \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \, dF(x) . \tag{10.4}$$

Для вычисления ω_n^2 по реальной выборке можно использовать формулу:

$$n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[F(x_{(i)}) - \frac{2i-1}{2n} \right]^2 . \tag{10.5}$$

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 $n \to \infty$) 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, [16], 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 года «Об одном новом подтверждении законов Менделя» в [43]. Мы воспроизведем изложение этой работы по брошюре В.Н.Тутубалина [74].

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 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 $n_1,\ n_2, \ldots,\ n_k$ численности проявления рецессивного признака — $\mu_1,\ \mu_2, \ldots,\ \mu_k$, 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 $\mu_1^*,\ \mu_2^*, \ldots,\ \mu_k^*$, можно смотреть (если модель Менделя верна) как на выборку, теоретическое распределение которой есть стандартный нормальный закон.

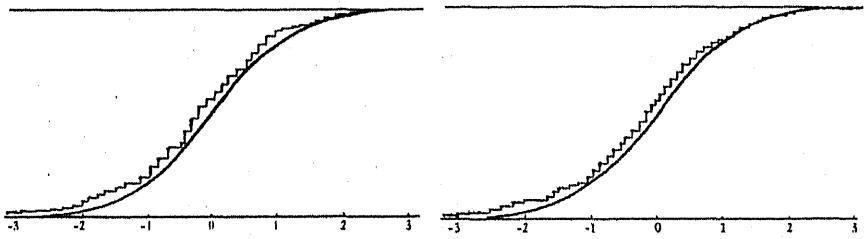


Рис. 10.1. Эмпирическая и теоретическая функции распределения: слева — для первой выборки (k = 98), справа — для второй выборки (k = 123)

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 параграфе 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 наблюденные 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, $A_1, \ldots, A_r$. Пусть $p_1, \ldots, p_r$ — вероятности этих исходов, причем $p_1 + \cdots + p_r = 1$. Обозначим через $m_1, \ldots, m_r$ количества опытов, заканчивающихся, соответственно, исходами $A_1, \ldots, A_r$. (Ясно, что $m_1 + \ldots + m_r = n$.) Введем случайную величину

$$\chi^2 = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i}.$$

Тогда справедливо следующее утверждение: при $n \to \infty$ 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 $p_1, \ldots, p_r$ приняли определенные зна-

чения $p_1^o, \ldots, p_r^o$. Далее будем называть это гипотезой H:

$$H: \; p_1 = p_1^o, p_2 = p_2^o, \ldots, p_r = p_r^o,$$

Рассмотрим статистику:

$$X^2 = \sum_{i=1}^{r} \frac{(m_i - np_i^o)^2}{np_i^o} = n \sum_{i=1}^{r} \left(\frac{m_i}{n} - p_i^o\right)^2 / p_i^o. \qquad (10.6)$$

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 $\left(\frac{m_1}{n}, \ldots, \frac{m_r}{n}\right)$ и вектором вероятностей $(p_1^o, \ldots, p_r^o)$. От евклидового 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 $n \to \infty$ указывает теорема К.Пирсона. Чтобы понять, что происходит с (10.6), когда H неверна, заметим, что по закону больших чисел $m_i/n \to p_i$ при $n \to \infty$, для $i = 1, \ldots, r$. Поэтому при $n \to \infty$:

$$\sum_{i=1}^{r} \left(\frac{m_i}{n} - p_i^o\right)^2 / p_i^o \to \sum_{i=1}^{r} (p_i - p_i^o)^2 / p_i^o.$$

Эта величина равна 0, только если $p_i = p_i^o$ для всех i. Поэтому если H неверна, то $X^2 \to \infty$ (при $n \to \infty$).

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 наблюденное значение X^2 превосходит критическое значение, которое в данном случае можно взять из таблиц распределения хи-квадрат, Иначе говоря, вероятность $P(\chi^2 \geqslant X^2)$ — 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 $p_1, \ldots, p_r$. Поэтому нельзя сказать, к примеру, что ста наблюдений будет достаточно, поскольку не только n должно быть велико, но и произведения $np_1, \ldots, np_r$ (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 $np_i \geqslant 10$. 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 наблюденную 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 $x_1, \ldots, x_n$ — 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 $x_1, \ldots, x_n$. Для этого можно использовать разные методы оценивания (см. главу 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 $x_1, \ldots, x_n$ 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*:

$$\hat{D}_n = \sup_x |\, F(x) - F(x, \hat{\theta}_n) \,|\,. \tag{10.7}$$

Аналогично, *модифицированная статистика омега-квадрат* есть:

$$\widehat{\omega}_n^2 = \int_{-\infty}^{+\infty} [F_n(x) - F(x, \hat{\theta}_n)]^2 \, dF(x, \hat{\theta}_n)\,. \tag{10.8}$$

Свойства. Свойства статистик $\hat{D}_n$ и $\widehat{\omega}_n^2$ во многом повторяют отмеченные ранее свойства статистик D_n и ω_n^2. В частности, $\sqrt{n}\hat{D}_n$ и $n\widehat{\omega}_n^2$ неограниченно возрастают, если проверяемая гипотеза неверна. Поэтому эту гипотезу следует отвергнуть, если наблюденное значение $\sqrt{n}\hat{D}_n$ (или $n\widehat{\omega}_n^2$, 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 $\widehat{\omega}_n^2$ — 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 $\widehat{\omega}_n^2$ 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 (смотри, например, [36]). Большинство из них рассчитаны методом случайных испытаний (методом Монте-Карло). Автор большинства этих расчетов М.Стефенс

(M.Stephens) заметил, что зависимость результатов от объема выборки резко уменьшается, если вместо $\hat{D}_n$, $\widehat{\omega}_n^2$ 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
$\widehat{\omega}_n^2$:	$\widehat{\omega}_n^2\left(1+\frac{0.5}{n}\right)$	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
$\widehat{\omega}_n^2$:	$\widehat{\omega}_n^2\left(1+\frac{0.16}{n}\right)$	0.149	0.177	0.224	0.273	0.337

Приближенные формулы. Предельное (при $n \to \infty$) распределение $n\widehat{\omega}_n^2$ 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,

$$\lim_{n\to\infty} P\left\{\sqrt{n}\sup\left|F_n(x)-\Phi\left(\frac{x-\overline{x}}{s}\right)\right|>z\right\}\simeq 2\sqrt{\frac{2\pi}{\pi-2}}\,\exp\left\{-\frac{2\pi}{\pi-2}z^2\right\}$$

для больших $z > 0$ (т.е. для $z \to \infty$).

Если же математическое ожидание известно и равно, скажем, a, то по выборке приходится оценивать только дисперсию. В этом случае для больших $z > 0$

$$\lim_{n\to\infty} P\left\{\sqrt{n}\sup\left|F_n(x)-\Phi\left(\frac{x-a}{s}\right)\right|>z\right\}\simeq\frac{2\sqrt{6}}{3}e^{-2z^2}.$$

Эти приближенные формулы дают хорошие результаты для малых вероятностей и больших объемов выборок, то есть для вероятностей, начиная примерно с 0.20 (и меньше) и для объемов n, начиная примерно с 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 9.2.

Теорема Фишера. Пусть n — число независимых повторений опыта, который может заканчиваться одним из r (r — произвольное натуральное число) элементарных исходов, скажем, $A_1, \ldots, A_r$. Пусть вероятности этих элементарных исходов известны с точностью до некоторого неопределенного, скажем, k-мерного параметра $\theta = (\theta_1, \ldots, \theta_k)$. Тогда эти вероятности являются функциями от θ: $P(A_i) = p_i(\theta)$. Мы будем предполагать, что функции $p_1(\theta), \ldots, p_r(\theta)$ заданы, дифференцируемы, $\sum_{i=1}^{r} p_i(\theta) = 1$ для всякого θ, а параметр θ изменяется в ограниченной области пространства. Тогда при $n \to \infty$ статистика:

$$X^2 = \min_{\theta} \sum_{i=1}^{r} \frac{[m_i - np_i(\theta)]^2}{np_i(\theta)} \tag{10.9}$$

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

$$X^2 = \sum_{i=1}^{r} \frac{[m_i - np_i(\hat{\theta}_n)]^2}{np_i(\hat{\theta}_n)}, \tag{10.10}$$

где $\hat{\theta}_n$ — оценка наибольшего правдоподобия для параметра θ, найденная по частотам $m_1, \ldots, m_r$. 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, $i = 1, \ldots, r$, и это не отразится на асимптотическом распределении X^2. Есть и другие возможности. Много интересного об этом можно узнать в книге С.Рао [63].

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

$$H : P(A_1) = p_1(\theta), \ldots, P(A_r) = p_r(\theta),$$

где $p_1(\cdot), \ldots, p_r(\cdot)$ — 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 наблюденным частотам $m_1, \ldots, m_r$ 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 $P(\chi^2 \geqslant X^2)$. 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 приходиться заменять требованием, чтобы не были малы наблюдаемые частоты $m_1, \ldots, m_r$.) 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 простран-

ства 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 $x_1, \ldots, x_n$ 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 непересекающихся 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 $\Phi\left(\frac{x-\overline{x}}{s}\right)$ истинной функции распределения $\Phi\left(\frac{x-a}{\sigma}\right)$.

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

$$\frac{k}{r} = \Phi\left(\frac{x-\overline{x}}{s}\right), \qquad k = 1, \ldots, r-1.$$

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 $m_1, \ldots, m_r$, 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 $m_1, \ldots, m_r$, и только их. Никакой другой информацией пользоваться нельзя! Нельзя, например, использовать $\overline{x}$, s^2 в качестве оценок a и σ^2, по которым затем вычислять $p_i(\theta)$. Причина та, что есте-

ственные оценки $\overline{x}$, 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 $\overline{x}$, s^2. Последние два графика задают область нахождения функции распределения X^2 при $r = 21$.

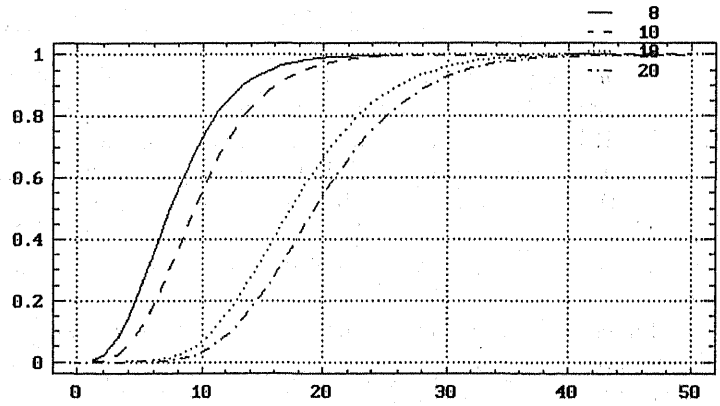


Рис. 10.2. Функции распределения хи-квадрат с 8, 10, 18 и 20 степенями свободы

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 главе 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*

$$\frac{M(\xi - M\xi)^3}{(D\xi)^{\frac{3}{2}}} \tag{10.11}$$

равен нулю. Поэтому *коэффициент асимметрии выборки*

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3} \tag{10.12}$$

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 нормального 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,

$$D\xi/M\xi = 1, \tag{10.13}$$

так как для распределения Пуассона $D\xi = M\xi = \lambda$, где λ — параметр распределения. Поэтому если выборка $x_1, \ldots, x_n$ извлечена из пуассоновской генеральной совокупности, то отношение

$$S^2/\overline{x}, \quad \text{где } S^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/n, \tag{10.14}$$

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 соотношение 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 не достаточно. То же самое можно сказать о (10.13): это необходимое, но не достаточное 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 $n \to \infty$ случайные величины $S^2 - D\xi$ и $\overline{x} - M\xi$ стремятся к 0 (закон больших чисел). Поэтому для пуассоновской выборки:

$$\frac{S^2}{\overline{x}} = \frac{D\xi + (S^2 - D\xi)}{M\xi + (\overline{x} - M\xi)} = \frac{D\xi}{M\xi}\,\frac{1 + \frac{S^2 - D\xi}{D\xi}}{1 + \frac{\overline{x} - M\xi}{M\xi}} =$$

$$\left(1 + \frac{S^2 - D\xi}{D\xi}\right)\left(1 - \frac{\overline{x} - M\xi}{M\xi} + \ldots\right).$$

Многоточие заменяет случайную величину, убывающую как n^{-1}. Раскрыв скобки, получаем, что:

$$\frac{S^2}{\overline{x}} = 1 + \frac{S^2 - D\xi}{D\xi} - \frac{\overline{x} - M\xi}{M\xi} + \cdots = 1 + \frac{1}{\lambda}(S^2 - \overline{x}) + \ldots,$$

Исследуем при $n \to \infty$ поведение выражения $\frac{S^2 - \overline{x}}{\lambda}$, главной случайной составляющей дроби $S^2/\overline{x}$. 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 $S^2 - \overline{x}$ появляется:

$$\frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \lambda)^2 - x_i\right].$$

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) $S^2/\overline{x}$ 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 $S^2/\overline{x}$ в случае справедливости гипотезы:

$$\left|\sqrt{n}\,\frac{S^2/\overline{x}-1}{\lambda\sqrt{2}}\right|<u_{0.995}\,, \tag{10.15}$$

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 главе 4, для больших выборок наилучшей является оценка наибольшего правдоподобия, которая для пуассоновского распределения равна $\overline{x}$. Следовательно, надо проверить по выборке, выполняется ли соотношение:

$$\left|\sqrt{n}\,\frac{S^2/\overline{x}-1}{\overline{x}\sqrt{2}}\right|<u_{0.995}=2.58,\ \text{т.е.}\ \left|\sqrt{n}\,\frac{S^2-\overline{x}}{(\overline{x})^2}\right|<3.6.4. \tag{10.16}$$

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 [16].

10.8. Критерии согласия в пакетах STADIA и STATGRAPHICS

В этом параграфе мы покажем, как процедуры проверки согласия могут быть реализованы в статистических пакетах. 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.

10.8.1. Пакет STADIA

Пример 10.1к. Проверим согласие распределения выборки диаметров головок заклепок (табл. 1.1) с нормальным распределением, используя критерий Колмогорова. Проведем аналогичные расчеты для «цензурированной» выборки.

Подготовка данных. Данные этого примера уже рассматривались в примерах 1.1к и 5.1к. Там же описан ввод (загрузка) данных в редактор базы данных пакета в переменную d (рис. 1.16). Как отмечалось в примере 5.1к, указанная выборка содержит одно резко выделяющееся наблюдение, равное 14.56. Удалив это значение, поместим оставшиеся данные в переменную dc.

Выбор процедуры. В меню Статистические методы (рис. 1.17) в разделе Распределения и частоты выберем пункт U = Согласие распределений.

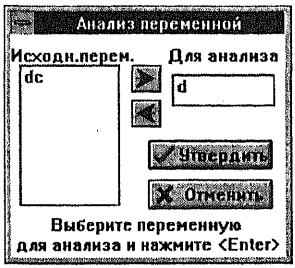


Рис. 10.3. Окно выбора переменной для анализа

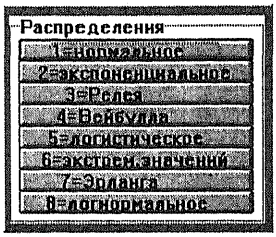


Рис. 10.4. Запрос типа распределения для проверки согласия

Заполнение полей ввода данных. В появившемся окне **Анализ переменной** (рис. 10.3) задайте переменную для анализа. Для этого в поле **Исходн. перем.** выделите мышью переменную **d** и, нажав кнопку запроса со стрелкой вправо, перенесите ее в поле **Для анализа.** Затем нажмите кнопку запроса [Утвердить]. На экране появится запрос типа распределения вероятностей (рис. 10.4). Нажмите в нем кнопку **1 = нормальное** (можно нажать клавишу [1]).

Результаты. Выдача результатов процедуры (рис. 10.5) содержит в строке **Распределение нормальное: 13.42, 0.1345** оценки среднего и стандартного отклонения выборки, а также значение статистик Колмогорова и омега-квадрат для сложной гипотезы, их уровни значимости и объем выборки в графе **степ. своб.** Сравнивая полученные уровни значимости с 5%, система выдает заключение **Гипотеза 1: Распределение отличается от теоретического** для каждого из указанных выше критериев.

```
СОГЛАСИЕ РАСПРЕДЕЛЕНИЙ.  Файл: diamz.std
Распределение нормальное: 13.422,  0.13445

Колмогоров=0.07175,  Значимость=0.016229,  степ.своб = 200
   Гипотеза 1: <Распределение отличается от теоретического>
Омега-квадрат=0.28287,  Значимость=0.00035817,  степ.своб = 200
   Гипотеза 1: <Распределение отличается от теоретического>
```

Рис. 10.5. Результаты проверки согласия для исходных данных

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 (сравните значения оценок стандартного отклонения на рис. 10.5 и 10.7), и тем самым повлияло на значения подобранной, согласно гипотезе, функции нормального распределения $F(x, \hat{\theta})$, где вектор $\theta = (\overline{x}, s^2)$. Эффект этого влияния хорошо виден на рис. 10.6 (левая часть), где приведены графики эмпирической и подобранной ги-

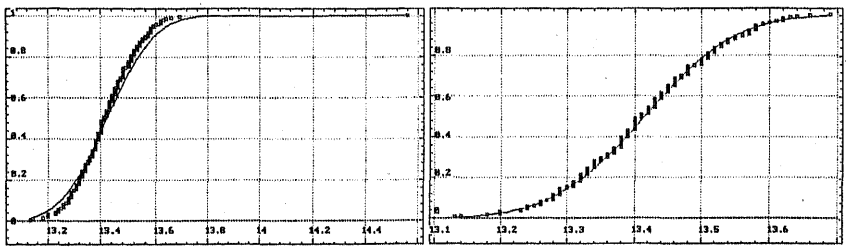


Рис. 10.6. Пакет STADIA. Графики эмпирической и подобранной гипотетической функции распределения: слева — исходные данные; справа — «цензурированные» данные

потетической функций распределения для исходных данных. Указанные графики выводятся при ответе [Y] на запрос программы **Вывести график.**

Проведем расчеты значений статистик Колмогорова и омега-квадрат для «цензурированных» данных. На рис. 10.7 приведены результаты процедуры в этом случае.

```
СОГЛАСИЕ РАСПРЕДЕЛЕНИЙ.  Файл: diamz.std
Распределение нормальное: 13.416, 0.10765

Колмогоров=0.045714, Значимость=0.47567, степ.своб = 199
   Гипотеза 0: <Распределение не отличается от теоретического>
Омега-квадрат=0.036583, Значимость=0.82347, степ.своб = 199
   Гипотеза 0: <Распределение не отличается от теоретического>
```

Рис. 10.7. Результаты проверки согласия для цензурированных данных

Как видно из полученных результатов, данные без резко выделяющегося значения («цензурированные»), не противоречат гипотезе о нормальности распределения. Графики эмпирической и подобранной гипотетической функций распределения для «цензурированных» данных приведены на рис. 10.6 (правая часть).

Комментарии. Для нормального распределения расчеты статистик Колмогорова и омега-квадрат включены также в процедуру **2=Гистограмма и нормальность.** Ее работа будет разобрана в примере 10.2к.

В следующем примере будет рассмотрена реализация критерия согласия хи-квадрат для сложной гипотезы. В качестве выборочных данных будет использован тот же массив диаметров головок заклепок.

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.

Подготовка данных — такая же, как в примерах 1.1к и 10.1к.

Выбор процедуры. В меню **Статистические методы** следует выбрать пункт **2 = Гистограмма/нормальность.** Работа этой процедуры, объединяющей

несколько различных задач, рассматривалась в примерах 1.2к и 1.3к при построении таблицы табуляции частот и гистограммы.

Заполнение полей ввода данных. На экране появится окно Анализ переменной (рис. 10.3), в котором следует выбрать переменную d для анализа. Далее последует запрос пакета о параметрах группировки данных (рис. 10.8). Зададим число интервалов группировки равным 15, левую границу группировки данных — 13.1 и правую границу — 14.6, как это показано на рис. 10.8.

Рис. 10.8. Задание интервалов группировки

Диапазон группировки наблюдений здесь выбран исходя из минимального (13.13) и максимального (14.56) элементов выборки. Число интервалов группировки выбрано так, чтобы длина интервала группировки составила 0.1. Напомним, что в выборке есть одно резко выделяющееся наблюдение, которое мы расцениваем как грубо ошибочное. Если мы хотим включить в обработку и его, то трудно говорить о каком-то оптимальном выборе разбиения данных. Все выборочные значения, кроме «грубого», сосредоточены в интервале (13.13, 13.69). Поэтому правая половина указанного диапазона группировки содержит всего одно наблюдение. Это влечет за собой образование (даже при малом числе интервалов группировки) таких интервалов, ожидаемая частота попадания в которые будет мала (такие интервалы будут располагаться на правом конце диапазона). Из-за этого возникает проблема аппроксимации распределения статистики критерия с помощью распределения хи-квадрат, о чем будет подробнее сказано ниже.

Результаты. Экран вывода результатов процедуры при введенных параметрах группировки представлен на рис. 10.9. (Описание формы экрана выдачи результатов дано в примере 1.2к.)

Как видно из таблицы, представленной на рис. 10.9, процедура увеличивает на единицу введенное число интервалов группировки за счет добавления справа от указанного правого конца диапазона бесконечного полуинтервала. Полученное значение статистики хи-квадрат столь велико, что даже при весьма приблизительном характере аппроксимации ее распределения нулевая гипотеза должна быть отвергнута. Конечно,

```
ГИСТОГРАММА И ТЕСТ НОРМАЛЬНОСТИ.  Файл: diamz.std

  X-лев   X-станд  Частота        %    Накопл.        %
   13.1   -2.3913        6        3          6        3
   13.2   -1.6475       24       12         30       15
   13.3   -0.9037       67     33.5         97     48.5
   13.4  -0.15991       58       29        155     77.5
   13.5   0.58387       36       18        191     95.5
   13.6    1.3277        8        4        199     99.5
   13.7    2.0714        0        0        199     99.5
   13.8    2.8152        0        0        199     99.5
   13.9     3.559        0        0        199     99.5
     14    4.3028        0        0        199     99.5
   14.1    5.0466        0        0        199     99.5
   14.2    5.7904        0        0        199     99.5
   14.3    6.5341        0        0        199     99.5
   14.4    7.2779        0        0        199     99.5
   14.5    8.0217        1      0.5        200      100
   14.6    8.7655
Колмогоров=0.07175, Значимость=0.016229, степ.своб = 200
   Гипотеза 1: <Распределение отличается от нормального>
Омега-квадрат=0.28287, Значимость=0.00035817, степ.своб = 200
   Гипотеза 1: <Распределение отличается от нормального>
Хи-квадрат=511.55, Значимость=0, степ.своб = 13
   Гипотеза 1: <Распределение отличается от нормального>
```

Рис. 10.9. Результаты проверки нормальности распределения

это происходит из-за присутствия аномального значения, о котором мы так много говорили.

Приведем результаты применения процедуры для «цензурированных» данных (рис. 10.10) при следующих параметрах: диапазон группировки (13.1, 13.7), число интервалов группировки — 6.

```
ГИСТОГРАММА И ТЕСТ НОРМАЛЬНОСТИ.  Файл: diamz.std

  X-лев   X-станд  Частота        %    Накопл.        %
   13.1   -2.9334        6   3.0151          6   3.0151
   13.2   -2.0045       24    12.06         30   15.075
   13.3   -1.0755       67   33.668         97   48.744
   13.4  -0.14658       58   29.146        155   77.889
   13.5   0.78237       36    18.09        191    95.98
   13.6    1.7113        8   4.0201        199      100
   13.7    2.6404

Колмогоров=0.045714, Значимость=0.47567, степ.своб = 199
   Гипотеза 0: <Распределение не отличается от нормального>
Омега-квадрат=0.0365832, Значимость=0.81358, степ.своб = 199
   Гипотеза 0: <Распределение не отличается от нормального>
Хи-квадрат=2.0531, Значимость=0.7259, степ.своб = 4
   Гипотеза 0: <Распределение не отличается от нормального>
```

Рис. 10.10. Результаты проверки нормальности распределения для цензурированных данных

Обратим внимание на то, что при составлении статистики хи-квадрат (при вычислении ожидаемых частот) в процедуре используются обычные оценки $\overline{x}$ и s^2. (По теории надо вычислять оценки параметров a и σ^2 по наблюденным частотам.) Поэтому истинный уровень значимости несколько отличается от указанного на экране 0.7259. Как отмечалось выше, аппроксимация распределения статистики в этом случае отлича-

ется от распределения хи-квадрат. Приближенный уровень значимости вычисленной статистики лежит между квантилями распределения хи-квадрат с $(r-3)$ и $(r-1)$ степенями свободы, где r — число интервалов группировки. То есть уровень значимости полученной статистики лежит в интервале (0.7259, 0.9148). Поэтому гипотезу о нормальном распределении (для «цензурированных» данных) следует принять.

Комментарии. Критерий согласия хи-квадрат для сложной гипотезы представлен в пакете только для нормального распределения.

10.8.2. Пакет STATGRAPHICS

Пример 10.1к. Проверим согласие распределения выборки диаметров головок заклепок (табл. 1.1) с нормальным распределением, используя критерий Колмогорова. Проведем аналогичные расчеты для «цензурированной» выборки.

Подготовка данных. Данные этого примера уже рассматривались в примерах 1.1к, и 5.1к. Они находятся в переменной d файла DIAMZ базы данных пакета. Пусть «цензурированные» данные находятся в переменной dc того же файла.

Выбор процедуры. В пакете не вычисляется уровень значимости критерия Колмогорова для сложной гипотезы. (Для ряда распределений, представленных в пакете, до сих пор теоретически неизвестно, как это сделать.) В этой ситуации мы попробуем выяснить, в каких случаях результаты критерия Колмогорова для простой гипотезы могут быть полезны при проверке сложной гипотезы.

В головном меню пакета выберем пункт H. Distribution Function. Его меню описано в пункте 2.7.2. Выберем в этом меню процедуру 1. Distribution Fitting.

Заполнение полей ввода данных. Экран ввода данных в эту процедуру представлен на рис. 4.7. Порядок его заполнения описан в примере 4.1к.

Результаты. После заполнения полей ввода и нажатия клавиши F6 на экран выводятся оценки параметров распределения (для нормального распределения — среднее значение и стандартное отклонение). Если мы готовимся проверить простую нулевую гипотезу, то их можно изменить согласно выдвинутой гипотезе. Если мы проверяем сложную гипотезу, этого делать не надо, так как для подбора гипотетического распределения используются полученные по выборке оценки. После нажатия клавиши F6 на экране появляется меню, представленное на рис. 10.11.

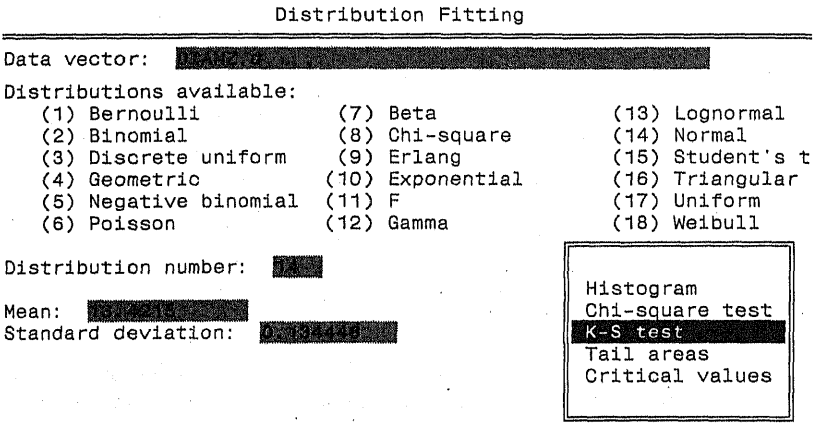


Рис. 10.11. Меню функций распределения для проверки согласия

В указанном меню надо выбрать процедуру K-S test (критерий Колмогорова-Смирнова). На рис. 10.12 представлены результаты полученных расчетов. Они включают значение статистик Колмогорова-Смирнова D_n^+ (Estimated KOLMOGOROV statistic DPLUS) и D_n^- (Estimated KOLMOGOROV statistic DMINUS), а так же значение статистики Колмогорова D_n (Estimated overall statistic DN) и минимальный уровень значимости последней статистики в случае простой гипотезы (Approximate significance level).

```
Estimated KOLMOGOROV statistic DPLUS = 0.0717387
Estimated KOLMOGOROV statistic DMINUS = 0.0612956
Estimated overall statistic DN = 0.0717387
Approximate significance level = 0.25474
```

Рис. 10.12. Результаты проверки согласия для исходных данных

Попытка использовать полученный уровень значимости для сложной гипотезы приведет в данном случае к ошибочному принятию нулевой гипотезы. Правильный уровень значимости статистики Колмогорова в случае сложной гипотезы равен 0.01605 (он был вычислен пакетом STADIA, см. рис. 10.5). Поэтому гипотезу о нормальности на самом деле надо отвергнуть.

Приведем результаты расчетов статистики Колмогорова для «цензурированных» данных (рис. 10.13).

```
Estimated KOLMOGOROV statistic DPLUS = 0.0457847
Estimated KOLMOGOROV statistic DMINUS = 0.0342026
Estimated overall statistic DN = 0.0457847
Approximate significance level = 0.999972
```

Рис. 10.13. Результаты проверки согласия для цензурированных данных

В этом случае вычисленный уровень значимости для простой гипотезы (0.999972) более чем вдвое превышает уровень значимости для сложной гипотезы (0.4584). Как видно, представленная в пакете процедура, в отличие от аналогичной процедуры пакета STADIA, дает ошибочные уровни значимости и для исходных данных, и для «цензурированных» данных. Поэтому полагаться на нее не следует.

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.

Комментарии. В документации пакета не делается различий между проверкой согласия для простой и сложной гипотез, что является серьезным упущением.

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.

Подготовка данных. Смотри примеры 1.1к и 10.1к.

Выбор процедуры. На первом этапе порядок действий совпадает с описанным в примере 10.1к. В меню рис. 10.11 надо выбрать процедуру Chi-square test (критерий хи-квадрат).

Заполнение полей ввода данных. На рис. 10.14 приведен экран ввода параметров табуляции разбираемой процедуры. Его подробное описание дано в примере 1.2к (рис. 1.26).

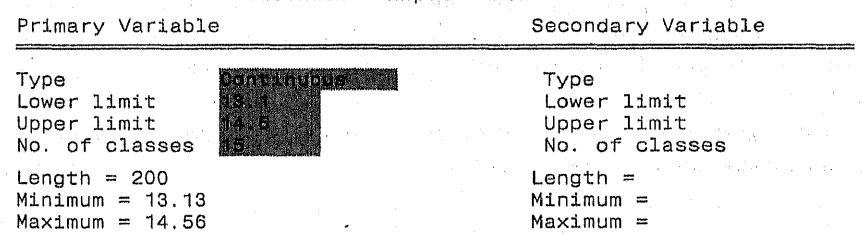


Рис. 10.14. Запрос параметров группировки для критерия хи-квадрат

В качестве параметров группировки выберем те же параметры, что и при анализе этих данных в пакете STADIA.

Chisquare Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chisquare
at or below		13.200	6	9.9	1.5655
	13.200	13.300	24	26.7	.2672
	13.300	13.400	67	50.7	5.2562
	13.400	13.500	58	56.8	.0264
	13.500	13.600	36	37.5	.0601
above	13.600		9	18.4	4.8245

Chisquare = 11.9999 with 3 d.f. Sig. level = 7.38343E-3

Рис. 10.15. Результаты проверки согласия по критерию хи-квадрат

Результаты. На рис. 10.15 приведен экран выдачи результатов для критерия хи-квадрат.

В двух первых столбцах таблицы результатов **Lower Limit** и **Upper Limit** указаны нижние и верхние границы интервалов группировки. В столбце **Observed Frequency** представлены наблюдаемые частоты, а в столбце **Expected Frequency** — частоты подобранного гипотетического распределения. Столбец **Chisquare** содержит значения слагаемых выражения (10.12) для каждого интервала группировки. Нижняя строка экрана выдачи результатов включает значение статистики хи-квадрат, число степеней свободы **d.f.** и уровень значимости **Sig. level**.

Обратим внимание на ряд существенных моментов в работе этой процедуры.

Во-первых, число интервалов группировки, указанное пользователем, корректируется с учетом обеспечения условий применимости аппроксимации распределения статистики с помощью распределения хи-квадрат. Так, вместо 15 введенных интервалов группировки (рис. 10.14), сформировано только 6. В данном случае все наблюдения, лежащие правее значения 13.6, были включены в один интервал группировки.

Во-вторых, для вычисления частот гипотетического распределения, так же как и в пакете STADIA, используются оценки $\overline{x}$ и s^2. Как отмечалось выше, уровень значимости подобной статистики лежит где-то между квантилями распределения хи-квадрат с $(r-3)$ и $(r-1)$ степенями свободы, где r — число интервалов группировки. То есть уровень значимости полученной статистики лежит в интервале (0.007383, 0.0348). Указать его точнее трудно. При этом, например, не ясно, следует ли отвергнуть гипотезу на однопроцентном уровне значимости.

Приведем результаты расчетов критерия хи-квадрат для «цензурированных» данных (рис. 10.16), используя в качестве областей группировки разбиение интервала (13.1, 13.7) на 6 равных частей.

Chisquare Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chisquare
at or below		13.300	30	28.1	.1336
	13.300	13.400	67	59.8	.8605
	13.400	13.500	58	67.9	1.4482
	13.500	13.600	36	34.5	.0624
above	13.600		8	8.7	.0506

Chisquare = 2.55532 with 2 d.f. Sig. level = 0.278689

Рис. 10.16. Результаты проверки согласия по критерию хи-квадрат для цензурированных данных

В этом случае интервал для уровня значимости статистики хи-квадрат есть (0.2787, 0.6348). В любом случае нет основания отвергать нулевую гипотезу.

Комментарии. Процедура позволяет получить значение статистики хи-квадрат для проверки простой гипотезы. Для этого в поля параметров семейства распределений (рис. 10.11) необходимо ввести значения, соответствующие простой гипотезе.

Для получения минимального уровня значимости полученной статистики хи-квадрат (рис. 10.15) необходимо обратиться к процедуре 3. Tail Area Probabilities пункта H. Distribition function головного меню пакета (пример 2.1к), указав число степеней свободы распределения хи-квадрат равным $(r-1)$, где r — число интервалов группировки, установленное самой процедурой.

Глава 11

Временные ряды: теоретические основы

11.1. Введение

Что такое временной ряд. Временной ряд — это последовательность чисел; его элементы — это значения некоторого протекающего во времени процесса. Они измерены в последовательные моменты времени, обычно через равные промежутки.

Как правило, составляющие временной ряд числа — *элементы временного ряда*, — нумеруют в соответствии с номером момента времени, к которому они относятся (например, x_1, x_2, x_3 и т.д.). Таким образом, порядок следования элементов временного ряда весьма существен.

Почти в каждой области знания встречаются явления, которые важно изучать в развитии во времени или пространстве. И почти всегда в закономерное течение явления вмешивается случай в виде случайных импульсов, случайных помех, случайных ошибок и т.д. Поэтому изучение временных рядов — это составная часть прикладной статистики (и довольно важная ее часть).

Расширения понятия временного ряда. Понятие временного ряда часто толкуют расширительно. Например, одновременно могут регистрироваться несколько характеристик упомянутого процесса. В этом случае говорят о *многомерных временных рядах*. Если измерения производятся непрерывно, говорят о временных рядах с непрерывным временем, или *случайных процессах*. Наконец, текущая переменная может иметь не временной, а какой-нибудь иной характер, например пространственный (тогда говорят о *случайных полях*).

Примеры временных рядов. Данные типа временных рядов широко распространены в самых различных областях человеческой деятельности. В экономике это ежедневные цены на акции, курсы валют, еженедельные и месячные объемы продаж, годовые объемы производства и т.п. В метеорологии типичными временными рядами являются ежедневная температура, месячные объемы осадков, в гидрологии — периодически измеряемые уровни воды в реках. В технике времен-

ные ряды возникают в результате отслеживания различных параметров технологических процессов.

На рис. 11.1 приведены примеры различных временных рядов (для наглядности последовательные измерения, составляющие временной ряд, на графиках соединены линиями).

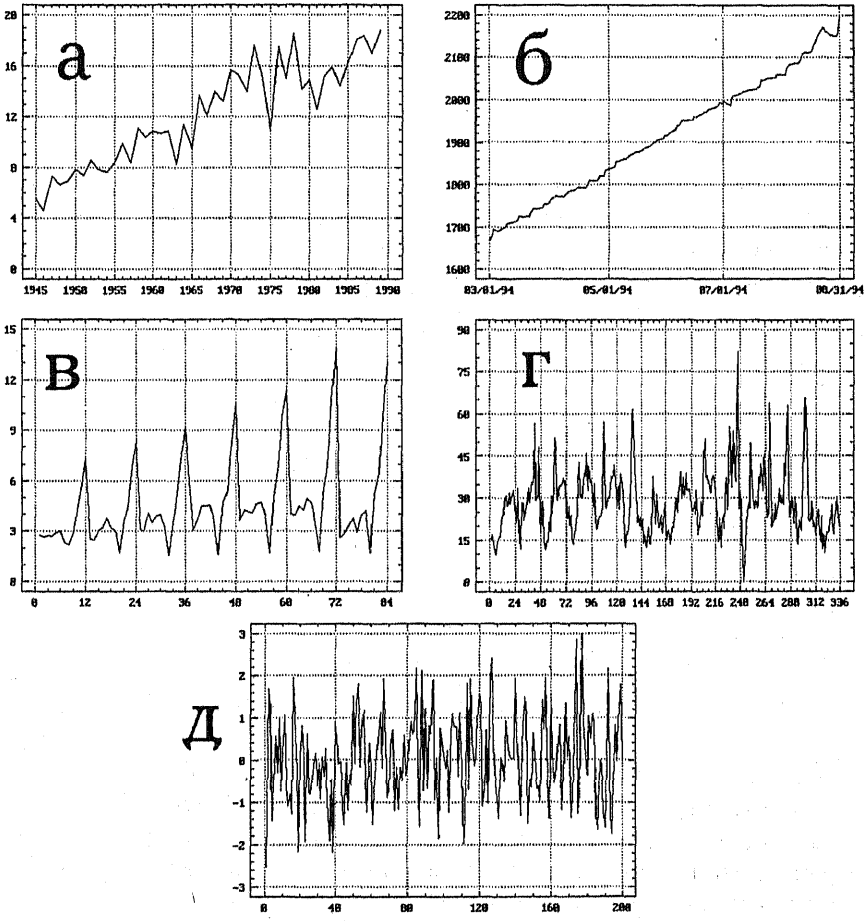


Рис. 11.1. а) – урожайность зерновых культур в СССР с 1945 по 1989 гг. в ц/га; б) – курс доллара на торгах ММВБ с 1.03 по 31.08 1994 г. в рублях; в) – ежемесячные продажи шампанского за 7 последовательных лет; г) – среднечасовая нагрузка телекоммуникационного канала Москва – Париж в течение 2-х недель (в Кбит/сек); д) – гауссовский «белый шум» с параметрами 0 и 1

Видно, что поведение временных рядов может быть весьма различным. Так, динамика урожайности зерновых в СССР (ряд а)) имеет

скорее всего линейный тренд, отклонения от которого можно считать независимыми случайными величинами. Курс доллара на торгах ММВБ весной-летом 1994 г. (ряд б)) также содержит линейный тренд, однако отклонения от него имеют более сложную статистическую структуру, чем в предыдущем случае. График ежемесячных продаж шампанского (ряд в)) содержит явно повторяющиеся годовые циклы с возрастающей амплитудой. Среднечасовая загрузка телекоммуникационного канала Москва-Париж в течение двух недель (одна неделя равна 168 часам) февраля 1996 г. (ряд г)) имеет ясные суточные циклы. Кроме суточных, этот ряд содержит и недельные циклы, но на приведенном графике они заметны мало, так как недостаточно длителен интервал наблюдения. Ряд д) создан датчиком нормальных случайных чисел на компьютере и служит примером чисто случайного процесса без внутренних закономерностей и зависимостей.

Измерение значений временного ряда. Чаще всего значения временного ряда получаются непосредственной записью значений некоторого процесса через определенные промежутки времени. Например, если ежесуточно в определенное время записывать показания термометра, то получится временной ряд со значениями температуры в том месте, в котором находится термометр.

Иногда значения элементов временного ряда получаются накапливанием некоторых данных за определенный промежуток времени (например, суммарное число посетителей магазина за день), усреднением (средняя температура за день) и т.д.

11.2. Анализ временных рядов и его разделы

Анализ временных рядов. Исследование временных рядов отличается от других задач анализа данных как кругом представляющих интерес вопросов, так и методами, применяемыми для исследования. Поэтому наука об исследовании временных рядов — *анализ временных рядов*, — образует самостоятельную и весьма обширную область статистики.

Разделы анализа временных рядов. Временные ряды, возникающие в различных предметных областях, имеют различную природу, поэтому для их изучения оказались эффективными разные методы. Исследователи придумывали и развивали многочисленные методы анализа, подходящие для изучения временных рядов в разных предметных областях. И в результате анализ временных рядов превратился в довольно

разветвленную науку. Вот только некоторые из видов временных рядов, исследование которых можно рассматривать как самостоятельный раздел теории анализа временных рядов:

- *стационарные случайные процессы*, то есть последовательности случайных величин, вероятностные свойства которых не изменяются во времени. Стационарные случайные процессы широко применяются в радиотехнике, теории связи, механике жидкости и газа, океанологии, метеорологии и т.д.;
- *диффузионные процессы* возникли при изучении процесса диффузии, то есть взаимопроникновения различных жидкостей или газов. Эти процессы используются при построении моделей непрерывных процессов, в которых существенна случайная составляющая;
- *точечные процессы* используют для описания таких явлений, как поступление вызовов или заявок на обслуживание, моментов несчастных случаев, стихийных и техногенных катастроф, каких-либо приметных явлений и т.п. Они широко применяются в таких разделах статистики, как теория очередей, теория массового обслуживания и т.д.

Всех этих обширных разделов анализа временных рядов в данной книге мы касаться не будем. Вместо этого мы хотим рассказать о тех прикладных аспектах анализа временных рядов, которые полезны и важны при решении практических задач в экономике, финансах, а также в различных гуманитарных науках. В частности, мы расскажем о методах подбора математической модели для описания временного ряда, об изучении взаимозависимостей временных рядов, выявления в них периодических и других составляющих, прогнозировании поведения временных рядов и т.д.

Как мы будем рассказывать об анализе временных рядов. В этой главе (главе 11) мы обсудим основные понятия статистической теории временных рядов. Мы расскажем о структуре временных рядов; о вероятностных предпосылках для их анализа; о детерминированных компонентах временных рядов и о причинах, их порождающих; о корреляционной структуре ряда; о случайной составляющей временного ряда и ее описании и т.д. В главе 12 обсуждаются вопросы прикладного анализа временных рядов, а в главе 13 мы опишем, как различные практические задачи анализа временных рядов решаются с помощью статистических пакетов Эвриста и SPSS. Наконец, в главе 14 мы вновь возвратимся к теории и расскажем о некоторых математических моделях временных рядов, важных для прикладного анализа.

Мы будем стараться вести изложение на уровне, доступном широкому читателю. Наш рассказ мы сопроводим обсуждением примеров (как тех, что привели на рис. 11.1, так и многих других).

Литература. Те читатели, которые захотят глубже изучить теоретические или прикладные аспекты анализа временных рядов, могут обратиться к литературе. По большинству разделов анализа временных рядов существует множество книг разной степени подробности и математической сложности. Одни из этих книг рассчитаны на математиков, другие — на практических работников и инженеров. Укажем некоторые наиболее известные из этих книг:

- в области теории временных рядов — [5], [75], [38], [96], [98], [18];
- в области общего прикладного анализа — [11], [12], [15], [39], [68], [9];
- в области прикладного спектрального анализа — [29], [58];
- в области радиотехнических приложений — [47], [24];
- в области экономических приложений — [101], [52] и др.

11.3. Цели, этапы и методы анализа временных рядов

Цели анализа временных рядов. При практическом изучении временных рядов исследователь на основании наблюденного отрезка временного ряда (конечной длины) должен сделать выводы о свойствах этого ряда и о вероятностном механизме, порождающем этот ряд. Чаще всего при изучении временных рядов ставятся следующие цели:

- краткое (сжатое) описание характерных особенностей ряда;
- подбор статистической модели (моделей), описывающей временной ряд;
- предсказание будущих значений на основе прошлых наблюдений;
- управление процессом, порождающим временной ряд.

На практике эти и подобные цели достижимы далеко не всегда и далеко не в полной мере. Часто этому препятствует недостаточный объем наблюдений (недостаточная длительность); еще чаще — изменяющаяся с течением времени статистическая структура временного ряда. Из-за этих изменений значение прошлых наблюдений обесценивается, и они уже не помогают предвидеть будущее.

Стадии анализа временных рядов. Обычно при практическом анализе временных рядов последовательно проходят следующие этапы:

- графическое представление и описание поведения временного ряда;
- выделение и удаление закономерных составляющих временного ряда, зависящих от времени: тренда, сезонных и циклических составляющих;
- выделение и удаление низко- или высокочастотных составляющих процесса (фильтрация);
- исследование случайной составляющей временного ряда, оставшейся после удаления перечисленных выше составляющих;
- построение (подбор) математической модели для описания случайной составляющей и проверка ее адекватности;
- прогнозирование будущего развития процесса, представленного временным рядом;
- исследование взаимодействий между различными временными рядами.

Методы анализа временных рядов. Для решения указанных выше (а также многих других) задач исследователями предложено большое количество различных методов. Отметим из них наиболее распространенные:

- *корреляционный анализ* позволяет выявить существенные периодические зависимости и их *лаги* (задержки) внутри одного процесса (автокорреляция) или между несколькими процессами (кросскорреляция);
- *спектральный анализ* позволяет находить периодические и квазипериодические составляющие временного ряда;
- *сглаживание и фильтрация* предназначены для преобразования временных рядов с целью удаления из них высокочастотных или сезонных колебаний;
- модели *авторегрессии и скользящего среднего* оказываются особенно полезными для описания и прогнозирования процессов, проявляющих однородные колебания вокруг среднего значения;
- *прогнозирование* позволяет на основе подобранной модели поведения временного ряда предсказывать его значения в будущем.

Как уже говорилось, в этой книге мы расскажем не обо всех из этих методов, а лишь о тех, которые наиболее важны для экономических и гуманитарных наук.

11.4. Детерминированная и случайная составляющие временного ряда

Следуя основной идее статистики, при анализе временного ряда видимую его изменчивость стараются разделить на закономерную и случайную составляющие. Закономерные изменения членов временного ряда следуют какому-то определенному правилу и поэтому предсказуемы. Эта составляющая x_t может быть вычислена при каждом t как некоторая функция от текущего момента t. Эта функция может зависеть, помимо t, также от некоторого набора параметров. Когда эти параметры неизвестны, их приходится оценивать по имеющимся наблюдениям — как, например, бывает в случае регрессии.

Изменчивость, оставшаяся необъясненной, иррегулярна и хаотична. Для ее описания необходим статистический подход (за неимением лучшего).

Определение. *Под закономерной (детерминированной) составляющей временного ряда $x_1, \ldots, x_n$ мы будем понимать числовую последовательность $d_1, \ldots, d_n$, элементы которой d_t вычисляются по определенному правилу как функция времени t.*

Детерминированная составляющая часто отражает действия каких-либо определенных факторов или причин. Так, у временных рядов из различных областей техники детерминированная составляющая обычно обязана своим возникновением действиям физических законов или условиям эксплуатации оборудования. Например, если значения временного ряда соответствуют положениям маятника в определенные моменты времени, то в качестве детерминированной компоненты ряда можно взять решение дифференциального уравнения движения маятника в эти моменты времени. В экономических и многих других приложениях математические модели изучаемых процессов нам обычно не известны, так что тенденции, отраженные в поведении временного ряда, нам приходится выявлять по наблюдаемым значениям временного ряда. Например, при изучении данных о месячном производстве молока в России (рис. 11.2) мы можем пытаться описать закономерную часть данного временного ряда в виде комбинации линейной функции (в течение последних лет производство молока, нашедшее отражение в статистической отчетности, постепенно уменьшалось) и периодической функции с периодом 12 месяцев. Эта периодическая компонента отражает влияние времени года на производство молока.

Для многих рядов в экономике и социальных науках причины, порождающие их закономерные составляющие, могут не быть столь ясными.

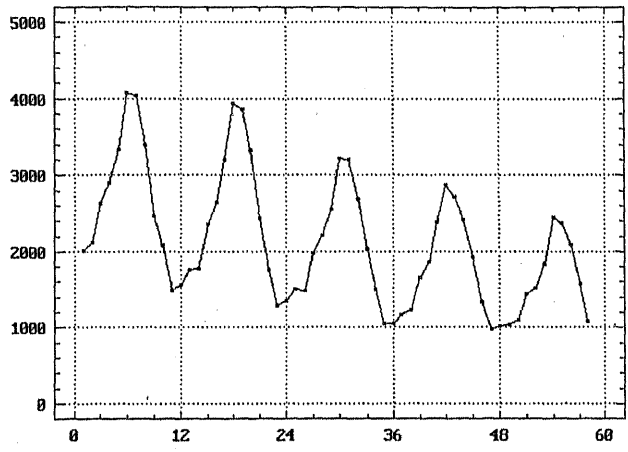


Рис. 11.2. Ежемесячное производство молока
в России с 01.1992 по 10.1996 (в тыс. тонн)

Тем не менее, их совокупное влияние может быть устойчивым в течение достаточно длительных промежутков времени. Это обеспечивает возможность прогноза для подобных временных рядов. Если мы полностью выявим закономерную составляющую в поведении временного ряда, то оставшаяся часть выглядит хаотично и непредсказуемо. Ее обычно именуют иррегулярной, или *случайной компонентой* временного ряда. Обозначим эту случайную компоненту через $\varepsilon_1, \ldots, \varepsilon_t, \ldots, \varepsilon_n$.

Для описания и анализа случайной компоненты временных рядов обычно используют понятия и методы теории вероятностей и математической статистики.

Аддитивная и мультипликативная модели. Формы разложения (декомпозиции) временного ряда на детерминированную и случайную компоненты могут различаться. Укажем наиболее простых из них.

Определение. *Аддитивной моделью временного ряда называется представление ряда в виде суммы детерминированной и случайной компонент, а именно:*

$$x_t = d_t + \varepsilon_t \qquad \textit{при} \quad t = 1, \ldots, n \qquad \textit{или} \qquad X = D + E. \qquad (11.1)$$

Определение. *Мультипликативной моделью временного ряда называется представление ряда в виде произведения детерминированной и случайных компонент, а именно:*

$$x_t = d_t \times \varepsilon_t \qquad \textit{при} \quad t = 1, \ldots, n \qquad \textit{или} \qquad X = D \times E$$

Мультипликативные модели часто бывают удобны при анализе экономических временных рядов.

Если в приведенном выше соотношении перейти к логарифмам, то мы вновь получим формулу (11.1) — но не для самих x_t, а для их логарифмов.

$$\log(x_t) = \log(d_t) + \log(\varepsilon_t) \qquad \text{при} \qquad t = 1, \ldots, n$$

Указанное соотношение объясняет распространенность логарифмических шкал при анализе экономических временных рядов.

11.5. Тренд, сезонная и циклическая компоненты

Способы описания детерминированных компонент временного ряда сильно зависят от области приложений. При выборе модели детерминированной компоненты должны прежде всего учитываться содержательные соображения, то есть те объективные факторы и закономерности, которые приводят к ее формированию.

В экономических (и многих других) приложениях в детерминированной компоненте временного ряда d_t обычно выделяют три составляющих части: тренд tr_t, сезонную компоненту s_t и циклическую компоненту c_t. Для простоты изложения мы рассмотрим только аддитивную модель временного ряда. Мы можем записать:

$$d_t = tr_t + s_t + c_t, \qquad \text{при} \quad t = 1, \ldots, n.$$

Замечание. В последнее время к указанным трем компонентам все чаще добавляют еще одну компоненту, именуемую интервенцией. Под интервенцией понимают существенное кратковременное воздействие на временной ряд. Примером интервенции могут служить события «черного вторника», когда курс доллара за день вырос почти на тысячу рублей. С анализом интервенций можно познакомиться в [9].

Термины «тренд», «сезонная компонента» и «циклическая компонента» не имеют однозначных общепринятых определений. Чаще всего расхождения относятся к определению тренда и циклической компоненты и связаны с различными традициями в разных науках. Мы определим их в виде, наиболее часто используемом в экономических приложениях.

Тренд. Анализ временного ряда обычно начинается с выделения именно этой компоненты. Ее присутствие или отсутствие наглядно показывает график временного ряда. Выделение тренда позволяет перейти к дальнейшей идентификации других компонент ряда.

Определение. *Трендом временного ряда tr_t при $t = 1, \ldots, n$ называют плавно изменяющуюся, не циклическую компоненту, описыва-*

ющую чистое влияние долговременных факторов, эффект которых сказывается постепенно.

В экономике к таким факторам можно отнести:

- изменение демографических характеристик популяции, включая рост населения, изменение структуры возрастного состава, изменение географического расселения и т.д.;
- технологическое и экономическое развитие;
- рост потребления и изменение его структуры.

Действие этих и им подобных факторов происходит постепенно, поэтому их вклад исследователи предпочитают описывать с помощью гладких кривых, просто задающихся в аналитическом виде. Мы опишем некоторые модели тренда в следующем параграфе.

Сезонная компонента. Сезонная компонента отражает присущую миру и человеческой деятельности повторяемость процессов во времени. Она часто присутствует в экономических, метеорологических и других временных рядах. Сезонная компонента чаще всего служит главным источником краткосрочных колебаний временного ряда, так что ее выделение заметно снижает вариацию остаточных компонент.

Определение. *Сезонная компонента s_t временного ряда при $t = 1, \ldots, n$ описывает поведение, изменяющееся регулярно в течение заданного периода (года, месяца, недели, дня и т.п.). Она состоит из последовательности почти повторяющихся циклов.*

Типичным примером сезонного эффекта является объем продаж в декабре каждого года в преддверии Рождества и нового года. В то же время пик объема продаж товаров для школьников приходится на начало нового учебного года. Объем перевозок пассажиров городским транспортом имеет два характерных пика утром и вечером, причем период вечернего пика продолжительней, а сам пик менее высокий. Сезонные эффекты присущи многим сферам человеческой активности: многие виды продукции имеют сезонный характер производства, потребление товаров также имеет ярко выраженную сезонность. На графике месячных объемов продаж шампанского в течение 7 лет (рис. 11.1в) видно, что пик реализации приходится на декабрь, а спад на жаркие летние месяцы.

В некоторых временных рядах сезонная компонента может иметь плавающий или изменяющийся характер. Классическим примером подобного эффекта является праздник Пасхи, сроки которого изменяются из года в год. Поэтому локальный пик объемов междугородных перевозок во время пасхальных каникул является плавающим сезонным эффектом.

Главная идея подхода к анализу сезонных компонент заключается в переходе от сравнения всех значений временного ряда между собой к сравнению значений через определенный период времени. Это позволяет заметно снизить оценку вариации временного ряда около своего среднего значения. Так, при изучении динамики месячных объемов продаж за несколько лет данные декабря одного года обычно сравнивают с данными декабря предыдущего года, а не с данными других месяцев рассматриваемого года. Методы анализа сезонных эффектов и выделения сезонных компонент рассмотрены в пункте 12.3.2.

Циклическая компонента занимает как бы промежуточное положение между закономерной и случайной составляющими временного ряда. Если тренд — это плавные изменения, проявляющиеся на больших временных промежутках, если сезонная компонента — это периодическая функция времени, ясно видимая, когда ее период много меньше общего времени наблюдений, то под циклической компонентой обычно подразумевают изменения временного ряда, достаточно плавные и заметные для того, чтобы не включать их в случайную составляющую, но такие, которые нельзя отнести ни к тренду, ни к периодической компоненте.

Определение. *Циклическая компонента c_t временного ряда описывает длительные периоды относительного подъема и спада. Она состоит из циклов, которые меняются по амплитуде и протяженности.*

Изучение циклической компоненты полезно для прогнозирования (особенно краткосрочного).

Замечание. Выделение в экономических временных рядах циклических компонент связано с тем, что экономическая активность не растет (или спадает) постоянными темпами. Она состоит из периодов относительных подъемов и спадов. Считается, что причиной циклических изменений в экономических показателях является взаимодействие спроса и предложения. Играют роль и другие факторы: рост и истощение ресурсов, увеличение размеров капитала, используемого в бизнесе, продолжительно действующие неблагоприятные (либо благоприятные) для тех или иных отраслей сельского хозяйства погодные условия, изменения в правительственной финансовой и налоговой политике и т.п. Влияние всех этих факторов приводит к тому, что циклическую компоненту крайне трудно идентифицировать формальными методами, исходя только из данных изучаемого ряда. Поэтому для ее анализа обычно приходится привлекать дополнительную информацию в виде других временных рядов, которые оказывают влияние на изучаемый ряд, например, учитывать информацию типа налоговых льгот, перенасыщенности рынка и т.п.

Методы определения циклической компоненты в экономических временных рядах и связанные с ней индексы деловой активности относят-

ся к эконометрии и выходят за рамки материала, рассматриваемого в данной книге. Более подробно об этом можно прочесть в [104].

11.6. Модели тренда

Простейшие модели тренда. Приведем модели трендов, наиболее часто используемые при анализе экономических временных рядов, а также во многих других областях. Во-первых, это простая линейная модель

$$tr_t = b_0 + b_1 \cdot t, \tag{11.2}$$

которая, несмотря на свою простоту, оказывается полезной во многих реальных задачах. Если нелинейный характер тренда очевиден, то может подойти одна из следующих моделей:

- полиномиальная: $tr_t = b_0 + b_1 t + b_2 t^2 + \cdots + b_n t^n$, где значение степени полинома n в практических задачах редко превышает 5;
- логарифмическая: $tr_t = \exp(b_0 + b_1 t)$. Эта модель чаще всего применяется для данных, имеющих тенденцию сохранять постоянные темпы прироста;
- логистическая: $tr_t = \frac{a}{1+b \cdot e^{-ct}}$;
- Гомперца: $\log(tr_t) = a - b \cdot r^t$, где $0 < r < 1$.

Две последние модели задают кривые тренда S-образной формы. Они соответствуют процессам с постепенно возрастающими темпами роста в начальной стадии и постепенно затухающими темпами роста в конце. Необходимость подобных моделей обусловлена невозможностью многих экономических процессов продолжительное время развиваться с постоянными темпами роста или по полиномиальным моделям, в связи с их довольно быстром ростом (или уменьшением).

Первое представление о возможном характере тренда дает графическое представление временного ряда. Так, график роста урожайности зерновых культур (рис. 11.1а) позволяет предположить наличие линейного тренда в этом временном ряде. Аналогичное предположение очевидно справедливо и для ряда роста курса доллара весной и летом 1994 г. (рис. 11.1б).

При прогнозировании тренд используют в первую очередь для долговременных прогнозов. Точность краткосрочных прогнозов, основанных только на подобранной кривой тренда, как правило, недостаточна. Методы выделения и удаления тренда подробно рассматриваются в пункте 12.3.1, а также в главе 8.

О временных рядах в технических приложениях. В технических приложениях мы часто знаем физические законы или технические ха-

рактеристики механизмов, генерирующих исследуемые временные ряды. Это, разумеется, существенно облегчает исследование. Мы рассмотрим только один тип моделей временных рядов, часто используемый в технических приложениях — *полигармоническую модель*. О других моделях временных рядов, возникающих в технических приложениях, Вы можете узнать в [47], [24].

Полигармоническая модель. Простейший вариант полигармонической модели временного ряда — это косинусоидальная модель:

$$x_t = a\cos(\omega t + \theta) + \varepsilon_t \tag{11.3}$$

Здесь детерминированная компонента представляет собой косинусоидальную функцию с амплитудой a, частотой ω, периодом $2\pi/\omega$ и фазой θ. Величины a, ω и θ в выражении (11.3) являются константами.

Комментарий. В технических приложениях часто рассматриваются модели типа (11.3), в которых амплитуда a является случайной величиной с нулевым средним или фаза θ является случайной равномерно распределенной величиной в интервале $(0, 2\pi)$. Такой подход, превращающий процесс (11.3) и ему подобные в стационарные процессы, часто обусловлен необходимостью обоснования возможности применения методов исследования стационарных процессов к процессам типа (11.3).

Круг данных, описываемых чисто косинусоидальной моделью (11.3), невелик. Во-первых, часто встречаются периодические зависимости, которые описываются не косинусоидальной, а более сложной функцией. Во-вторых, обычно в изучаемом процессе можно выделить не одну, а несколько периодических компонент с разными периодами.

Как известно из математического анализа, любую гладкую периодическую функцию $G(t)$ с периодом p (то есть функцию, для которой $G(t + kp) = G(t)$ для любого целого k) можно представить в виде ряда Фурье:

$$G(t) = \sum_{j=1}^{p} a_j \cos(j\omega t + \theta_j), \tag{11.4}$$

где $\omega = 2\pi/p$ называется основной (Найквистовой) частотой, a_j, θ_j — некоторые параметры. Частоты $j\omega$ называются гармониками основной частоты.

Функцию, являющуюся суммой нескольких периодических функций с разными периодами, можно задать в виде $G(t) = \sum_k G_k(t) = \sum_{j,\,k} a_{jk}\cos(j\omega_k t + \theta_{jk})$. Таким образом, получаем следующее обобщение модели (11.3).

Определение. *Говорят, что временной ряд описывается полигармонической моделью, если он представлен в виде:*

$$x_t = \sum_{j,k} a_{jk} \cos(j\omega_k t + \theta_{jk}) + \varepsilon_t \qquad (11.5)$$

где $\omega_k = 2\pi/p_k$, *а* ε_t *является белым шумом (см. п. 11.7).*

Пример ряда, описываемого полигармонической моделью, приведен на рис. 11.3.

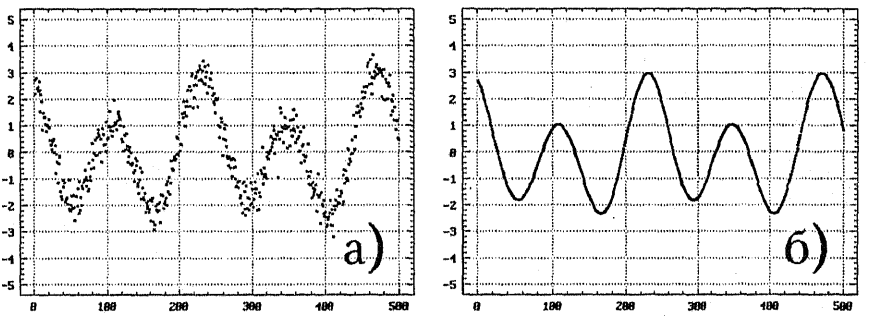


Рис. 11.3. а) — 500 значений ряда, описываемого полигармонической моделью $x_t = 2\cos(\frac{\pi}{60}t + \frac{\pi}{6}) + \cos(\frac{\pi}{120}t) + \varepsilon_t$, где ε_t — белый шум с дисперсией 0.16; б) — детерминированная компонента этого ряда

Если периоды p_k известны, то для определения величин a_{jk} и θ_{jk} можно использовать методы линейного регрессионного анализа. Если периоды p_k не известны, для их определения используют методы спектрального анализа. Мы практически не будем затрагивать этот вопрос в настоящей главе, так как он требует достаточно высокой математической подготовки читателей, и ограничимся лишь кратком рассказом об одном из его простейших случаев — периодограмме (см. п. 11.10). Наиболее полный обзор современного состояния методов прикладного спектрального анализа на русском языке дан в [58]. Этим методам посвящено много различной специальной литературы: [29], [39], [68], [38], [96], [47], [24].

Исторически анализ временных рядов из различных областей деятельности, включая экономику, начинался в конце прошлого и начале этого века именно с подбора полигармонических моделей для их описания. Однако с середины этого века стали появляться более простые модели и методы анализа временных рядов, включая линейные параметрические модели типа авторегрессии-скользящего среднего, на которых и будет в основном сосредоточено наше внимание.

11.7. Модели случайной компоненты

Прежде чем перейти к вопросам практического анализа временных рядов, кратко остановимся на математических основаниях этого анализа. При первом чтении этот параграф можно пропустить (тогда к нему время от времени придется возвращаться впоследствии).

Случайные процессы. Практический опыт показывает, что обычно временной ряд не удается полностью описать одной лишь детерминированной компонентой. В нем, как правило, присутствует и нерегулярная, случайная компонента. Ее поведение нельзя точно предсказать заранее. Для ее описания приходится привлекать понятия из теории вероятностей.

Для описания нерегулярной компоненты и всего временного ряда в целом используют понятие *случайного (стохастического) процесса* или случайной последовательности (как процесса от целочисленного аргумента). Ниже будут приведены некоторые сведения из теории случайных процессов, необходимые для понимания процедур прикладного анализа временных рядов. Более подробное изложение математической теории случайных процессов можно найти в [18], [75], [38], [96].

Определение. *Случайным процессом $X(t)$, заданном на множестве T, называют функцию от t, значения которой при каждом $t \in T$ является случайной величиной.*

Выделяются случайные процессы с непрерывным временем (когда T — интервал на числовой оси, например) и с дискретным временем (когда T — натуральный ряд или его часть, например). Последние чаще называют случайными последовательностями.

Если T — конечное множество, то случайный процесс — это просто совокупность случайных величин. Для статистического описания такой совокупности надо указать распределение вероятностей в конечномерном пространстве. Для этого можно использовать многомерную функцию распределения или плотности, если распределение непрерывное.

Если T — бесконечное множество, то для описания бесконечной совокупности случайных величин (которые в этом случае и составляют случайный процесс) применяется следующая конструкция.

Определение. *Говорят, что случайный процесс $X(t)$ задан, если для каждого t из T определена функция распределения величины $X(t)$:*

$$F_t(x) = P(X(t) \leqslant x), \tag{11.6}$$

для каждой пары элементов t_1, t_2 из T определена функция распределения двумерной случайной величины $(X(t_1), X(t_2))$

$$F_{t_1,t_2}(x_1, x_2) = P(X(t_1) \leqslant x_1, X(t_2) \leqslant x_2), \tag{11.7}$$

и вообще для любого конечного числа элементов $t_1, t_2, \ldots, t_n$ из множества T определена n-мерная функция распределения величины $(X(t_1), X(t_2) \ldots, X(t_n))$

$$F_{t_1,t_2,\ldots,t_n}(x_1, x_2, \ldots, x_n) = P(X(t_1) \leqslant x_1, X(t_2) \leqslant x_2, \ldots, X(t_n) \leqslant x_n) \tag{11.8}$$

При этом распределения (11.6)–(11.8) должны быть согласованы в том смысле, что «старшие» распределения определяют «младшие». Например,

$$F_{t_1,t_2,\ldots,t_n}(x_1, x_2, \ldots, x_n) = \lim_{x_{n+1}\to\infty} F_{t_1,t_2,\ldots,t_n,t_{n+1}}(x_1, x_2, \ldots, x_n, x_{n+1})$$

Функции (11.6) – (11.8) называют конечномерными распределениями случайного процесса.

На практике общее определение случайного процесса используется редко. Чаще случайные процессы задают с помощью предположений типа независимости приращений, марковского свойства траекторий и т.д. Примеры подобных определений будут даны чуть позже.

Гауссовские случайные процессы. Важным классом случайных процессов являются *нормальные (гауссовские) случайные процессы.* Все конечномерные распределения этих процессов являются нормальными. (Определения одномерного и двумерного нормального распределений даны в пунктах 2.4 и 2.5. Аналогичным образом можно определить многомерные нормальные распределения.) Для полного описания нормальных случайных процессов достаточно указать его двумерные распределения. Если эти распределения должным образом согласованы, то с их помощью можно задать любые конечномерные распределения вида (11.8). Это обстоятельство играет важную роль в прикладном анализе гауссовских процессов, позволяя ограничиться исследованием математического ожидания и корреляционной функции процесса.

Белый шум. Математически простейшей моделью случайной компоненты временного ряда является последовательность независимых случайных величин. Независимость двух случайных величин была определена ранее (смотри, например, п. 1.6). Аналогично определяется и независимость произвольного числа случайных величин. С помощью функций распределения независимость последовательности случайных величин определяется так:

Определение. *Пусть T — множество типа $t = 0, 1, 2, \ldots$ или $t = 0, \pm 1, \pm 2, \ldots$ Случайный процесс $X(t)$ называется последовательностью независимо распределенных случайных величин, если для любых наборов чисел $t_1, t_2, \ldots, t_n$*

$$F_{t_1,t_2,\ldots,t_n}(x_1, x_2, \ldots, x_n) = F_{t_1}(x_1) \cdot F_{t_2}(x_2) \cdot \ldots \cdot F_{t_n}(x_n) \tag{11.9}$$

Из (11.9) следует, что для последовательности независимых случайных величин все ее конечномерные распределения определяются с помощью одномерных распределений (11.6).

Определение. *Белым шумом называют временной ряд (случайный процесс) с нулевым средним, если составляющие его случайные величины $X(t)$ независимы и распределены одинаково (при всех t).*

Это так называемый *белый шум в узком смысле*. Белый шум в широком смысле будет определен позже, после определения свойства стационарности в широком смысле. В определение белого шума часто включают предположение о нормальности распределения величин $X(t)$. Другими словами, *гауссовский белый шум* — это последовательность независимых нормально распределенных случайных величин со средним 0 и общей дисперсией (скажем, σ^2).

Последовательности независимых случайных величин далеко не всегда адекватно описывают случайные компоненты временных рядов. Теорией и практикой для описания случайных последовательностей выработаны и более сложные модели. Некоторые из них мы упомянем ниже, а более подробно рассмотрим в дальнейшем.

Процессы скользящего среднего. Для этих процессов часто употребляют аббревиатуру MA — от английского moving average (движущееся среднее). Это сокращение стандартно используется в англоязычной литературе и статистических пакетах.

Пусть $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n, \ldots$ — независимые одинаково распределенные случайные величины (белый шум).

Определение. *Процессом скользящего среднего (первого порядка) со средним μ (сокращенно MA(1)) называют процесс $X(t)$:*

$$X(t) = \varepsilon_t + \theta\varepsilon_{t-1} + \mu, \tag{11.10}$$

где θ — некоторый числовой коэффициент, а μ — константа.

Заметим, что у процесса скользящего среднего (11.10) статистически зависимыми оказываются только соседние величины $X(t-1)$ и $X(t)$. Значения процесса, разделенные промежутком времени 2 и более, статистически независимы, ибо в их формировании участвуют разные слагаемые ε_t. По этой причине процессы скользящего среднего являются непосредственным и простейшим обобщением процессов белого шума.

Описание процессов скользящего среднего второго и более высоких порядков, а также свойств этих процессов, будет дано в гл. 14.

Процессы авторегрессии. Для них часто употребляют аббревиатуру AR — от английского autoregression.

Определение. *Процессом авторегрессии (первого порядка) со средним значением μ (сокращенно AR(1)) называют случайный процесс $X(t)$, удовлетворяющий соотношению:*

$$X(t) - \mu = \phi \cdot (X(t-1) - \mu) + \varepsilon_t, \tag{11.11}$$

где ϕ и μ — некоторые числа.

Члены процесса авторегрессии, разделенные промежутком времени $h > 0$, не становятся независимыми, как бы ни было велико h. Однако зависимость между ними быстро убывает с ростом h, если $|\phi| < 1$. Именно такие процессы авторегрессии обычно встречаются в прикладных задачах.

Процессы авторегрессии порядка 2 и выше будут определены в главе 14. Там же мы обсудим их свойства и области приложений.

Марковское свойство. Поведение многих процессов в будущем определяется только их состоянием в настоящем и воздействиями на процесс, которые будут оказываться в будущем. А предыдущее развитие процесса (то есть его состояние до настоящего времени) при этом несущественно. Такие процессы называются *марковскими*. Дадим этому понятию более строгое определение.

Пусть t, $t \in T$ — произвольный момент времени, который мы назовем «настоящим». Пусть A — произвольное событие, выраженное через случайные величины $X(s)$, где $s \leqslant t-1$. Это событие, относящееся к прошлому последовательности $X(\cdot)$. Пусть B — произвольное событие, относящееся к будущему процесса $X(\cdot)$, т.е. событие B выражается через случайные величины $X(s)$, где $s \geqslant t+1$. Рассмотрим условные вероятности событий A, B и AB при фиксированном значении $X(t)$. Обозначим эти условные вероятности через $P(A|X(t))$, $P(B|X(t))$ и $P(AB|X(t))$.

Определение. *Случайная последовательность $X(t)$, $t \in T$ называется марковской, если для любых A, B и t*

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

Нередко марковскому свойству последовательности $X(\cdot)$ дают несколько иное определение (впрочем, эквивалентное приведенному).

Определение. *Случайная последовательность $X(t)$, $t \in T$ называется марковской, если для любых A, B и t*

$$P(B|X(t), A) = P(B|X(t)).$$

В обычных обстоятельствах нет возможности проверить, обладает или нет наблюдаемый временной ряд этим свойством. Марковское

свойство для временного ряда обычно постулируют, когда физическая природа ряда дает для того основания.

В статистическом анализе марковское свойство процесса редко используется непосредственно. Обычно оно служит для вывода уравнений, описывающих изменения во времени каких-либо его средних характеристик (например, математического ожидания). Среди математических моделей временных рядов, которых мы далее касаемся, марковским свойством обладает процесс авторегрессии первого порядка (см. п. 14.1). Процесс авторегрессии $X(t)$ произвольного порядка $p \geqslant 1$ тоже можно представить как марковский, если его состоянием в момент t считать набор $(X(t), X(t-1), \ldots, X(t-p-1))$.

Стационарность. В теоретических исследованиях и практических задачах важную роль играют последовательности случайных величин, вероятностные свойства которых не изменяются во времени. Такие случайные последовательности называют стационарными. Их можно использовать для описания временных рядов, течение которых стабилизировалось и происходит в неизменных условиях.

Определение. *Случайный процесс $X(t)$ называется стационарным, если для любых n, $t_1, t_2, \ldots, t_n$ и τ распределения случайных величин $(X(t_1), \ldots, X(t_n))$ и $(X(t_1+\tau), \ldots, X(t_n+\tau))$ одинаковы.*

Это означает, что функции конечномерных распределений (11.8) не меняются при сдвиге времени, т.е.

$$F_{t_1+\tau,\, t_2+\tau \ldots, t_n+\tau}(x_1, \ldots, x_n) = F_{t_1, t_2, \ldots, t_n}(x_1, x_2, \ldots, x_n).$$

В частности, образующие стационарную случайную последовательность случайные величины $X(1), X(2), \ldots, X(t), \ldots$ распределены одинаково (но независимыми они, вообще говоря, не являются).

Этот вид стационарности называют также *стационарностью в узком смысле*. Другой вид стационарности — *стационарность в широком смысле*, — мы введем после того, как для случайных последовательностей мы определим их числовые характеристики.

Определенный ранее процесс белого шума является стационарным (в узком смысле).

11.8. Числовые характеристики временных рядов

Числовые характеристики временных рядов вводятся в полной аналогии с числовыми характеристиками случайных величин (см. п. 1.5).

Математическое ожидание *(первый момент) случайного процесса* $X(t)$ *— это функция* $m(t)$*, такая, что для каждого* t *значение функции* $m(t)$ *является математическим ожиданием случайной величины* $X(t)$*:*

$$m(t) = MX(t).$$

Функцию $m(t)$ часто называют *средним значением* процесса $X(t)$. Она используется для описания систематического изменения процесса. Например, для случайного процесса, допускающего запись в виде аддитивной модели (11.1), среднее значение равно $tr_t + s_t + c_t$. Заметим, что под словом «усреднение» здесь понимается усреднение случайной величины $X(t)$ при неизменном t, а не усреднение по времени, хотя такое тоже бывает. Ниже мы более подробно коснемся этого вопроса.

Ковариационная функция *случайного процесса* $X(t)$ *(кратко* $\operatorname{cov}(X(t), X(s))$*) — это величина*

$$B(s,t) = \operatorname{cov}(X(t), X(s)) = M[(X(t) - m(t))(X(s) - m(s))].$$

Она является функцией пары переменных (t, s)*. Иногда ее именуют функцией вторых центральных моментов.*

Значение ковариационной функции при $t = s$ задает *дисперсию* случайного процесса $DX(t) = \operatorname{cov}(X(t), X(t))$. Квадратный корень из $\operatorname{cov}(X(t), X(t))$ называют *стандартным отклонением* $\sigma(t)$ случайного процесса $X(t)$:

$$\sigma(t) = \sqrt{\operatorname{cov}(X(t), X(t))}.$$

Корреляционная функция *случайного процесса* $X(t)$ *— это величина:*

$$\operatorname{corr}(X(t), X(s)) = \frac{\operatorname{cov}(X(t), X(s))}{\sigma(t)\sigma(s)}.$$

Как и ковариационная функция, корреляционная функция также зависит от пары переменных (t, s).

При фиксированных t и s $\operatorname{corr}(X(t), X(s))$ по определению является коэффициентом корреляции (см. п. 1.6) случайных величин $X(t)$ и $X(s)$, и для него выполняются свойства 1—4 п. 1.6. Из определения $\operatorname{cov}(X(t), X(s))$ и $\operatorname{corr}(X(t), X(s))$ следует их симметрия относительно t и s:

$$\operatorname{cov}(X(t), X(s)) = \operatorname{cov}(X(s), X(t)),$$
$$\operatorname{corr}(X(t), X(s)) = \operatorname{corr}(X(s), X(t))$$

Заметим, что функции $m(t)$, $B(s,t)$ могут и не существовать: как мы знаем, не всегда случайные величины имеют математическое ожи-

дание и дисперсию. Но в статистической практике такие временные ряды, для которых $m(t)$ и $B(s,t)$ не существуют, встречаются редко. Поэтому в дальнейшем к средним значениям временных рядов и их ковариационной или корреляционной функциям мы будем обращаться без особых оговорок.

Ковариационная и корреляционная функции играют важную роль в теоретическом и в практическом анализе случайных процессов и временных рядов. Ниже мы обсудим их свойства, а также способы оценивания этих функций по наблюдениям (см. п. 11.10). А сейчас вернемся к свойству стационарности (см. п. 11.7) и с помощью функций $m(t)$ и $B(s,t)$ дадим ему другое определение.

Из определения стационарности, данного выше, следует, что для любых s, t и любого τ:

$$m(t+\tau) = m(t), \qquad B(s+\tau, t+\tau) = B(s, t). \tag{11.12}$$

Положив $\tau = -t$, мы получаем, что

$$m(t) = m(0), \qquad B(s, t) = B(s-t, 0).$$

Отсюда следует, что у стационарного процесса функции $m(t)$ и $\sigma(t)$ постоянны, а ковариационная функция $B(s,t)$ реально зависит не от пары (s, t), как в общем случае, а от $|s-t|$. Точно так же можно убедиться, что и корреляционная функция стационарного процесса является функцией $|s-t|$.

Рассмотрим $t = s+k$, $k > 0$. Положим по определению

$$r(k) = \operatorname{corr}(X(t), X(s)) = \operatorname{corr}(X(t), X(t+k)).$$

Автокорреляционная функция. *Автокорреляционной функцией стационарного процесса $X(t)$ называют функцию $r(k) = \operatorname{corr}(X(t), X(t+k))$, где $k > 0$ — целое число.*

Величину k часто называют *задержкой*, или *лагом*. Она указывает расстояние между членами временного ряда, для которых вычисляется коэффициент корреляции.

11.9. Процессы, стационарные в широком смысле

Вообще говоря, выполнение свойства (11.12) не гарантирует того, что процесс $X(t)$ является стационарным в смысле приведенного выше определения стационарности в узком смысле. Тем не менее, свойство (11.12) определенно отражает некую неизменность во времени свойств процесса $X(t)$. Поэтому принято следующее

Определение. *Случайный процесс $X(t)$ называется стационарным в широком смысле, если его среднее значение $m(t)$ постоянно, а ковариационная функция $B(s, t)$ зависит только от расстояния между аргументами, т.е. от $|t - s|$.*

Свойство стационарности в широком смысле играет важную роль при нахождении оценок числовых характеристик временных рядов (см. п. 11.10).

Белый шум в широком смысле. Аналогичное определение можно дать для белого шума:

Определение. *Временной ряд (случайный процесс) $X(t)$ называют белым шумом (в широком смысле), если для любого t выполняется $MX(t) = 0$ и*

$$\operatorname{cov}(X(s), X(t)) = \begin{cases} \sigma^2, & \text{когда} \quad s = t, \\ 0, & \text{когда} \quad s \neq t. \end{cases}$$

Из этого определения видно, что этот белый шум является стационарным (в широком смысле) случайным процессом. На практике различие между двумя видами белого шума (в широком и в узком смысле) не всегда проводятся четко. В дальнейшем, говоря о белом шуме в связи с прикладными исследованиями, мы чаще всего будем иметь в виду белый шум именно в только что введенном широком смысле.

Хотя на практике процессы белого шума в чистом виде встречаются не часто, они играют фундаментальную роль как в теории, так и в прикладном анализе временных рядов. Типичным для такого анализа является, например, процесс «выбеливания» временного ряда, т.е. исключения из него тренда, циклической, сезонной и прочих компонент, так чтобы остаток статистически не отличался от процесса белого шума.

Гауссовские процессы. Ясно, что стационарный в узком смысле случайный процесс является одновременно и стационарным в широком смысле, если существуют функции двух первых моментов. Уже отмечалось, что обратное, вообще говоря, неверно.

Одно из исключений составляют нормальные, или гауссовские случайные процессы, то есть процессы, конечномерные распределения которых (11.8) являются гауссовскими. Для гауссовских процессов любые конечномерные распределения определяются через функции $m(t)$ и $B(s, t)$. Поэтому гауссовские процессы, стационарные в широком смысле, одновременно являются стационарными в узком смысле. Это весьма важное и полезное обстоятельство, так как на практике проверка стационарности в узком смысле не осуществима. Судить о стационарности в широком смысле значительно проще. Для этого существуют раз-

личные статистические критерии, базирующиеся на одной реализации случайного процесса. Наиболее важные из таких критериев основаны на выборочных оценках автокорреляционной функции и спектральной плотности (см. [96]).

Замечание. Для некоторых протекающих во времени процессов модель гауссовского случайного процесса дает приемлемое по качеству описание. К сожалению, в полном объеме проверить по наблюдениям, верна ли эта модель, невозможно. Поэтому гауссовский случайный процесс представляет собой, в первую очередь, удобный математический объект.

Преобразование процесса в стационарный. Наиболее распространенным случаем нарушения стационарности на практике является изменение среднего значения $m(t)$ с изменением времени t. В тех случаях, когда $m(t)$ удается тем или другим способом оценить, преобразование $Y(t) = X(t) - m(t)$ превращает процесс в стационарный. Далее $Y(t)$ изучают как стационарный, используя для этого его специфические свойства.

11.10. Оценки числовых характеристик временных рядов

В каждый фиксированный момент времени t случайный процесс $X(t)$ является случайной величиной. Следовательно, для построения оценок его моментов $m(t)$ и $B(s, t)$ теоретически можно использовать те же методы, что и для обычных случайных величин. Напомним (см. п. 1.8.1 и п. 4.3), что при этом требуется некоторая совокупность независимых реализаций этой случайной величины X, полученных при повторении опыта в неизменных условиях. Другими словами, нужна выборка $x_1, \ldots, x_k$. Применение этой методики к случайному процессу $X(t)$ требует от нас набора реализаций (траекторий) этого процесса $x_1(t), \ldots, x_k(t)$.

В технических приложениях возможности для независимых повторений опыта иногда имеются. Скажем, изучая колебания напряжения в электрических сетях в течение суток, мы можем считать временные ряды, полученные в разные сутки, независимыми реализациями одного случайного процесса. Для большей уверенности, что повторения наблюдений произведены в неизменных условиях, можно сопоставлять данные за определенные дни недели (за вторники, например), отдельно по разным сезонам и т.д.

Однако в экономических, социальных, демографических и подобных процессах мы обычно имеем дело с единственной траекторией развития, повторить которую невозможно. Поэтому при изучении статистических

свойств таких процессов приходится обходиться этой самой единственной реализацией. Зато длина ее может расти. Значительная часть математических результатов о временных рядах относится к стационарным рядам, наблюдаемым на растущем интервале времени $(0, T)$. Они формулируются в виде предельных теорем при $T \to \infty$. Многие физические, технические и естественнонаучные приложения статистической теории нуждаются именно в такой постановке проблемы.

Впрочем, для упомянутых экономических, социальных и т.п. временных рядов эти результаты дают не очень много. Во-первых, эти ряды обычно довольно коротки. Во-вторых, они не стационарны, так как условия, в которых они протекают, изменяются с течением времени. Поэтому наблюдения даже из относительно недавнего прошлого порой мало что говорят о современных тенденциях.

По единственной реализации процесса $X(t)$ мы не можем составить оценки для его среднего, дисперсии, ковариации и т.д., как мы сделали бы это, располагая выборкой. Но некоторые похожие средние величины составить можно.

Оценка среднего значения. Имея ряд $x(t_1), \ldots, x(t_n)$ последовательных наблюдений случайного процесса $X(t)$, можно составить *«среднее по реализации»*: $\overline{m} = \frac{1}{n}\sum_{i=1}^{n} x(t_i)$. В теории временных рядов для наблюденных значений $x(t_1), \ldots, x(t_n)$ используют более короткую форму записи $x_1, \ldots, x_n$. В этих обозначениях среднее по реализации есть

$$\overline{m} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{11.13}$$

Оказывается, при некоторых условиях это среднее может служить оценкой математического ожидания процесса $X(t)$. Первым из таких условий является стационарность случайного процесса $X(t)$ (в широком смысле). Поскольку для стационарного процесса все моменты времени равноправны и его числовые характеристики неизменны во времени, в качестве оценки $m(t) = m$ естественно рассмотреть именно $\overline{m}$. Легко убедиться, что $\overline{m}$ как оценка m для стационарных процессов является несмещенной, т.е. $M\overline{m} = m$.

Рассмотрим вопрос о точности этой оценки. Естественно, хотелось бы, чтобы эта оценка $\overline{m}$ приближалась к неизвестному истинному значению с ростом числа наблюдений n, то есть была бы состоятельной (см. п. 4.5). Так как отклонение оценки от истинного значения можно описать с помощью ее дисперсии, то для состоятельности достаточно, чтобы

$$D\overline{m} \to 0 \qquad \text{при} \quad n \to \infty. \tag{11.14}$$

К сожалению, одна лишь стационарность случайного процесса не обеспечивает выполнения (11.14). Простейший и отчасти вырожденный пример стационарного процесса, для которого не выполняется свойство (11.14), устроен следующим образом. Рассмотрим стационарный процесс, для которого с вероятностью единица $X(t) = X(1)$ для любого t. Ясно, что траектории этого процесса являются константами. При этом:

$$\overline{m} = x_1 \qquad D\overline{m} = const.$$

В более общем случае типичным примером невыполнения условия (11.11) являются смеси, т.е. процессы, у которых различные участки траекторий сформированы при разных условиях. Более подробно модель таких процессов описана в [96].

Хоть мы и говорили о том, что предельные теоремы математической теории мало полезны для интересующей нас области приложений, все же приведем одно из достаточных условий для выполнения (11.14).

Теорема Слуцкого. *Для стационарного в широком смысле случайного процесса $X(t)$ оценка его среднего значения (11.13) состоятельна тогда и только тогда, когда:*

$$\frac{1}{n}\sum_{t=0}^{n-1} r_t \to 0 \qquad при \quad n \to \infty, \tag{11.15}$$

где r_t — автокорреляционная функция процесса.

Мы не будем более подробно обсуждать этот результат. Обратим внимание лишь на то, что для выполнения (11.15) достаточно, чтобы $r_t \to 0$ при $t \to \infty$. Последнее замечание позволяет на практике судить о том, можно ли использовать осреднение по одной реализации для получения состоятельных оценок его характеристик. Таким образом теорема Слуцкого подчеркивает важность анализа поведения автокорреляционной функции случайного процесса.

Еще два замечания о точности приближения оценки $\overline{m}$ к истинному значению. Первое из них касается скорости сходимости $\overline{m}$ к m. Можно показать, что стандартное отклонение $\overline{m}$ при больших n пропорционально $1/\sqrt{n}$, то есть увеличение точности оценки обратно пропорционально квадратному корню из объема наблюдений. Второе замечание относится к случаю, когда объема временного ряда недостаточно для получения достаточно точной оценки среднего значения. Определим величину T в виде:

$$T = \sum_{k=0}^{\infty} r_k$$

считая, что указанная сумма конечна. Величина T называется *временем корреляции* и дает представление о порядке величины промежутков времени τ, на которых сохраняется заметная корреляция между $X(t)$ и $X(t+\tau)$. Если объем n рассматриваемой реализации временного ряда меньше T, то оценка $\overline{m}$ считается весьма неточной. Введенная величина T позволяет также указать более точную скорость сходимости $\overline{m}$ к m. А именно, эта скорость пропорциональна $\sqrt{T/n}$.

Выборочная автокорреляционная функция. В главе 9 (п. 9.5.1) была подробно разобрана оценка коэффициента корреляции (9.17) пары случайных величин, построенная по выборкам. Методика получения оценок значений автокорреляционной функции $r(k)$ во многом напоминает случай двух выборок. Разберем ее устройство на оценке $r(1)$ — корреляции между соседними членами временного ряда X_t и X_{t+1}. (Напомним, что большие буквы X мы используем для обозначения случайного процесса, а малые буквы x — для обозначения реализации этого случайного процесса.)

Образуем из временного ряда $x_1, x_2, \ldots, x_n$ совокупность из $n-1$ пар: $(x_1, x_2), (x_2, x_3), \ldots, (x_{n-1}, x_n)$. Первый элемент каждой пары, в силу стационарности, мы можем рассматривать как реализацию случайной величины X_t, а второй — как реализацию случайной величины X_{t+1}. Тогда, согласно (9.17) оценка коэффициента корреляции между X_t и X_{t+1} может быть записана в виде:

$$\overline{r}_1 = \frac{\sum_{t=1}^{n-1}(x_t - \overline{x}_{(1)})(x_{t+1} - \overline{x}_{(2)})}{\sqrt{\left[\sum_{t=1}^{n-1}(x_t - \overline{x}_{(1)})^2 \sum_{t=1}^{n-1}(x_{t+1} - \overline{x}_{(2)})^2\right]}}, \tag{11.16}$$

где

$$\overline{x}_{(1)} = \sum_{t=1}^{n-1} x_t/(n-1), \qquad \overline{x}_{(2)} = \sum_{t=2}^{n} x_t/(n-1),$$

соответственно оценки средних значений величин X_t и X_{t+1}

При больших значениях n, учитывая что $\overline{x}_{(1)} \approx \overline{x}_{(2)} \approx \overline{x}$ и $n/(n-1) \approx 1$, выражение (11.16) часто заменяют гораздо более простым:

$$\overline{r}_1 = \frac{\sum_{t=1}^{n-1}(x_t - \overline{x})(x_{t+1} - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2}. \tag{11.17}$$

Аналогичным образом может быть определена оценка корреляции между X_t и X_{t+k} или k-го члена автокорреляционной функции r_k:

$$\overline{r}_k = \frac{\sum_{t=1}^{n-k}(x_t - \overline{x})(x_{t+k} - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2}. \tag{11.18}$$

Обратим внимание читателя, что точность приближения (11.18) заметно снижается с ростом лага k, как в силу ухудшения точности использованных выше замен, так и в силу уменьшения числа наблюдений используемых для вычисления оценки $\overline{r}_k$ Поэтому на практике обычно ограничиваются изучением небольшого числа первых членов автокорреляционной функции. Вряд ли имеет смысл рассматривать оценки r_k при $k > n/4$.

Функцию $\overline{r}_k$ аргумента k при $k = 1, 2, \ldots$ называют *выборочной автокорреляционной функцией* или, если не возникает недоразумений, просто автокорреляционной функцией. (При $k = 0$ $\overline{r}_k$ по определению равно 1 и это значение обычно исключают из рассмотрения как не несущее никакой информации.) В англоязычной литературе эту функцию также называют *сериальной корреляцией*. График выборочной автокорреляционной функции называют *коррелограммой*. На этом графике (см. рис. 11.3) кроме значений самой функции, обычно указывают доверительные пределы этой функции в предположении, что значения автокорреляционной функции равны 0 для всех $k \neq 0$. Более подробно об интерпретации графика выборочной автокорреляционной функции будет рассказано ниже при рассмотрении роли коррелограммы в практическом анализе временных рядов (см. п. 12.4.2).

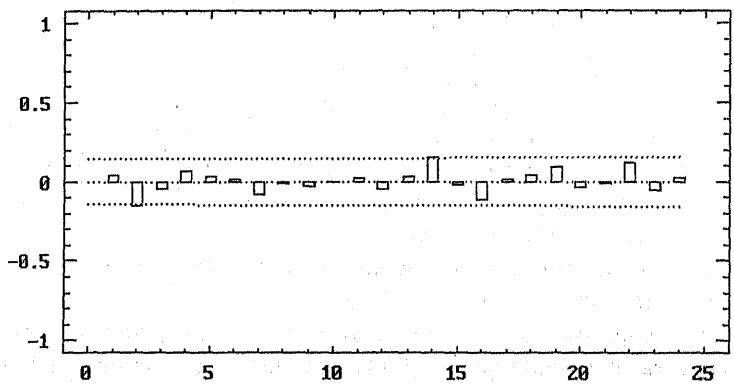


Рис. 11.3. Коррелограмма с доверительными интервалами (при равенстве нулю автокорреляционной функции для всех $k \neq 0$)

Свойства. Изучение свойств выборочных оценок автокорреляционной функции временного ряда — в общем случае довольно сложная и до конца не решенная задача.

М.Бартлетом в 1946 г. для случая бесконечного дискретного временного ряда ($-\infty < t < \infty$) было указано выражение дисперсии оценки

$\overline{r}_k$ для гауссовского процесса:

$$D\overline{r}_k = \frac{1}{n}\sum_{t=-\infty}^{t=\infty}\left[r_t^2 + r_{t-k}r_{t+k} - 4r_k r_t r_{t+k} + 2r_t^2 r_k^2\right]. \qquad (11.19)$$

Этот результат показывает, что мы не можем оценить по конечному отрезку временного ряда дисперсию оценки $\overline{r}_k$, так как она зависит от бесконечного неизвестного числа автокорреляций r_t. Поэтому на практике приходится довольствоваться лишь приближениями для выражения (11.19).

Другая проблема изучения свойств совокупности оценок $\overline{r}_k$ связана с тем, что оценки с различным лагом k коррелированы между собой. Это заметно затрудняет интерпретацию коррелограммы. Не касаясь более подробно этих и других проблем (см. [29], [38]), укажем свойства оценок $\overline{r}_k$ для наиболее простого и практически важного случая. А именно, рассмотрим свойства оценок автокорреляций для временного ряда, являющегося стационарной последовательностью независимых нормально распределенных случайных величин или, другими словами, гауссовским белым шумом (см. п. 11.7). В этом случае для любых k, не равных нулю, по определению $r_k = 0$. Таким образом, все слагаемые, стоящие под знаком суммы в выражении (11.19), равны нулю, кроме $r_0^2 = 1$. Отсюда дисперсия $\overline{r}_k$ равна:

$$D\overline{r}_k = \frac{1}{n}$$

Обратим внимание на то, что оценка $\overline{r}_k$ в форме (11.18) является смещенной. Можно показать, что $M\overline{r}_k \approx -\frac{1}{n}$, однако величина этого смещения стремится к нулю с ростом объема изучаемого ряда и не столь существенна в прикладном анализе.

Другим важным свойством оценки $\overline{r}_k$ является ее асимптотическая нормальность при $n \to \infty$.

Таким образом, для каждого отдельного значения $\overline{r}_k$ мы можем указать приблизительный 95% доверительный интервал в виде: $-1/n \pm 2/\sqrt{n}$. Границы этого доверительного интервала обычно наносятся на график коррелограммы и называются *доверительной трубкой*. Они в определенной мере позволяют судить о том, насколько изучаемый процесс напоминает белый шум. Указание 95% доверительных границ для каждого коэффициента автокорреляционной функции в отдельности не означает, что с 95% вероятностью все рассматриваемые оценки $\overline{r}_k$ одновременно попадают в доверительную трубку. Так, рассматривая 20 первых оценок $\overline{r}_k$ для гауссовского белого шума, довольно часто можно наблюдать, что одна или две из оценок выходят за границы довери-

тельной трубки. Это обстоятельство также затрудняет интерпретацию коррелограммы.

Периодограмма. Завершая рассказ об оценках основных числовых характеристик временного ряда, кратко остановимся на *периодограмме* — характеристике, особенно полезной для анализа временных рядов, допускающих представление в виде полигармонических моделей (11.5).

Многие временные ряды, возникающие в физических и технических приложениях, удобно рассматривать не во временной области значений аргумента, а в частотной. Этот переход можно совершить с помощью периодограммы. Ее назначение — обнаружение периодических составляющих в рассматриваемом ряде. Первое представление о наличии таких составляющих может дать обычный график. Если стационарный временной ряд на графике ведет себя более гладко и регулярно, чем гауссовский белый шум (см. рис. 11.1д), то можно предположить, что в нем есть периодические составляющие.

В настоящее время определение периодограммы часто использует понятие спектральной плотности, но так как это понятие не рассматривается в настоящей главе, мы дадим определение периодограммы в том виде, как оно было предложено А.Шустером в 1898 г.

Пусть x_t — временной ряд с нулевым средним, а t пробегает целые числа от 1 до n. Рассмотрим ковариацию ряда x_t с рядами $\cos(2\pi t/\lambda)$ и $\sin(2\pi t/\lambda)$, где λ — некоторая фиксированная величина, обычно именуемая периодом, или длиной волны. Пусть:

$$A = \frac{2}{n}\sum_{t=1}^{n} x_t \cos\frac{2\pi t}{\lambda}, \qquad B = \frac{2}{n}\sum_{t=1}^{n} x_t \sin\frac{2\pi t}{\lambda}.$$

Введем величину $S^2(\lambda)$:

$$S^2(\lambda) = A^2(\lambda) + B^2(\lambda)$$

Определение. *График зависимости $S^2(\lambda)$ от длины волны λ называется периодограммой.*

По замыслу, функция $S^2(\lambda)$ должна принимать большие значения (иметь локальные максимумы — пики) для тех значений λ, которые являются периодами для имеющихся у ряда x_t периодических составляющих. Практически это далеко не так, и часть максимумов $S^2(\lambda)$ к реальным периодам ряда x_t не имеет отношения. Вообще анализ периодограммы очень часто ведет к ложным выводам, и потому к нему надо подходить с осторожностью. Эти вопросы подробно освещены в литературе по спектральному анализу временных рядов. (Смотри, в частности, критический анализ в [58] и в гл. 4 книги [75].)

Глава 12

Временные ряды: практический анализ

12.1. Порядок анализа временных рядов

Кратко опишем общий порядок прикладного статистического анализа временных рядов. Обычно целью такого анализа является построение математической модели ряда, с помощью которой можно объяснить поведение ряда и осуществить прогноз его дальнейшего поведения.

Построение и изучение графика. Анализ временного ряда обычно начинается с построения и изучения его графика. Если нестационарность временного ряда очевидна, то первым делом надо выделить и удалить нестационарную составляющую ряда. Методы, используемые для этого, описаны в п. 12.3. Процесс удаления тренда и других компонент ряда, приводящих к нарушению стационарности, может проходить в несколько этапов. На каждом из них рассматривается ряд остатков, полученный в результате вычитания из исходного ряда подобранной модели тренда, или результат разностных и других преобразований ряда. Кроме графиков, признаками нестационарности временного ряда могут служить не стремящаяся к нулю автокорреляционная функция (за исключением очень больших значений лагов) и наличие ярко выраженных пиков на низких частотах в периодограмме.

Подбор модели для временного ряда. После того, как исходный процесс максимально приближен к стационарному, можно приступить к подбору различных моделей полученного процесса. Цель этого этапа — описание и учет в дальнейшем анализе корреляционной структуры рассматриваемого процесса. При этом на практике чаще всего используются два типа моделей: параметрические модели авторегрессии-скользящего среднего (ARMA-модели) и полигармонические модели (см. п. 11.6). ARMA-модели мы будем рассматриваить в главе 14, а описание способов подбора полигармонических моделей можно найти в книгах [29], [58], [68], [96].

Модель может считаться подобранной, если остаточная компонента ряда является процессом типа белого шума (см. п. 11.7, 11.9). После подбора модели обычно выполняются:

- оценка дисперсии остатков, которая в дальнейшем может быть использована для построения доверительных интервалов прогноза;
- анализ остатков с целью проверки адекватности модели.

Прогнозирование или интерполяция. Последним этапом анализа временного ряда может быть прогнозирование его будущих (экстраполяция) или восстановление пропущенных (интерполяция) значений и указания точности этого прогноза на базе подобранной модели. Обратим внимание, что хорошо подобрать математическую модель удается не для всякого временного ряда. Нередко бывает и так, что для описания подходят сразу несколько моделей. Неоднозначность подбора модели может наблюдаться как на этапе выделения детерминированной компоненты ряда, так и при выборе структуры ряда остатков. Поэтому исследователи довольно часто прибегают к методу нескольких прогнозов, сделанных с помощью разных моделей.

Методы анализа. Перечислим основные группы статистических приемов, используемых для анализа временных рядов:

- графические методы представления временных рядов и их сопутствующих числовых характеристик;
- методы сведения к стационарным процессам;
- методы исследования внутренних связей между элементами временных рядов.

Ниже будет подробно рассказано о каждой из этих групп методов.

12.2. Графические методы анализа временных рядов

Зачем нужны графические методы. В выборочных исследованиях простейшие числовые характеристики описательной статистики (среднее, медиана, дисперсия, стандартное отклонение, коэффициенты асимметрии и эксцесса) обычно дают достаточно информативное представление о выборке. Графические методы представления и анализа выборок при этом играют лишь вспомогательную роль, позволяя лучше понять локализацию и концентрацию данных, их закон распределения.

Роль графических методов при анализе временных рядов совершенно иная. Дело в том, что табличное представление временного ряда и описательные статистики чаще всего не позволяют понять характер процесса, в то время как по графику временного ряда можно сделать

довольно много выводов. В дальнейшем они могут быть проверены и уточнены с помощью расчетов.

Человеческий глаз довольно уверенно определяет по графику временного ряда:

- наличие тренда и его характер;
- наличие сезонных и циклических компонент;
- степень плавности или прерывистости изменений последовательных значений ряда после устранения тренда. По этому показателю можно судить о характере и величине корреляции между соседними элементами ряда.

Так графический анализ ряда обычно задает направление его дальнейшего анализа.

Построение и изучение графика. Построение графика временного ряда — совсем не такая простая задача, как это кажется на первый взгляд. Современный уровень анализа временных рядов предполагает использование той или иной компьютерной программы для построения их графиков и всего последующего анализа. Ряд полезных рекомендаций при построении графика вручную даны в [76]. Большинство статистических пакетов и электронных таблиц снабжено теми или иными методами настройки на оптимальное представление временного ряда, но даже при их использовании могут возникать различные проблемы, например:

- из-за ограниченности разрешающей способности экранов компьютеров размеры выводимых графиков могут быть также ограничены;
- при больших объемах анализируемых рядов точки на экране, изображающие наблюдения временного ряда, могут превратиться в сплошную черную полосу.

Для борьбы с этими затруднениями используются различные способы. Наличие в графической процедуре режима «лупы» или «увеличения» позволяет изобразить более крупно выбранную часть ряда, однако при этом становится трудно судить о характере поведения ряда на всем анализируемом интервале. Приходится распечатывать графики для отдельных частей ряда и состыковывать их вместе, чтобы увидеть картину поведения ряда в целом. Иногда для улучшения воспроизведения длинных рядов используется *прореживание*, то есть выбор и отображение на графике каждой второй, пятой, десятой и т.д. точки временного ряда. Эта процедура позволяет сохранить целостное представление ряда и полезна для обнаружения трендов. На практике полезно сочетание

обеих процедур: разбиения ряда на части и прореживания, так как они позволяют подметить разные черты в поведении временного ряда.

Еще одну проблему при воспроизведении графиков создают *выбросы* — наблюдения, в несколько раз превышающие по величине большинство остальных значений ряда. Их присутствие тоже приводит к неразличимости колебаний временного ряда, так как масштаб изображения программа автоматически подбирает так, чтобы все наблюдения поместились на экране. Выбор другого масштаба на оси ординат устраняет эту проблему, но резко отличающиеся наблюдения при этом остаются за границами экрана.

Дадим еще несколько полезных советов по построению и оформлению графика временного ряда:

- внимательно следите за масштабом представления данных по каждой из осей, так как они, как правило, различаются. Многие программы автоматически используют экспоненциальную форму записи для обозначения делений осей, например 0.241E+03 вместо 241, что не всегда удобно и оправданно;
- не забывайте указывать, какие величины отображает каждая из осей и их единицы измерения. Это особенно важно при представлении экономических временных рядов, где наряду с равномерными шкалами часто используются логарифмические шкалы;
- точки на графике временного ряда обычно соединяют отрезками прямых линий. Однако в некоторых ситуациях эти линии могут вносить существенное искажение в представление о поведении ряда. Поэтому полезно построить график временного ряда с линиями между точками и без них, и внимательно изучить оба этих графика;
- наличие не слишком густой координатной сетки облегчает восприятие графиков.

Вспомогательные графики. При анализе временных рядов часто используются вспомогательные графики для числовых характеристик ряда:

- график выборочной автокорреляционной функции (коррелограммы) с доверительной зоной (трубкой) для нулевой автокорреляционной функции;
- график выборочной частной автокорреляционной функции (см. п. 14.3) с доверительной зоной для нулевой частной автокорреляционной функции;
- график периодограммы.

Первые два из этих графиков позволяют судить о связи (зависимости) соседних значений временного ряда, они используются при подборе параметрических моделей авторегрессии-скользящего среднего. График периодограммы позволяет судить о наличии гармонических составляющих во временном ряде. Эти графики и свойства соответствующих функций рассмотрены в параграфе 12.4.

Учитывая, что многие методы анализа временных рядов рассчитаны на работу с рядами с нормально распределенной случайной компонентой, в процедуры анализа временных рядов обычно включают различные графики на нормальной вероятностной бумаге. Самый распространенный из них подробно описан в главе 5.

12.3. Методы сведения к стационарности

После изучения графика временного ряда обычно пробуют выделить во временном ряде тренд, сезонные и периодические компоненты. После их исключения временной ряд должен стать стационарным. Кроме того, для облегчения дальнейшего анализа иногда используются преобразования значений временного ряда (точнее, той шкалы, в которой они измерены) — это позволяет приблизить распределение значений временного ряда к нормальному или сделать дисперсию этих значений более постоянной (иначе говоря, стабилизировать дисперсию).

В п. 12.3.1 мы рассмотрим методы оценки и удаления тренда, а в пп. 12.3.2—12.3.4 — оценки и удаления сезонных эффектов и циклических компонент временного ряда. В п. 12.3.5 рассматриваются преобразования шкалы измерений временного ряда — переход к логарифмической шкале и преобразование Бокса-Кокса.

12.3.1. Выделение тренда

Метод наименьших квадратов. Для оценки и удаления трендов из временных рядов чаще всего используется метод наименьших квадратов. Этот метод подробно обсуждался в гл. 8 при рассмотрении задач линейного регрессионного анализа.

Говоря языком регрессионного анализа, значения временного ряда x_t рассматривают как отклик (зависимую переменную), а время t — как фактор, влияющий на отклик (независимую переменную):

$$x_{t_i} = f(t_i, \theta) + \varepsilon_i\,, \qquad i = 1, \ldots, n$$

где f — функция тренда (она обычно предполагается гладкой), θ — неизвестные нам параметры (параметры модели временного ряда), а ε_i —

независимые и одинаково распределенные случайные величины, распределение которых мы предполагаем нормальным. Метод наименьших квадратов состоит в том, что мы выбираем функцию тренда так, чтобы

$$\sum_{i=1}^{n}[x_{t_i} - f(t_i, \theta)]^2 \to \min_{\theta}$$

Для временных рядов типично, что статистические предпосылки регрессионного анализа, как они перечислены в (8.2), (8.3), выполняются не полностью. Это особенно касается предположения о независимости случайных отклонений. Для временных рядов характерна именно взаимная зависимость его членов (по крайней мере, не далеко отстоящих по времени). Тем не менее, оценки тренда и в этих условиях обычно оказываются разумными, если выбрана адекватная модель тренда и если среди наблюдений нет больших выбросов. Упомянутые выше нарушения предпосылок регрессионного анализа сказываются не столько на значениях оценок, сколько на их статистических свойствах. Так, при наличии заметной зависимости между членами временного ряда оценки дисперсии, основанные на остаточной сумме квадратов (8.10), дают неправильные результаты. Неправильными оказываются и доверительные интервалы для коэффициентов модели, и т.д. В лучшем случае их можно рассматривать как очень приближенные.

Это положение может быть частично исправлено, если применять модифицированные алгоритмы метода наименьших квадратов, такие как взвешенный метод наименьших квадратов [37], или метод наименьших квадратов для коррелированных наблюдений [20]. Однако для этих методов требуется дополнительная информация о том, как меняется дисперсия наблюдений или их корреляция. Если же такая информация недоступна, исследователям приходится применять классический метод наименьших квадратов, несмотря на указанные недостатки.

Пример 1. Проиллюстрируем применение метода наименьших квадратов для данных об урожайности зерновых в СССР, представленных в табл. 1.2 и на рис. 11.1а. Визуальное изучения графика данных позволяет предположить, что тренд этого ряда может быть задан в виде прямой линии $tr_t = b_0 + b_1 \cdot t$. С помощью метода наименьших квадратов по формулам, аналогичным (8.7) и (8.8), находим, что

$$\hat{b}_0 = \overline{x} \qquad (\text{где } \overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t), \tag{12.1}$$

$$\hat{b}_1 = \frac{\sum_{t=1}^{n}(x_t - \overline{x})\,(t - \frac{t(t+1)}{2})}{\sum_{t=1}^{n}(t - \frac{t(t+1)}{2})^2}. \tag{12.2}$$

В формуле (12.2) в качестве независимой переменной фигурирует время t.

Для данных рассматриваемого ряда $\hat{b}_0 = 5.868$, $\hat{b}_1 = 0.275$. При этом коэффициент $\hat{b}_0$ показывает среднюю урожайность зерновых в начальный (1945 г.) момент времени рассматриваемого ряда, а коэффициент $\hat{b}_1$ дает оценку среднегодового прироста урожайности. Подробная таблица результатов процедуры выделения тренда, а также дальнейший анализ ряда, приведены в главе 13, где рассматривается решение этой задачи с помощью статистических пакетов ЭВРИСТА и SPSS.

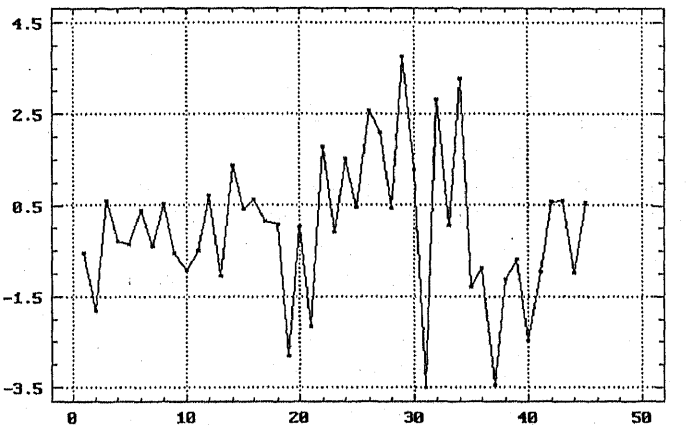


Рис. 12.1. График ряда остатков, полученный в результате удаления из ряда урожайности зерновых в СССР линейного тренда

На рис. 12.1 дан график остатков временного ряда после удаления из него подобранной модели тренда $tr_t = 5.868 + 0.275 \cdot t$. Дальнейший анализ полученного ряда (см. главу 13) показывает, что его уже можно рассматривать как последовательность независимых случайных величин. Более того, для описания остатков можно применять гауссовскую модель, согласно которой их совокупность можно рассматривать как выборку из некоторой нормальной совокупности (с нулевым средним). Последнее означает, что на базе подобранной модели тренда и модели случайной составляющей (независимые ошибки) можно осуществлять прогноз будущих значений ряда и строить доверительную зону для прогноза, используя оценку (8.11) дисперсии остатков.

Пример 2. Поведение случайной компоненты, которое мы наблюдали в примере 1 — это скорее исключение, чем правило. Чтобы убедиться в этом, рассмотрим поведение случайной компоненты у курса доллара весной и летом 1994 г. (рис. 11.1б). График этого ряда позволяет предположить, что его тренд также описывается простой линейной за-

висимостью. Найдя с помощью метода наименьших квадратов значения оценок коэффициентов $\hat{b}_0 = 1675.33$ и $\hat{b}_1 = 3.722$, и, вычтя значения тренда из рассматриваемого временного ряда, получим остаточную компоненту. Ее график приведен на рис. 12.2.

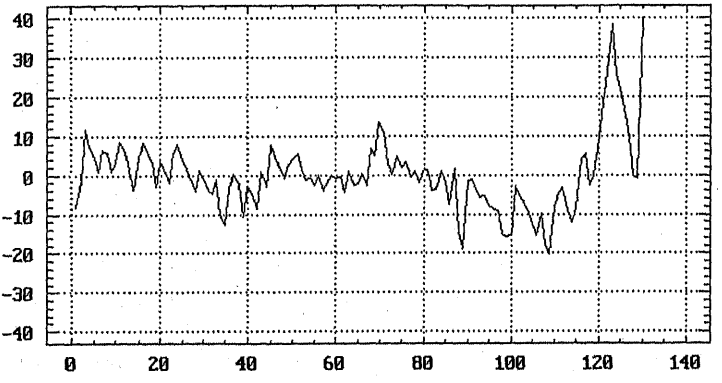


Рис. 12.2. График ряда остатков, полученный в результате удаления из ряда курса доллара линейного тренда

Даже визуальный анализ показывает, что остатки ведут себя не как последовательность независимых одинаково распределенных случайных величин (сравните рис. 12.2, например, с графиком гауссовского белого шума на рис. 11.1д). Действительно, приведенные на рис. 12.3 графики выборочной автокорреляционной функции и выборочной частной автокорреляционной функции (см. п. 14.3) показывают, что соседние значения этого ряда сильно зависимы при значениях лага от 1 до 5. При дальнейшем увеличении значения лага зависимость исчезает. Как следует интерпретировать графики указанных функций и что они означают, мы подробно расскажем ниже в п. 12.4.

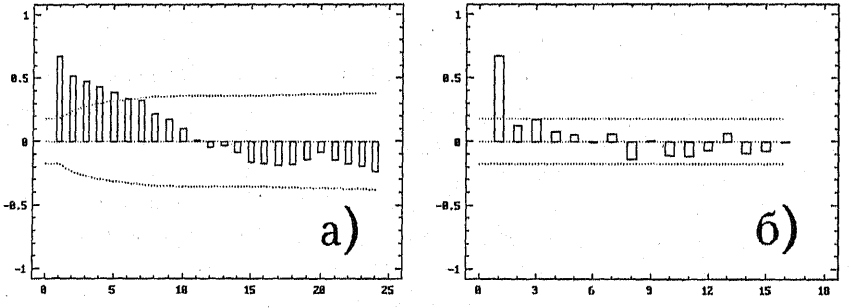


Рис. 12.3. а) – выборочная автокорреляционная функция ряда остатков, полученный в результате удаления из ряда курса доллара линейного тренда; б) – выборочная частная автокорреляционная функция того же ряда

Замечание. Как будет показано ниже в п. 12.4, такое поведение выборочной автокорреляционной и частной автокорреляционной функций характерно для процессов авторегрессии первого порядка (см. п. 11.7 и 14.1). У этих процессов значение в момент времени t формируется из их значения в предыдущий момент времени с некоторым весовым коэффициентом (в данном случае этот коэффициент равняется примерно 0.7) и независимой случайной добавки — белого шума.

Простые разностные операторы. Наряду с методом наименьших квадратов, для удаления тренда можно использовать и ряд других методов. Одним из них является метод перехода от исходного ряда к ряду разностей соседних значений ряда. В более общем виде эта идея описывается с помощью применения к ряду разностных операторов различных порядков. Эти методы сведения временного ряда к стационарному являются частным случаем общего метода, предложенного Дж.Боксом и Г.Дженкинсом в 1970 году [15]. В целом, мы относимся к разностным методам критически, но считаем нужным упомянуть о них. Они часто обсуждается в литературе и представлены во многих статистических пакетах.

Определение. *Процедура перехода от ряда* x_t *при* $t = 1, \ldots, n$ *к ряду* $y_t = x_t - x_{t-1} = \nabla x_t$ *при* $t = 2, \ldots, n$ *называется взятием первых разностей, а оператор* ∇ *называется простым разностным оператором первого порядка.*

Заметим, что длина ряда первых разностей y_t на единицу меньше, чем длина исходного ряда x_t. Покажем, как действует разностный оператор на временном ряде x_t, содержащем простой линейный тренд $tr_t = b_0 + b_1 \cdot t$:

$$y_t = \nabla x_t = x_t - x_{t-1} =$$
$$= b_0 + b_1 t + \varepsilon_t - b_0 - b_1(t-1) - \varepsilon_{t-1} = b_1 + \varepsilon_t - \varepsilon_{t-1} \qquad (12.3)$$

Из (12.3) видно, что в отличие от ряда x_t, преобразованный ряд y_t уже не содержит тренда, однако структура случайной компоненты в нем уже другая. Так, если ε_t была последовательностью независимых случайных величин, то последовательность $\varepsilon_t - \varepsilon_{t-1}$, $t = 2, \ldots, n$, этим свойством уже не обладает. Корреляция между соседними членами этой последовательности равна -0.5.

Итак, удалить линейный тренд из временного ряда можно разными способами: с помощью метода наименьших квадратов или с помощью простого разностного оператора первого порядка.

На рис. 12.4 приведен график ряда, полученного в результате применения разностного оператора ∇ к ряду урожайности зерновых. Дальнейшие исследования показывают, что в данном случае проще анализи-

ровать ряд остатков, полученный после удаления линейного тренда методом наименьших квадратов (рис. 12.1), чем ряд первых разностей. А в общем случае, к сожалению, нельзя сказать, какой из этих двух методов удаления тренда предпочтительней. Все зависит от заранее неизвестной структуры случайной компоненты временного ряда ε_t. Так, для временного ряда с независимыми приращениями проще анализировать ряд его первых разностей. Он будет представлять из себя просто белый шум.

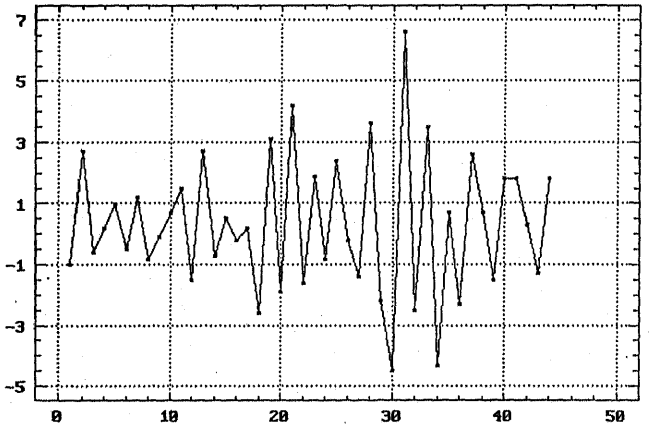


Рис. 12.4. Ряд первых разностей для урожайности зерновых

Аналогичным образом можно ввести разностный оператор второго и более высоких порядков. Так, простой разностный оператор второго порядка преобразует ряд x_t к ряду y_t, где

$$y_t = \nabla^2 x_t = \nabla(\nabla x_t) = \nabla(x_t - x_{t-1}) = \nabla x_t - \nabla x_{t-1} = \\ = x_t - 2x_{t-1} + x_{t-2}.$$

Часто для записи разностных операторов используют оператор B «сдвига назад»: $Bx_t = x_{t-1}$. При этом

$$\nabla x_t = (1-B)x_t, \quad \nabla^2 x_t = (1-B)^2 x_t, \quad \nabla^k x_t = (1-B)^k x_t.$$

Ясно, что длина ряда $\nabla^k x_t$ на k единиц меньше длины исходного ряда.

Простые разностные операторы более высоких порядков позволяют удалять из ряда полиномиальные тренды соответствующих порядков.

Возможно, разностные операторы действительно пригодны для удаления трендов, особенно если не видна подходящая аналитическая модель тренда. Беда же метода разностных операторов в том, что не всегда ясно, как приложить к исходному временному ряду результаты статистического анализа его разностей. В частности, это относится к законам распределения ошибок. К тому же эти разности могут иметь (и

часто имеют) гораздо более сложную статистическую структуру, нежели исходный ряд. Рассмотрите, например, первые разности для процесса авторегрессии первого порядка. (Об авторегрессии см. п. 14.1.)

12.3.2. Выделение сезонных эффектов

Многие временные ряды, особенно экономические, содержат *сезонные компоненты*. Сезонные компоненты ряда могут как представлять интерес сами по себе, так и выступать в роли мешающего фактора. В обоих случаях задача исследователя — выделить и устранить их из ряда.

Для этого есть несколько способов. Их выбор обычно определяется моделью подбираемого временного ряда. Ниже мы рассмотрим две наиболее распространенных модели описания экономических временных рядов. Первая из них включает в себя тренд (tr_t), сезонную (s_t) и случайную (ε_t) компоненты:

$$x_t = tr_t + s_t + \varepsilon_t \tag{12.4}$$

Вторая модель, кроме перечисленных выше компонент, включает еще и циклическую компоненту (c_t):

$$x_t = tr_t + s_t + c_t + \varepsilon_t \tag{12.5}$$

Циклическая компонента c_t в экономических временных рядах отражает периоды роста и спада экономической активности различной амплитуды и продолжительности. (Более подробно о каждой из компонент модели временного ряда рассказано в п. 11.5.)

Сезонные эффекты на фоне тренда. Предположим, что рассматриваемый временной ряд $x_1, \ldots, x_n$ может быть описан аддитивной моделью (12.4). Пусть p — период последовательности s_t, так что $s_t = s_{t+p}$ для всякого t. Наша задача — оценить значения s_t по наблюдениям x_t при том, что величина p известна.

Для этого сначала мы должны оценить тренд tr_t. Это можно сделать с помощью метода наименьших квадратов или его модификаций. Обозначим через $\hat{tr}_t$ полученную оценку тренда. Обычно она выражается в виде некоторой достаточно гладкой функции зависящей от времени t и одного или нескольких неизвестных параметров. Оценки этих параметров и дает метод наименьших квадратов. Наиболее распространенные функции тренда приведены в п. 11.6.

Затем для каждого сезона i, $1 \leqslant i \leqslant p$, рассмотрим все относящиеся к нему разности

$$x_i - \hat{tr}_i, \quad x_{i+p} - \hat{tr}_{i+p}, \quad \ldots, \quad x_{i+mp} - \hat{tr}_{i+mp}. \tag{12.6}$$

(для простоты изложения мы предполагаем, что в рассматриваемом ряде содержится целое число периодов, т.е. $n = (m+1)p$.) Каждое из этих отклонений x_i от $\hat{tr}_i$ можно рассматривать как результат влияния сезонных изменений. Усреднение этих разностей дает нам оценку сезонной компоненты s_i. В качестве простейшей оценки можно взять простое среднее, т.е. положить

$$\hat{s}_i = \frac{1}{m+1}\sum_{l=0}^{m}(x_{i+lp} - \hat{tr}_{i+lp}) \qquad \text{для} \qquad i = 1, \ldots, p \tag{12.7}$$

В качестве других оценок $\hat{s}_i$ можно взять взвешенное среднее, цензурированное среднее, медиану и т.д. Перечисленные средние уменьшают влияние резко выделяющихся наблюдений.

Часто бывает желательно, чтобы сумма сезонных эффектов равнялась нулю. Тогда переходят к скорректированным оценкам сезонных эффектов в виде

$$s_i^* = \hat{s}_i - \frac{1}{p}\sum_{i=1}^{p}\hat{s}_i. \tag{12.8}$$

В практических задачах распространена ситуация, когда сезонные колебания пропорциональны среднему значению процесса в рассматриваемый момент времени. Для описания подобных данных можно использовать одну из следующих моделей:

$$x_t = tr_t \cdot s_t + \varepsilon_t$$
$$x_t = tr_t \cdot s_t \cdot \varepsilon_t.$$

Первая из них является смешанной мультипликативно-аддитивной моделью, вторая — мультипликативной моделью временного ряда. Для модели $x_t = tr_t \cdot s_t + \varepsilon_t$ при оценке сезонных эффектов вместо совокупности (12.6) рассматривают совокупность (12.9) частных от деления x_{i+lp} на $\hat{tr}_{i+lp}$, выраженных в процентах.

$$\frac{x_{i+lp}}{\hat{tr}_{i+lp}} \cdot 100\% \qquad \text{при} \quad l = 0, 1, 2, \ldots, m \tag{12.9}$$

В этом случае оценкой сезонной компоненты или *сезонным индексом* называют величину:

$$\hat{s}_i = \frac{1}{m+1}\sum_{l=0}^{m}\left(\frac{x_{i+lp}}{\hat{tr}_{i+lp}} \cdot 100\%\right) \qquad \text{где} \quad 1 \leqslant i \leqslant p \tag{12.10}$$

Так же, как и в случае аддитивной модели, вместо среднего арифметического в правой части (12.10) может фигурировать взвешенное или

цензурированное среднее, медиана или другие более устойчивые к грубым выбросам оценки. Сезонные индексы (12.10) особенно популярны при анализе экономических временных рядов. Оценка сезонного индекса для мультипликативной модели будет рассмотрена ниже в более общей ситуации.

На практике считается, что оценки сезонных эффектов недостаточно точны, если число периодов в исследуемом сезонном временном ряде меньше пяти-шести. Это означает, например, что при рассмотрении месячных данных для достаточно точной оценки сезонных эффектов необходимы, как минимум, наблюдения за пять-шесть лет.

Удаление сезонной компоненты. Получив оценки сезонных эффектов (12.7), в аддитивной модели легко провести удаление этих эффектов из рассматриваемого ряда, вычитая их из начальных значений ряда. Подобная процедура часто носит название *сезонного выравнивания ряда* или *сезонной коррекции ряда*. Еще одно название этой процедуры — *сезонная декомпозиция*. Для мультипликативно-аддитивной модели эта процедура сводится к делению значений исходного ряда на соответствующие сезонные индексы и умножению на 100%.

Проиллюстрируем оценку сезонных индексов и их использование при прогнозировании на основе данных о производстве молока в России.

Пример. В таблице 12.1 и на рис. 12.5 приведены величины месячного производства молока (в тыс. тонн) в России с января 1992 г. по октябрь 1996 г. (по данным ЦСУ Госкомстата России).

Таблица 12.1

Производство молока в России с января 1992 г. по октябрь 1996 г. (тыс. тонн в месяц)

Месяц \ год	1992	1993	1994	1995	1996
январь	2015	1759	1510	1172	1038
февраль	2123	1773	1484	1226	1104
март	2624	2361	1988	1651	1439
апрель	2891	2649	2211	1859	1521
май	3335	3203	2559	2392	1827
июнь	4071	3936	3209	2864	2446
июль	4040	3861	3204	2714	2369
август	3392	3321	2687	2420	2081
сентябрь	2467	2438	2031	1925	1577
октябрь	2092	1760	1506	1338	1081
ноябрь	1494	1299	1050	984	
декабрь	1562	1345	1054	1020	

График ряда показывает, что производство молока имеет тенденцию к сокращению, обусловленную сокращением поголовья молочного ста-

да, и подвержено сильным сезонным колебаниям с максимумом производства в летние месяцы и минимумом — в зимние. При этом величина сезонных колебаний пропорциональна среднему уровню производства.

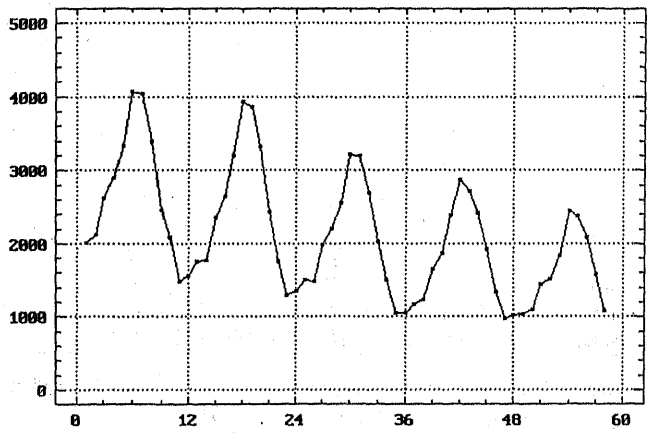


Рис. 12.5. Ежемесячное производство молока
в России с 01.1992 по 10.1996 (в тыс. тонн)

Оценим сезонные индексы этого ряда и проведем выравнивание ряда с учетом сезонности. Для описания тренда используем линейную модель $tr_t = a + b \cdot t$, где $t = 1, 2, \ldots, 58$. Оценки неизвестных коэффициентов a и b методом наименьших квадратов есть: $\hat{a} = 2841.1$ и $\hat{b} = -23.63$. Таким образом, в каждой точке t можно вычислить $\hat{tr}_t = \hat{a} + \hat{b} \cdot t$. Подобранная модель тренда описывает общую тенденцию поведения ряда. Но сделать на базе этой модели достаточно точный прогноз ежемесячного производства молока в следующем году нельзя, учитывая большую сезонную изменчивость ряда. Для построения месячного прогноза необходимо оценить сезонные эффекты, или сезонные индексы.

На графике 12.5 видно, что величина сезонных колебаний пропорциональна среднему уровню производства. Поэтому для описания сезонных колебаний следует использовать мультипликативно-аддитивную или мультипликативную модель. Воспользуемся первой из этих моделей. Для получения оценок сезонных индексов используем формулы (12.9) и (12.10).

В таблице 12.2 приведены в процентах значения отношений $x_t/\hat{tr}_t$ для каждого месяца t. Обратим внимание на то, что полученные для каждого месяца индексы в таблице 12.2 довольно устойчивы. Так, производство молока в июне в среднем на 155% превышает среднегодовой уровень, а в октябре — составляет только 75% от него. Для получения сезонных индексов производства молока (12.10) для каждого

Таблица 12.2

Значения отношений $x_t/\hat{tr}_t$ для временного ряда с данными о производстве молока в России (в %)

Месяц \ год	1992	1993	1994	1995	1996
январь	71.52	69.42	67.10	59.59	61.67
февраль	75.99	70.63	66.65	63.09	66.52
март	94.72	94.95	90.23	86.01	87.96
апрель	105.26	107.55	101.45	98.05	94.34
май	122.48	131.30	118.70	127.76	115.00
июнь	150.82	162.93	150.50	154.93	156.29
июль	150.99	161.41	151.95	148.71	153.69
август	127.90	140.22	128.88	134.34	137.107
сентябрь	93.86	103.97	98.53	108.28	105.54
октябрь	80.31	75.82	73.91	76.28	73.51
ноябрь	57.88	56.54	52.13	56.86	
декабрь	61.07	59.15	52.95	59.75	

месяца следует провести усреднение данных по строкам таблицы 12.2. Полученный результат приведен в таблице 12.3.

Таблица 12.3

Сезонные индексы производство молока в России (в %)

Месяц	Индекс по всем данным	Индекс по данным 1992–1995 гг.
январь	65.86	66.96
февраль	68.58	69.23
март	90.77	91.84
апрель	101.33	103.64
май	123.05	126.01
июнь	155.09	156.18
июль	153.35	154.83
август	133.69	134.46
сентябрь	102.04	102.64
октябрь	75.97	77.73
ноябрь	55.85	56.80
декабрь	58.23	59.31

Для проведения сезонного выравнивания каждое значения исходного ряда следует разделить на соответствующий ему сезонный индекс и умножить полученный результат на 100%. Полученный результат приведен на рис. 12.6. Как видно из графика, выровненный ряд имеет ярко выраженную тенденцию линейного убывания.

Замечание. Для прогнозирования поведения рассмотренного ряда могут быть применены и другие методы. В частности, можно описывать этот ряд моделью, использующей простые и сезонные разностные операторы. Рассматривая

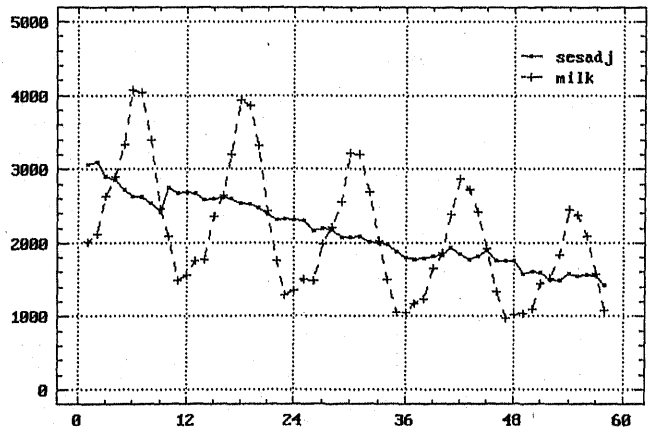


Рис. 12.6. Сезонное выравнивание ряда производства молока в России

этот пример, как иллюстрацию вычисления сезонных индексов, мы не касаемся в нем вопросов выбора наилучшей модели. Эта обширная тема, требующая определенной подготовки, выходит за рамки данной книги.

Прогнозирование. Посмотрим к каким результатам привела бы эта методика, если бы мы хотели получить прогноз на 1996 г, по данным 1992–1995 гг. Учитывая устойчивое поведение сезонного индекса в данной задаче, мы проведем его оценку по данным за 4 года. (В других задачах такой объем данных может быть недостаточным.) Повторим описанные выше действия для данных 1992–1995 гг. Подобранная модель линейного тренда $\hat{tr}_t$ имеет вид:

$$\hat{tr}_t = 2899.9 - 26.64 \cdot t \qquad (12.11)$$

Ее коэффициенты в целом не сильно отличаются от коэффициентов, полученных выше по всем данным. (Следует учитывать, что оценка $\hat{b}$ по всем данным несколько завышена. Это связано с отсутствием данных последних двух месяцев 1996 г., которые с учетом сезонности являются обычно самыми низкими в году. Здесь уместно заметить, что использование метода наименьших квадратов для подбор модели тренда сезонных рядов с незавершенными циклами, как это было сделано выше, обычно влечет за собой подобные смещения оценок. Поэтому лучше этого избегать или использовать устойчивые методы оценивания.)

Оценки сезонных индексов для данных 1992–1995 гг., рассчитанные по (12.9), приведены в таблице 12.4. Сравнение данных таблиц 12.2 и 12.4 показывает, что они хорошо согласуются между собой. Для получения сезонного индекса $\hat{s}_t$ усредним по строкам данные таблицы 12.4. Полученный результат приведен в третьем столбце табл. 12.3. Из этой

Таблица 12.4

Месяц \ год	1992	1993	1994	1995
январь	70.13	68.88	67.59	61.22
февраль	74.58	70.16	67.23	64.95
март	93.05	94.43	91.16	88.71
апрель	103.50	107.09	102.64	101.34
май	120.54	130.89	120.29	132.32
июнь	148.57	162.62	152.75	160.80
июль	148.89	161.29	154.47	154.69
август	126.25	140.30	131.23	140.06
сентябрь	92.74	104.17	100.50	113.16
октябрь	79.44	76.06	75.52	79.90
ноябрь	57.31	56.79	53.37	59.71
декабрь	60.54	59.49	54.30	62.91

таблицы видно хорошее согласие сезонных индексов, что говорит об устойчивости этого показателя.

Для осуществления прогноза ряда на 10 месяцев 1996 г. следует сначала рассчитать предварительный ежемесячный прогноз $prog_t$ по подобранной модели тренда (12.11). А именно, вычислить значения $\hat{tr}_t$ для следующих 10-ти значений t, то есть для $t = 49, 50, 51, \ldots, 58$ Результаты этого прогноза приведены во втором столбце таблицы 12.5. Для получения окончательного прогноза ряда надо скорректировать предварительный прогноз с помощью полученных сезонных индексов, вычислив $(prog_t \cdot s_t)/100$ для указанных выше значений t. Результаты этой процедуры приведены в третьем столбце табл. 12.5. В четвертом столбце таблицы приведены реальные данные за 10 месяцев 1996 г. Наглядное сравнение прогноза с реальными данными дано на рис. 12.7, где пунктирной линией обозначены реальные данные за 1995–1996 гг., а сплошной линией — построенный прогноз на 1996 г.

В табл. 12.5 и на рис. 12.7 видно хорошее согласие прогноза с реальными данными в первые 5 месяцев. Относительная погрешность прогноза здесь не превышает 3%. В последующие 4 месяца прогноз ниже реальных данных на 6–10%. Относительная ошибка прогноза в последний, 10-ый месяц не превышает 3%. Заметим, что наибольшие различия прогноза с реальными данными наблюдаются в летние месяцы, сезонный индекс которых подвержен наибольшим колебаниям (см. табл. 12.2 и 12.4). Эти колебания из года в год имеют порядок 10–20%. С учетом этого, можно признать в целом хорошее согласие прогноза с реальными данными.

Аналогичным образом можно осуществить прогноз на 1997 г. по данным 1992–1996 гг. Говорить о достоверности подобного прогноза можно лишь при сохранении общих тенденций, наблюдаемых в предыду-

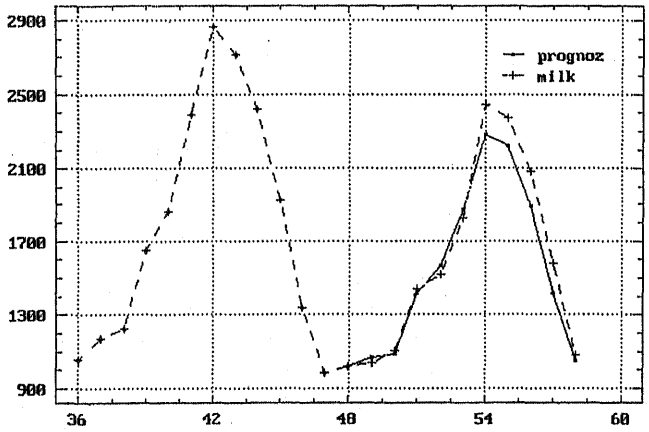


Рис. 12.7. Производство молока в России за 1995–1996 гг. и прогноз на 1996 г. (в тыс. тонн)

Таблица 12.5

Прогноз производства молока в России на 1996 г. (в тыс. тонн) и его сравнение с реальными данными

Месяц	Тренд	Прогноз на 1996 г.	Реальные данные
январь	1595	1068	1038
февраль	1568	1085	1104
март	1541	1415	1439
апрель	1515	1569	1521
май	1488	1875	1827
июнь	1461	2282	2446
июль	1435	2221	2369
август	1408	1893	2081
сентябрь	1381	1418	1577
октябрь	1355	1053	1081

щие годы. Для более аргументированных прогнозов необходимо привлечение дополнительной информации (например, о тенденциях изменения поголовья молочного стада).

12.3.3. Метод скользящих средних

При наличии в ряде циклической компоненты расчет сезонных эффектов несколько отличается от описанного выше. В этом случае для выяснения сезонных вкладов в виде (12.6) или (12.9) необходимо оценить не только тренд, но и циклическую компоненту. Проще всего одновременно оценить тренд и циклическую компоненту можно с помо-

щью скользящего среднего. Этот метод полезен и тогда, когда модель тренда не ясна. Рассмотрим его подробнее.

О методе скользящих средних. Метод скользящих средних — один из самых старых и широко известных способов сглаживания временного ряда. Он основан на переходе от начальных значений ряда к их средним значениям на интервале времени, длина которого выбрана заранее. При этом сам выбранный интервал времени скользит вдоль ряда.

Получаемый таким образом ряд скользящих средних ведет себя гораздо более гладко, чем исходный ряд, за счет усреднения отклонений исходного ряда. Таким образом эта процедура дает представление об общей тенденции поведения ряда. Ее применение особенно полезно для рядов с сезонными колебаниями и неясным характером тренда. В частности, переход к ряду скользящих средних может быть использован для выявления сезонной компоненты (или сезонного индекса) временного ряда.

Вид средних. Применяя метод скользящих средних, можно использовать различные виды усреднения значений ряда: среднее арифметическое (простое или с некоторыми весами), медианы и др. К сглаживанию с помощью медианы (медианное сглаживание) прибегают тогда, когда среди наблюдений есть выбросы (резко выделяющиеся данные).

Примеры для обсуждения. Мы дадим формальные определения метода скользящих средних, используя для их иллюстрации два следующих примера. В первом из них величина интервала сглаживания равна 7, по числу дней недели. Во втором примере величина интервала сглаживания равна 12, что соответствует двенадцати месяцам года. Это типичные интервалы сглаживания в экономических временных рядах. Для ежеквартальных данных подходящим может оказаться сглаживание с интервалом 4, для почасовых данных, собираемых круглосуточно, сглаживание с интервалом 24, и т.д. Вообще говоря, величину интервала сглаживания целесообразно выбирать равным или кратным периоду сезонности. При этом каждый интервал вычисления скользящего среднего будет содержать данные, отвечающие всему периоду (периодам) сезонности.

Пример 1. На рис. 12.8а приведен среднесуточный трафик (величина загрузки) телекоммуникационного канала Париж-Москва сети Internet за четыре последовательных недели (февраль 1996 г.). Этот график характеризует интенсивность (в килобитах в секунду) получения информации западными пользователями с российских компьютерных серверов по указанному каналу. Из графика видно, что в отдельные дни недели (субботу и воскресенье) происходит уменьшение загрузки кана-

ла, в другие дни загрузка повышается. Кроме того, вероятно, имеет место плавный рост объема загрузки с начала месяца к его концу. Таким образом, можно предположить, что рассматриваемый временной ряд имеет тренд и сезонную компоненту с периодом сезонности $p = 7$ дней.

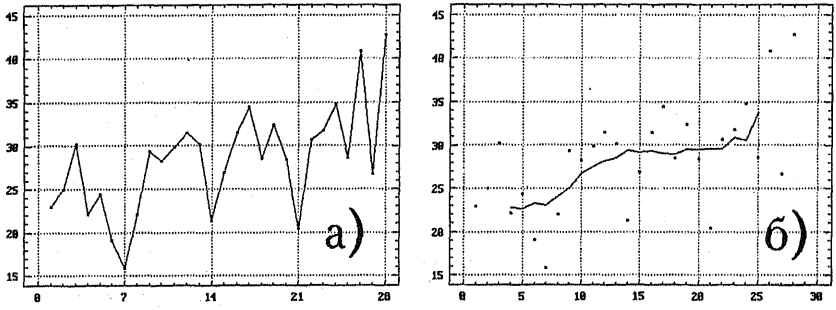


Рис. 12.8. Среднесуточный трафик в Кбит/сек
телекоммуникационного канала Париж-Москва в феврале 1996 г.:
а) исходный ряд; б) исходный ряд и его скользящее среднее

Пример 2. На рис. 12.9а приведен график ежемесячных продаж шампанского за ряд лет. На графике отчетливо прослеживаются сезонные колебания с пиками в декабре каждого года и спадами в летние месяцы. Период сезонности этих данных равен 12.

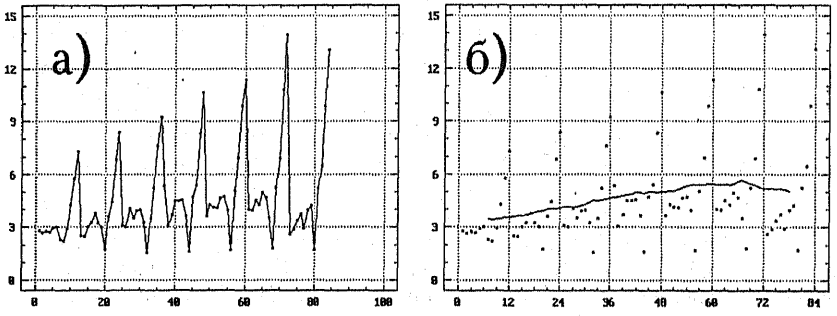


Рис. 12.9. Месячный объем продаж шампанского за ряд лет:
а) исходный ряд; б) исходный ряд и его скользящее среднее

Вычисление скользящего среднего. Дадим формальное определение скользящего среднего сначала для интервалов сглаживания, длина которых выражается нечетными числами. Причина, по которой четные и нечетные длины рассматриваются порознь, выяснится чуть ниже. Пусть $p = 2m + 1$. Обозначим через $\hat{x}_t$ результат усреднения элементов ряда

$$x_{t-m}, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_{t+m}.$$

Если обсуждаемое среднее есть среднее арифметическое, то

$$\hat{x}_t = \frac{1}{2m+1}(x_{t-m} + \ldots + x_{t-1} + x_t + x_{t+1} + \cdots + x_{t+m}).$$

Для медианного сглаживания

$$\hat{x}_t = \mathrm{med}(x_{t-m}, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_{t+m}).$$

Для четных $p = 2m$ определение несколько сложнее. Причина в том, что вычисленное по аналогичным формулам (как среднее арифметическое, медиана и т.д.) усредненное значение нельзя сопоставить какому-либо определенному моменту времени t. Например, среднее арифметическое $\frac{1}{2m}\sum_{t=1}^{t=2m}(x_t)$ следовало бы сопоставить моменту времени $t = (2m+1)/2$, но такого момента во временном ряде нет. А это сильно осложняет дальнейшее выделение сезонных эффектов.

Поэтому при четном интервале сглаживания $2m$ в усреднении задействуют не $2m$, а $2m+1$ значений временного ряда, но значения на краях интервала сглаживания берут с весами $1/2$. Так, при использовании для усреднения среднего арифметического получается следующая формула:

$$\hat{x}_l = \frac{1}{2m}\left(\frac{1}{2}x_{l-m} + x_{l-m+1} + \cdots + x_{l+m-1} + \frac{1}{2}x_{l+m}\right) \qquad (12.12)$$

Выражение (12.12) задает величину простого скользящего среднего $\hat{x}_l$ для $l = m+1, m+2, \ldots, n-m$ при четной величине интервала сглаживания $p = 2m$.

Свойства скользящего среднего. Скользящее среднее, сглаживая исходный ряд, дает представление об общей тенденции поведения ряда — его тренде и циклической компоненте. Сделаем несколько замечаний о его свойствах.

1. При применении метода скользящих средних выбор величины интервала сглаживания должен делаться из содержательных соображений и привязываться к периоду сезонности для сезонных данных. Если процедура скользящего среднего используется для сглаживания несезонного ряда, то чаще всего величину интервала сглаживания выбирают равной трем, пяти или семи. Чем больше интервал усреднения, тем более гладкий вид имеет график скользящих средних.

2. Соседние члены ряда скользящих средних сильно коррелированы, так как в их формировании участвуют одни и те же члены исходного ряда. Это может приводить, к тому, что ряд скользящих средних может содержать циклические компоненты, отсутствующие в исходном ряде. Это явление носит название *эффекта Слуцкого-Юла* (см. [39], [38]).

3. В качестве метода усреднения, кроме упомянутых выше среднего арифметического и медианы, можно рассматривать *взвешенные скользящие средние*, когда значения исходного ряда суммируется с определенными весами. Подобные процедуры целесообразны, если изменение временного ряда во времени носит явно нелинейный характер. Мы не будем более касаться этого вопроса. Он подробно изложен, например, в [38].

Оценка сезонных компонент. Предположим, что наблюдаемый временной ряд имеет структуру $x_t = tr_t + c_t + s_t + \varepsilon_t$, где $tr_t + c_t$ — тренд и циклическая составляющая, s_t — сезонная составляющая, а ε_t — случайная составляющая ряда. Пусть p — период последовательности s_t, так что $s_t = s_{t+p}$ для всякого t. Пусть величина p нам известна. Мы хотим оценить значения s_t по наблюдениям x_t.

Порядок оценки сезонных компонент в этом случае, в целом, аналогичен рассмотренному в п. 12.3.2. Только вместо оценки тренда методом наименьших квадратов мы будем использовать скользящее среднее в качестве совместной оценки тренда и циклической компоненты. Обозначим через $\hat{x}_t$ скользящее среднее с периодом p, построенное по ряду x_t. Для упрощения обозначений начнем нумерацию величин $\hat{x}_t$ с единицы, так что ряд из скользящих средних есть: $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k$. Соответственно изменим нумерацию исходного ряда так, чтобы величине $\hat{x}_t$ соответствовал член x_t. (При этом приходится отбросить $[p/2]$ первых членов исходного ряда, для которых значения $\hat{x}_t$ не определены. Здесь через $[p/2]$ обозначена целая часть от деления p пополам.)

Ради простоты предположим, что $k = (m+1)p$, где m — положительное целое число. (Обратим внимание, что общая длина n исходного ряда при этом равна $n = (m+2)p$ при четном p и $n = (m+2)p - 1$ при нечетном p.) Для каждого сезона i, $1 \leqslant i \leqslant p$, рассмотрим все относящиеся к нему разности

$$x_i - \hat{x}_i,\ x_{i+p} - \hat{x}_{i+p},\ \ldots,\ x_{i+mp} - \hat{x}_{i+mp}. \tag{12.13}$$

Каждое из этих отклонений x_i от $\hat{x}_i$ можно рассматривать как результат влияния сезонных изменений. Усреднение этих разностей дает нам оценку сезонной компоненты s_i. В качестве простейшей оценки можно взять простое среднее, т.е. положить

$$\hat{s}_i = \frac{1}{m+1} \sum_{l=1}^{m+1} (x_{i+lp} - \hat{x}_{i+lp}) \qquad \text{для} \qquad i = 1, \ldots, p \tag{12.14}$$

Как и выше (см. 12.7), вместо простого среднего можно взять взвешенное среднее, цензурированное среднее, медиану и т.д., для уменьшения влияния резко выделяющихся наблюдений.

Для мультипликативной модели временного ряда, когда $x_t = tr_t \cdot c_t \cdot s_t \cdot \varepsilon_t$ целесообразно перейти к логарифмам $y_t = \log x_t$. Тогда $y_t = d_t + g_t + r_t + \delta_t$, где $d_t = \log tr_t$, $g_t = \log c_t$, $r_t = \log s_t$, $\delta_t = \log \varepsilon_t$. К ряду y_t можно применить изложенную выше методику, начиная с вычисления скользящих средних и кончая составлением оценки $\hat{r}_i$ для r_i. Оценкой для исходной величины $s_i = e^{r_i}$ будет служить $e^{\hat{r}_i}$, если

$\log x$ — натуральный логарифм x, либо $\hat{s}_i = 10^{\hat{r}_i}$, если наши логарифмы десятичные.

Удаление сезонной компоненты. Оно проводится так же, как и в разобранном выше случае. Для аддитивной модели удаление сезонной компоненты сводится к вычитанию оцененной сезонной компоненты из исходного ряда. Для мультипликативной модели эта процедура заключается в делении значений исходного ряда на соответствующие сезонные индексы.

Пример оценки и удаления сезонных компонент с помощью скользящего среднего рассмотрен ниже в главе 13 (пример 13.2к). Этот пример решается с помощью компьютерных программ SPSS и Эвриста.

12.3.4. Сезонные разностные операторы

Еще один способ удаления сезонных компонент из ряда основан на использовании специальных разностных операторов, которые называются *сезонными*. Использование этих операторов особенно распространено в линейных моделях временных рядов типа авторегрессии-скользящего среднего (см. главу 14).

Пусть $x_1, \ldots, x_n$ — реализация временного ряда, а p — период его сезонности.

Определение. *Процедура перехода от ряда x_t (при $t = 1, \ldots, n$) к ряду $y_t = x_t - x_{t-p} = \nabla_p x_t$ (при $t = p+1, \ldots, n$) называется взятием первой сезонной разности, а оператор ∇_p называется сезонным разностным оператором с периодом p.*

Преобразование $x_t - x_{t-p}$ может быть также записано с помощью оператора сдвига назад B в виде:

$$y_t = x_t - x_{t-p} = (1 - B^p)x_t$$

На рис. 12.10 изображен результат применения сезонного оператора ∇_{12} к ряду месячных продаж шампанского за 7 лет. Длина полученного ряда сократилась на 12. Разброс значений полученного ряда существенно сократился и в нем уже не просматриваются периодические колебания.

Для этого ряда теперь можно попытаться подобрать, например, линейную параметрическую модель типа авторегрессии-скользящего среднего (см. гл. 14). В случае успешного подбора модели можно осуществить прогноз для ряда разностей. Этот прогноз может быть пересчитан и для исходного ряда.

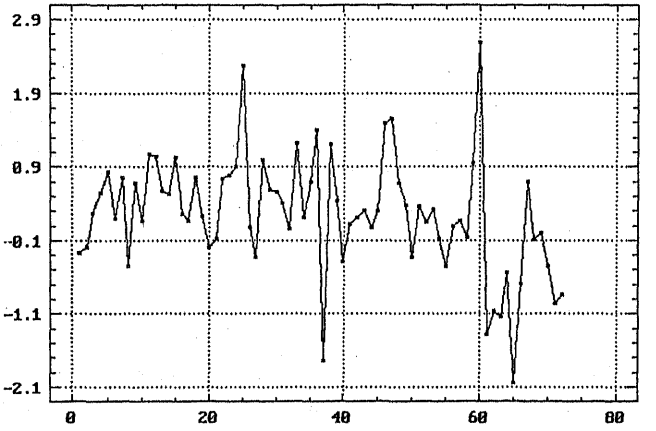


Рис. 12.10. Ряд сезонных разностей для продаж шампанского

Другим способом устранения сезонности может служить метод наименьших квадратов, в котором используется полигармоническая модель (11.5) для описания периодически повторяющихся сезонных эффектов. Сравнивая эти два подхода устранения сезонных эффектов, заметим, что метод сезонных разностей значительно проще и наглядней.

Сезонные операторы более высоких порядков. Как и в случае с простыми разностными операторами (см. п. 12.3.1), иногда бывают полезны сезонные операторы более высоких порядков. Так, сезонный оператор второго порядка с периодом p задается соотношением:

$$\nabla_p^2 x_t = \nabla_p(\nabla_p x_t) = \nabla_p(x_t - x_{t-p}) = x_t - 2x_{t-p} + x_{t-2p}$$

или, с помощью оператора сдвига назад B:

$$\nabla_p^2 x_t = (1 - B^p)^2 x_t = (1 - 2B^p + B^{2p}) x_t.$$

Смешанные разностные операторы. Выше указывалось, что простые и сезонные разностные операторы могут быть использованы соответственно для удаления тренда и сезонной компоненты из временного ряда. Если временной ряд одновременно содержит обе эти компоненты, то их удаление возможно с помощью последовательного применения простых и сезонных операторов. Нетрудно убедиться, что порядок применения этих операторов не существен:

$$\nabla\nabla_p x_t = \nabla(x_t - x_{t-p}) = (x_t - x_{t-1}) - x_{t-p} - x_{t-p-1}) = \nabla_p \nabla x_t.$$

Замечание. Существуют и другие методики оценивания и учета сезонных эффектов. Часть из них опирается на совместное использование методов однофакторного анализа и анализа временных рядов. Другие используют обобщенные сезонные модели процессов авторегрессии-скользящего среднего. Мы не будем останавливаться на этих вопросах.

12.3.5. Преобразование шкалы

К преобразованиям значений временного ряда (точнее — к преобразованиям той шкалы, в которой измерены значения временного ряда) прибегают обычно по двум причинам: либо для того, чтобы приблизить распределение к нормальному (например, избавиться от его скошенности), либо для того, чтобы сделать дисперсию временного ряда более постоянной (иными словами, стабилизировать дисперсию временного ряда).

Пусть переменная x употребляется для записи значений временного ряда. Рассмотрим преобразование x в y по правилу $y = f(x)$, где f обозначает некоторую определенную функцию. (Обычно f — монотонная функция; тогда от значений y можно однозначно вернуться к значениям x.) Применяя преобразование f к каждому члену ряда x_t, мы получим новый временной ряд $y_t = f(x_t)$.

Логарифмическое преобразование. Чаще других используемое преобразование — логарифмическое, когда

$$y = \log x, \qquad \text{либо} \qquad y = \log(x + c),$$

где c — некоторая постоянная величина, выбор которой находится в распоряжении исследователя. При логарифмическом преобразовании

$$y_t = \log(x_t + c).$$

Логарифмическое преобразование можно применять только к положительным величинам. В тех случаях, когда часть членов ряда x_t отрицательна, перед переходом к логарифмам ко всем членам ряда прибавляют постоянную c, добиваясь того, чтобы $x_t + c > 0$ при всех t.

Посмотрим, как действует логарифмическое преобразование на практике. Скошенные (асимметричные) распределения довольно часто появляются в экономической статистике. Типичным примером являются данные о душевом доходе: лиц с небольшими и средними доходами гораздо больше, чем лиц с высокими доходами. А этих последних значительно больше, чем лиц с очень высокими доходами. Примерная гистограмма распределения доходов приведена на рис. 12.11а. Прологарифмируем данные о доходах и вновь построим гистограмму. Она приведена на рис. 12.11б. Видно, что в логарифмической шкале распределение доходов близко к нормальному (гауссовскому).

Логарифмическое преобразование может оказаться полезным и при некоторых нарушениях стационарности наблюдаемого ряда. Допустим, что мы наблюдаем процесс $x_t = b_t \cdot z_t$, где z_t — стационарный ряд, а b_t — некоторая положительная неслучайная последовательность. Обозначив

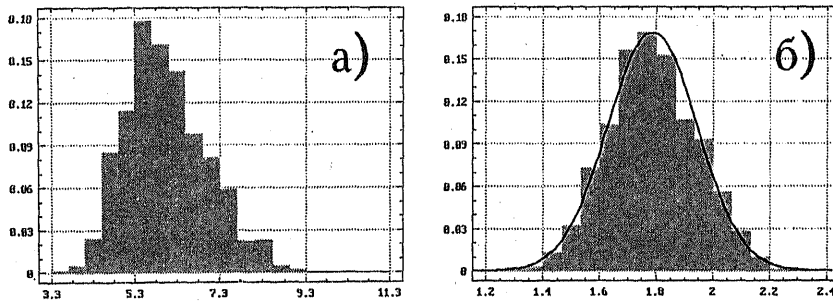


Рис. 12.11. Гистограмма данных о доходах: а) исходная шкала, б) логарифмическая шкала (для наглядности на график (б) наложена функция плотности нормального распределения)

Dz_t через σ^2, получим, что $Dx_t = \sigma^2 b_t^2$ изменяется во времени. Переход к логарифмической шкале $y_t = \log x_t$ дает

$$y_t = \log b_t + \log z_t.$$

При этом ряд $\log z_t$ — стационарный, его дисперсия во времени не изменяется. Это позволяет применить метод наименьших квадратов для выделения тренда $\log b_t$ из ряда y_t.

Примером временного ряда, дисперсия которого изменяется со временем, является ряд продаж шампанского (рис. 11.1.в). На рис. 12.12 приведены данные о продажах шампанского в логарифмической шкале. Видно, что логарифмирование устранило рост размаха сезонных колебаний значений ряда.

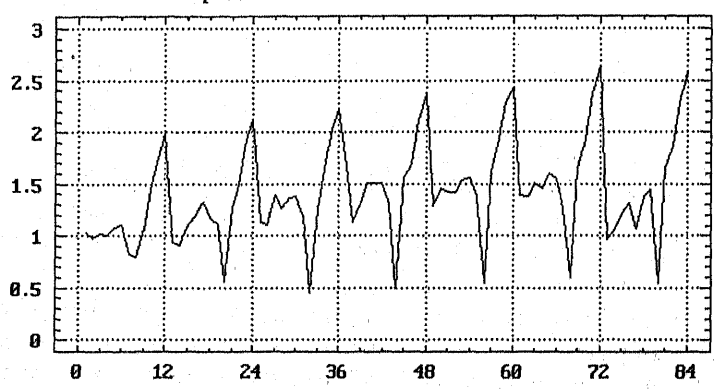


Рис. 12.12. Данные месячных продаж шампанского в логарифмической шкале

Преобразование Бокса-Кокса. Логарифмическое преобразование является частным случаем некоторого семейства преобразований, которое ввели Дж.Бокс и Д.Кокс в 1964 г. [97]. С тех пор эти преобразования приобрели популярность. Преобразования, образующие это семейство,

зависят от параметра λ, $\lambda \geqslant 0$. Если вернуться к формуле преобразований $y = f(x)$, то можно сказать, что теперь $y = f(x, \lambda)$, где значение $\lambda \geqslant 0$ исследователь может выбрать по своему усмотрению. Бокс и Кокс предложили следующую формулу

$$f(x, \lambda) = \begin{cases} (x_t^{\lambda} - 1)/\lambda & \text{при} \quad \lambda > 0 \\ \log x_t & \text{при} \quad \lambda = 0 \end{cases} \tag{12.15}$$

Нетрудно убедиться, что при фиксированном λ функция $f(x, \lambda)$ монотонно возрастает с ростом x, и что $f(x, \lambda)$ непрерывна не только по x, но и по λ, если $\lambda \geqslant 0$.

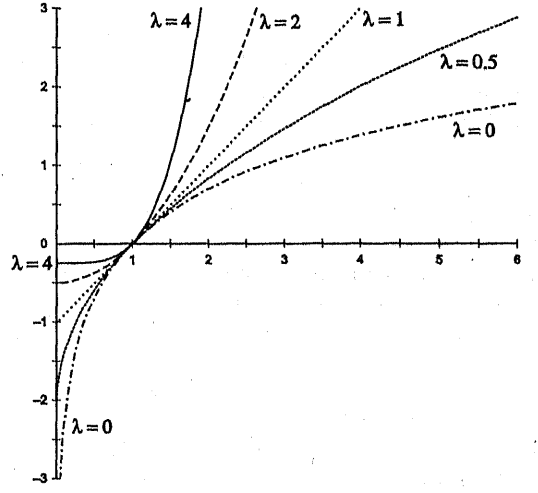


Рис. 12.13. Характер преобразования Бокса-Кокса при различных значениях параметра λ

Как видно из рис. 12.13, преобразование Бокса-Кокса при $\lambda < 1$ растягивает расстояния между малыми значениями и сжимает его между большими по величине значениями данных. При $\lambda > 1$ наблюдается обратная картина.

Следует заметить, что применение преобразования Бокса-Кокса к временным рядам может порождать определенные трудности в их дальнейшем анализе. Дело в том, что показатель степени λ существенно влияет на корреляционную функцию процесса и способен значительно усложнить дальнейший подбор модели ряда.

Ряды, имеющие отрицательные значения. Подобно логарифмическому, преобразование Бокса-Кокса можно применять только к положительным числам. Если часть членов ряда x_t отрицательна, прежде чем применить к ряду

преобразование Бокса-Кокса, ко всем членам ряда прибавляют постоянную c. Члены преобразованного ряда получают по формуле

$$y_t = \frac{(x_t + c)^\lambda - 1}{\lambda}$$

если выбранное $\lambda > 0$. Для $\lambda = 0$ преобразование Бокса-Кокса действует как уже упомянутое логарифмическое: $y_t = \log(x_t + c)$.

12.4. Методы исследования структуры стационарного временного ряда

12.4.1. Цели и методы анализа

Цели анализа. В предыдущих параграфах этой главы мы рассматривали методы выделения из временного ряда детерминированной компоненты — тренда, сезонной и циклической компонент. После удаления детерминированной компоненты временной ряд должен свестись к стационарному процессу. Так что следующим шагом после выделения детерминированной компоненты должен быть анализ остатков, то есть изучение ряда, полученного из исходного временного ряда после исключения детерминированной компоненты. При этом могут ставиться следующие цели.

1. Описание ряда с помощью той или иной модели, которая отражает зависимость между его соседними элементами. На базе построенной модели можно осуществлять прогноз будущего поведения ряда.
2. Уточнение оценки дисперсии временного ряда. Эта оценка важна для прогнозирования, так как исходя из нее вычисляется ширина доверительной трубки прогноза. Привычные оценки дисперсии, которые мы использовали в регресионном анализе (глава 8), — например, нормированная сумма квадратов отклонений элементов реализации от их среднего, — рассчитаны на независимые случайные величины. Для статистически зависимых данных такие оценки дисперсии временного ряда могут как сильно превышать истинное значение σ^2, так и быть значительно меньше.
3. Проверка стационарности остатков (при нестационарности подбор детерминированной компоненты нуждается в уточнении).

Методы анализа. В качестве модели стационарных временных рядов чаще всего используются процессы авторегрессии, скользящего среднего и их комбинации. Этим моделям посвящена глава 14.

А для проверки стационарности ряда остатков и оценки его дисперсии на практике чаще всего используются выборочная автокорреляционная (коррелограмма, см. п. 11.10) и частная автокорреляционная функция. В пп. 12.4.2 и 12.4.3 мы рассмотрим методы интерпретации графиков этих функций.

Замечания. 1. Для выяснения статистических зависимостей между элементами временного ряда может также быть использована периодограмма (см. п. 11.10).

2. Методы исследования структуры стационарного временного ряда по одной реализации наиболее успешно и полно разработаны для нормально распределенных процессов. Это объясняется тем, что у этих процессов из стационарности в широком смысле, которая поддается определенной проверке, следует стационарность в узком смысле, которая практически не поддается проверке (см. п. 11.3.2).

12.4.2. Интерпретация графика коррелограммы

Анализ коррелограммы — это порой довольно непростая задача. О причинах возникающих при этом трудностей уже говорилось в п. 11.10. Здесь мы кратко остановимся на типичном поведении коррелограммы для некоторых классов временных рядов.

Для начала рассмотрим поведение коррелограммы для некоторых нестационарных рядов. В этом случае следует помнить, что коррелограмма практически не несет никакой информации о статистической зависимости или независимости членов временного ряда, однако она может отражать причины нарушения стационарности. Именно с этой точки зрения мы и рассматриваем два следующих примера.

Наличие тренда. Для временного ряда, содержащего тренд, коррелограмма не стремится к нулю с ростом значения лага k. Ее характерное поведение изображено на рис. 12.14, где кореллограмма построена для ряда урожайности зерновых (рис. 11.1а).

Наличие сезонных колебаний. Для ряда с сезонными колебаниями коррелограмма также будет содержать периодические всплески, соответствующие периоду сезонных колебаний. Это позволяет устанавливать предполагаемый период сезонности. Однако, как было сказано в п. 11.10, отдельные редкие выхода графика коррелограммы за границы доверительной трубки могут наблюдаться и у белого шума. Типичное поведение коррелограммы для ряда с сезонными колебаниями приведено на рис. 12.15, где она построена для данных месячных продаж шампанского в логарифмической шкале (рис. 12.12) после удаления из них линейного тренда.

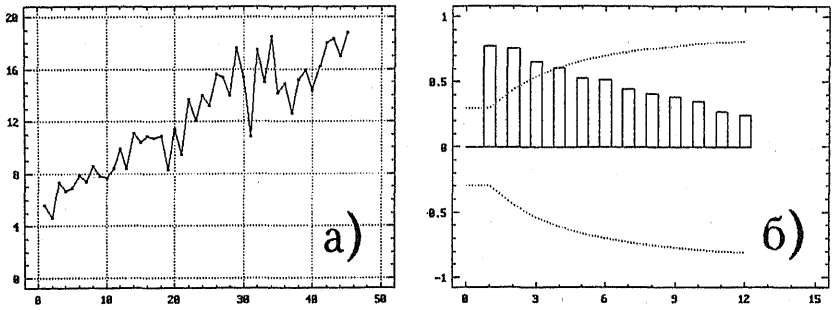


Рис. 12.14. Коррелограмма ряда урожайности зерновых:
а) исходный временной ряд; б) его коррелограмма

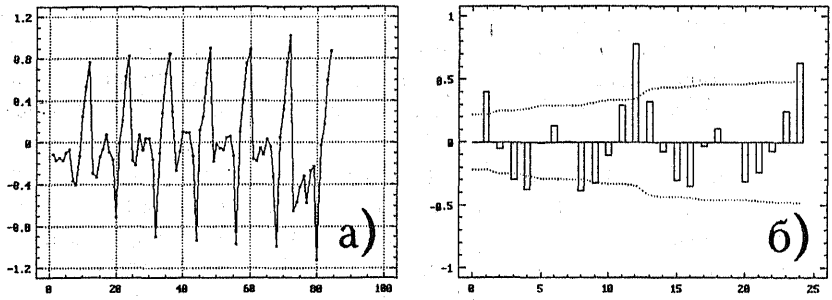


Рис. 12.15. Коррелограмма ряда месячных продаж шампанского
в логарифмической шкале (после удаления линейного тренда):
а) преобразованный ряд продаж шампанского; б) его коррелограмма

Перейдем к рассмотрению коррелограмм стационарных случайных процессов. В этом случае коррелограмма показывает коррелированность значений временного ряда при различных расстояниях между ними.

Коррелограмма белого шума. Как указывалось выше, автокорреляционная функция r_k белого шума равна нулю для всех $k \neq 0$. На рис. 12.16 изображена типичная коррелограмма белого шума. Как указывалось в п. 11.10, для гауссовского белого шума можно указать 95% доверительный интервал для каждого конкретного значения $\overline{r}_k$ в виде $-1/n \pm 2/\sqrt{n}$. Он изображен на графике коррелограммы пунктирными линиями. Если выборочные оценки корреляционной функции попадают в указанные доверительные интервалы, то можно предположить, что значения процесса являются белым шумом. Однако, как уже говорилось, довольно часто одно или несколько значений выборочной автокорреляционной функции белого шума могут выходить из указанных пределов. Особенно часто этот эффект можно наблюдать при наличии относительно небольшого числа наблюдений.

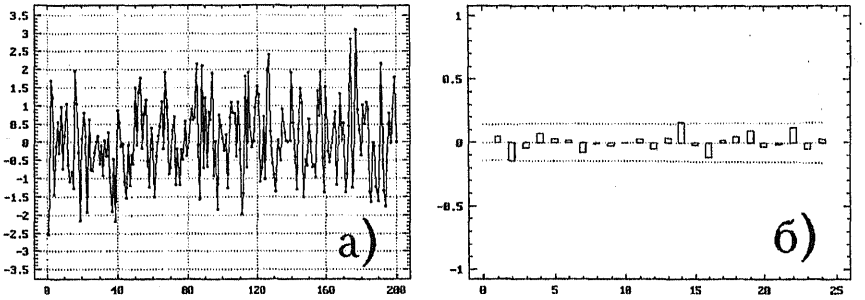


Рис. 12.16. Коррелограмма белого шума: а) исходный ряд; б) его коррелограмма

Коррелограмма процессов скользящего среднего. Траектории многих стационарных случайных процессов выглядят гораздо более гладко, чем траектории белого шума. Это связано с наличием положительной корреляции между двумя или несколькими соседними членами подобных рядов. Если же корреляция между соседними членами ряда отрицательна, то траектории подобных процессов будут более изломанными, чем траектории белого шума. Простейшим примером процессов, у которых зависимы одно или несколько соседних значений, являются процессы скользящего среднего. Определение процесса скользящего среднего первого порядка было дано в п. 11.7. Более подробно свойства этих процессов рассматриваются в п. 14.4. Здесь мы приведем вид типичных графиков этих процессов и их автокорреляционных функций.

Пусть $\varepsilon_1, \ldots, \varepsilon_n$ — гауссовский белый шум. Обозначим через $X(t)$ процесс скользящего среднего первого порядка (кратко MA(1)) с коэффициентом θ и средним равным нулю. Согласно (11.10):

$$X(t) = \varepsilon_t + \theta\varepsilon_{t-1}.$$

Нетрудно убедиться, что у этого процесса зависят между собой только соседние значения $X(t)$ и $X(t-1)$. При этом их корреляция r_1 равна:

$$r_1 = \frac{\theta}{1+\theta^2}$$

На рис. 12.17 приведены графики ста значений реализации процесса скользящего среднего с коэффициентом $\theta = 0.75$ и его коррелограммы. На рис. 12.18 приведены аналогичные графики при $\theta = -0.75$.

На графиках видно, что хотя полученные оценки значений r_k при $k = 2, 3\ldots$ не равны нулю, они значимо не отличаются от нулевых значений, так как попадают в 95% доверительный интервал, который построен в предположении равенства нулю соответствующих значений автокорреляционной функции.

Для процессов скользящего среднего второго порядка, как будет показано ниже в п. 14.4, отличаются от нуля только значения r_1 и r_2,

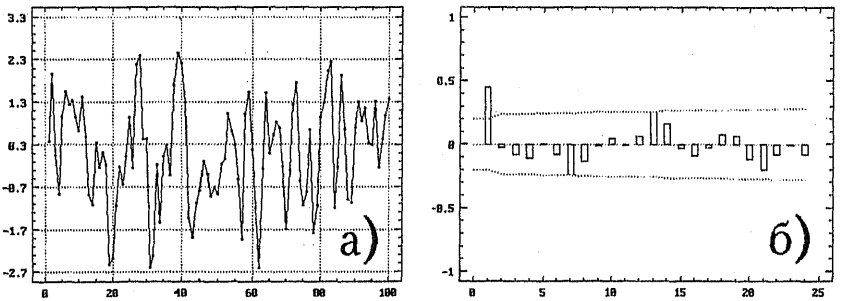


Рис. 12.17. Коррелограмма MA(1) процесса при $\theta = 0.75$: а) исходный ряд; б) его коррелограмма

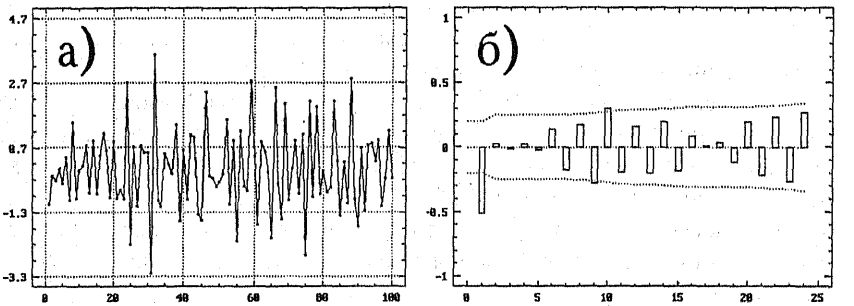


Рис. 12.18. Коррелограмма MA(1) процесса при $\theta = -0.75$: а) исходный ряд; б) его коррелограмма

а все последующие значения r_k при $k = 3, 4, \ldots$ равны нулю. Наконец, для процессов скользящего среднего порядка q отличны от нуля только первые q значений автокорреляционной функции. Строя графики коррелограмм для подобных процессов, мы можем на основании указанного свойства сделать предварительный вывод о возможном порядке процесса скользящего среднего, который может быть использован для описания наблюденного ряда.

Указанное правило хорошо, если подобранный порядок модели скользящего среднего невелик, скажем от одного до четырех-пяти. Однако на практике часто встречаются стационарные процессы с автокорреляционной функцией заметно отличной от нуля даже при больших задержках. Следуя сформулированному правилу, их можно пытаться описать процессами скользящего среднего высоких порядков. Это приводит к большому числу коэффициентов процесса скользящего среднего, которые подлежат дальнейшей оценке. При этом точность этих оценок заметно снижается. Практическая ценность таких многопараметрических моделей скользящего среднего невелика. В этой ситуации лучше попытаться описать временной ряд с помощью модели авторегрессии.

Если и эта попытка не увенчается успехом — перейти к комбинированным моделям авторегрессии-скользящего среднего. Проиллюстрируем типичное поведение автокорреляционной функции и ее оценки для процессов авторегрессии.

Коррелограмма процессов авторегрессии. Пусть, как и прежде, $\varepsilon_1, \ldots, \varepsilon_n$ — гауссовский белый шум. Напомним, что простейший процесс авторегрессии первого порядка $X(t)$ с нулевым средним задается соотношением:

$$X(t) = \phi X(t-1) + \varepsilon_t, \tag{12.16}$$

где ε_t не зависит от $X(t-1)$. Как указывалось в п. 11.7, члены даже этого простейшего процесса не становятся независимыми с ростом промежутка времени между ними. Однако при определенных условиях на коэффициенты эта зависимость быстро убывает.

В общем случае свойства этих процессов подробно разбираются нами ниже в пп. 14.1—14.3. Здесь же мы приведем два типичных графика поведения выборочных автокорреляционных функций этих процессов. Как будет показано ниже в п. 14.2 и 14.3 автокорреляционные функции этих процессов с ростом лага либо просто экспоненциально затухают либо представляют из себя экспоненциально затухающие синусоидальные волны.

На рис. 12.19 приведены графики ста значений реализации процесса авторегрессии второго порядка:

$$X(t) = \phi_1 X(t-1) + \phi_2 X(t-2) + \varepsilon_t. \tag{12.17}$$

при $\phi_1 = 0.7$ и $\phi_2 = 0.25$. Здесь автокорреляционная функция процесса и соответственно коррелограмма экспоненциально затухают с ростом лага.

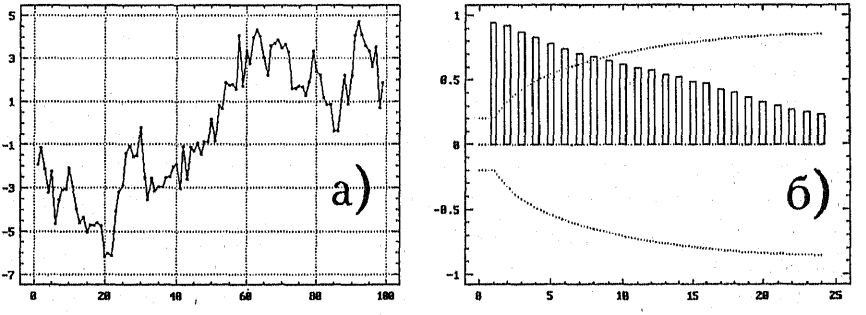


Рис. 12.19. Коррелограмма AR(2) процесса при $\phi_1 = 0.7$, $\phi_2 = 0.25$: а) исходный ряд; б) его коррелограмма

На рис. 12.20 приведены графики ста значений реализации AR(2) процесса (12.17) при $\phi_1 = 0.7$ и $\phi_2 = -0.25$. Автокорреляционная

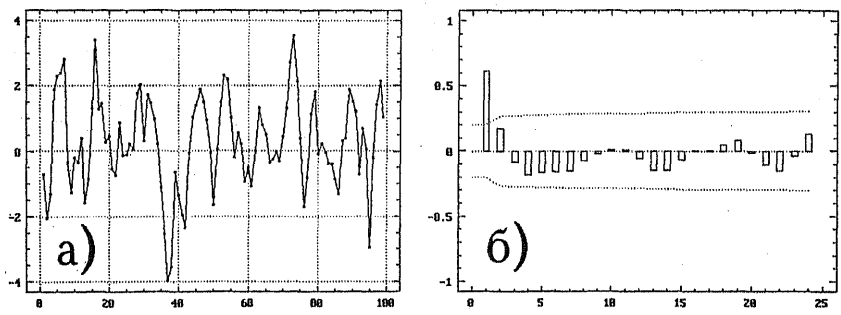


Рис. 12.20. Коррелограмма AR(2) процесса при $\phi_1 = 0.7$, $\phi_2 = -0.25$: а) исходный ряд; б) его коррелограмма

функция этого процесса и соответственно коррелограмма ведут себя с ростом лага как экспоненциально затухающая синусоида.

Замечание. Вид выборочных автокорреляционных функций процесса (12.16) для различных значений ϕ приведен на рис. 14.1 главы 14. На этом рисунке видно, что при $\phi = \pm 0.75$ значения оценок автокорреляционной функции $\overline{r}_k$ довольно сильно отличаются от нуля даже при $k = 15$.

12.4.3. Интерпретация графика частной автокорреляционной функции

Выборочная частная автокорреляционная функция. Для того, чтобы по полученной реализации процесса подобрать модель авторегрессии, необходимо предварительно указать возможный порядок этой модели. Приведенные примеры авторегрессионных процессов показывают, что непосредственно из вида выборочной автокорреляционной функции этот вывод сделать довольно трудно. Эту задачу значительно облегчает специально преобразованная автокорреляционная функция. Она называется *частной автокорреляционной функцией*. Расскажем, как следует интерпретировать поведение этой функции.

Замечание. Формальное определение частной автокорреляционной функции мы отложим до главы 14, так как это определение довольно сложно и основано на моделях авторегрессии. Однако при практическом анализе временных рядов это определение и не нужно — графики автокорреляционной функции и частной автокорреляционной функции строит статистическая программа, а человеку нужно лишь знать, как их интерпретировать.

Обозначения. Для краткости мы будем использовать сокращения: АКФ — автокорреляционная функция, ЧАКФ — частная автокорреляционная функция.

Обозначим через ϕ_{kk} значения ЧАКФ для каждого значения лага $k = 1, 2, \ldots$. Оценку этой функции по реализации временного ряда будем

обозначать через $\hat{\phi}_{kk}$ при $k = 0, 1, 2, \ldots$. Значения функций ϕ_{kk} и $\hat{\phi}_{kk}$ для каждого значения k по абсолютной величине меньше единицы.

Процессы авторегрессии первого порядка. Как показано ниже в п. 14.3, для процессов авторегрессии первого порядка отлично от нуля только значение ϕ_{11}, а все остальные значения этой функции равны нулю. Для выборочной частной автокорреляционной функции $\hat{\phi}_{kk}$ это означает, что все ее значения, начиная со второго, должны значимо не отличаться от нуля, то есть попадать в соответствующий доверительный интервал.

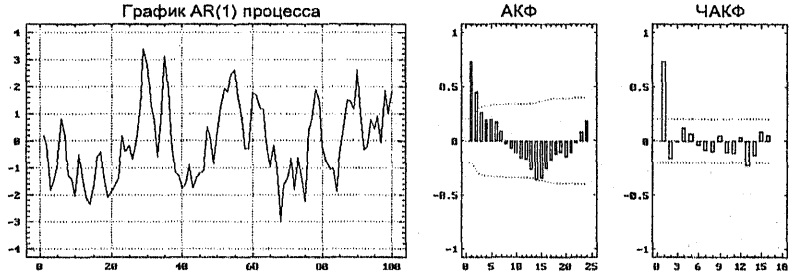


Рис. 12.21. График AR(1) процесса (значение ϕ равно 0.75, его выборочная автокорреляционная функция (АКФ) и выборочная частная автокорреляционная функция (ЧАКФ)

Это и видно на рис. 12.21. Здесь значения выборочной ЧАКФ (в отличие от выборочной АКФ), начиная со второго, малы и значимо неотличимы от нуля. Таким образом, по поведению выборочной ЧАКФ легче выяснить вид модели временного ряда.

Замечание. Стоит заметить, что при малых значениях коэффициента ϕ для того, чтобы отличить процесс от белого шума, требуется достаточно много наблюдений. Так, из рис. 12.22, на котором представлен AR(1) процесс с $\phi = -0.25$, видно, что для этого оказалось недостаточно ста наблюдений. Все значения выборочных автокорреляционной и частной автокорреляционной функций здесь попали в доверительные трубки для предположения, что процесс является белым шумом.

Процессы авторегрессии второго порядка. Для процесса авторегрессии второго порядка отличны от нуля только первые два значения ЧАКФ. На рис. 12.23а и 12.23б изображены выборочные ЧАКФ процессов, представленных на рис. 12.19 и 12.20.

Процессы скользящего среднего. Для процессов скользящего среднего (MA-процессов), в отличие от процессов авторегрессии, ЧАКФ при больших значениях лага k не обращается в ноль, а экспоненциально убывает. Мы не будем останавливаться на этом подробнее. Просто проиллюстрируем поведение выборочной ЧАКФ для MA(1) процессов,

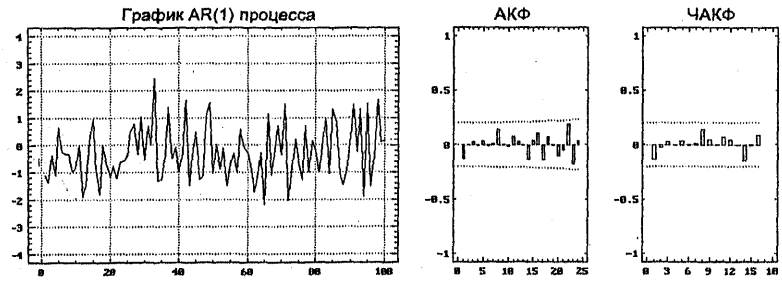


Рис. 12.22. График AR(1) процесса (значение ϕ равно -0.25), его выборочная АКФ и выборочная ЧАКФ

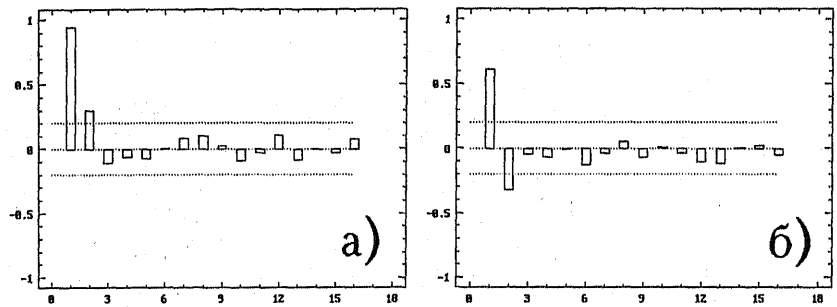


Рис. 12.23. Выборочная ЧАКФ AR(2) процессов:
а) $\phi_1 = 0.7$, $\phi_2 = 0.25$; б) $\phi_1 = 0.7$, $\phi_2 = -0.25$

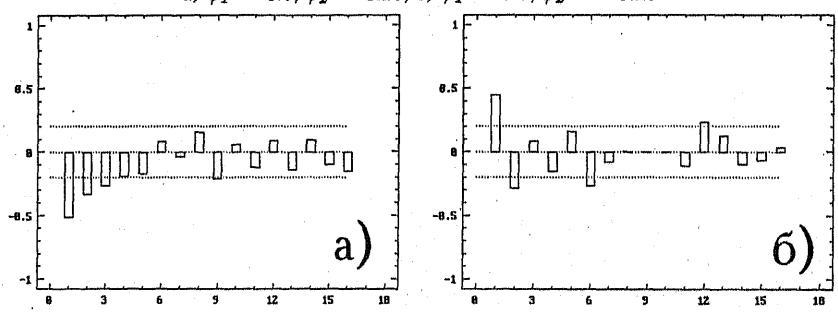


Рис. 12.24. Выборочная ЧАКФ MAR(1) процессов: а) $\theta = 0.75$; б) $\theta = -0.75$

приведенных на рис. 12.17 и 12.18. На рис. 12.24 приведены выборочные ЧАКФ MA(1) процессов при $\theta = 0.75$ и $\theta = -0.75$.

Глава 13

Анализ временных рядов на компьютере

13.1. О выборе пакетов для описания в этой книге

Методы анализа временных рядов широко представлены во многих универсальных статистических пакетах, включая разобранные в предыдущих главах STADIA и STATGRAPHICS. Но анализ временных рядов — это очень специфическая область статистики, отличающаяся по кругу задач и методов их решения, а также по кругу пользователей, применяющих эти методы. Поэтому для анализа временных рядов имеются также и специализированные статистические пакеты. В этой главе мы рассмотрим способы решения рассмотренных выше задач в специализированном статистическом пакете ЭВРИСТА и универсальном статистическом пакете SPSS. Выбор данных пакетов обусловлен следующими причинами.

Пакет ЭВРИСТА является одним из лучших специализированных отечественных пакетов для анализа временных рядов. Его функциональные возможности значительно шире стандартных процедур анализа временных рядов универсальных статистических пакетов. Пакет постоянно совершенствуется и пополняется, он хорошо зарекомендовал себя во многих организациях, в том числе активно работающих на финансовом рынке. Более подробная информация об этом пакете дана в приложениях 1 и 3, а также в [9].

Универсальный пакет SPSS занимает одно из первых мест в мире среди программ статистической обработки данных (см. приложения 1 и 3). Отечественным специалистам ранние версии SPSS в основном были известны как мощный инструмент обработки социологических и психологических данных. В связи с этим мы решили показать этот пакет с менее известной его стороны. Другими аргументами выбора этого пакета было желание показать более подробно Windows-интерфейс современного статистического пакета, а также познакомить пользователя с англоязычной терминологией в области анализа временных рядов. И, наконец, нам хотелось дать пользователям представление о более широком круге статистических пакетов.

13.2. Анализ временных рядов в SPSS

13.2.1. Обзор возможностей

Возможности в области анализа временных рядов. Пользователи, знакомые с ранними и неполными версиями пакета SPSS, чаще всего имеют совершенно неадекватное представление о возможностях этого пакета в области анализа временных рядов. Во-первых, ранние версии SPSS использовались, в основном, специалистами в области психологии, биологии и социологии, где задачи анализа временных рядов менее характерны. Во-вторых, современные версии пакета имеют модульную структуру (см. приложения 1 и 3), в которой анализ временных рядов выделен в отдельный модуль SPSS Trends. При отсутствии этого модуля возможности в области анализа временных рядов будут ограничены только процедурами регрессионного анализа базового модуля пакета SPSS Base.

Кратко перечислим основные процедуры пакета SPSS в области анализа временных рядов:

1. Regression (Регрессионный анализ) — позволяет выделять и удалять широкий набор моделей трендов;

2. ARIMA (Модели авторегрессии–проинтегрированного скользящего среднего) — вычисляет оценки параметров для сезонных и несезонных моделей, а также строит доверительные интервалы для прогноза. Процедура допускает пропущенные значения во временных рядах и выполняет анализ интервенций;

3. EXSMOOTH (Экспоненциальное сглаживание) — включает широкий круг методов экспоненциального сглаживания для сезонных и несезонных рядов с трендом;

4. SEASON (Сезонные составляющие) — оценивает мультипликативные или аддитивные сезонные составляющие для сезонных временных рядов;

5. SPECTRA (Спектральный анализ) — производит разложение временного ряда на гармонические составляющие. Вычисляет и выводит на график одномерную и двумерную периодограмму и оценку спектральной плотности. Позволяет использовать различные спектральные окна;

6. AREG (Авторегрессионный анализ) — оценивает регрессионную модель, когда ошибки близких по времени значений ряда коррелируют между собой;

7. X11 ARIMA — оценивает сезонные факторы для процессов типа авторегрессии–проинтегрированного скользящего среднего.

Пакет также выполняет широкий круг других процедур, например, генерацию временных рядов, вычисление оценок автокорреляционной и частной автокорреляционной функции, построение различных типов графиков временных рядов и т.д.

Командный макроязык и система меню. Прежде чем начать разбор примеров в пакете SPSS, сделаем одно важное замечание. Пакет SPSS обладает развитым командным макроязыком, позволяющим создавать командные файлы, полностью описывающие все этапы анализа. Только в последних Windows-версиях пакета появилась возможность проводить почти все процедуры ввода, редактирования и анализа данных в режиме меню-ориентированного интерфейса с диалоговыми окнами. Мы ограничим свой рассказ, ориентированный на начинающих пользователей пакета, только работой с этим интерфейсом. Заодно будет проиллюстрирован довольно типичный Windows-интерфейс современных статистических пакетов. Однако, работая в SPSS, следует помнить, что при решении задачи с использованием меню-ориентированного интерфейса одновременно происходит создание командного файла решаемой задачи. Одно из удобств и достоинств командного языка заключается в том, что при решении однотипных задач нет необходимости каждый раз заполнять поля ввода и настраивать режимы работы процедур. Можно просто запускать однажды сформированный командный файл. При этом можно практически ничего не знать о самом командном языке SPSS.

13.2.2. Подбор тренда и прогнозирование

Рассмотрим эти задачи на следующем примере.

Пример 13.1к. Для данных урожайности зерновых культур в СССР подобрать модель тренда с помощью процедур регрессионного анализа и построить на базе подобранной модели прогноз на несколько лет вперед.

Подготовка данных. Пусть данные таблицы 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

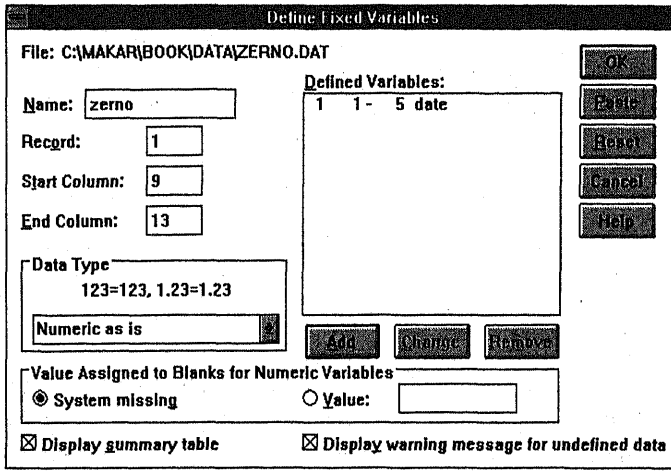


Рис. 13.1. Окно определения переменных фиксированного формата процедуры загрузки текстовых файлов в SPSS

Define Fixed Variables (Определить переменные фиксированного формата), см. рис. 13.1.

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. 13.2).

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. 13.2. Таблица SPSS с данными об урожайности зерновых

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 International 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 временного ряда 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 временного ряда в пункте меню Graphs (Диаграммы) можно выбрать один из двух возможных типов процедур: Simple Scatterplot (Простой график рассеивания) или Sequence (График последовательности). В этой задаче будет рассмотрена работа процедуры Simple Scatterplot. Ее диалоговое окно приведено на рис. 13.3.

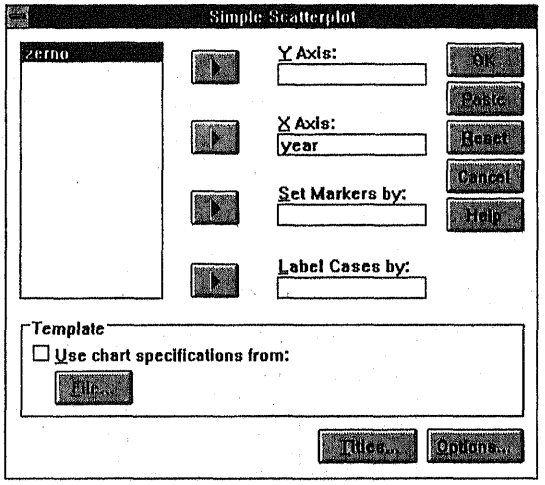


Рис. 13.3. Диалоговое окно процедуры Simple Scatterplot в SPSS

Выделяя щелчком мыши требуемые переменные и нажимая соответствующие кнопки запроса ▶, присвоим значения переменных date и zerno осям X и Y соответственно. Подобный простой способ выбора переменных используется практически во всех процедурах SPSS и в ряде других статистических пакетов под Windows (STADIA, STATGRAPHICS). На рис. 13.4 изображен полученный результат (после небольшого дополнительного оформления).

Рис. 13.4. График урожайности зерновых в SPSS

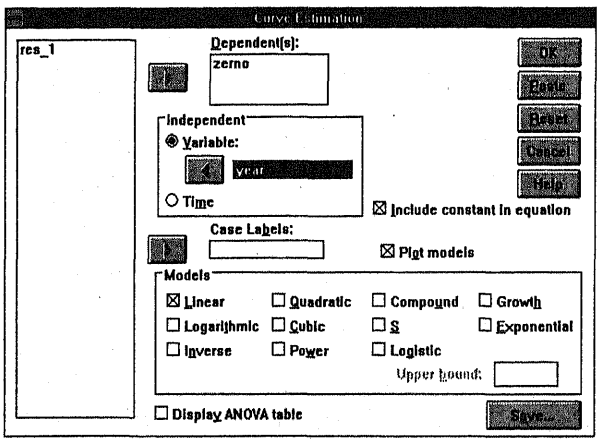


Рис. 13.5. Диалоговое окно процедуры Curve Estimation в SPSS

Выбор процедуры. На графике рис. 13.4 видно, что анализируемые данные содержат линейный тренд. Для его идентификации следует выбрать в меню пакета пункт **Statistics** (Статистика), 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 кривых трендов. Поэтому мы продолжим свой анализ с помощью процедуры **Curve Estimation**, диалоговое окно которой приведено на рис. 13.5.

Задание параметров процедуры. В окне рис. 13.5 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 примоугольнике 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 тренда. Дадим формулы моделей тренда, приведенных в прямоугольнике Models на рис. 13.5. Пусть x — независимая переменная или время, b_i и u — константы (параметры моделей). Тогда формулы моделей можно записать так:

Linear (линейная): $y = b_0 + b_1 x$;
Logarithmic (логарифмическая): $y = b_0 + b_1 \ln(x)$;
Inverse (обратная): $y = b_0 + (b_1/x)$;
Quadratic (квадратичная): $y = b_0 + b_1 x + b_2 x^2$;
Cubic (кубическая): $y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$;
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): $y = e^{(b_0 + b_1 x)}$ или $\ln(y) = b_0 + b_1 x$;
Exponential (экспоненциальная): $y = b_0 \cdot (e^{b_1 x})$ или $\ln(y) = \ln(b_0) + b_1 \cdot x$.

Замечание. Кнопка [Save] диалогового окна (рис. 13.5) 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. 13.6). Одновременно график ряда с подобранной кривой тренда будет помещен в окно Chart Carousel (рис. 13.7)

Результаты расчетов (см. рис. 13.6) показывают, что линейная модель тренда объясняет примерно 83% общей вариации данных, а полученные оценки коэффициентов модели значимо отличаются от нуля.

```
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. 13.6. Результаты работы процедуры Curve Estimation в окне выдачи результатов в SPSS

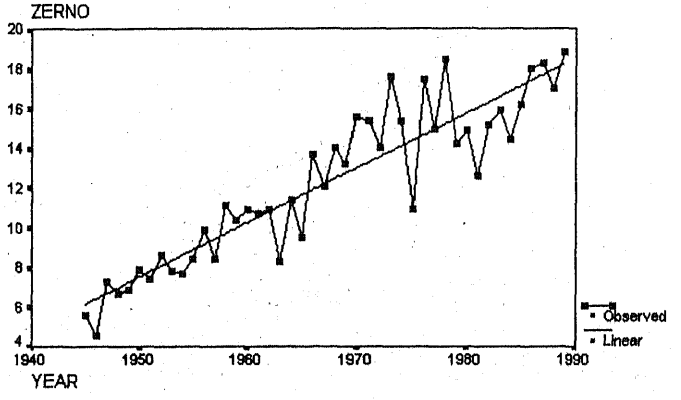


Рис. 13.7. Окно Chart Carousel с результатами работы процедуры Curve Estimation в SPSS

В частности, значение коэффициента *B* при переменной year (то есть средний прирост урожайности за год), равен примерно 0.275 (ц/га).

Анализ остатков. Дальнейший анализ модели связан с исследованием остатков. Выясним два вопроса: можно ли считать остатки некоррелированными и насколько их распределение согласуется с нормальным.

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 коррелированности остатков. Выше упоминалось, что одним из результатов работы процедуры Curve Estimation является создание новой переменной err_1, в которой хранятся остатки подобранной модели. Для выяснения коррелированности остатков вычислим оценки их автокорреляционной функции. Это можно сделать, например, вызвав процедуру Autocorrelations из пункта Time Series меню Graphs. Диалоговое окно процедуры приведено на рис. 13.8.

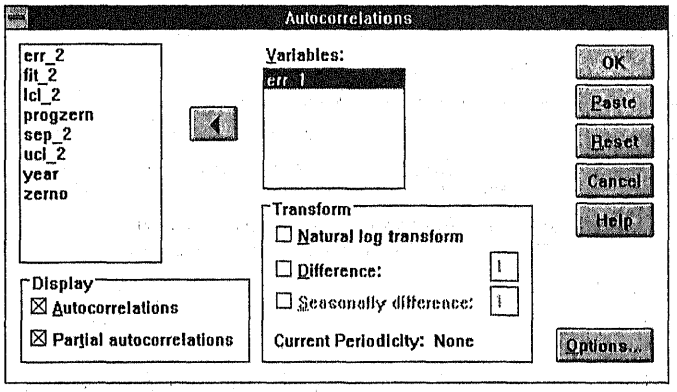


Рис. 13.8. Диалоговое окно процедуры Autocorrelations в SPSS

В окно Variables (переменные) поместим переменную err_1 со значениями остатков. С помощью кнопки Options зададим максимальное число лагов Maximum Numbers of Lags равным 10, учитывая небольшую длину изучаемого ряда. Затем нажмем кнопку OK. Программа выведет график автокорреляционной функции (коррелограмму), см. рис. 13.9. Из графика видно, что у нас нет основания считать остатки коррелированными, так как полученные оценки лежат внутри доверительного интервала для нулевых значений автокорреляционной функции.

Результаты расчетов процедуры помещаются в окно Output (см. рис. 13.10). Здесь приводятся значения оценок автокорреляционной функции, их стандартные ошибки и доверительные интервалы.

Замечание. Процедура Autocorrelations позволяет вычислять автокорреляционную и частную автокорреляционную функцию не только для исходного ряда, но и для прологарифмированного ряда или ряда простых и сезонных разностей. Все указанные настройки задаются в окне Transform (преобразования).

Еще одна проверка коррелированности остатков. Другой способ проверки некоррелированности остатков — статистика Дурбина-Уотсона. Она предназначена для проверки нулевой гипотезы H_0 о том, что все автокорреляции $r_k = 0$, против альтернативы: $r_k = r^k$ при $r \neq 0$ и $|r| < 1$. Эта альтернатива соответствует предположению, что ошибки

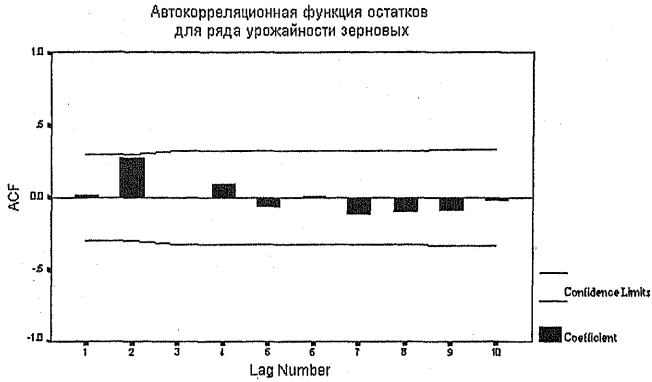


Рис. 13.9. График автокорреляционной функции остатков для ряда урожайности зерновых в SPSS

Lag	Corr.	Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	.015	.149	. * .	.011	.916
2	.272	.149	. *****.	3.641	.162
3	.004	.160	. * .	3.642	.303
4	.096	.160	. **.	4.120	.390
5	-.059	.161	. *.	4.306	.506
6	.011	.161	. * .	4.312	.634
7	-.107	.162	. **.	4.951	.666
8	-.097	.163	. **.	5.494	.704
9	-.087	.164	. **.	5.940	.746
10	-.017	.165	. * .	5.958	.819

Plot Symbols: Autocorrelations * Two Standard Error Limits .
Standard errors are based on the Bartlett (MA) approximation.

Total cases: 52 Computable first lags: 44

Рис. 13.10. Результаты расчетов процедуры Autocorrelations в окне вывода результатов в SPSS

образуют процесс авторегрессии первого порядка с коэффициентом r. В SPSS эту статистику вычисляет процедура Linear Regression меню Statistics. Для остатков ряда урожайности зерновых эта статистика равна 1.96419. Таблицы процентных точек этой статистики приведены в документации модуля SPSS Trends (см. также [31]). В данном случае нулевая гипотеза должна быть отвергнута для двусторонних альтернатив, если значение статистики Дарбина-Уотсона попадает в одну из областей $(-\infty, 1.48)$ и $(2.52, \infty)$. Полученное значение статистики в эти области не попадает. Таким образом по результатам двух тестов у нас нет оснований считать остатки коррелированными.

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 (гра-

фик на нормальной вероятностной бумаге). В диалоговом окне этой процедуры (рис. 13.11) необходимо указать в поле Variables обрабатываемую переменную.

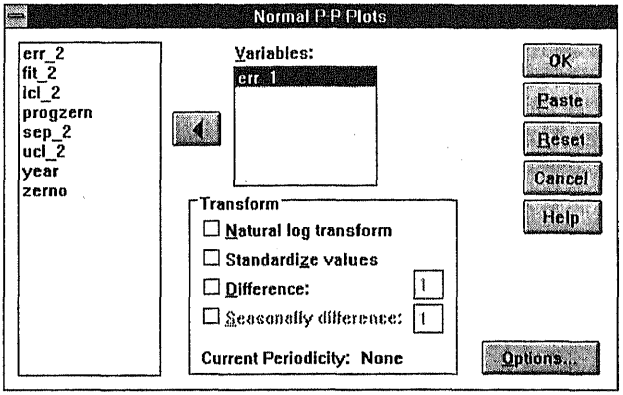


Рис. 13.11. Диалоговое окно процедуры Normal P-P в SPSS

Результаты работы процедуры приведены на рис. 13.12. 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 полученных результатов. Поэтому результаты, полученные с помощью метода наименьших квадратов, можно считать приемлемыми.

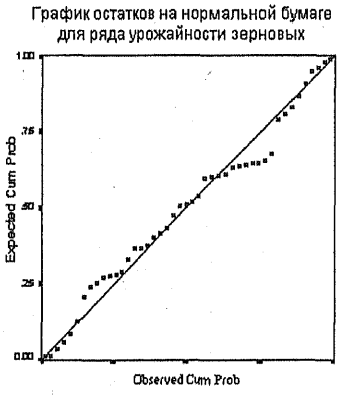


Рис. 13.12. Результаты процедуры Normal P-P в SPSS

Замечания. 1. Процедура Normal P-P (см. рис. 13.11) 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.

13.2.3. Устранение сезонной компоненты

Рассмотрим эту задачу на следующем примере.

Пример 13.2к. Для данных месячных продаж шампанского построить оценки сезонной компоненты и провести сезонную коррекцию ряда.

Подготовка данных. Пусть данные о месячных продажах шампанского находятся в файле bubbly.sav в формате пакета SPSS в виде переменной bubbly. Для загрузки данных в электронную таблицу SPSS с помощью меню в пункте File выберите команду Open и в выведенном подменю укажите тип открываемого файла Data. В запросе открытия файла выберите файл bubbly.sav. Содержимое этого файла появится в электронной таблице пакета.

Для анализа сезонных временных рядов SPSS требует, кроме ввода самого ряда, дополнительного задания специальной переменной, которая указывает структуру периодичности ряда или его календарную структуру. Без этого анализ сезонных эффектов в SPSS будет недоступен. Для задания указанной переменной в меню Data выберите процедуру Define Dates (Определение календарных дат). Ее диалоговое окно приведено на рис. 13.13.

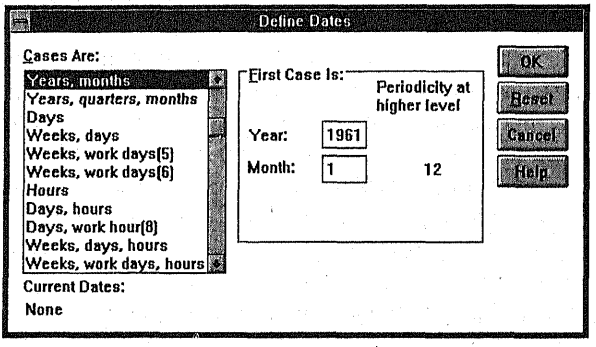


Рис. 13.13. *Диалоговое окно задания типа календарных дат в SPSS*

В поле Cases Are: (наблюдения являются) приведен обширный список различных типов календарных переменных. Из них нам следует выбрать

Years, months (годы, месяцы), так как изучаемые данные являются результатом месячных наблюдений за ряд лет. В окне Firth Case Is: (первое наблюдение является) для выбранного типа переменных указывается первый год и месяц наблюдения, как это показано на рис. 13.13. В результате этой операции SPSS создаст три новых переменных year_, month_ и date_ и поместит их в электронную таблицу (см. рис. 13.14).

1:bubbly 2.815

	bubbly	year_	month_	date_	var
1	2.815	1961	1	JAN 1961	
2	2.672	1961	2	FEB 1961	
3	2.755	1961	3	MAR 196	
4	2.721	1961	4	APR 1961	
5	2.946	1961	5	MAY 196	
6	3.036	1961	6	JUN 1961	
7	2.282	1961	7	JUL 1961	
8	2.212	1961	8	AUG 1961	
9	2.922	1961	9	SEP 1961	
10	4.301	1961	10	OCT 1961	
11	5.764	1961	11	NOV 1961	
12	7.312	1961	12	DEC 1961	
13	2.541	1962	1	JAN 1962	
14	2.475	1962	2	FEB 1962	

Рис. 13.14. Электронная таблица SPSS с вновь созданными календарными переменными

Выбор процедуры. Для оценки сезонных эффектов выберите в меню программы пункт Statistics, в появившемся подменю — пункт Time Series (временные ряды), и затем еще в одном появившемся подменю — пункт Seasonal Decomposition. Диалоговое окно выбранной нами процедуры приведено на рис. 13.15. Оно требует указания анализируемой переменной bubbly в окне Variable(s), типа модели временного ряда: Multiplicative (мультипликативная) или Additive (аддитивная), а также указания типа взвешивания точек в скользящем среднем Moving Average Weight. Учитывая, что сезонная вариация данных растет с ростом времени, выбираем мультипликативную модель временного ряда. Так как величина периода — 12 месяцев, — четная, указываем в типе взвешивания: Endpoints weighted by .5 (веса крайних точек равны 0.5). Заполнив поля ввода, как показано на рис. 13.15, нажмем [OK] для запуска процедуры.

Результаты. Процедура сезонной декомпозиции создает четыре новые переменные:

- saf_1 — оценки сезонных эффектов;

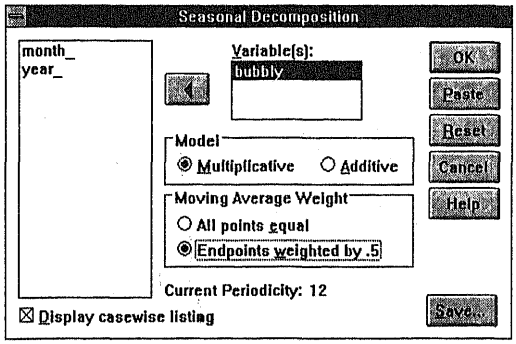


Рис. 13.15. Диалоговое окно процедуры сезонной декомпозиции в SPSS

1:bubbly 2.815

	bubbly	year_	month_	date_	err_1	sas_1	sac_1	stc_1
1	2.815	1961	1	JAN 1961	1.01944	3.80088	.74062	3.72839
2	2.672	1961	2	FEB 1961	1.03829	3.77009	.70874	3.63107
3	2.755	1961	3	MAR 196	.96677	3.32225	.82926	3.43644
4	2.721	1961	4	APR 1961	.97110	3.22230	.84443	3.31819
5	2.946	1961	5	MAY 196	.97586	3.17550	.92773	3.25406
6	3.036	1961	6	JUN 1961	.96168	3.43123	.88481	3.56796
7	2.282	1961	7	JUL 1961	.81518	3.13072	.72891	3.84054
8	2.212	1961	8	AUG 1961	1.43449	5.98323	.36970	4.17097
9	2.922	1961	9	SEP 1961	.81350	3.16832	.92226	3.89470
10	4.301	1961	10	OCT 1961	.96689	3.55972	1.20824	3.68161
11	5.764	1961	11	NOV 1961	.97774	3.33071	1.73056	3.40654
12	7.312	1961	12	DEC 1961	1.00868	3.47404	2.10476	3.44413
13	2.541	1962	1	JAN 1962	.98934	3.43091	.74062	3.46787
14	2.475	1962	2	FEB 1962	.98246	3.49213	.70874	3.55446

Рис. 13.16. Электронная таблица SPSS с результатами работы процедуры сезонной декомпозиции

- sas_1 — скорректированный с учетом сезонных эффектов временной ряд;
- stc_1 — тренд и циклическая компонента скорректированного временного ряда;
- err_1 — иррегулярная компонента скорректированного временного ряда (иными словами — случайная ошибка).

Эти переменные загружаются в электронную таблицу (рис. 13.16). Кроме того, значения этих переменных, а также вычисленные скользящие средние и промежуточные результаты расчетов сезонных индексов, выводятся в окно Output (рис. 13.17).

Results of SEASON procedure for variable BUBBLY.
Multiplicative Model. Centered MA method. Period = 12.

DATE_	BUBBLY	Moving averages	Ratios (* 100)	Seasonal factors (* 100)	Seasonally adjusted series	Smoothed trend-cycle	Irregular component
JAN 1961	2.815	.	.	74.062	3.801	3.728	1.019
FEB 1961	2.672	.	.	70.874	3.770	3.631	1.038
MAR 1961	2.755	.	.	82.926	3.322	3.436	.967
APR 1961	2.721	.	.	84.443	3.222	3.318	.971
MAY 1961	2.946	.	.	92.773	3.176	3.254	.976
JUN 1961	3.036	.	.	88.481	3.431	3.568	.962
JUL 1961	2.282	3.467	65.825	72.891	3.131	3.841	.815
AUG 1961	2.212	3.447	64.169	36.970	5.983	4.171	1.434
SEP 1961	2.922	3.450	84.685	92.226	3.168	3.895	.813
OCT 1961	4.301	3.485	123.428	120.824	3.560	3.682	.967
NOV 1961	5.764	3.542	162.737	173.056	3.331	3.407	.978
DEC 1961	7.312	3.585	203.985	210.476	3.474	3.444	1.009
JAN 1962	2.541	3.624	70.121	74.062	3.431	3.468	.989
FEB 1962	2.475	3.636	68.070	70.874	3.492	3.554	.982
MAR 1962	3.031	3.645	83.152	82.926	3.655	3.687	.991
APR 1962	3.266	3.680	88.741	84.443	3.868	3.800	1.018
MAY 1962	3.776	3.732	101.170	92.773	4.070	3.895	1.045
JUN 1962	3.230	3.821	84.541	88.481	3.650	4.003	.912

Рис. 13.17. Результаты расчетов процедуры сезонной декомпозиции

13.3. Анализ временных рядов в пакете ЭВРИСТА

13.3.1. Общие сведения о пакете

Основные возможности. Кратко перечислим основные функциональные возможности Windows-версии пакета Эвриста в области анализа временных рядов и в смежных областях. Меню статистических методов пакета включает следующие процедуры:

1. Предварительный анализ — вычисляет основные описательные статистики ряда, осуществляет различные преобразования ряда (удаление среднего, преобразование Бокса-Кокса, сезонные и несезонные разности), проверяет гипотезы о случайности и нормальности выборки (критерии повторных точек, хи-квадрат, Колмогорова-Смирнова) и др.;

2. Анализ тренда — оценивает тренд методом простой линейной или нелинейной регрессии, методом скользящего среднего, а также удаляет тренд из временного ряда и предоставляет другие полезные сервисные возможности;

3. Прогнозирование — строит прогнозы с помощью сглаживания методом Брауна и сезонного сглаживания Хольта-Уинтерса, а также строит прогнозы на базе подобранных моделей простой и полиномиальной регрессии и моделей типа авторегрессии-скользящего среднего;

4. **Спектральный анализ** — оценивает спектральную плотность и автокорреляционную функцию с помощью различных периодограмм;

5. **АРСС модели** — идентифицирует сезонную и несезонную модель авторегрессии-скользящего среднего (АРСС), используя выборочные автокорреляционную и частную автокорреляционную функцию. Автоматически подбирает порядок этих моделей с помощью различных критериев, подгоняет сезонную модель АРСС и показывает соответствующую ей спектральную плотность, автокорреляционную функцию или таблицу параметров и др.;

6. **Кепстральный анализ** — строит оценку кепстра для исследования ряда на наличие в нем эхо-эффекта, а также удаляет эхо в кепстральной или временной области;

7. **Кросс-спектр** — оценивает параметрическую и непараметрическую кросс-спектральную плотность, кросс-корреляционную функцию и когерентность, а также модели передаточных функций;

8. **Регрессионные модели** — оценивает линейные регрессионные модели в различных режимах, включая шаговую регрессию;

9. **Факторный анализ** — оценивает модель факторного анализа методом главных компонент;

10. **Анализ интервенций** — оценивает параметры динамической модели интервенции и удаляет интервенции из ряда. Порядок и вид интервенции могут задаваться вручную или вычисляться автоматически;

11. **Гармонический анализ** — оценивает параметры гармонической модели и их статистические характеристики;

12. **Моделирование данных** — осуществляет моделирование сезонных и несезонных временных рядов типа авторегрессии-скользящего среднего, а также функции интервенции.

Этот далеко не полный список возможностей пакета наглядно иллюстрирует многообразие методов и инструментов анализа временных рядов. К сожалению, в этой книге мы не можем рассмотреть большую часть из этих возможностей. Мы лишь опишем, как обсуждавшиеся выше задачи можно решить с помощью пакета Эвриста.

Интерфейс пакета. Сделаем предварительно несколько общих замечаний об интерфейсе Windows-версии этого пакета. Общий вид интерфейса пакета приведен на рис. 13.18. Он включает в себя строку меню, строку пиктограмм для выполнения наиболее часто используемых команд и одно или несколько рабочих окон, в которые помещаются данные и результаты. Интерфейс пакета полностью соответствует стандарту интерфейсов прикладных программ в среде Windows, так что

работа с программой по форме практически не отличается от работы с другими Windows-программами.

На рис. 13.18 справа в рабочем окне приведена электронная таблица с загруженными данными примера 13.1к, а слева — график этих данных.

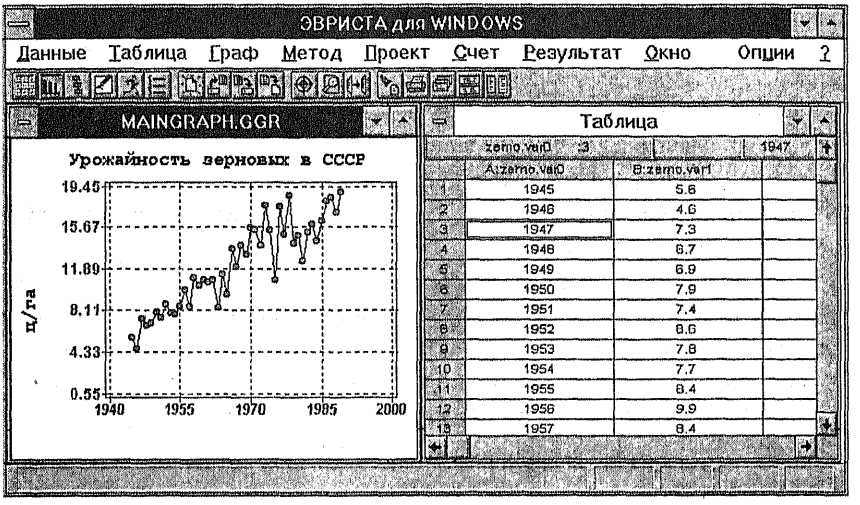


Рис. 13.18. Общий вид интерфейса Windows-версии пакета Эвриста

О дальнейшем изложении. Задачи, которые будут рассмотрены ниже, в пакете Эвриста можно решить различными способами. Так, вычисление линейной и полиномиальной регрессии в пакете реализуют процедуры Анализ тренда, Прогнозирование, Регрессионные модели. У каждой из них есть свои особенности. Рассказывая о том или ином способе решения задачи, будем стремиться к тому, чтобы показать различные возможности пакета, а не только к выбору кратчайшего пути к цели.

13.3.2. Подбор тренда и прогнозирование

Пример 13.1к. Для данных урожайности зерновых культур в СССР подобрать модель тренда с помощью процедур регрессионного анализа, провести анализ остатков и построить на базе подобранной модели прогноз на несколько лет вперед.

Ввод исходных данных. Пусть данные таблицы 1.2 находятся в текстовом (ASCII) файле zerno.dat в виде двух столбцов, первый из которых содержит значение года, а второй — значение урожайности. Для загрузки этих данных в пакет Эвриста выберем в пункте меню Данные (рис. 13.18) подпункт Импорт. На экране появится диалоговое окно импорта данных (рис. 13.19).

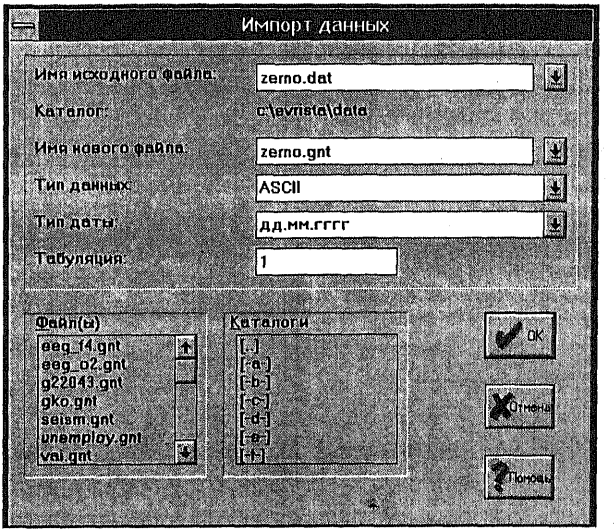


Рис. 13.19. Диалоговое окно импорта данных

В нем необходимо указать имя импортируемого файла и каталог, в котором он находится. Файлы данных пакета Эвриста должны иметь расширение gnt, поэтому в качестве имени нового файла указан файл zerno.gnt. Пакет предусматривает импорт из формата ASCII, dBase-III, Эвриста (DOS). Тип формата импортируемого файла указывается в поле **Тип данных**. Поле **Тип даты** позволяет загружать в пакет данные, записанные в различных календарных форматах (например, пятнадцатое февраля 1997 г. может быть записано в виде 15.02.1997 или 15/02/97). В случае наших данных настройка этого поля несущественна. После заполнения необходимых полей щелкните мышью по кнопке [OK] для выполнения импорта. Результатом работы этой процедуры будет появление файла zerno.gnt, содержащего две переменные: var0 и var1, которые соответствуют первому и второму столбцу исходного файла.

Замечание. Если исходные данные уже находятся в какой-нибудь Windows-программе, например в Excel, Quattro Pro, Lotus и т.п., то загрузка их в Эвристу может быть осуществлена стандартным образом через буфер обмена (Clipboard).

Задание рабочих переменных. Для редактирования и обработки данных в пакете Эвриста надо предварительно указать рабочие переменные. Для этого в пункте меню **Данные** выберем подпункт **Открыть**. На экране появится диалоговое окно этой процедуры (рис. 13.20). В нем следует указать имя файла zerno.gnt. Все переменные указанного файла появятся в окне **Переменные**. Из них с помощью мыши выбираются **Рабочие переменные**. После этого надо нажать в окне 13.20 кнопку [OK].

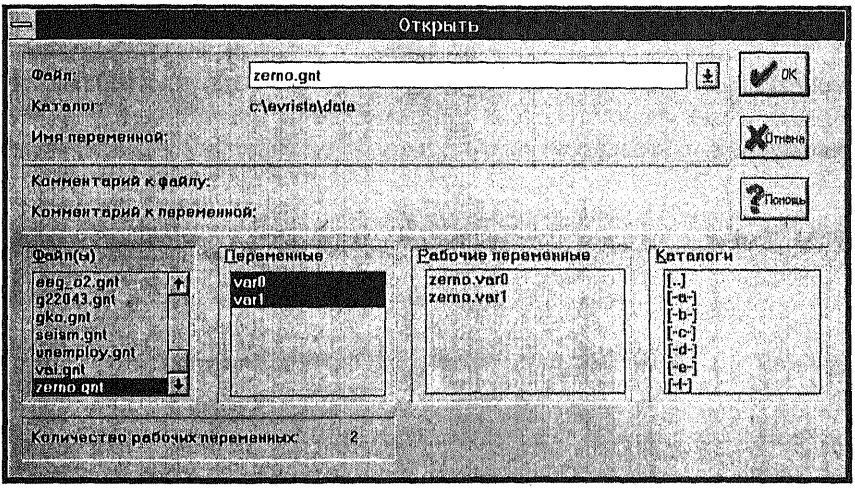


Рис. 13.20. Диалоговое окно открытия данных

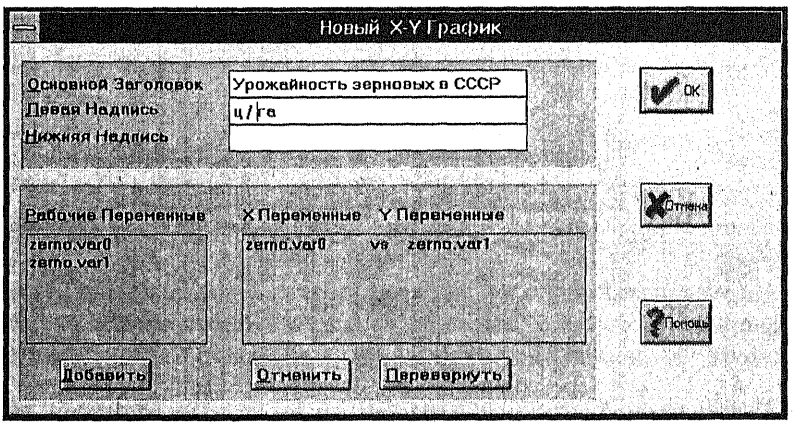


Рис. 13.21. Диалоговое окно создания нового X–Y графика

Вывод на экран данных и графика. Задав рабочие переменные, их можно посмотреть, отредактировать в электронной таблице и вывести на график. Для этого в пункте меню **Таблица** необходимо выбрать пункт **Открыть**. На экране появится электронная таблица с выбранными рабочими переменными (см. рис. 13.18). Одновременно можно открыть и графическое окно с рабочими данными. Для этого в пункте меню **Граф** следует выбрать пункт **Создать график**, в нем подпункт **Создать новый** и выбрать тип графика. (В данном случае это график зависимости урожайности от времени или, кратко, X–Y график.) Диалоговое окно этой процедуры приведено на рис. 13.21.

Заполнение диалогового окна включает задание переменных по осям X и Y и элементов оформления графика. Результаты этой процедуры приведены слева на рис. 13.18. Пакет также включает многочисленные средства для редактирования графика: выбор типа линий и точек, шрифтов и их размеров и т.п.

Выбор процедуры. Для оценки и удаления тренда из временного ряда в пакете следует в пункте меню **Метод** выбрать процедуру **Анализ тренда**, а в выведенном подменю — пункт **Полиномиальная Регрессия**. На экран выводится диалоговое окно, показанное на рис. 13.22. Оно предполагает задание степени подбираемого полинома, уровня доверия для доверительных интервалов, а также задание форм выдачи результатов.

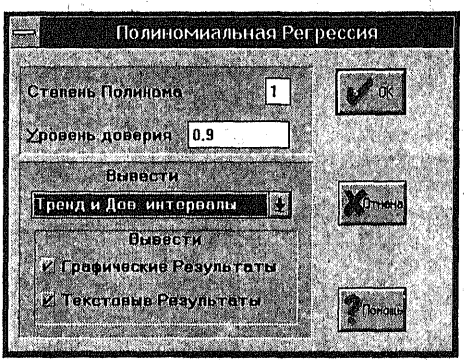


Рис. 13.22. Диалоговое окно процедуры полиномиальной регрессии

После заполнения полей диалогового окна для выполнения расчетов следует в пункте меню **Счет** выбрать опцию **Выполнить метод** или просто щелкнуть мышкой по пиктограмме на панели управления пакета.

Результаты. Текстовая форма выдачи результатов процедуры полиномиальной регрессии содержит оценки параметров регрессионной модели, их статистические характеристики (включая уровни значимости) и таблицу дисперсионного анализа (таблицу разложения вариации). Эти результаты приведены на рис. 13.23.

Выбранная графическая форма представления результатов процедуры (тренд и доверительные интервалы) приведена на рис. 13.24.

Анализ остатков. Для получения остатков временного ряда необходимо после подбора модели тренда еще раз запустить на выполнение процедуру полиномиальной регрессии. Это можно сделать с помощью меню, как было описано выше, или просто нажать клавиши Alt F9 для повторения последней процедуры. При этом в поле **Вывести** диалого-

Модель= a0+a1t^1

Переменная zerno.var1

Параметр	Оценка	Станд.Ошибка	Т-Значение	Р-Значение
a0	6.143091778	0.4719433357	13.01658761	0
a1	0.2751119857	0.01847328547	14.89242324	0

Источник	Сумма Квадратов	Степ. Свободы	Среднее Знач.
Модель	7267.382119	2	3633.691059
Ошибка	111.377769	43	2.590180675
Общая	7378.759888	45	163.9724419

Рис. 13.23. Расчеты, выдаваемые процедурой регрессионного анализа

Рис. 13.24. Графические результаты процедуры полиномиальной регрессии

Рис. 13.25. График остатков в процедуре полиномиальной регрессии

вого окна (рис. 13.22) надо указать режим **Вывести остатки**. Полученный график остатков приведен на рис. 13.25.

Для дальнейшего анализа остатков их следует сохранить в виде отдельной переменной (при этом они будут одновременно добавлены в электронную таблицу). Для этого в пункте меню **Результат** следует выбрать подпункт **Сохранить и добавить в таблицу**. Диалоговое окно этой

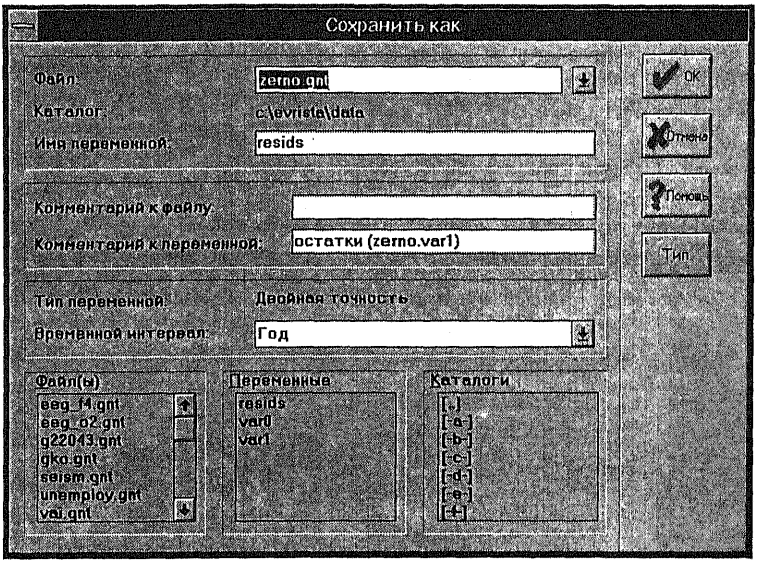


Рис. 13.26. Диалоговое окно сохранения остатков

Рис. 13.27. График автокорреляционной функции для остатков ряда урожайности зерновых

процедуры приведено на рис. 13.26. В нем необходимо указать имя файла (zerno.gnt) и имя переменной, например resids, и нажать [OK].

Проверка коррелированности остатков. В пакете существует несколько различных способов выяснения коррелированности остатков. Мы построим для этого лишь выборочную автокорреляционную функцию. Для этого пункте меню Метод можно выбрать подпункт Предварительный анализ и в нем процедуру Автокорреляционная функция. В диалоговом окне этой процедуры надо указать максимальное число задержек (лагов) автокорреляционной функции и запустить процедуру на счет, как это было указано выше. Результат этой процедуры приведен на рис. 13.27.

Обратим внимание, что график автокорреляционной функции включает нулевую задержку, в которой она по определению равна единице. Это сделано для улучшения визуального восприятия графика. На графике отсутствуют традиционные доверительные интервалы для нулевой гипотезы о некоррелированности значений ряда для всех задержек. Поэтому вывод о значимом отличии полученных оценок автокорреляций от нуля следует делать на основании формулы $-1/n \pm 2/\sqrt{n}$ для примерных 95% границ доверительного интервала нулевой автокорреляционной функции, где n — объем временного ряда (см. п. 11.10). Для ряда урожайности зерновых $n = 45$ и размах доверительного интервала составляет примерно ± 0.3. Полученные оценки выборочной автокорреляционной функции в целом укладываются в эти пределы.

Замечание. Следует помнить, что при таких объемах анализируемых данных статистические методы могут выявить только явные нарушения исходных предположений. Если же эти нарушения слабые, то формальные критерии чаще всего выявить их не в состоянии, в силу нехватки информации. Об этом всегда надо помнить при анализе относительно коротких временных рядов. Поэтому в случае анализа остатков ряда зерновых мы вынуждены заключить, что не обнаружено явных нарушений предположения о некоррелированности. А для более точных выводов у нас нет информации.

Для дальнейшего анализа остатков в пакете может быть использована процедура **Тесты на случайность** из раздела **Предварительный анализ**. Мы не будем останавливаться на этом вопросе.

Прогнозирование. Для осуществления прогноза на базе линейной (или полиномиальной) модели линейной регрессии следует в меню **Метод** выбрать пункт **Прогнозирование** и в нем выбрать подпункт **Прогноз полиномиального тренда**. В открывшемся диалоговом окне процедуры **Прогнозирование по Полиномиальной регрессии** (рис. 13.28) следует указать степень полинома, уровень доверия для линии прогноза и количество шагов прогноза. Результат работы этой процедуры приведен рис. 13.29.

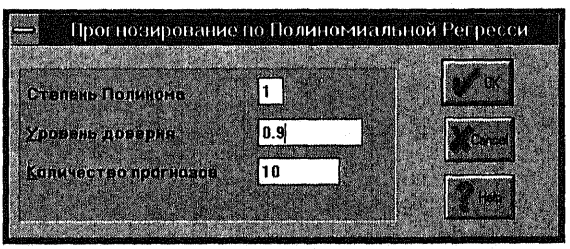


Рис. 13.28. Диалоговое окно процедуры полиномиального прогнозирования

Результаты прогнозирования могут быть сохранены в численном виде в отдельной переменной и помещены в электронную таблицу. К

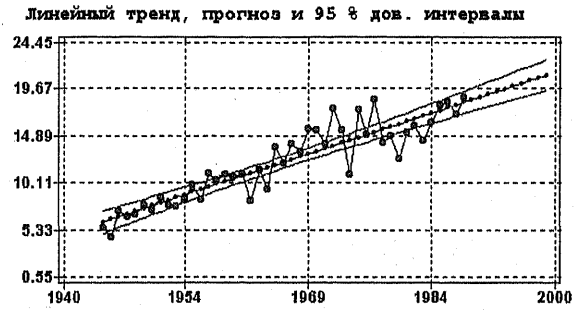


Рис. 13.29. Прогноз поведения ряда урожайности зерновых

сожалению, в пакете отсутствует построение доверительных интервалов для самих будущих значений временного ряда.

13.2.3. Устранение сезонной компоненты

Пример 13.2к. Для ряда производства молока в России оценим сезонные индексы и проведем сезонное выравнивание ряда.

Подготовка данных. Пусть данные о производстве молока в России (таблица 12.1) находятся в текстовом (ASCII) файле в виде одного столбца. Способ загрузки подобных данных в пакет и выбор их в качестве рабочей переменной подробно описан в предыдущем примере. Будем считать, что загруженные данные находятся в файле milk.gnt в переменной milk.

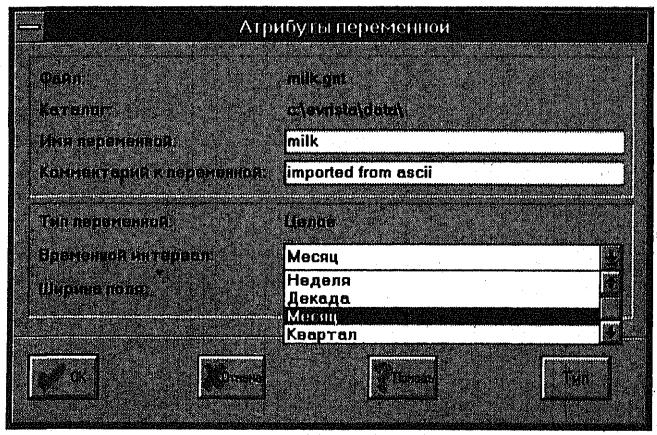


Рис. 13.30. Диалоговое окно процедуры Атрибуты переменной

Для удобства дальнейшего анализа и выдачи результатов определим дополнительно календарную структуру исходных данных. А именно,

укажем, что мы имеем дело с ежемесячными данными, начиная с января 1992 года. Для этого в пункте меню Таблица надо выбрать пункт Открыть и в его подменю выбрать пункт Атрибуты переменной. Диалоговое окно этой процедуры показано на рис. 13.30. В нем следует указать файл milk.gnt и имя переменной milk, а в поле Временной интервал выбрать значение Месяц. Затем надо щелкнуть мышью по кнопке [Тип] диалогового окна для указания стартовых календарных значений исходного ряда. При этом откроется окно Параметры временного ряда (рис. 13.31), в котором указываются день, месяц и год первого значения ряда (первое января 1992 года) и временной интервал между наблюдениями (1 месяц). Для запоминания календарных атрибутов переменной необходимо нажать функциональную клавишу [F2].

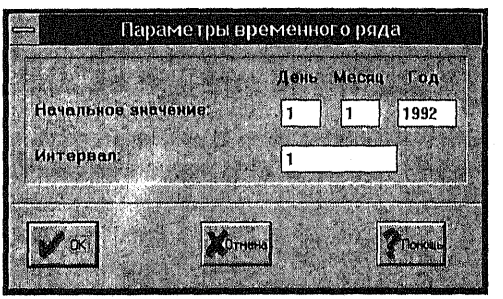


Рис. 13.31. Окно задания параметров ряда

График ряда производства молока в России приведен на рис. 13.32. (Порядок вывода данных на график описан в предыдущей задаче. В данном случае в качестве типа графика был выбран не X–Y график, а график временного ряда, а значения по оси ординат были прочитаны из календарной структуры переменной milk, созданной выше.)

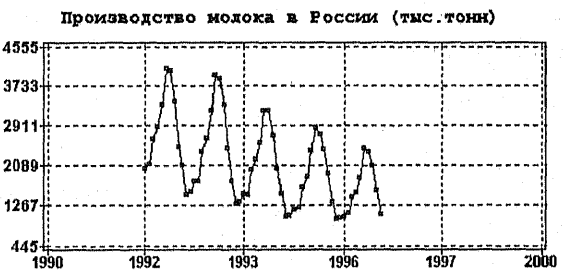


Рис. 13.32. График ряда производства молока

Выбор процедуры. Для вычисления оценок сезонной компоненты ряда следует в пункте меню Метод выбрать подпункт Предварительный анализ и в нем процедуру Сезонная Компонента. При этом возникает диалоговое ок-

но, приведенное на рис. 13.33. В нем указываем: период сезонности — 12 месяцев, тип сезонной компоненты — мультипликативная или аддитивная, и форму вывода результатов. После заполнения всех параметров и нажатия OK метод следует запустить на счет с помощью соответствующего пункта меню или пиктограммы панели управления пакета.

Рис. 13.33. Диалоговое окно процедуры Сезонной Компоненты

Порядок работы процедуры выделения сезонной компоненты в пакете Эвриста включает переход от мультипликативной модели к аддитивной, выделение тренда и циклической компоненты методом скользящих средних, затем — вычисление сезонных оценок в аддитивной модели и возврат к мультипликативной модели исходного ряда. Для удаления тренда с помощью метода наименьших квадратов можно воспользоваться процедурой Анализ тренда, разобранной выше, а потом применить к ряду остатков процедуру оценки сезонной компоненты.

Результаты. На рис. 13.34 приведены оценки сезонных индексов в процентах. По оси абсцисс приведены номера месяцев года (от 1 до 12), а по оси ординат значения сезонных индексов. Обратим внимание, что полученные сезонные индексы немного отличаются от тех, что приведены в таблице 13.3, так как для их вычисления использовались разные модификации этого метода.

Рис. 13.34. Сезонные индексы для ряда производства молока

Для получения сезонно-выровненного ряда следует повторить запуск процедуры Сезонная компонента, изменив форму выдачи результатов на режим «сезонность устранена». Результаты работы этой процедуры приведены на графике (рис. 13.35).

Рис. 13.35. График ряда производства молока после удаления сезонности

13.2.4. Подбор модели авторегрессии и построение прогноза

Пример 13.3к. Для ряда среднесуточного трафика (количества переданных данных) компьютерного телекоммуникационного канала Москва-Новосибирск сети RUNNet подберем и оценим модель авторегрессии, а затем осуществим прогноз на базе этой модели.

Подготовка данных. Пусть данные среднесуточного трафика за 10 последовательных недель находятся в переменной var0 файла trafik.gnt. (Различные способы загрузки данных в пакет Эвриста, их редактирование и вывод на график обсуждались в примере 13.1к.) График данных трафика приведен на рис. 13.36.

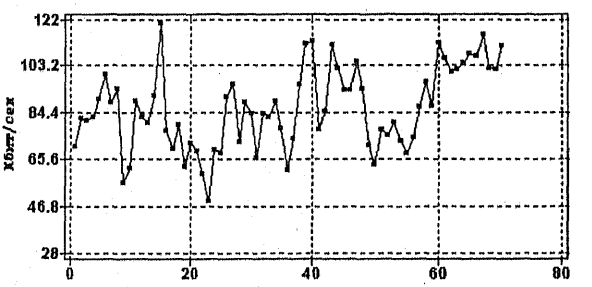


Рис. 13.36. Среднесуточный трафик в Кбит/сек

Выбор процедуры. Визуальный анализ трафика и результаты предварительных расчетов показывают, что ряд содержит локально-

линейный тренд. (В этом можно убедиться проведя оценку тренда методом наименьших квадратов.) Устраним этот тренд с помощью простого разностного оператора первого порядка. Для этого в пункте **Предварительный анализ** меню **Метод** выберем процедуру **Разности**. В диалоговом окне этой процедуры (которое мы не приводим) следует указать единственный параметр — порядок разности, то есть 1 для наших данных. Ряд первых разностей запишем в переменную **differ** файла **trafik** и выберем ее как рабочую для дальнейшего анализа. (Описание этих операций в пакете Эвриста дано в примере 13.1к.) Вид графика первых разностей (рис. 13.37) позволяет заключить, что в полученном ряде отсутствует тренд и его можно рассматривать как стационарный.

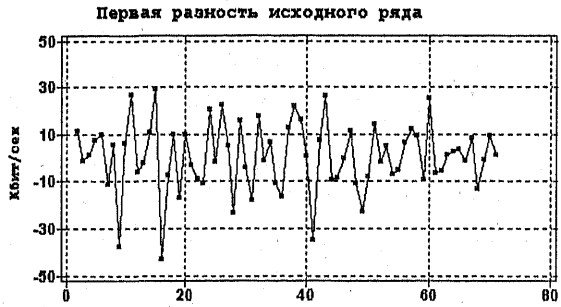


Рис. 13.37. Ряд первых разностей трафика в Кбит/сек

Подбор порядка модели авторегрессии. Для предварительного подбора порядка модели авторегрессии построим выборочную частную автокорреляционную функцию (ЧАКФ) ряда первых разностей и изучим ее значения. Для этого в пункте **АРСС Модели** меню **Метод** выберем процедуру **Идентификация**. Эта процедура осуществляет построение выборочных автокорреляционной и частной автокорреляционной функций для указанного числа задержек (лагов). График полученной частной автокорреляционной функции приведен на рис. 13.38.

Замечание. В пакете Эвриста график ЧАКФ (рис. 13.38) имеет следующие особенности: включение значения ЧАКФ для нулевой задержки (которое по определению равно 1) для лучшего визуального восприятия графика и отсутствие границ доверительных интервалов для ЧАКФ. Последнее неудобство компенсируется наличием в пакете нескольких эффективных тестов для определения порядка АР-модели, о которых мы скажем чуть ниже.

Для предварительного выбора порядка АР-модели по ЧАКФ следует найти такую задержку p, после которой все значения ЧАКФ по абсолютной величине меньше $2/\sqrt{n}$, где n — размер временного ряда (это правило объясняется далее в п. 14.3). Указанное значение p и может рассматриваться как предварительный порядок АР-модели. В рассма-

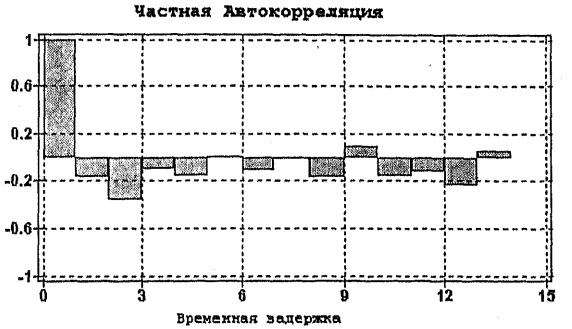


Рис. 13.38. График частной автокорреляционной функции

триваемом примере размер исходного ряда равен 70. Соответственно размер ряда первых разностей $n = 69$. Таким образом $2/\sqrt{n} = 0.2407$. Значение ЧАКФ для второй задержки явно превышает этот уровень, а все последующие значения ЧАКФ укладываются в указанные границы. Таким образом, в качестве предварительного порядка модели авторегрессии можно рассматривать значение 2.

Оценка коэффициентов модели. Для оценки коэффициентов АР(2) модели в пункте **АРСС Модели** меню **Метод** выбираем процедуру **АР-Модель**. В диалоговом окне этой процедуры (рис. 13.39) необходимо указать один из четырех возможных алгоритмов оценки параметров модели и режим выбора порядка модели. При надобности можно заказать оценки некоторых дополнительных характеристик модели, например спектральной плотности.

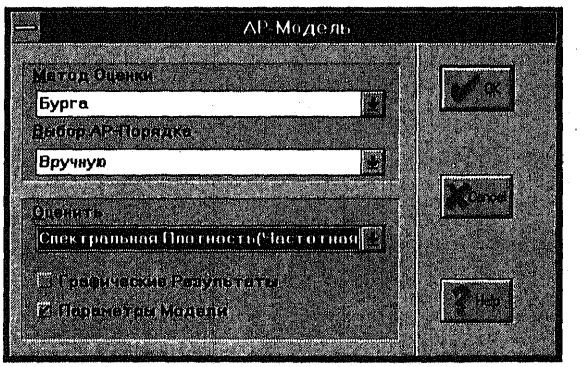


Рис. 13.39. Диалоговое окно процедуры оценки АР-Модели

Алгоритмы оценки параметров модели, реализованные в пакете, можно разбить на две группы. К первой относятся методы Левинсона-Дурбина и Бурга. Они заключаются в решении системы уравнений

Юла-Уолкера (см. п. 14.3) с использованием рекурсии для оценки параметров модели. Вторая группа алгоритмов представляет различные модификации метода наименьших квадратов. Для получения более подробной информации об особенностях этих методов пользователь может обратиться в гипертекстовой статистический справочник пакета, нажав кнопку Help в диалоговом окне (рис. 13.39).

Возможны два различных режима оценки порядка модели: автоматический и вручную. При выборе последнего режима пакет проводит экспертную оценку порядка модели с помощью различных критериев (Парзена, Акаике, Хеннана-Куина) и выдает их результаты в окно задания порядка модели (рис. 13.40). В этом окне необходимо указать выбранный вами порядок. Заметим, что различные статистические критерии подтвердили наш выбор, сделанный на базе изучения частной автокорреляционной функции.

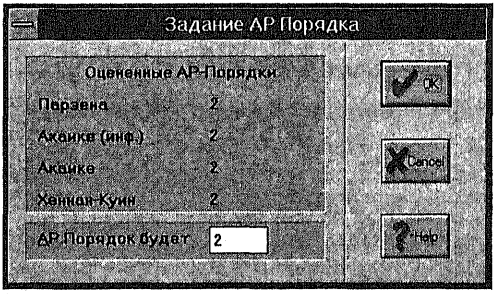


Рис. 13.40. Окно задания порядка AP-Модели

Результаты. Результаты выполнения процедуры (рис. 13.41) помещаются в специальное текстовое окно. Эти результаты включают значения оценок параметров модели и их статистические характеристики, включая уровни значимости для гипотезы о равенстве коэффициентов модели нулю, а также ряд других характеристик.

```
                    Метод :Бурга
                Переменная trafik.differ

        Дисперсия AP-Шума =182.2837271
        (Дисперсия Шума ) / (Дисперсия Ряда) =0.8634783279
  Параметр          Оценка        Станд.Ошибка   Т-Значение      Р-Значение
AP(1) =        0.2288382407    0.1107349887    2.066539613     0.02126643282
AP(2) =        0.3618942282    0.1107349887    3.268110942     0.0008447498217

        Сумма кв. Остатков =12202.3536
      Хи-Квадрат Тест автокорреляции Остатков =5.140072066
         с числом степеней свободы = 6
      Остатки есть Белый Шум с вероятностью не более чем =0.5259776017
```

Рис. 13.41. Окно результатов оценки AP-Модели

Полученные результаты показывают, что ряд первых разностей может быть описан моделью:

$$X(t) = -0.229X(t-1) - 0.362X(t-2) + \varepsilon_t$$

Прогнозирование. Для использования построенной модели при прогнозировании будущего поведения ряда необходимо записать оцененную модель в специальный файл на диске. Для этого служит команда **Сохранить модель** в меню **Результат**. Пусть название этого файла будет **traffic.mod**.

Выберем исходный ряд в качестве рабочей переменной (описание этой процедуры дано в примере 13.1к). В меню **Метод** выберем пункт **Прогнозирование**, а в его подменю — процедуру **АРСС Прогнозирование**. В диалоговом окне этой процедуры (рис. 13.42) укажем число шагов прогнозирования, скажем 5, время начала прогноза и уровень доверия. Так как прогноз осуществляется на базе модели ряда первых разностей, то в поле порядок разностей указана 1. Наконец в поле имени АРСС модели указываем имя файла **traffic.mod**, в котором была сохранена подобранная модель для ряда первых разностей.

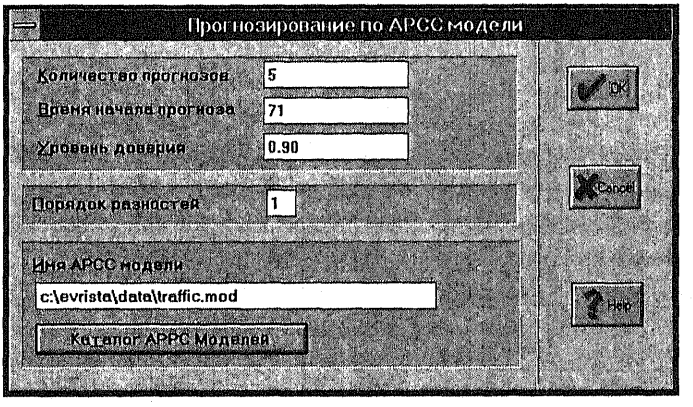


Рис. 13.42. Диалоговое окно процедуры прогнозирования АРСС модели

Результат работы процедуры в графическом виде показан на рис. 13.43. Он показывает, что ожидается рост загрузки канала. Доверительные интервалы прогноза довольно быстро расширяются, так как дисперсия белого шума в подобранной модели (см. рис. 13.43) довольно велика. На рис. 13.44 приведено сравнение полученного прогноза (пунктирная линия с отмеченными точками прогноза) с реальными данными (сплошная линия) за две последующие недели. В целом подобное качество прогноза для телекоммуникационных каналов можно считать удовлетворительным.

Рис. 13.43. Результаты процедуры прогнозирования АРСС модели

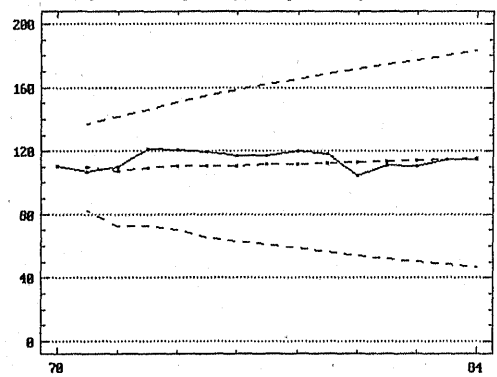


Рис. 13.44. Сравнение прогноза с реальными данными за две недели

Значение прогноза и его доверительные границы могут быть сохранены в численном виде в отдельных переменных и просмотрены в электронной таблице пакета.

Автоматизация расчетов. Разобранные примеры показывают, что анализ и прогнозирование поведения временного ряда состоит из нескольких этапов, включающих выполнение различных статистических и сервисных процедур. В пакете Эвриста нет необходимости повторять эти шаги каждый раз после пополнения ряда исходных данных. Проведенное исследование можно оформить в виде «Статистического проекта», в котором сохраняется последовательность всех необходимых для анализа процедур. Сформированный «Статистический проект» затем может быть выполнен полностью или частично для нового (пополненного) временного ряда. При этом будут построены все необходимые графики, выполнены и оформлены результаты расчетов, сохранены подобранные модели рядов и осуществлен прогноз. Процедура «Статистический проект» особенно удобна для анализа экономических данных, которые регулярно пополняются текущими наблюдениями.

Глава 14

Линейные модели временных рядов

В этой главе мы более подробно рассмотрим некоторые уже упомянутые ранее математические модели для случайных компонент временных рядов. Это процессы авторегрессии, скользящего среднего и их комбинации. Эти модели называют линейными, так как определяющие их соотношения для элементов временного ряда и случайных ошибок выражаются с помощью линейных операций над ними: сложения-вычитания и умножения-деления на действительные числа.

В первой части этой главы мы подробно рассмотрим процессы авторегрессии. Затем, уже более коротко, мы расскажем о процессах скользящего среднего. И, наконец, бегло поговорим об их комбинациях и обобщениях.

14.1. Авторегрессия первого порядка AR(1)

Рассмотрим процесс $X(t)$, значения которого в момент времени t формируется как комбинация значений этого процесса в предшествующий момент $t-1$ и некоторой случайной составляющей ε_t, независимой от значения $X(t-1)$.

Процессы такого типа могут описывать как экономические, так и технологические временные ряды. Мы предположим, что ε_t — это процесс белого шума, т.е. что в разные моменты t случайные величины ε_t независимы и одинаково распределены, причем $M\varepsilon_t = 0$, $D\varepsilon_t = \sigma^2$. Часто дополнительно предполагают, что ε_t распределены по нормальному закону.

Определение. *Случайный процесс $X(t)$ называют процессом авторегрессии первого порядка (коротко AR(1)), если для него выполняется соотношение*

$$X(t) = \phi X(t-1) + \varepsilon_t \qquad (14.1)$$

где ϕ — некоторая константа.

С помощью соотношения (14.1) можно задать значение процесса $X(t)$ в любой момент времени $t > t_0$ через значения процесса ε_t, если известна величина $X(t_0)$ в момент t_0.

В дальнейшем мы будем рассматривать только стационарные процессы авторегрессии. Это условие накладывает определенные ограничения на параметр ϕ. Они скоро выяснятся.

Числовые характеристики стационарного процесса авторегрессии. Пусть

$$m = MX(t), \quad b_k = \operatorname{cov}(X(t), X(t+k)), \quad r_k = \operatorname{corr}(X(t), X(t+k)).$$

Взяв математическое ожидание от обеих частей (14.1), получим, что $m = \phi \cdot m$. Отсюда следует, что $m = 0$, если $\phi \neq 1$. Взяв дисперсию от обеих частей (14.1), получим, что $DX(t) = \phi^2 \cdot DX(t-1) + \sigma^2$. Отсюда следует (учитывая, что $DX(t) = DX(t-1)$), что

$$|\phi| < 1, \qquad b_0 = \frac{\sigma^2}{1-\phi^2} = DX(t)$$

Таким образом, для стационарного процесса AR(1) получаем, что $|\phi| < 1$ и для любых t и k

$$MX(t) \equiv 0, \qquad b_k = MX(t)X(t+k).$$

Похожим приемом можно вычислить b_k при $k = 1, 2, \ldots$. Чтобы вычислить b_1, умножим (14.1) на $X(t-1)$ и возьмем математическое ожидание. Получаем, что $MX(t)X(t-1) = \phi \cdot MX(t-1)^2 + M\varepsilon_t X(t-1)$. Так как $X(t-1)$ и ε_t независимы, то $M\varepsilon_t X(t-1) = M\varepsilon_t \cdot MX(t-1) = 0$. Поэтому $b_1 = \phi DX(t-1)$, т.е.

$$b_1 = \phi \frac{\sigma^2}{1-\phi^2}.$$

Для вычисления b_2 заметим, что, согласно (14.1) $X(t-1) = \phi X(t-2) + \varepsilon_{t-1}$, а потому $X(t) = \phi \cdot (\phi X(t-2) + \varepsilon_{t-1}) + \varepsilon_t$. Последнее равенство умножим на $X(t-2)$ и возьмем математическое ожидание. Вычисляя, как выше, найдем, что

$$b_2 = \phi^2 \cdot DX(t-2) = \phi^2 \frac{\sigma^2}{1-\phi^2}.$$

Аналогичным образом вычисляем b_3 (здесь соотношение (14.1) надо применить дважды). Получаем, что $b_3 = \phi^3 \frac{\sigma^2}{1-\phi^2}$. Действуя таким образом и далее, найдем для любого k, что

$$b_k = \phi^k \frac{\sigma^2}{1-\phi^2}.$$

Из этих соотношений следует, что

$$r_k = \phi^k \tag{14.2}$$

Таким образом, автокорреляционная функция AR(1) процессов экспоненциально убывает с ростом лага k.

На рис. 14.1 приведены графики AR(1) процессов с дисперсией белого шума $\sigma^2 = 1$ и различными значениями коэффициента $\phi = 0.75; 0.25; -0.25; -0.75$ (каждый график построен по ста значениям ряда), а также вид выборочных автокорреляционных и частных автокорреляционных функций. (Определение частной автокорреляционной функции будет дано чуть ниже.)

Обратим внимание, что чем ближе значение ϕ к единице, тем более гладко ведет себя траектория процесса AR(1) по сравнению с траекторией белого шума. И наоборот, чем ближе значение ϕ к минус единице, тем более изломанно (пилообразно) ведет себя траектория.

Стационарный процесс авторегрессии первого порядка с ненулевым средним μ определяется соотношением:

$$X(t) - \mu = \phi \cdot (X(t-1) - \mu) + \varepsilon_t. \tag{14.3}$$

Здесь $MX(t) = \mu$.

Оценивание параметров процесса авторегрессии (14.3) по наблюдениям траектории $x_1, x_2, \ldots, x_n$.

Учитывая стационарность процесса $X(t)$, в качестве оценки μ можно взять среднее по траектории: $\hat{\mu} = \overline{x}$, где $\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$. Еще ранее для ϕ мы получили, что

$$\phi = r_1$$

Заменяя r_1 его оценкой по траектории (11.17), получаем для ϕ оценку:

$$\hat{\phi} = \overline{r}_1 = \frac{\sum_{t=1}^{n-1}(x_t - \overline{x})(x_{t+1} - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2}.$$

Наконец, уже известное соотношение $DX(t) = \frac{\sigma^2}{1-\phi^2}$ позволяет оценить и σ^2. Для этого можно воспользоваться стандартной оценкой дисперсии $DX(t)$ стационарного процесса:

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

Отсюда

$$\hat{\sigma}^2 = (1 - \overline{r}_1^2)\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

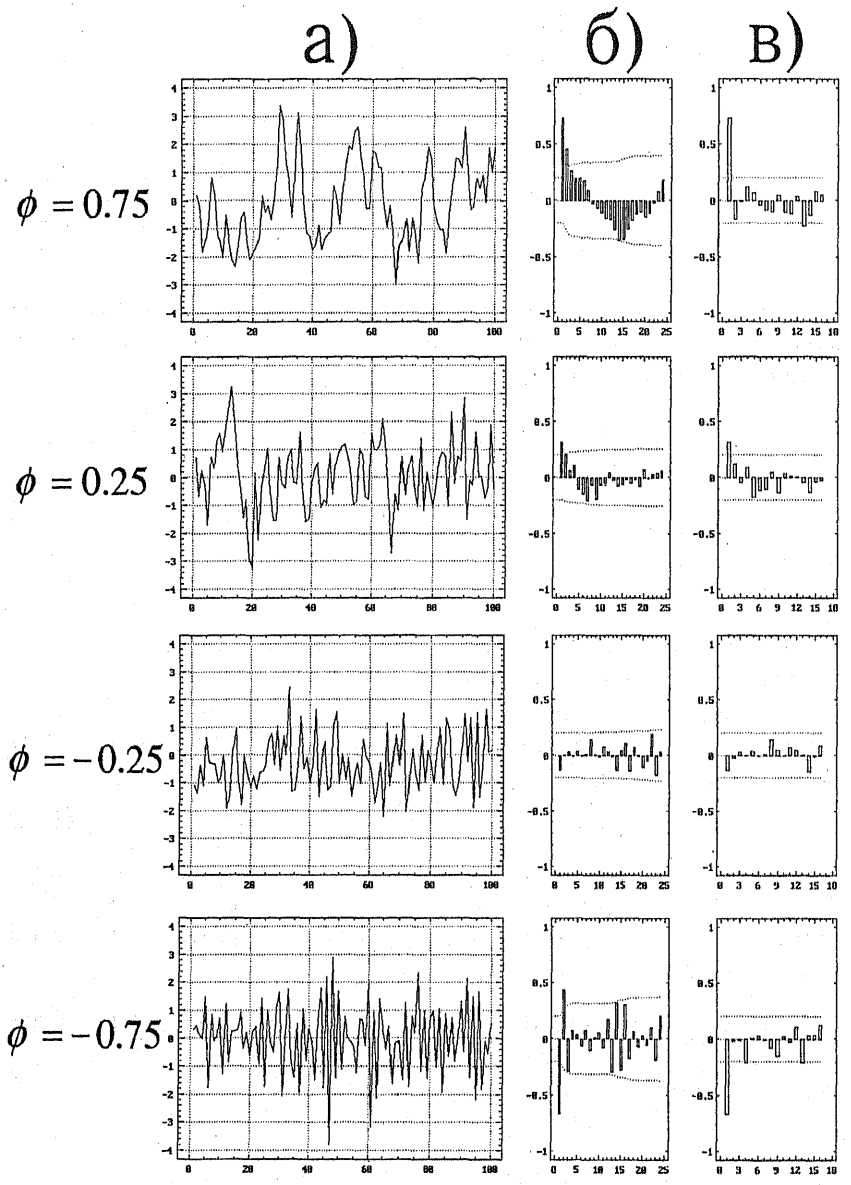


Рис. 14.1. Графики AR(1) процессов и их выборочных автокорреляционных и частных автокорреляционных функций для различных значений коэффициента ϕ. а) график исходного ряда; б) график выборочной автокорреляционной функции; в) график частной автокорреляционной функции

14.2. Авторегрессия второго порядка AR(2)

Текущее значение процесса AR(2) в момент t формируется как линейная комбинация его значений в предыдущие моменты $(t-1)$ и $(t-2)$, и независимой от них случайной величины ε_t. Как и ранее, процесс ε_t будем считать белым шумом. Процессы AR(2) обладают большей «памятью», чем процессы AR(1).

Определение. *Случайный процесс $X(t)$ называют процессом авторегрессии второго порядка (коротко AR(2)), если для $X(t)$ выполняется соотношение*

$$X(t) = \phi_1 X(t-1) + \phi_2 X(t-2) + \varepsilon_t \qquad (14.4)$$

где ϕ_1 и ϕ_2 — некоторые константы.

С помощью соотношения (14.4) значения $X(t)$ можно определить в любой момент $t > t_0$ через посредство последовательности ε_t и значений $X(t)$ в моменты t_0 и $t_0 - 1$.

Условие стационарности. Так же, как это было для AR(1), из условия стационарности $X(t)$ вытекает, что $MX(t) = 0$.

Условие стационарности накладывает также определенные ограничения на параметры ϕ_1, ϕ_2. Ниже будет показано, что для стационарного процесса AR(2):

$$\phi_1 + \phi_2 < 1, \qquad \phi_1 - \phi_2 > -1, \qquad \phi_2 > -1. \qquad (14.5)$$

Ограничения (14.5) задают на плоскости (ϕ_1, ϕ_2) треугольную область. Верно и обратное: если точка с координатами (ϕ_1, ϕ_2) попадает внутрь этого треугольника, то с помощью (14.4) можно задать стационарный процесс AR(2) с параметрами (ϕ_1, ϕ_2).

Числовые характеристики и их оценки. Уравнения Юла-Уолкера. Пусть $b_k = \operatorname{cov}(X(t), X(t-k))$. Для стационарного процесса AR(2) с нулевым средним $b_k = MX(t)X(t-k)$ для любого t. С использованием (14.4) для b_1 выводим соотношения

$$b_1 = MX(t)X(t-1) = MX(t-1) \cdot (\phi_1 X(t-1) + \phi_2 X(t-2) + \varepsilon_t) = \\ = \phi_1 b_0 + \phi_2 b_1$$

Вычисляя $\operatorname{cov}(X(t), X(t-2))$, таким же образом получим, что

$$b_2 = \phi_1 b_1 + \phi_2 b_0.$$

Для автокорреляционной функции $r_k = \frac{b_k}{b_0}$ эти равенства дают

$$r_1 = \phi_1 + \phi_2 r_1$$
$$r_2 = \phi_1 r_1 + \phi_2 \tag{14.6}$$

Соотношения (14.6) называют *уравнениями Юла-Уолкера*. Они связывают параметры процесса AR(2) со значениями его автокорреляционной функции:

$$\phi_1 = \frac{r_1 - r_1 r_2}{1 - r_1^2}, \qquad \phi_2 = \frac{r_2 - r_1^2}{1 - r_1^2}. \tag{14.7}$$

Аналогичным путем для произвольного целого k получаем соотношение:

$$r_k = \phi_1 r_{k-1} + \phi_2 r_{k-2}. \tag{14.8}$$

Рассмотрим это соотношение как уравнение, и найдем все последовательности, скажем u_k, которые ему удовлетворяют. Решения уравнения (14.8) связаны с корнями квадратного уравнения (его называют характеристическим)

$$\lambda^2 = \phi_1 \cdot \lambda + \phi_2. \tag{14.9}$$

Пусть λ_1, λ_2 — корни (14.9), которые сейчас предположим различными. Случай $\lambda_1 = \lambda_2$ рассмотрим позже. Легко проверить, что последовательности $u_k = \lambda_1^k$, $u_k = \lambda_2^k$ удовлетворяют (14.8). Более того, нетрудно доказать, что любое решение (14.8) является их линейной комбинацией, т.е. любое решение (14.8) имеет вид:

$$u_k = a_1 \lambda_1^k + a_2 \lambda_2^k, \tag{14.10}$$

где a_1, a_2 — произвольные числа.

Теперь рассмотрим случай, когда уравнение (14.9) имеет кратный корень $\lambda = \lambda_1 = \lambda_2$. Легко проверить, что в этом случае линейно-независимыми решениями (14.8) служат последовательности $u_k = \lambda^k$ и $u_k = k\lambda^{k-1}$. Поэтому общее решение (14.8) в случае кратного корня (14.9) имеет вид

$$u_k = a_1 \lambda^k + a_2 k \lambda^{k-1}. \tag{14.11}$$

Заметим, что последовательности (14.10) и (14.11) неограниченно возрастают с ростом k, если хотя бы одно из чисел $|\lambda_1|$, $|\lambda_2|$ превосходит 1. Поскольку r_k — коэффициент корреляции, и не может превосходить по модулю 1, необходимо, чтобы $|\lambda_1| \leqslant 1$, $|\lambda_2| \leqslant 1$. Более аккуратный анализ показывает, что если $X(t)$ — стационарная последовательность, не являющаяся постоянной, то

$$|\lambda_1| < 1, \qquad |\lambda_2| < 1. \tag{14.12}$$

Последнее условие — не только необходимое следствие стационарности $X(t)$, но и достаточное: если выполнено (14.12), то существует стационарная последовательность $X(t)$, удовлетворяющая (14.8).

Формулы (14.10), (14.11) указывают общее решение уравнения (14.8). Чтобы полностью задать автокорреляционную функцию r_k стационарного процесса AR(2), надо еще правильно подобрать значения неопределенных коэффициентов a_1, a_2.

Начнем со случая, когда корни λ_1, λ_2 — действительные числа. В этом случае надо взять такие действительные числа a_1, a_2, чтобы выполнялись соотношения (см. (14.6))

$$r_0 = 1, \qquad r_1 = \phi_1 + \phi_2 r_1. \tag{14.13}$$

При таком выборе a_1, a_2 формулы (14.10), (14.11) дают явное выражение для r_k при любом $k = 0, 1, 2, \ldots$.

Корни уравнения (14.9) могут быть и комплексными (комплексно-сопряженными) числами. В этом случае надо дополнительно позаботиться о том, чтобы формула (14.10) при всяком k определяла бы действительное значение для автокорреляции r_k, $k = 0, 1, \ldots$. Для этого числа a_1, a_2 следует взять тоже комплексными и сопряженными. При таком выборе выражение (14.10) преобразуется так, что в нем участвуют только действительные числа и действительные функции переменного k:

$$u_k = a \cdot b^k \sin(2\pi k f + \omega). \tag{14.14}$$

Действительные числа b и f определяются значениями λ_1, λ_2. Роль неопределенных параметров в (14.14), которые надо подбирать, играют a и ω. Для того, чтобы получить окончательную формулу для r_k, их надо выбрать с помощью условия (14.13). Видно, что формула (14.14) задает экспоненциально затухающую синусоиду. Условие стационарности (14.12) можно выписать в явном виде через значения коэффициентов ϕ_1 и ϕ_2 AR(2) процесса. Для этого надо записать значения λ_1 и λ_2 в виде корней квадратного уравнения (14.9) через ϕ_1 и ϕ_2. Решение получаемых таким образом неравенств приводит к указанным в (14.5) условиям для ϕ_1 и ϕ_2.

Определение. *Процесс авторегрессии второго порядка с ненулевым средним μ определяют соотношением*

$$X(t) - \mu = \phi_1 \cdot (X(t-1) - \mu) + \phi_2 \cdot (X(t-2) - \mu) + \varepsilon_t$$

Здесь $MX(t) = \mu$.

Оценивание параметров процесса AR(2) по наблюденной траектории $x_1, x_2, \ldots, x_n$. Учитывая стационарность рассматриваемого процесса, в качестве оценки μ можно взять $\hat{\mu} = \overline{x}$, где $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Оценки ϕ_1 и ϕ_2 можно получить из (14.7), заменяя истинные значения r_1, r_2 их выборочными оценками $\overline{r}_1$, $\overline{r}_2$:

$$\hat{\phi}_1 = \frac{\overline{r}_1 - \overline{r}_1\overline{r}_2}{1-\overline{r}_1^2}, \qquad \hat{\phi}_2 = \frac{\overline{r}_2 - \overline{r}_1^2}{1-\overline{r}_1^2}.$$

Для оценки дисперсии белого шума σ^2 может быть использована остаточная сумма квадратов S, а именно:

$$S(\hat{\mu}, \hat{\phi}_1, \hat{\phi}_2) = \sum_{t=3}^{n} \left[(x_t - \hat{\mu}) - \hat{\phi}_1 (x_{t-1} - \hat{\mu}) - \hat{\phi}_2 (x_{t-2} - \hat{\mu}) \right]^2 \quad (14.15)$$

Откуда получаем:

$$\hat{\sigma}^2 = \frac{1}{n-5} S(\hat{\mu}, \hat{\phi}_1, \hat{\phi}_2) \quad (14.16)$$

где значение знаменателя $(n-5)$ в (14.16) получено уменьшением исходного числа слагаемых $n-2$ в (14.15) на 3, за счет оценки параметров μ, ϕ_1 и ϕ_2.

14.3. Авторегрессия порядка p — AR(p)

Выше для простейших моделей авторегрессии были довольно подробно выведены и разобраны их свойства. В этом пункте мы приведем без доказательства сводку основных результатов, касающихся AR(p) процессов.

Определение. *Случайный процесс $X(t)$ со средним значением μ называется процессом авторегрессии порядка p или кратко AR(p), если для него выполняется соотношение:*

$$X(t) - \mu = \phi_1(X(t-1) - \mu) + \phi_2(X(t-2) - \mu) + \cdots + \\ + \phi_p(X(t-p) - \mu) + \varepsilon_t \quad (14.17)$$

Поведение автокорреляционной функции AR(p) процесса. По аналогии с тем, как мы поступали с AR(2) процессом, рассмотрим корреляцию между $X(t)$ и $X(t-k)$. Получаем:

$$r_k = \phi_1 r_{k-1} + \phi_2 r_{k-2} + \cdots + \phi_p r_{k-p}, \qquad k > 0. \quad (14.18)$$

Укажем общее решение уравнения (14.18) относительно r_k. Оно задается с помощью корней характеристического уравнения:

$$\lambda^p = \phi_1 \lambda^{p-1} + \phi_2 \lambda^{p-2} + \cdots + \phi_p$$

Пусть $\lambda_1, \ldots, \lambda_p$ — корни этого уравнения, которые мы предполагаем различными. Так же, как и в случае AR(2) процесса общее решение

системы разностных уравнений (14.18) относительно r_k может быть записано в виде:

$$r_k = a_1\lambda_1^k + a_2\lambda_2^k + \cdots + a_p\lambda_p^k.$$

Из требования стационарности AR(p) процесса вытекает, что все $|\lambda_i| < 1$.

Рассмотрим возможное поведение автокорреляционной функции в случае несовпадающих корней λ_i. При этом возможны два случая.

1. Корень λ_i вещественный. При этом член $a_i\lambda_i^k$ экспоненциально затухает с ростом k.
2. Пара корней λ_i, λ_j — комплексно-сопряженные числа. Как и в случае AR(2), они вносят в r_k слагаемые типа $ab^k\sin(2\pi fk+\omega)$, которые являются экспоненциально затухающими синусоидами.

Таким образом, в общем случае автокорреляционная функция стационарного AR процесса является суммой затухающих экспонент и затухающих синусоидальных волн.

Оценка коэффициентов AR(p) процесса. Рассмотрим выражение (14.18) для значений $k = 1, 2, \ldots, p$. При этом мы получим систему уравнений Юла-Уолкера (аналогичную (14.6) для AR(2) процесса).

$$\begin{cases} r_1 = \phi_1 + \phi_2 r_1 + \cdots + \phi_p r_{p-1}; \\ r_2 = \phi_1 r_1 + \phi_2 + \cdots + \phi_p r_{p-2}; \\ \quad \ldots \\ r_p = \phi_1 r_{p-1} + \phi_2 r_{p-2} + \cdots + \phi_p. \end{cases} \tag{14.19}$$

Решая эту систему относительно неизвестных значений параметров $\phi_1, \phi_2, \ldots, \phi_p$ и подставляя вместо неизвестных значений $r_1, r_2, \ldots, r_p$ их оценки $\overline{r}_1, \overline{r}_2, \ldots, \overline{r}_p$ по наблюденному временному ряду, получаем искомые оценки коэффициентов AR(p) модели.

Частная автокорреляционная функция (ЧАКФ). ЧАКФ полезна, когда по наблюденному отрезку временного ряда мы пытаемся подобрать для его описания подходящую ARMA-модель. Подобно автокорреляционной функции, ЧАКФ определяется для каждого натурального k и представляет собой бесконечную последовательность. Ее элементы мы обозначим через ϕ_{kk}, $k = 1, 2, \ldots$. Определение ЧАКФ и ее значений ϕ_{kk} тесно связано с AR(p) моделями.

Дадим определение ϕ_{pp} для произвольного p. Систему уравнений Юла-Уолкера (14.19) можно формально рассмотреть как систему уравнений, связывающих неизвестные $\phi_1, \ldots, \phi_p$ со значениями автокорреляции $r_1, \ldots, r_p$. Эта система — линейная; при заданных $r_1, \ldots, r_p$ она

легко может быть решена численно. Пусть $\phi_{1p}, \phi_{2p}, \ldots, \phi_{pp}$ — решение системы (14.19). Из этого набора чисел нам нужно всего одно число, а именно ϕ_{pp}. По определению, мы полагаем ϕ_{pp} значением ЧАКФ при $k = p$.

С уравнениями Юла-Уолкера и их решениями для $p = 1, 2$ мы уже встречались в п. 14.1 и 14.2. По результатам этих разделов мы можем найти ϕ_{kk} при $k = 1, 2$:

$$\phi_{11} = r_1, \qquad \phi_{22} = \frac{r_2 - r_1^2}{1 - r_1^2}. \tag{14.20}$$

Формальное определение ЧАКФ дано. Посмотрим, каковы ее свойства. Рассмотрим для примера стационарный процесс авторегрессий первого порядка (14.1). Согласно (14.2), в этом случае $r_1 = \phi, r_2 = \phi^2, \ldots$, причем $|\phi| < 1$. По определению ЧАКФ, здесь $\phi_{11} = \phi$. Чтобы найти ϕ_{22}, надо рассмотреть систему Юла-Уолкера (14.19) при $p = 2$ и ее решение ϕ_{12}, ϕ_{22}. С учетом (14.2), получаем, что ϕ_{12}, ϕ_{22} удовлетворяют условиям

$$\phi = \phi_{12} + \phi_{22} \cdot \phi, \qquad \phi^2 = \phi_{12} \cdot \phi + \phi_{22}.$$

Умножим первое уравнение на ϕ и вычтем из второго. Получим, что $\phi_{22} \cdot \phi^2 = \phi_{22}$. Так как $|\phi| < 1$, то это равенство возможно лишь при $\phi_{22} = 0$. Подобным способом находим, что для AR(1)

$$\phi_{kk} = 0 \qquad \text{для всякого} \quad k \geqslant 2. \tag{14.21}$$

Обратно, если выполняется (14.21), то процесс является процессом авторегрессии первого порядка.

Свойства. Приведем без доказательства некоторые свойства частной автокорреляционной функции.

1. Для любого k $|\phi_{kk}| < 1$.
2. При $k \to \infty$ имеет место $\phi_{kk} \to 0$.
3. Если рассматриваемый стационарный процесс является AR(p) процессом, то все $\phi_{kk} = 0$ при $k > p$.

Оценивание ЧАКФ. Для того, чтобы получить оценки ϕ_{kk} по реализации $x_1, \ldots, x_n$, следует для каждого k решить соответствующую систему уравнений Юла-Уолкера (14.19), в которой значения автокорреляционной функции заменены их выборочными оценками $\overline{r}_1, \ldots, \overline{r}_k$. На практике в статистических пакетах для вычисления оценок ϕ_{kk} используется специальные рекурсивные процедуры (см. [68]), позволяющие быстро осуществить вычисления оценок. Мы не будем подробнее останавливаться на этом вопросе. Последовательность оценок $\hat{\phi}_{kk}$ называют *выборочной частной автокорреляционной функцией.*

Укажем некоторые статистические свойства оценок $\hat{\phi}_{kk}$ при условии, что они построены по реализации AR(p) процесса. При $k > p$

$$M\hat{\phi}_{kk} \approx 0, \qquad D\hat{\phi}_{kk} \approx 1/n. \tag{14.21}$$

Указанные аппроксимации справедливы, если k много меньше длины реализации n. Это свойство оценок позволяет использовать выборочную частную автокорреляционную функцию для подбора порядка p модели процесса авторегрессии.

Подбор порядка p модели AR(p) процесса. Правило предварительного выбора порядка модели AR(p) процесса с использованием выборочной частной автокорреляционной функции звучит так. В качестве предварительного порядка модели AR(p) можно рассматривать такое число p, начиная с которого все последующие оценки выборочной частной автокорреляционной функции отклоняются от нуля не более чем на $\pm 2/\sqrt{n}$. То есть

$$|\hat{\phi}_{kk}| < 2/\sqrt{n}, \qquad \text{для всех} \quad k > p.$$

Окончательный подбор порядка модели AR(p) процесса связан со статистической значимостью полученных коэффициентов модели и детальным изучением поведения остатков, получаемых вычитанием из исходного ряда $x_1, \ldots, x_n$ значений подобранной AR(p) модели $\hat{x}_i$. Пусть $\hat{\phi}_1, \ldots, \hat{\phi}_p$ — оценки коэффициентов подобранной модели. Для удобства записи формул обозначим первые p значений реализации $x_1, \ldots, x_n$ через $\hat{x}_1, \ldots, \hat{x}_p$. Тогда подобранное значение AR(p) с номером $p+1$ можно записать в виде:

$$\hat{x}_{p+1} = \hat{\phi}_1\hat{x}_p + \hat{\phi}_2\hat{x}_{p-1} + \cdots + \hat{\phi}_p\hat{x}_1 \tag{14.22}$$

Подобранное значение с номером $p+2$ имеет вид:

$$\hat{x}_{p+2} = \hat{\phi}_1\hat{x}_{p+1} + \hat{\phi}_2\hat{x}_p + \cdots + \hat{\phi}_p\hat{x}_2 \tag{14.23}$$

где значение $\hat{x}_{p+1}$ в (14.23) вычислено с помощью (14.22). Продолжая этот итеративный процесс, можно получить все значения $\hat{x}_i$ при $i = p+1, \ldots, n$, а также спрогнозировать дальнейшее поведение процесса, то есть вычислить значения $\hat{x}_{n+1}, \hat{x}_{n+2}$ и т.д. Если полученные остатки $x_i - \hat{x}_i$ для $i = p+1, \ldots, n$ ведут себя как белый шум, то процесс подбора модели можно считать завершенным. В противном случае, следует изменить порядок подбираемой модели или перейти к более сложным комбинированным моделям авторегрессии-скользящего среднего.

14.4. Процессы скользящего среднего MA(q)

Аббревиатура MA в заголовке образована от английского названия этих процессов: moving average. Данное сокращение стандартно используется для этих процессов в литературе и статистических пакетах.

Начнем с примера. Пусть, как и ранее, ε_t обозначает процесс белого шума, $M\varepsilon_t = 0$, $D\varepsilon_t = \sigma^2$. Белый шум ε_t можно понимать как в широком, так и в узком смысле. Соответственно в широком либо узком смысле окажутся стационарными далее вводимые случайные процессы $X(t)$. Рассмотрим временной ряд, заданный соотношением

$$X(t) = \varepsilon_t + \varepsilon_{t-1} \tag{14.24}$$

Очевидно, что $X(t)$ — стационарный процесс, причем $MX(t) = 0$, $DX(t) = 2\sigma^2$. Ясно, что траектории $X(t)$ будут более гладкими, чем траектории белого шума ε_t, так как корреляция между соседними членами процесса $X(t)$ положительна:

$$\operatorname{corr}(X(t), X(t+1)) = 0.5.$$

Корреляция между более удаленными членами при этом равна 0:

$$\operatorname{corr}(X(t), X(t+k)) = 0 \qquad \text{для} \quad |k| \geqslant 2.$$

Процесс (14.24) — простой пример процессов скользящего среднего. Дадим общее определение этих процессов.

Определение. *Случайный процесс $X(t)$ называется процессом скользящего среднего порядка q (кратко MA(q)), если для него выполняется соотношение:*

$$X(t) = \varepsilon_t + \theta_1\varepsilon_{t-1} + \dots \theta_q\varepsilon_{t-q}. \tag{14.25}$$

Свойства. Очевидно, что MA(q) (14.25) — стационарный случайный процесс,

$$MX(t) = 0, \qquad DX(t) = \sigma^2(1 + \theta_1^2 + \dots + \theta_q^2).$$

Используя (14.25), нетрудно подсчитать, что для $|k| \leqslant q$

$$\operatorname{cov}(X(t), X(t+k)) = \sigma^2(\theta_k + \theta_1\theta_{k+1} + \dots + \theta_{q-k}\theta_q), \tag{14.26}$$

и что для $|k| > q$ выполняется $\operatorname{cov}(X(t), X(t+k)) = 0$. Из этого последнего свойства следует, что автокорреляция r_k обращается в нуль вне некоторого конечного участка:

$$r_k = 0 \qquad \text{для} \qquad |k| > q.$$

Это свойство автокорреляции хорошо различимо на ее графике. Оно позволяет уверенно различать процессы скользящего среднего,

основываясь на графике выборочной автокорреляционной функции $\overline{r}_k$, если наблюдаемая траектория процесса достаточно велика.

К сожалению, оценивание коэффициентов θ_j в (14.25) по наблюдаемому участку траектории — довольно сложная в теоретическом и вычислительном отношении задача. Ниже мы излагаем одно из возможных ее решений. К сожалению, при этом приходится обращаться к несколько более сложным математическим средствам, чем мы обходились до сих пор.

Оценивание неизвестных параметров $\theta_1, \ldots, \theta_q$ процесса скользящего среднего MA(q) по наблюденному отрезку его траектории может быть проведено по методу наименьших квадратов. Этот метод нам хорошо знаком по обработке независимых наблюдений, например, в схеме линейной регрессии. Для статистических наблюдений, связанных взаимной зависимостью, каковыми являются временные ряды, метод наименьших квадратов приходится применять в соответственно измененном и обобщенном виде.

Рассмотрим сначала схему MA(1). Пусть θ означает (единственный) неизвестный параметр. Пусть $X(1), X(2), \ldots, X(n)$ обозначает отрезок временного ряда. Рассмотрим эту совокупность как n-мерный случайный вектор. Соотношение (14.26) позволяет немедленно выписать матрицу ковариаций этого вектора. Обозначив ее через $\Delta = \Delta(\theta)$, находим что

$$\Delta = \begin{pmatrix} 1+\theta^2 & \theta & 0 & & \ldots & 0 & 0 \\ \theta & 1+\theta^2 & \theta & 0 & \ldots & 0 & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & 0 & 0 & & \ldots & \theta & 1+\theta^2 \end{pmatrix}$$

(Матрица Δ — квадратная матрица размера $n \times n$; элементы ее главной диагонали равны $1+\theta^2$; элементы диагоналей, примыкающих к главной сверху и снизу равны θ; прочие элементы матрицы Δ равны 0.)

Пусть $x_1, \ldots, x_n$ суть наблюденные значения упомянутых $X(1), \ldots, X(n)$. Обозначим через Z их совокупность, записанную в виде вектора столбца

$$Z = (x_1, \ldots, x_n)^T.$$

По этому вектору Z нам необходимо сделать вывод о том, каково истинное значение θ в формуле (14.25), с помощью которой отрезок белого шума $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_n$ перешел в Z.

Метод наименьших квадратов, относящийся к взаимно зависимым случайным величинам, в данном случае состоит в выборе в качестве оценки неизвестного θ такого числа $\hat{\theta}$, которое доставляет минимальное значение квадратичной формы $Z^T [\Delta(\theta)]^{-1} Z$:

$$Z^T [\Delta(\theta)]^{-1} Z \to \min_{\theta} \tag{14.27}$$

Здесь $[\Delta(\theta)]^{-1}$ обозначает матрицу, обратную по отношению к $\Delta(\theta)$. Явные выражения для элементов матрицы $[\Delta(\theta)]^{-1}$ указать довольно сложно, да и они не приведут нас к аналитическому решению полученной экстремальной задачи.

Указанная задача на минимум решается численно, когда известен вектор Z. Статистические пакеты, предназначенные для анализа временных рядов, как правило, содержат для этого специальные программы.

Можно указать несколько более явные формулы для оценивания θ, если ввести некоторые упрощающие предположения. Эти упрощения мало влияют на конечный результат (на количественное значение оценки), когда число наблюдений n не слишком мало. Рассмотрим сначала случай $|\theta| < 1$. Предположим, что $\varepsilon_0 = 0$. При этом предположении временной ряд $X(t)$, задаваемый соотношением (14.25), не является стационарным. Однако распределение его элементов $X(t)$ с ростом t быстро сближается с распределением для стационарного ряда MA(1). Это и обеспечивает количественную близость той оценки, которая сейчас будет выведена, с уже упомянутой оценкой $\hat{\theta}$.

В случае $\varepsilon_0 = 0$ вектор $X = (X(1), X(2), \ldots, X(n))^T$, удовлетворяющий (14.25), получается из вектора $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ с помощью линейного преобразования с матрицей, скажем

$$A = A(\theta) = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ \theta & 1 & 0 & \ldots & 0 & 0 \\ 0 & \theta & 1 & \ldots & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & & \ldots & 0 & \theta & 1 \end{pmatrix}$$

(Матрица A — квадратная матрица размера $n \times n$; ее элементы на главной диагонали равны 1; элементы диагонали, снизу примыкающей к главной равны θ; прочие элементы матрицы A равны 0.) С помощью матрицы A можно записать, что $X = A\varepsilon$ и что матрица ковариаций X равна AA^T.

Применяя к $Z = (x_1, \ldots, x_n)^T$ упомянутый обобщенный метод наименьших квадратов, приходим к следующей задаче

$$Z^T \left[A(\theta)A^T(\theta)\right]^{-1} Z \to \min_{\theta} \tag{14.28}$$

В данном случае $(AA^T)^{-1} = (A^T)^{-1}A^{-1}$, ибо A — квадратная матрица; обратную к ней мы обозначили как A^{-1}. Поэтому решение (14.28) совпадает с решением более простой экстремальной задачи

$$|A^{-1}(\theta)Z|^2 \to \min_{\theta}. \tag{14.29}$$

Явное выражение для $[A(\theta)]^{-1}$ легко указать:

$$A^{-1} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ -\theta & 1 & 0 & \ldots & 0 & 0 \\ \theta^2 & \theta & 1 & \ldots & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ (-\theta)^{n-1} & (-\theta)^{n-2} & & \ldots & -\theta & 1 \end{pmatrix}$$

Поэтому (14.29) можно записать более явно, как

$$\min_{\theta:\,|\theta|<1} \left[x_1^2 + (x_2 - \theta x_1)^2 + (x_3 - \theta x_2 + \theta^2 x_1)^2 + \ldots \right.$$
$$\left. \cdots + (x_n - \theta x_{n-1} + \theta^2 x_{n-2} + \cdots + (-\theta)^{n-2} x_2 + (-\theta)^{n-1} x_1)^2\right].$$

Решение этой последней задачи также может быть найдено только численными методами.

Для $|\theta| > 1$ можно провести аналогичное упрощение (14.27). Для этого надо положить $\varepsilon_n = 0$ и явно выразить $\varepsilon_{n-1}, \varepsilon_{n-2}, \ldots, \varepsilon_1, \varepsilon_0$ через $x_n, x_{n-1}, \ldots, x_1$. Мы не будем проводить этих выкладок.

Без принципиальных изменений указанный обобщенный метод наименьших квадратов можно применять к процессу MA(q) — с тем отличием, что матрица ковариаций в этом случае будет зависеть не от одного параметра, как это было для MA(1), а от q параметров $\theta_1, \ldots, \theta_q$. Оценки для $\theta_1, \ldots, \theta_q$ тоже можно найти только численно, с помощью уже упомянутых вычислительных программ. Существуют и другие методы оценивания для схемы MA(q), но здесь не место для их обсуждения.

14.5. Комбинированные процессы авторегрессии-скользящего среднего ARMA(p, q)

Происхождение аббревиатуры ARMA очевидно: она соединяет сокращения AR и MA, нам уже известные. Числа p и q указывают порядок процесса.

Определение. *Случайный процесс $X(t)$ называется процессом авторегрессии-скользящего среднего порядков p и q соответственно (кратко ARMA(p, q)), если для него выполняется соотношение:*

$$X(t) = \sum_{i=1}^{p} \phi_i X(t-i) + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}, \qquad (14.30)$$

где ε_t — процесс белого шума, $M\varepsilon_t = 0$, $D\varepsilon_t = \sigma^2$

В согласии с параграфом 14.3, процесс (14.30) может быть стационарным, только если все корни характеристического многочлена

$$\lambda^p = \phi_1 \lambda^{p-1} + \cdots + \phi_{p-1}\lambda + \phi_p$$

по абсолютному значению меньше единицы.

Формулы, выражающие автоковариацию и автокорреляцию стационарного случайного процесса (14.30) через коэффициенты $\phi_1, \ldots, \phi_p$ и $\theta_1, \ldots, \theta_q$, выглядят сложно и мы их не приводим. Скажем только, что для $k > q$ автокорреляция r_k процесса (14.30) удовлетворяет тем же уравнениям Юла-Уолкера, что уже были получены для процесса AR(p):

$$r_k = \sum_{i=1}^{p} \phi_i r_{k-i} \quad \text{для всех} \quad k > q.$$

Поэтому при больших k автокорреляция r_k процесса ARMA(p, q) приобретает такую же форму, как и у процесса AR(p).

Оценивание. Прежде чем приступить к оцениванию параметров в (14.30) по наблюдаемому участку траектории X, надо прежде выбрать порядок (p, q) модели ARMA(p, q). Для такого выбора редко когда есть теоретические основания. Обычно решение принимают, руководствуясь формой выборочной автокорреляционной функции $\overline{r}_k$, выборочной частной автокорреляционной функцией $\hat{\phi}_{kk}$ и естественным стремлением иметь модель наиболее простого вида.

Но даже после выбора порядка модели оценивание коэффициентов $\phi_1, \ldots, \phi_p, \theta_1, \ldots \theta_q$ представляет сложную задачу. По счастью, многие статистические пакеты прикладных программ содержат алгоритмы для ее решения. Мы не будем касаться подробностей.

Заметим, что задача оценивания не всегда разрешима. Рассмотрим, например, процесс ARMA(1, 1), где

$$X(t) - \phi X(t-1) = \varepsilon_t + \theta\varepsilon_{t-1}.$$

Здесь, если $\phi = -\theta$, решением $X(t)$ служит ε_t, и значения ϕ и θ не оказывают на процесс $X(t)$ никакого влияния. Поэтому они и не могут быть определены по траектории.

14.6. Линейные модели и операторы сдвига

Короткие формулы для записи перечисленных ранее линейных моделей и их обобщений можно получить, если ввести так называемый оператор сдвига. Точнее — оператор сдвига назад, который мы обозначим через B. Этот оператор действует на множестве последовательностей конечных или бесконечных. Проще всего определить его для бесконечных в обе стороны последовательностей. Пусть u, v — две такие последовательности:

$$u = (\ldots, u_{-2}, u_{-1}, u_0, u_1, \ldots),$$
$$v = (\ldots, v_{-2}, v_{-1}, v_0, v_1, \ldots).$$

Мы скажем, что $v = Bu$ (последовательность v получена из u путем действия на u оператором B), если $v_t = u_{t-1}$ для всякого целого t. Тем же соотношением можно определить, как оператор B действует на последовательность $u = (u_1, u_2, \ldots, u_n, \ldots)$ или на конечную последовательность $u = (u_1, u_2, \ldots, u_N)$. При этом оператор B применяется

ко всем элементам, за исключением первого, так как значение Bu_1 не определено. В частности,

$$B(u_2, u_3, \ldots, u_N) = (u_1, u_2, \ldots, u_{N-1}).$$

Следует помнить, что после каждого действия оператора B на конечную последовательность число ее членов уменьшается на единицу.

Можно определить и степени оператора B. Например,

$$B^2u = B(B(u)).$$

Это означает, что если $v = B^2u$, то для всякого целого t: $v_t = u_{t-2}$. Можно говорить и об операторах, представляющих собой многочлены от B. Пусть

$$P(B) = a_0 + a_1B + a_2B^2 + \cdots + a_pB^p.$$

По определению, $v = P(B)u$, если:

$$v_t = a_0u_t + a_1u_{t-1} + a_2u_{t-2} + \cdots + a_pu_{t-p}.$$

С помощью оператора B и многочленов от B процессы AR, MA и ARMA можно определить короткими и выразительными формулами. Пусть $\varepsilon = (\ldots, \varepsilon_{-2}, \varepsilon_{-1}, \varepsilon_0, \varepsilon_1, \ldots)$ — процесс белого шума, $X = (\ldots, X(-2), X(-1), X(0), X(1), \ldots)$ — один из упомянутых процессов. Тогда определяющее AR(p) соотношение (14.17) можно представить в виде

$$P(B)X = \varepsilon, \quad \text{где} \quad P(B) = 1 - \phi_1B - \phi_2B^2 - \cdots - \phi_pB^p.$$

Соотношение (14.25), определяющее процесс скользящего среднего MA(q), теперь выглядит так

$$X = Q(B)\varepsilon, \qquad \text{где} \quad Q(B) = 1 + \theta_1B + \cdots + \theta_qB^q.$$

Наконец, для процесса ARMA(p, q) определяющее соотношение (14.30) есть

$$P(B)X = Q(B)\varepsilon.$$

Оператор B позволяет коротко выражать и некоторые другие преобразования, совершаемые над временными рядами. Например, с помощью B можно записать переход от последовательности к ее первым разностям или разностям более высоких порядков. Первыми разностями последовательности u называют последовательность v, если $v_t = u_t - u_{t-1}$. С помощью оператора B этот переход от u к v выглядит так:

$$v = (1 - B)u,$$

считая, что в этой формуле 1 обозначает оператор, оставляющий последовательность неизменной (тождественный оператор): $1u = u$.

Переход от u к последовательности вторых разностей, скажем, z, определяют как переход от u к $v = (1 - B)u$ и затем как переход от v к $z = (1 - B)v$. Ясно, что конечный результат можно записать как

$$z = (1 - B)^2 u.$$

Таким же образом переход к разностям произвольного порядка r можно записать как $(1 - B)^r u$.

В главе 12 мы упоминали о том, что для временных рядов переход от наблюдений к разностям первого или более высокого порядков иногда практикуют, чтобы удалить тренд.

Перечисленные выше операторы иногда применяют одновременно для выражения более сложных моделей. Таковы, например, модели ARIMA(p, r, q) (сокращение от английского названия autoregression integrated moving average) порядков p, r, q. Определяющее эти процессы соотношение таково:

$$P(B)(1 - B)^r X = Q(B)\varepsilon. \tag{14.31}$$

Многочлены $P(B)$ степени p и $Q(B)$ степени q были введены выше.

Если случайный процесс X, удовлетворяющий (14.31), стационарен, то корни характеристического многочлена

$$\lambda^p - a_1\lambda^{p-1} - \cdots - a_{p-1}\lambda - a_p = 0$$

по абсолютному значению не превосходят 1. И обратно: при этом условии существует стационарное решение для (14.31).

Глава 15

Многомерный анализ и другие статистические методы

15.1. 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.

15.2. 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 глава 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 [4] и [7]. 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 [4], [68], [79].

15.3. Факторный анализ

При исследовании сложных объектов и систем (например, в психологии, биологии, социологии 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 бо́льшую 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 главах 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 отражающие 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 [8], [79], [86].

15.4. 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 [8], [79].

15.5. 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 моет 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 [35], [68], [79].

15.6. 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 [45], [68], [70].

15.7. 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 [56], [71].

15.8. Использование статистических пакетов

В пакетах STADIA, SPSS и STATGRAPHICS представлены все перечисленные выше методы, хотя реализации их в этих пакетах отличаются. Например, для кластерного анализа и шкалирования обеспечивается различный набор возможных расстояний, стратегий объединения объектов в кластеры и методов шкалирования.

В документации и во встроенном справочнике системы STADIA читатель сможет найти дополнительные пояснения по назначению и методике применения статистических методов, описанных в этой главе.

Приложение 1

Средства анализа данных на персональных компьютерах

П1.1. Введение

Для успешного функционирования в условиях жесткой конкуренции западные фирмы, банки, страховые компании и т.д. нуждаются в тщательном анализе имеющейся информации и получении из нее надежных и обоснованных выводов. Поэтому потребность в средствах статистического анализа данных на Западе очень велика, что и послужило причиной для развития рынка статистических программ, на котором предлагаются более тысячи программ. Различные по объему и качеству реализованной статистики, области возможного применения, пользовательскому интерфейсу, цене, требованиям к оборудованию и т.п., они отражают многообразие потребностей обработки данных в различных областях человеческой деятельности.

Даже справочники, содержащие очень краткие описания пакетов, составляют солидные тома (см., например, [99], [109]). В этих справочниках содержатся описания назначения пакетов, требования к техническим характеристикам компьютера, дополнительные сервисные возможности пакетов, цены и адреса фирм-поставщиков. Информацию о новых версиях пакетов можно найти в популярных компьютерных журналах и газетах типа «PC Magazine», «PC World», «BYTE», «PC Week» и др. Некоторые рекомендации по выбору статистических пакетов периодически публикует «Мир ПК» ([19], [46], [53], [88], [34]).

Число статистических пакетов, получивших распространение в России, тоже достаточно велико (несколько десятков) и спрос на них заметно возрос в середине 90-х годов. Из зарубежных пакетов это STATGRAPHICS, SPSS, SYSTAT, BMDP, SAS, CSS, STATISTICA, S-plus, и др. (кстати, большинство из этих пакетов занимают по качеству лидирующее места в мире). Из отечественных можно назвать такие пакеты, как STADIA, ЭВРИСТА, МЕЗОЗАВР, ОЛИМП:СтатЭксперт, Статистик-Консультант, САНИ, КЛАСС-МАСТЕР и др. Проблема выбора наиболее подходящего пакета для данной категории пользователей, круга решаемых задач, типа и возможностей компьютеров и т.д., весьма непроста.

Ниже мы расскажем о принципах выбора статистических пакетов и о характеристиках пакетов, используемых в России. Специальное внимание будет уделено версиям статистических пакетов под управлением среды Windows. В приложении 2 мы постараемся более подробно описать характеристики пакетов STADIA и STATGRAPHICS, на которых мы иллюстрировали статистические методы в главах 1–10 этой книги.

П1.2. Виды статистических пакетов

Основную часть имеющихся статистических пакетов составляют специализированные пакеты и пакеты общего назначения.

Специализированные пакеты обычно содержат методы из одного-двух разделов статистики или методы, используемые в конкретной предметной области (контроль качества промышленной продукции, расчет страховых сумм и т.д.). Чаще всего встречаются пакеты для анализа временных рядов (например, Эвриста, МЕЗОЗАВР, ОЛИМП:СтатЭксперт, Forecast Expert), регрессионного и факторного анализа, кластерного анализа, многомерного шкалирования. Обычно такие пакеты содержат весьма полный набор традиционных методов в своей области, а иногда включают также и оригинальные методы и алгоритмы, созданные разработчиками пакета. Как правило, пакет и его документация ориентированы на специалистов, хорошо знакомых с соответствующими методами. Применять такие пакеты целесообразно в тех случаях, когда требуется систематически решать задачи из той области, для которой предназначен специализированный пакет, а возможностей пакетов общего назначения недостаточно.

Пакеты общего назначения. Особое место на рынке занимают так называемые *статистические пакеты общего назначения*. Отсутствие прямой ориентации на специфическую предметную область, широкий диапазон статистических методов, дружелюбный интерфейс пользователя привлекает в них не только начинающих пользователей, но и специалистов. Универсальность этих пакетов особенно полезна:

- на начальных этапах обработки, когда речь идет о подборе статистической модели или метода анализа данных;
- когда поведение статистических данных выходит за рамки использовавшейся ранее модели;
- в процессе обучения основам статистики.

Именно пакеты общего назначения составляют большинство продаваемых на рынке статистических программ. К таким пакетам относятся

системы STADIA и STATGRAPHICS, рассмотренные в этой книге, а также пакеты SPSS, SYSTAT, S-plus и др.

Неполные пакеты общего назначения. Некоторое хождение на рынке статистических программ (особенно в нашей стране) имеют пакеты, которые можно было бы назвать неполными пакетами общего назначения. Чаще всего они содержат простейшие методы описательной статистики и некоторые методы из двух-трех других разделов статистики. Как правило, это либо недоработанные первые версии вновь создаваемых пакетов, либо вынесенные на рынок программы для внутреннего, узкоспециализированного использования. Последние, кроме ограниченности статистических методов, обычно характеризуются недоработанными интерфейсами, скудностью сервисных возможностей. Отличительной чертой таких пакетов является отсутствие или слабая методическая проработка документации.

По-видимому, использование неполных пакетов общего назначения вряд ли может быть целесообразным, так как при практической работе почти наверняка (и, скорее всего, очень быстро) потребуются те методы, которые разработчики не смогли включить в пакет. Образно выражаясь, неполный пакет общего назначения похож на автомобиль, рассчитанный, скажем, на работу при температуре только от 15 до 20 градусов — иногда его можно использовать, а очень часто нельзя.

П1.3. Возможности табличных процессоров и баз данных

Вследствие большой популярности (к сожалению, имеется в виду популярность на Западе) статистических методов обработки данных соответствующие средства стали включаться в табличные процессоры общего назначения (например, в Excel, Lotus 1–2–3 и т.д.), а также в некоторые базы данных. Наиболее часто в таких пакетах встречаются средства описательной статистики, методы регрессионного анализа, средства анализа временных рядов, сглаживания и прогнозирования.

Несмотря на полезность этих средств, мы хотим самым серьезным образом предостеречь читателя от чрезмерного доверия к ним. Речь, разумеется не идет о том, что в табличном процессоре или в базе данных неверно считается среднее или дисперсия — формулы для вычисления простейших статистик, естественно, в них заложены правильные. Однако для более сложных задач типа проверки согласия или регрессионного анализа табличные процессоры и базы данных очень часто содержат грубейшие ошибки, приводящие к неправильности делаемых ими выводов. Это не удивительно — при создании этих программ статистические методы обычно включаются как некое очередное украшение, наравне со встраиванием в них двадцатого или тридцатого типа графиков и пятисотой или шестисотой встроенной функции. Поэтому обычно про-

граммирование статистических методов для таких программ сводится к переписыванию из какого-либо справочника по статистике соответствующих формул без учета их предназначения и границ применимости, что и приводит к указанным выше последствиям.

Непрофессионализм в статистике создателей подобных программ способен сказаться и во многом другом. Например, проведя регрессионный анализ, Вы можете получить совсем не те результаты из-за того, что где-то в матрице данных случайно забыли ввести одно число, а программа не исключила соответствующее наблюдение из обработки, не выдала сообщение об ошибке, а просто посчитала пропущенное число нулевым — просто потому, что таковы были заложенные в нее «соглашения». Ясно, что возможность подобных ситуаций требует крайней осторожности при использовании статистических методов, заложенных в табличные процессоры и базы данных.

Таким образом, надежнее не использовать продвинутые статистические возможности табличных процессоров и баз данных, а экспортировать анализируемые данные и обрабатывать их с помощью статистических пакетов. Если же это неудобно, то следует сравнить на одних и тех же наборах данных результаты вычислений той статистической процедуры табличного процессора или базы данных, которую Вы хотите использовать, и аналогичной процедуры статистического пакета. Если результаты для нескольких наборов данных в обоих случаях совпадают, то пользоваться статистической процедурой табличного процессора или базы данных можно. Однако при этом следует тщательно следить за правильностью подготовки исходных данных, так как обычно при каких-либо ошибках в этом случае Вы получите не сообщение об ошибке, а неправильный результат.

Далее мы будем рассматривать наиболее распространенные и универсальные статистические средства — статистические пакеты общего назначения.

П1.4. Требования к статистическим пакетам общего назначения

Для того, чтобы статистический пакет общего назначения был удобен и эффективен в работе, он должен удовлетворять многочисленным и весьма жестким требованиям. В частности, необходимо, чтобы он:

- содержал достаточно полный набор стандартных статистических методов;
- был достаточно прост для быстрого освоения и использования;

- отвечал высоким требованиям к вводу, преобразованиям и организации хранения данных, а также к обмену с широко распространенными базами данных (Excel, dBase и т.п.);
- имел широкий набор средств графического представления данных и результатов обработки: картинка порой отражает суть дела лучше, чем любые статистические показатели;
- предоставлял удобные возможности для включения в отчеты таблиц исходных данных, графиков, промежуточных и окончательных результатов обработки;
- имел подробную документацию, доступную для начинающих и информативную для специалистов-статистиков.

О том, в какой мере этим требованиям отвечают последние версии различных статистических пакетов, будет рассказано ниже в П1.6 и П1.7.

Наконец, немаловажное значение имеет цена пакета. Профессиональные западные статистические пакеты (SPSS, SAS, BMDP и т.д.) обычно стоят от 2 до 10 тысяч долларов и более. Эти пакеты позволяют обрабатывать гигантские объемы данных, включают средства описания задач на встроенном языке и дают возможность построения на их основе систем обработки информации для целых предприятий.

Пакеты, рассчитанные на массового пользователя, стоят дешевле — обычно 500–1500 долларов. Эти пакеты отличаются от профессиональных прежде всего ориентацией на индивидуального пользователя: преимущественно диалоговым режимом работы, наличием ограничений по объему обрабатываемых данных и т.д. Имеются и более дешевые пакеты (200–300 долларов и ниже), но они обычно обладают весьма скромными возможностями.

Отечественные статистические пакеты стоят существенно дешевле, как правило, их цена составляет от 200 до 500 долларов.

Информация о стоимости отечественных и зарубежных пакетов, представленных на российском рынке, дана в приложении 3.

П1.5. Состояние и особенности российского рынка

За время, прошедшее с момента подготовки первого издания этой книги, на российском рынке статистического программного обеспечения произошли существенные изменения. Они коснулись как общего качества и разнообразия предлагаемой продукции и услуг, так и состояния самого рынка и работающих на нем фирм.

О спросе на статистические программы. Середина 90-х годов характеризовалась заметным ростом спроса на прикладные программы статистической обработки данных. Выделим несколько наиболее значимых факторов этого процесса. Динамичное развитие финансового рынка в России привело к значительному спросу на работы по анализу и прогнозу его состояния. В банках и других организациях, работающих на финансовом рынке, оперативно создавались аналитические отделы, привлекавшие на работу специалистов-статистиков. Не меньший интерес вызывали социологические исследования, особенно в преддверии выборов разных уровней. Многие региональные администрации активизировали эколого-эпидемологические исследования, которые часто используют статистические методы. Постепенно восстанавливалась платежеспособность традиционных для России пользователей статистических пакетов — научно-исследовательских организаций. Спрос на специалистов в области прикладного анализа данных не замедлил сказаться на содержании и качестве программ подготовки специалистов в высших учебных заведениях. Многие ведущие вузы страны стали активно использовать отечественные и зарубежные статистические пакеты при обучении студентов. Так, пакет STADIA используется в более чем 30 различных вузах России.

Положение отечественных разработчиков и распространителей статистических пакетов, довольно тяжелое в начале 90-х годов из-за резкого падения платежеспособного спроса и отсутствия (или прекращения действия) других источников финансирования, начало постепенно укрепляться в середине 90-х. Тем не менее, крупнейшие фирмы по разработке и распространению пакетов оставались на рынке и по мере возможности вели создание новых версий своих программных продуктов. Среди них, в первую очередь, следует отметить НПО «Информатика и компьютеры» (пакеты STADIA, CONAN и SIGN), «СТАТ-ДИАЛОГ» (пакеты «МЕЗОЗАВР», «КЛАСС-МАСТЕР», «САНИ»), Центр статистических исследований МГУ (пакеты «Эвриста», «Сократ», «Ананасс»), «Росэкспертиза» (пакет «ОЛИМП:СтатЭксперт»).

Производители западных пакетов. Другой особенностью рынка статистического программного обеспечения в России стало четко обозначенное присутствие на нем представителей крупнейших международных корпораций, разрабатывающих и распространяющих статистические пакеты. В первую очередь это относится к корпорации SPSS, которой принадлежит около 40–45% продаж на мировом рынке статистических пакетов. Созданное в середине 1995 года на базе Института социологии представительство этой фирмы в России — «Статистиче-

ские системы и сервис», — проводит активную политику по распространению пакета SPSS, включающую проведение регулярных презентаций, подготовку материалов для средств массовой информации, создание русифицированной документации пакета SPSS, проведение обучения и консультаций по прикладной статистике для пользователей пакета.

Активизировалась деятельность представительства Manugistics Inc. — Санкт-Петербургской фирмы ИнфоСтрой, распространяющей в России пакет STATGRAPHICS. Ранние (до 3.0) версии этого пакета, созданные в 1983–1988 годах, по-видимому, остаются самыми распространенными в России в силу их незащищенности от нелегального копирования и возможности добиться некоторых результатов без изучения документации. Для демонстрации возможностей пакета ИнфоСтрой активно использует доклады на научных конференциях, семинарах и научных школах по теоретической и прикладной статистике [33], а также публикации в массовых и специальных изданиях ([34], [27] и др.). Фирма поддерживает работу «горячей линии» по использованию пакета в прикладных исследованиях.

Созданный в начале 90-х годов сразу для среды Windows американской фирмой StatSoft Inc. пакет STATISTICA/w в России активно распространяет фирма СофтЛайн.

Менее активно действует на российском рынке SAS Institute Inc. (его представительство открылось в Москве в 1996 году). В настоящее время этот производитель известного статистического пакета SAS главный упор в своей деятельности делает на работы по созданию и внедрению комплексных интегрированных систем доставки информации и поддержки принятия решений на уровне предприятия. При этом сами статистические методы являются лишь составной частью современных версий SAS и не распространяются отдельно.

Координаты фирм распространителей статистического программного обеспечения и более подробная информация о назначении и ценах пакетов дана в приложении 3.

Особенности распространения статистических программ в России. Следует отметить определенную специфику распространения статистических программ в России. В основном этой деятельностью занимаются непосредственно указанные выше специализированные фирмы. Это связано с необходимостью проведения квалифицированных предпродажных консультаций, так как многие покупатели стремятся предварительно убедиться на практике в том, что приобретаемый продукт способен качественно решить их задачу. При этом такие фирмы, как НПО «Информатика и компьютеры», Центр прикладных статистических исследований МГУ, «Статистические системы и сервис», по зака-

зу покупателей проводят консультации и обучение персонала методам прикладной статистики и работе с пакетами.

Распространители программного обеспечения общего назначения (продукции фирм Microsoft, Symantec, Borland и т.д.) за распространение статистических программ не берутся, так как не могут обеспечить надлежащее консультирование пользователей. А попытки организации распространения статистических пакетов через специализированные книжные магазины и компьютерные фирмы обычно приводят лишь к дискредитации этой продукции, чему авторы не раз были свидетелями.

Улучшение качества статистических программ. В последнее время качество предлагаемых на рынке пакетов в целом заметно возросло. В первую очередь это связано с появлением у многих пакетов версий под Windows. Так, довольно невзрачная графика DOS-версии пакета STADIA стала соответствовать мировым стандартам в Windows-версии. Непростой для освоения макроязык DOS-версии пакета SPSS почти полностью заменен удобным меню-ориентированным интерфейсом в Windows-версии этого пакета. Сложная процедура настройки и редактирования графики в ранних DOS-версиях STATGRAPHICS значительно упростилась в Windows-версии. Более подробно о статистических пакетах под Windows будет сказано ниже в п. П1.6.

Другой характерной чертой современных пакетов является существенное улучшение разделов документации, посвященных сути статистических процедур. Роль этой части документации статистических пакетов, на наш взгляд, для российских пользователей столь велика, что ниже этой теме посвящен отдельный пункт П1.7.

Сравнение отечественных и зарубежных программ. Сравнивая в целом отечественные и зарубежные пакеты, заметим, что с появлением Windows-версий российских пакетов они перестали уступать, а порой превзошли западные по качеству интерфейса и сервиса. К этому надо добавить наличие в ряде пакетов мощных оригинальных статистических процедур, таких как знаковые методы оценивания в регрессионных, авторегрессионных и факторных задачах (пакет SIGN), анализ временных структур (пакет ЭВРИСТА) и т.д. Главной и наиболее отличительной чертой российских пакетов от западных является их ориентация на отечественных пользователей. Следствием этого является их простота освоения, продуманный интерфейс, более содержательная контекстная подсказка по сути статистических методов. Одной из причин подобного различия является ориентация создателей зарубежных пакетов на западную культурно-информационную среду, отличающуюся от российской, по крайней мере, в следующих аспектах:

- наличием значительно более высокой статистической подготовки у пользователей, которая закладывается обязательным изучением прикладной статистики и методов анализа данных практически во всех университетах, школах бизнеса и технических колледжах, а в ряде стран и в старших классах школ;
- наличием многочисленной специальной и популярной литературы по анализу данных, которую можно без труда найти в любом ближайшем книжном магазине;
- наличием многочисленных консультационных фирм, где по телефону за несколько минут можно получить исчерпывающую консультацию по применению вычислительных методов, а при необходимости — заказать решение и более сложных проблем.

Все перечисленное оставило весьма существенный отпечаток на западных статистических пакетах середины 80-х — начала 90-х годов. В частности, обычно они лишены развитых средств оперативной помощи, подсказок и интерпретаций выводов, а их документация нередко отличается запутанностью и необозримым объемом (SAS) или же отсутствием необходимых сведений типа списка формул, по которым можно проверить корректность производимых выводов. Приятным исключением здесь является пакет SPSS, документация которого всегда представляла собой своеобразный эталон понятного и систематического учебника по использованию статистических методов.

Другой особенностью многих западных пакетов является ориентация на командный язык, как один из главных инструментов пользователя. Наличие подобного языка значительно расширяет возможности пакетов, позволяя пользователю реализовывать нестандартные подходы к обработке данных и добавлять новые процедуры. Доля и функции командного языка в различных пакетах могут варьироваться. Так в довольно популярном в научно-исследовательских центрах на Западе пакете S-plus (версия для Windows) командный язык играет доминирующую роль. Даже для проведения простейших стандартных вычислений пользователь должен составить программу из нескольких строк на специальном языке. Весьма характерен командный язык и для пакета SPSS, имеется он и в пакетах STATGRAPHICS и STATISTICA.

Таким образом, для использования западных статистических пакетов пользователь должен обладать высокой квалификацией в статистике, а часто и в программировании. Он должен быть готовым к тщательному изучению объемистой и не всегда ясно написанной документации, а также к добыванию малодоступной западной статистической литературы.

Широкое распространение современных версий западных статистических пакетов в России сдерживается прежде всего их высокой ценой — от 700–800 до нескольких тысяч долларов, а также отсутствием русифицированных версий или хотя бы документации на русском языке (исключением является SPSS, часть документации которого выпущена на русском языке), что особенно важно при невысокой общей статистической квалификации пользователей и отсутствии доступа к западной литературе по современным прикладным методам анализа данных. Чаще всего в подобной ситуации пользователь даже приблизительно не может сказать, для чего предназначена та или иная процедура анализа. Мы более подробно остановимся на этом вопросе в П1.7, рассказывая о документации современных статистических пакетов.

В отличие от западных, многие отечественные пакеты в гораздо большей степени подходят для нужд среднего российского пользователя. Здесь основные операции обычно сразу обозримы из головных меню, а рутинные процедуры выполняются с минимумом действий и разветвлений по принципу: «прямым путем — к понятному результату». Вся сопутствующая информация содержится в самой программной системе, включая справочник и интерпретатор выводов. Так, скажем, устроены наиболее популярные отечественные статистические системы STADIA, Эвриста, Мезозавр, ОЛИМП:СтатЭксперт, Статистик-Консультант.

Наиболее развитой системой контекстной экранной помощи, включающей объемный справочник-гипертекст и экспертную систему по выбору метода статистического анализа, обладает пакет STADIA. Здесь каждый числовой статистический вывод сопровождается короткой и понятной интерпретацией (впрочем, более искушенный в статистике пользователь может сделать интерпретацию результатов сам, благо все данные для этого также выводятся на экран). В пакете Мезозавр реализована оригинальная система экспертной оценки сложных моделей временных рядов. Система Эвриста включает ряд уникальных методов анализа финансовых рынков и выделяется живо и изобретательно написанной документацией, которая читается как захватывающее повествование о возможностях статистических методов. Пакет ОЛИМП:СтатЭксперт, ориентированный в первую очередь на экономистов, сам подбирает лучшую модель временного ряда из довольно обширного класса моделей. Оригинальной и довольно мощной системой экспертной поддержки при выборе регрессионных и факторных моделей обладает пакет Статистик-Консультант.

Все эти пакеты аккумулируют передовой опыт российской науки, что не удивительно: их создавали ведущие специалисты Академии наук и Московского государственного университета. Они стабильно распро-

страняются и эксплуатируются сотнями пользователей на протяжении целого ряда последних лет. За это время их основные процедуры и операции тщательно верифицированы и отшлифованы. Методы анализа сгруппированы в пунктах меню по содержательному принципу, а не по малозначащим для пользователя фамилиям авторов, как это имеет место во многих западных пакетах.

Явным недостатком отечественного рынка является малочисленность российских пакетов. Они не могут удовлетворить всех специфических запросов пользователей. Так российские социологи и психологи наряду с отечественными пакетами вынуждены использовать пакет SPSS, предоставляющий очень удобные возможности для ввода и обработки весьма специфичных данных социологических анкет. А ихтиологи используют привычный и удобный для них западный пакет NTSYS.

Замечание. Периодически на научных конференциях выставках и семинарах можно встретить и другие российские программы анализа данных. Их разработка ведется в учебных и научных центрах прежде всего для конкретных собственных нужд. Например, в Томском государственном университете разработан и используется для обучения студентов пакет MATSTAT. Такие пакеты, как правило, либо вовсе не выходят на рынок, либо не выдерживают жестких условий выживания на рынке наукоемких разработок и вскоре исчезают из широкой маркетинговой сферы. Кстати, самоокупаемость западных статистических продуктов (кроме нескольких ведущих) также недостаточно высока, поэтому большинство из них создаются не компьютерными фирмами, а университетами за счет различных дотаций.

Публикации и семинары. Развитию и становлению рынка статистических пакетов способствовало увеличение числа публикаций на эту тему такими журналами, как «Мир ПК», «Рынок ценных бумаг», русским изданием газеты «PC Week» и др. Заметную роль в освещении прикладных исследований играет журнал «Обозрение прикладной и промышленной математики», выпускаемый научным издательством «ТВП». Это издательство также осуществляет распространение в России статистической литературы ведущих западных издательств.

Свою положительную роль в этом процессе играют регулярные тематические семинары, среди которых особо следует выделить всероссийские научно-практические семинары: «Аналитика в государственных учреждениях», проводимый Программно-аналитическим управлением Администрации Президента Российской Федерации, «Анализ и прогноз финансовых рынков», проводимый Центром «Прикладная прогностика» совместно с ЦБ России на базе института Мировой экономики и международных отношений, а также регулярные семинары на базе кафедры теории вероятностей механико-математического факультета МГУ и Центрального экономико-математического института РАН.

Книги по статистическим пакетам. На Западе весьма распространены монографии и учебные пособия по прикладной статистике, и во многих из них изложение материала строится на базе того или другого пакета. К примеру, только в 1994 г. в известных издательствах вышло 3 книги по пакету S-plus [100], [107], [108]. Подобные книги стали появляться и на русском языке [9], [27], но они, к сожалению, издаются очень небольшим тиражом и недоступны широкому кругу читателей.

П1.6. Статистические пакеты в среде Windows

Середина 90-х годов характеризуется интенсивным созданием статистических пакетов, работающих в среде Windows. Windows-версии появились у традиционных производителей статистических пакетов, многие из которых начинали свою деятельность с создания программ и библиотек для больших машин (SPSS, SYSTAT, STADIA и др.). Множество новых пакетов было разработано именно для среды Windows (STATISTICA, Статистик-Консультант и др.).

На наш взгляд, стремительное развитие статистических пакетов для Windows связано не только с ростом популярности среды Windows среди пользователей, но и с рядом существенных потребностей самих статистических пакетов. Вкратце напомним основные достоинства среды Windows для пользователей и разработчиков:

- единый пользовательский интерфейс, дающий разработчику прикладной программы стандартные функции для реализации окон, меню, запросов, списков и т.д. Это приводит к значительной унификации интерфейсов различных программ;
- доступность всей оперативной памяти;
- средства обмена данными, включающие буфер обмена данными (clipboard), динамический обмен данными (DDE) и механизм связи и внедрения объектов (OLE). Эти средства существенно помогают при решении сложных задач, требующих использования более чем одной программы.
- независимость программы от внешних устройств, что дает пользователю уверенность в том, что программа будет работать со всеми устройствами, поддерживаемыми Windows;
- богатые шрифтовые возможности;
- возможность организации встроенных справочников.

Покажем, как указанные возможности отразились на качестве версий статистических пакетов под Windows.

Интерфейс. Интерфейс статистического пакета неизбежно объединяет в себе электронные таблицы, текстовый и графический редакторы, а также многочисленные процедуры ввода данных и параметров, различные для разных статистических методов. Таким образом, по сложности и разносторонности он превосходит интерфейсы большинства других программ.

Работая в DOS, пользователи при переходе с одного статистического пакета (скажем, STATGRAPHICS) на другой (например, SPSS) вынуждены были с нуля изучать загадочный интерфейс нового пакета. В Windows-версиях статистических пакетов произошла унификация большинства процедур интерфейса. Работа с данными, графиками и текстами стала выполняться примерно так же, как в других Windows-программах. Это позволило намного быстрее осваивать эти пакеты тем, кто уже хоть немного поработал с Windows-программами. На рис. П1.1—П1.2, в качестве примера, приведен вид электронных таблиц статистических пакетов STADIA и SPSS.

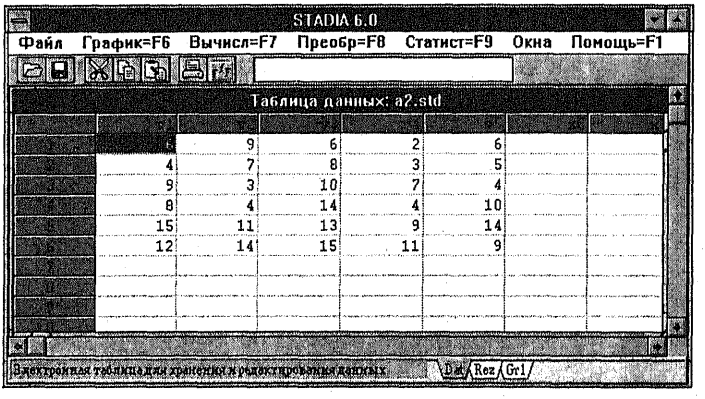


Рис. П1.1. Вид электронной таблицы в пакете STADIA 6.0

Заметной унификации подверглись пункты меню панели управления пакетов. Здесь пользователь практически всегда обнаружит заголовки: Файл (File), Преобразования (Transform), Статистика (Statistics), График (Graphs), Окна (Window), Помощь (Help). Открывая в любом из этих пакетов пункт меню Файл, пользователь найдет привычные для Windows-программ опции Новый, Сохранить, Печать, Принтер, Выход и т.д. Конечно, отдельные детали меняются от пакета к пакету, но общего стало значительно больше, чем различий.

Объемы обрабатываемых данных. Расширение возможностей работы с памятью в Windows привело к тому, что в ряде пакетов программные ограничения на объемы обрабатываемых данных или сняты

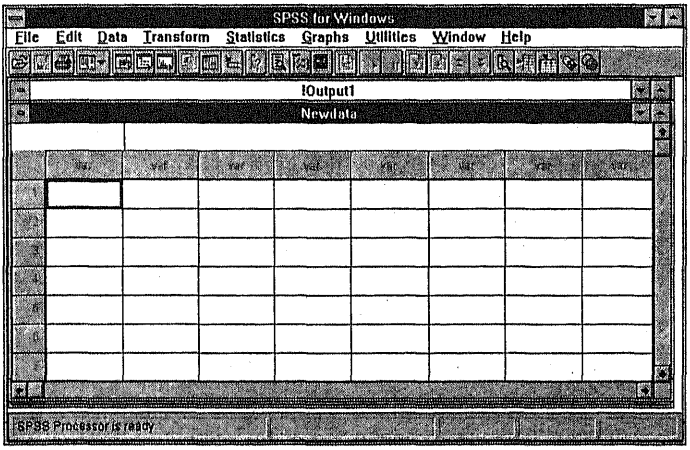


Рис. П1.2. Вид электронной таблицы в пакете SPSS

вообще (SPSS), или заметно ослаблены (STADIA 6.0, STATGRAPHICS Plus for Windows).

Многооконность. Возможности многооконной работы в Windows как нельзя лучше отвечают потребностям статистического анализа. Ведь решение статистической задачи часто требует возвращения на предыдущие этапы для коррекции данных или изменения стратегии обработки. Например, при изучении графика остатков или графика автокорреляционной функции временного ряда обычно возникает необходимость уточнения подбираемой модели, при выявлении неоднородности совокупности — необходимость разбить ее на части или удалить тренд. Графическое или расчетное обнаружение грубых, нехарактерных наблюдений влечет за собой коррекцию данных или выбор процедур, устойчивых к подобным эффектам. Подобные, порой многократные, возвраты назад типичны при решении задач статистического анализа.

Для упрощения этой работы большинство статистических пакетов предлагают возможности многооконного режима. Они могут выводить на экран окно данных, окно графиков, окно статистических процедур, окно итоговых результатов. В отдельных пакетах к ним могут добавляться и другие окна (например, окно команд в SPSS). В одних пакетах эти окна могут присутствовать на экране одновременно, в других — вызываться по очереди с помощью меню или «закладок» (как листы рабочей книги табличного процессора). При этом каждое окно сохраняет информацию, введенную на последнем шаге. Однако стоит заметить, что чрезмерное увлечение окнами, когда результат каждой промежуточной обработки выводится в отдельное окно (как это принято в пакете STATISTICA/w), засоряет экран и затрудняет работу.

Качественное улучшение графического редактирования и вывода. Ясное и четкое представление результатов является одним из важнейших элементов статистической обработки данных. Это особенно важно, если на основе статистического анализа руководство должно принимать принципиальные решения, а также если результаты статистического анализа должны быть доведены в доступной форме до массовой (не профессиональной) аудитории. Практически все DOS-версии статистических пакетов можно было подвергнуть критике за те или иные недостатки графического вывода (например, крайне сложное интерактивное редактирование графика, невозможность или трудность настройки его различных элементов и т.д.). Ситуация заметно изменилась в Windows-версиях статистических пакетов. Здесь настройка элементов графического вывода обычно сводится к указанию мышью требуемого объекта на графике и заданию его атрибутов (размера, толщины, цвета, шрифта и т.п.) в открывающихся меню.

Использование шрифтов. Шрифтовое разнообразие среды Windows позволило заметно улучшить оформление выводимых графиков, а также содержание и внешний вид гипертекстовых справочников по методам прикладной статистики, позволяя в привычном виде воспроизводить специальные математические символы и формулы (см. п. П1.7).

П1.7. Документация статистических пакетов

Особенности документации статистических пакетов. Документация статистических пакетов существенно отличается от документаций других широко используемых программных средств. Это связано с тем, что кроме общего описания порядка установки и эксплуатации, она должна содержать информацию о каждом из многочисленных и порой весьма специфических методов статистики, содержащихся в пакете. В хорошей документации для каждой статистической процедуры должны быть подробно описаны назначение процедуры, порядок заполнения полей ввода данных, выбора параметров, протокол выдачи результатов, приведены формулы для рассчитываемых величин, указаны ограничения метода, даны ссылки на первоисточники. Очень удобно, если использование статистических процедур в документации иллюстрируется на содержательных примерах.

Привычка многих отечественных пользователей осваивать программы без изучения документации или в лучшем случае по многочисленным переводным и отечественным книгам не приносит желаемых результа-

тов, когда они сталкиваются со статистическими пакетами. Дело в том, что выяснить назначение и свойства многих статистических процедур большинству отечественных пользователей попросту негде. Ведь прикладные методы статистики в последние десятилетия довольно интенсивно развивались. Например, были созданы робастные (устойчивые) методы оценивания, методы, свободные от распределения и т.д. Много нового за это время появилось в анализе временных рядов и многомерных данных. Эти методы, многие из которых реализованы в современных статистических пакетах, значительно раздвинули границы применимости статистики по сравнению с классическими методами. А базовая подготовка специалистов в наших вузах по-прежнему включает лишь классические методы анализа, выработанные в первой половине XX века и ранее (хотя в последние два-три года здесь наметились некоторые изменения, см. [54], [55]). Нет никаких сведений о современных статистических методах и практически во всех книгах по статистике, доступных на русском языке. Поэтому даже выпускники кафедр теории вероятностей и математической статистики ведущих университетов страны, в силу узкой и в основном теоретической специализации своего образования, оказываются беспомощными, сталкиваясь с массой новых незнакомых терминов и критериев.

Пример. На наш взгляд, даже среди специалистов невелико число тех, кто может сразу ответить, для чего предназначены тесты Шапиро-Уилкса и Лильефорса (они представлены в разделе *Разведочный анализ* базового модуля пакета SPSS и в разделе *Описательная статистика* пакета STATISTICA/w). Между тем, речь идет о модификации критерия Колмогорова-Смирнова для проверки нормальности распределения в случае сложной гипотезы, когда неизвестные параметры распределения оцениваются по выборке (см. главу 10). Именно эти тесты необходимо использовать в большинстве задач регрессионного и факторного анализа и анализа временных рядов, чтобы на основе анализа остатков сделать выводы об адекватности модели и возможности применения выбранного метода.

Таким образом, документация пакета и «горячая линия» фирм распространителей — это практически единственные источники информации о работе незнакомых статистических процедур. Следует отметить, что у многих известных западных пакетов в последние два-три года та часть документации, в которой описываются конкретные статистические процедуры (назначение процедур, подготовка данных, параметры, применяемые формулы, интерпретация результатов и т.д.), заметно улучшилась. Это особенно важно для отечественных пользователей.

Структура документации. Документация современных статистических пакетов обычно довольно объемна и состоит из нескольких частей. В качестве примера дадим описание структуры документации пакета STATGRAPHICS Plus v. 7 for DOS. В нее входят следующие руководства:

- *Single User Installation Guide* — установка системы (объем 100 стр.);
- *Quickstart Guide* — установка и начало работы с системой (объем 300 стр.);
- *User Manual* — описание общих процедур работы с системой (объем 450 стр.);
- *Reference Manual* — описание назначения всех статистических процедур (объем 800 стр.);
- *Examples Manual* — описание заполнения полей ввода и примеры для всех статистических процедур (объем 600 стр.).

Естественно, часть информации в этих руководствах повторяется, иногда она рассматривается под разными углами зрения. Однако в целом документация пакета хорошо структурирована и дает возможность быстро отыскать необходимую информацию.

Документация отечественных статистических пакетов значительно компактней, однако по своему содержанию может не уступать своим лучшим зарубежным аналогам (пакеты STADIA, Эвриста, Мезозавр, SIGN). В документации этих пакетов дано описание назначения процедур, большинство необходимых математических определений и, главное, содержательный разбор примеров. Эти документации стремятся к максимальной простоте изложения материала (иногда при этом приходится жертвовать строгостью изложения). Но не следует путать документацию пакета с учебником по прикладной статистике.

Хорошо оформленная документация пакетов ОЛИМП:СтатЭксперт и FORECAST EXPERT, на наш взгляд, слишком компактна. Попытка изложить основные понятия и сведения о рассматриваемых в этих пакетах моделях на 30 страницах страдает неизбежной математизацией текста и частыми недоговорками. Впрочем, ориентация этих пакетов на автоматический подбор наилучшей модели отчасти может оправдывать подобную документацию.

П1.8. Встроенный справочник и экспертная поддержка

Наряду с документацией, задачу освоения пакета помогает решать встроенный гипертекстовый справочник. И здесь среда Windows предоставляет неоценимые возможности по сравнению с DOS. Возможность включать специальные символы, формулы и графики, открывать дополнительные окна и т.д., позволяет сделать встроенный справочник значительно более удобным и информативным. Встроенными гипертекстовыми справочниками по статистическим методам анализа данных оснащены практически все упоминаемые нами пакеты. При этом в ряде пакетов присутствует режим помощи выбора метода анализа данных.

Примеры. Проиллюстрируем выбор метода анализа данных в пакете STADIA 6.0. Пусть мы хотим проанализировать связанные данные, измеренные в порядковой шкале (подобные данные весьма распространены в социологических и психологических исследованиях). В оглавлении справочника (рис. П1.3) на-

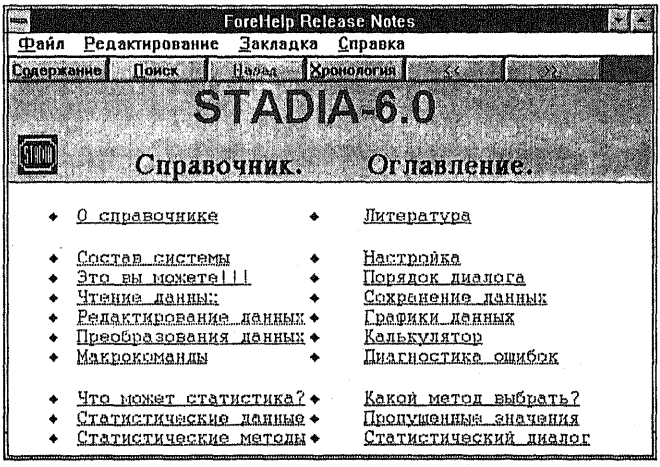


Рис. П1.3. Оглавление гипертекстового справочника пакета STADIA 6.0

ходим раздел **Какой метод выбрать?**. Содержание окна этого раздела приведено на рис. П1.4. Здесь пользователю предлагается уточнить тип своих данных. Выбираем **Связанные выборки**. (О типах статистических данных можно прочесть, щелкнув надпись **Статистические данные**.) Выведенная страница справочника **Связанные выборки** говорит, что данные этого типа могут быть измерены в различных шкалах: номинальной, порядковой и количественной, и просит сделать дальнейшие уточнения. В результате подобной цепочки уточнений будет указан конкретный метод анализа **Непараметрические коэффициенты корреляции** и подробно описан порядок работы и результаты этой процедуры (см. рис. П1.5).

Довольно похоже устроена процедура подбора метода анализа данных в пакете SPSS. Правда, на некотором этапе выбора пользователю будет предложено подряд прочитать назначение 15 различных типов процедур анализа данных и выбрать из них требуемый. Весьма подробный и обстоятельный встроенный справочник пакета STATISTICA/w, на наш взгляд, неудачно структурирован. Так, например, в его разделе выбора метода анализа данных нам не удалось обнаружить совета по поиску методов оценки связи данных, измеренных в порядковой шкале. (Как выяснилось, справочник относит этот тип данных к **Continuos Variable** (непрерывным или количественным переменным), что вводит в заблуждение исследователя.)

Другой тип экспертной поддержки в статистических пакетах заключается в автоматическом комментировании программой полученных результатов. Так, пакет STADIA для большинства статистических процедур выдает заключения типа: принять или отвергнуть нулевую гипотезу, адекватна или не адекватна подобранная модель. Наиболее мощно из известных нам пакетов такая поддержка реализована в последних версиях пакета STATGRAPHICS. Процедура StatAdvisor этого пакета после обработки ваших данных генерирует текст отчета, в котором делает содержательные выводы из полученных результатов.

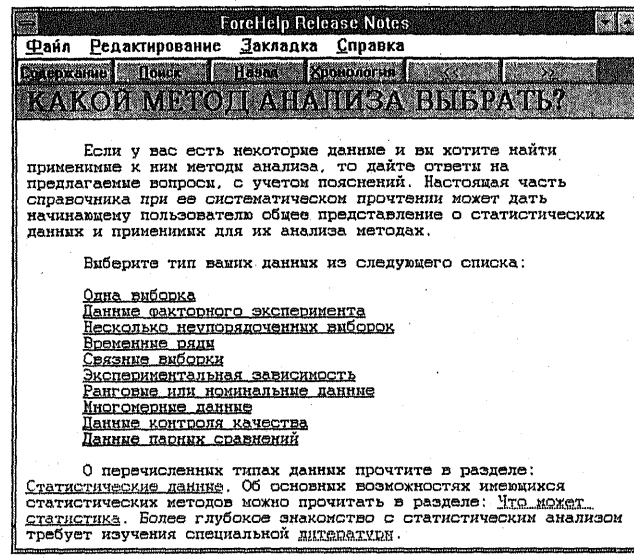


Рис. П1.4. Окно подбора метода анализа данных в пакете STADIA 6.0

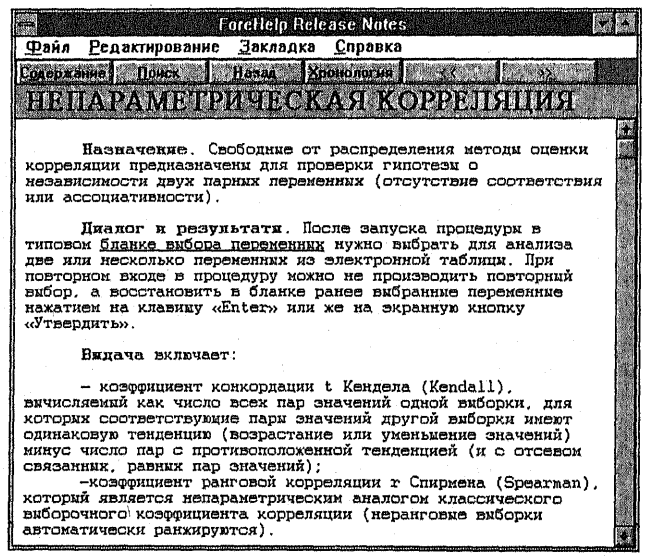


Рис. П1.5. Описание подобранной статистической процедуры в пакете STADIA 6.0

Таким образом, большинство современных статистических пакетов стремится стать доступнее и удобнее, совершенствуя свои документации, гипертекстовые справочники, экспертные системы, обучающие программы и т.п. И это дает результат — статистические пакеты ис-

пользуются все более широким кругом пользователей, они все больше применяются для обучения математической и прикладной статистике в вузах как на Западе, так и в России.

П1.9. Делая выбор

Наилучший выбор статистического пакета для анализа данных зависит от характера решаемых задач, объема и специфики обрабатываемых данных, квалификации пользователей, имеющегося оборудования и т.д.

Процедуру выбора лучше всего начать с телефонного звонка в фирму производителя и распространителя. (Информация о том, где можно приобрести статистические пакеты, приведена в Приложении 3.) При этом Вы не только получите необходимую информацию, но и проверите уровень работы «горячей линии». Последнее весьма важно, так как найти ответы на вопросы, возникшие в ходе работы, бывает совсем не просто. (Наша практика показывает, что пользователи «пиратских» копий программ порой годами не могут найти без документации ответа на вопрос, как задать и сохранить модель нелинейной регрессии в STATGRAPHICS или ввести сезонный фактор в модель временного ряда в SPSS.) Причины этого явления мы подробно рассматривали в п. П1.7.

У большинства упоминаемых нами статистических пакетов существуют демонстрационные версии. Некоторые из них (STADIA 6.0, Олимп:СтатЭксперт, SYSTAT) являются работающими программами с сильными ограничениями на объемы обрабатываемых данных и отключением некоторых сервисных функций. Демонстрационная версия STADIA 6.0 может быть приобретена (за очень небольшую плату) вместе с полной документацией к пакету STADIA 6.0.

Для пользователей, имеющих дело со сверхбольшими объемами данных или узкоспециальными методами анализа, пока нет альтернативы использованию профессиональных западных пакетов. Среди интерактивных пакетов такого рода наибольшими возможностями обладает пакет SAS.

Объемы обрабатываемых данных в пакете SPSS ограничиваются только величиной памяти вашего компьютера. Этот пакет также весьма удобен для работы с данными сложной структуры, когда необходимо делать их всевозможные срезы, как, например, в комплексном социологическом исследовании.

При создания собственной системы обработки данных можно воспользоваться библиотекой подпрограмм IMSL, содержащей сотни тщательно и квалифицированно составленных программ на Фортране и Си, которые Вы сможете встроить в собственную разработку. Библиотека

IMSL содержит также и программы по многим другим разделам численного анализа (линейная алгебра, оптимизация, дифференциальные уравнения и т.д.).

Универсальные пакеты. Если Вам необходимо обработать данные умеренных объемов (несколько тысяч наблюдений) стандартными статистическими методами, подойдет универсальный пакет. В подобных пакетах (STADIA, SPSS, STATGRAPHICS, SYSTAT, STATISTICA) достаточно полно представлены статистические методы обработки из всех областей анализа. Их простое перечисление (с модификациями) занимает несколько страниц и содержит более сотни наименований.

Работа с зарубежным пакетом потребует большей квалификации в статистике, тщательного изучения объемистой документации или специального обучения пользователей. Отчасти это компенсируется возможностью настройки ряда из этих пакетов (STATGRAPHICS, SPSS) на узкоспециализированную задачу, которая решается регулярно по мере обновления статистических данных. На наш взгляд, эти пакеты могут быть рекомендованы организациям, имеющим специалистов в области прикладного статистического анализа, которые вдобавок достаточно хорошо владеют основами программирования.

Работа с отечественными пакетами требует менее высокой квалификации пользователей, да и стоят эти пакеты существенно дешевле.

Пакеты анализа временных рядов. Учитывая особую популярность этих пакетов на отечественном рынке, сделаем ряд дополнительных замечаний по их выбору. Эти пакеты можно разбить на две группы. В первой из них (Forecast Expert, Олимп:СтатЭксперт) делается упор на автоматический или почти автоматический подбор модели временного ряда из заданного класса моделей. Это позволяет пользователю не вдумываться в результаты предварительного анализа и не требует от него специальных знаний из области временных рядов. Подобный режим работы полезен как для экспресс-анализа, так и для сравнения с результатами подбора модели вручную. Однако этот способ обработки может приводить к излишне усложненным моделям, а в некоторых случаях — и к прямым ошибкам.

Пакеты второй группы (Эвриста, Мезозавр) тоже содержат алгоритмы подбора оптимальных моделей. Но их главной чертой является широкий набор инструментов предварительного и окончательного анализа данных и возможность их пошагового применения. При этом пользователь сам задает стратегию анализа ряда. Последнее подразумевает его высокую квалификацию.

Приложение 2

Возможности пакетов STADIA и STATGRAPHICS

П2.1. Введение

Для иллюстрации большинства описываемых в книге методов статистического анализа мы использовали наиболее популярные в нашей стране пакеты общего назначения STADIA и STATGRAPHICS. Хотя в настоящее время официально распространяется 7-ая версия пакета STATGRAPHICS для DOS и 2-ая его версия для Windows (см. приложения 1, 3), мы решили оставить в книге описание ранней версии 3.0, как наиболее распространенной (судя по отзывам наших коллег) в стране. Для нас также был немаловажен достаточно простой интерфейс этих пакетов, что позволило сосредоточить внимание читателей не на особенностях их использования, а на постановках задач, формах представления данных и результатов и интерпретации результатов.

На наш взгляд, ознакомиться с примерами применения пакетов STADIA и STATGRAPHICS (равно как и пакетов SPSS и Эвриста, используемых при обсуждении временных рядов) будет полезно всем читателям, в том числе и пользователям других статистических пакетов (особенно зарубежных). Дело в том, что в большинстве статистических пакетов порядок ввода данных, а тем более, формы представления результатов и методы их интерпретации, определяются не прихотью разработчиков, а содержанием статистической процедуры и сложившимися традициями, а потому мало зависят от используемого пакета. Кроме того, среда Windows заметно унифицировала интерфейсы статистических пакетов. Поэтому описание методов работы с пакетом STADIA 6.0 может быть полезно и при использовании других статистических пакетов для Windows. А описание пакета STATGRAPHICS поможет освоению англоязычной прикладной статистической терминологии.

П2.2. О пакетах STADIA и STATGRAPHICS

Пакет STADIA. Первые версии статистической диалоговой системы STADIA (автор — А.П.Кулаичев) появились на рынке программной

продукции в 1989 г. Пакет ориентирован на массового пользователя, имеющего небольшой опыт как в статистическом анализе, так и в общении с персональным компьютером, но нуждающемся в быстром и удобном средстве оформления и обработки данных. Этот удивительно компактный пакет за счет удачного совмещения диалогового характера организации интерфейса с иерархической системой меню и продуманной контекстно-ориентированной помощью легко обучает пользователя решению возникающих перед ним задач, включая подбор статистических методов обработки данных и интерпретацию результатов анализа. Постоянное пополнение и совершенствование пакета в области статистических методов постепенно превращает его в довольно мощное средство анализа данных.

В 1996 г. начала распространяться версия этого пакета в среде Windows — STADIA 6.0. При сохранении простоты и доступности пакета, а также преемственности его интерфейса, в Windows-версии пакета кардинально улучшились графические возможности, стал более наглядным ввод данных в статистические процедуры, несколько расширились статистические возможности.

Пакет STATGRAPHICS (Statistical Graphics System) с середины 80-х годов разрабатывался корпорацией STSC (США). В настоящее время пакет распространяется Manugistics Inc. (США), имеющей свое представительство и в России (см. приложение 3). Пакет занимает одно из лидирующих мест в мире [106], и его ранние DOS-версии широко распространены в России (в основном, в виде нелегальных копий). Довольно подробное описание версии 3.0 можно найти в [27].

Кардинальная переработка пакета началась с версии 5.0. Более подробно эволюция пакета и его современное состояние описаны в [34], [33]. Отметим полноту представленных в пакете статистических методов, прекрасную двумерную и трехмерную графику, широкие возможности оперирования данными. С точки зрения массового отечественного пользователя существенным недостатком является то, что пакет рассчитан на специалистов, хорошо знакомых с концепциями применяемых процедур. Так, в довольно обширной документации ранних версий пакета практически отсутствуют как формулы расчета тех или иных характеристик, так и ссылки на конкретные первоисточники. Современная документация пакета значительно улучшилась и заметно увеличилась в объеме.

Далее мы приведем общие характеристики пакетов STADIA 6.0 и STATGRAPHICS 3.0.

П2.3. Статистические методы

В пакетах STADIA и STATGRAPHICS представлены практически все необходимые на практике статистические процедуры. Общее представление о них дают меню блока статистики системы STADIA (рис. П2.1) и головное меню пакета STATGRAPHICS (рис. П2.2).

Рис. П2.1. Система STADIA. Меню блока статистики

STATGRAPHICS Statistical Graphics System

DATA MANAGEMENT AND SYSTEM UTILITIES
- A. Data Management
- B. System Environment
- C. Report Writer and Graphics Replay
- D. Graphics Attributes

PLOTTING AND DESCRIPTIVE STATISTICS
- E. Plotting Functions
- F. Descriptive Methods
- G. Estimation and Testing
- H. Distribution Functions
- I. Exploratory Data Analysis

ANOVA AND REGRESSION ANALYSIS
- J. Analysis of Variance
- K. Regression Analysis

TIME SERIES PROCEDURES
- L. Forecasting
- M. Quality Control
- N. Smoothing
- O. Time Series Analysis

ADVANCED PROCEDURES
- P. Categorical Data Analysis
- Q. Multivariate Methods
- R. Nonparametric Methods
- S. Sampling
- T. Experimental Design

MATHEMATICAL AND USER PROCEDURES
- U. Mathematical Functions
- V. Supplementary Operations

Рис. П2.2. Пакет STATGRAPHICS. Головное меню

Многие из названных на рис. П2.1 и П2.2 статистических процедур рассмотрены в компьютерных разделах глав 1–10 этой книги. Поэтому сейчас мы скажем о возможностях пакетов только в тех задачах, которые остались за рамками книги, а так же обратим внимание на некоторые особенности в реализации процедур.

В области анализа временных рядов в обоих пакетах довольно полно реализованы процедуры корреляционного, кросскорреляционного и спектрального анализа. Представлены методы анализа моделей авторегрессии-скользящего среднего (АРСС или ARIMA модели). Подробное описание этих методов смотри в главах 12 и 14, а также в [11], [15], [29], [58].

В рассматриваемых пакетах содержатся также и методы контроля качества, широко используемые на производстве при контроле за технологическими процессами. Они включают различные типы контрольных карт, диаграмму Парето и пр. Этим вопросам посвящена обширная литература, например, [56], [71], [93]. Современные версии пакета STATGRAPHICS включают значительно расширенный модуль методов контроля качества, позволяющий эффективно использовать пакет на промышленных предприятиях.

В системе STADIA полнее, чем в STATGRAPHICS, реализованы методы дисперсионного анализа, кластерного анализа, многомерного шкалирования и критерии проверки согласия для сложных гипотез. В свою очередь, пакет STATGRAPHICS превосходит STADIA в таких разделах статистики, как планирование эксперимента, лог-линейный анализ, прогнозирование, использующее различные методы сглаживания, и др. Кроме того, из менее распространенных процедур в пакете STATGRAPHICS представлен блок методов разведочного анализа. Он включает различные методы анализа выборок, типа Box-and-Whisker Plot («ящик с усами»), Stem-and-Leaf Display («дерево с листьями») и др. Изложение этих методов можно найти в [76].

Оба пакета поддерживают обработку наблюдений с пропущенными значениями. Кроме того, в пакет STATGRAPHICS включен большой запас общематематических алгоритмов (решение систем линейных уравнений, численное дифференцирование и интегрирование и др.).

П2.4. Архитектура пакетов

Пакет STADIA. Общая архитектура пакета STADIA 6.0 является типичной для Windows-программы. В то же время интерфейс пользователя не перегружен редко используемыми процедурами и пиктограммами, а работа с окнами максимально упрощена. На рис. П2.3 приведен общий

вид интерфейса пакета. Панель управления пакета позволяет получить быстрый доступ к работе с файлами, графиками, преобразованиям данных в электронной таблице, единому меню статистических методов и статистическому и функциональному справочнику пакета.

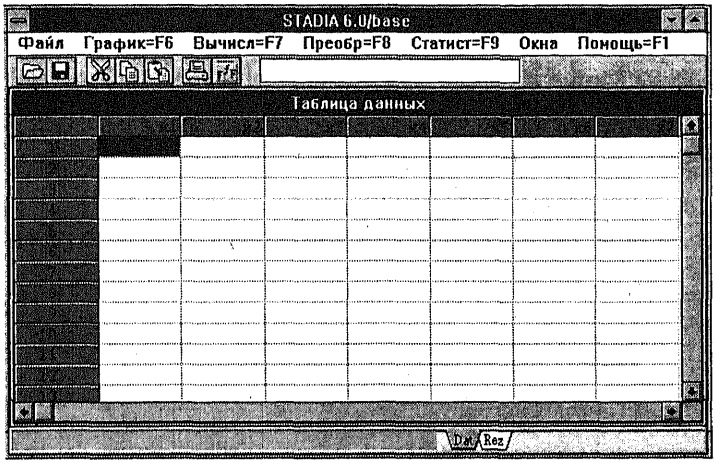


Рис. П2.3. Пакет STADIA. Общий вид интерфейса пользователя

В пакете существует три основных типа окон: окно для работы с данными в виде электронной таблицы, окна для работы с графиками и окно результатов. Для задания данных в статистических процедурах используются вспомогательные окна и меню. Общий порядок выполнения задачи в пакете сводится к загрузке или вводу файла данных в электронную таблицу пакета, вызову необходимой статистической процедуры из меню статистических методов (см. рис. П2.1), заполнению полей ввода данных и параметров процедуры и ее выполнению. Результаты работы процедуры помещаются в текстовом виде в окно результатов, а генерируемые в ходе работы процедуры графики — в отдельные графические окна. Переход из одного основного окна в другое осуществляется с помощью закладок, таких же, как закладки листов рабочей книги табличного процессора (например, Excel). Подобная идеология организации работы, учитывающая особенности порядка и специфики статистического анализа, весьма проста, логична и удобна в освоении даже неискушенными пользователями.

Пакет STATGRAPHICS. В основе организации пакета STATGRAPHICS лежит широко разветвленная иерархическая система меню. Головное меню пакета приведено на рис П2.2. Каждый пункт головного меню включает подменю. Их пунктам соответствуют конкретные статистические и другие процедуры.

Управление пакетом может осуществляться как с помощью меню, так и в командном режиме — заданием имени нужной процедуры (команды). Часть наиболее часто используемых сервисных команд присвоены функциональным клавишам F1—F10. Кроме того, пользователь имеет возможность использовать настраиваемые командные клавиши Ctrl F2 — Ctrl F5 для вызова требуемых ему процедур (см. п. П2.5).

Ввод данных в большинство процедур STATGRAPHICS сводится к заполнению полей ввода данных, выданных выбранной процедурой. Порой это требует хорошего знания статистических особенностей процедуры и использования своеобразных правил преобразования данных, принятых в пакете (см. п. П2.6).

Замечание. Некоторая громоздкость разветвленной системы меню пакета STATGRAPHICS отчасти компенсируется объединением в специальные подменю всех вспомогательных методов и процедур, которые могут иметь отношение к рассматриваемой задаче (например, в процедурах регрессионного и многофакторного анализа). Такое объединение бывает очень удобно для исчерпывающего анализа данных и проверки правильности выбранных методов обработки.

П2.5. Интерфейсы пользователя

Удобность и эффективность работы со статистическим пакетом во многом определяется дружественностью его интерфейса. Не ставя задачи сравнивать интерфейсы пакетов, реализованных в среде Windows и в среде DOS (у каждого из этих типов интерфейсов есть свои несомненные достоинства и недостатки), приведем их основные характеристики для разбираемых пакетов в таблице П2.1.

Таблица П2.1

Некоторые характеристики интерфейса пакетов

Характеристики	STADIA 6.0	STATGRAPHICS 3.0
Текущая подсказка	Есть	Есть
Помощь разъясняет	Суть процедуры и ее результаты	Порядок ввода данных
Функциональные клавиши	Действие зависит от режима программы	Контекстно-неориентированные
Применение комбинаций клавиш	Достаточно удобное	Трудное в освоении
Диагностика ошибок	Есть	Есть

Выполнение процедур в пакетах STADIA и STATGRAPHICS может быть осуществлено с помощью выбора из меню. Кроме того, в пакете STADIA выполнение процедуры осуществляется и при нажатии

клавиши, указанной в меню напротив имени процедуры (это позволяет опытным пользователям значительно ускорить свою работу). В пакете STATGRAPHICS возможен запуск процедуры с помощью ввода с клавиатуры ее имени, состоящего из нескольких символов. Так, загрузка редактора базы данных пакета может быть запущена вводом имени FILE или выбором из меню процедуры 2. File Operations пункта A. Data Management головного меню пакета.

Замечание. Даже имея опыт работы с пакетом STATGRAPHICS, порой приходится заниматься поиском необходимой процедуры в многочисленных подменю или в списке всех имен процедур. Часто возникают осложнения при заполнении многочисленных полей ввода данных и параметров процедур. Мы старались указывать возможные затруднения в компьютерных разделах каждой главы.

Панели статуса и управления. Важную роль в организации интерфейсов пакетов STADIA и STATGRAPHICS играет использование панелей статуса и управления. Панель управления пакета STADIA 6.0 приведена в верхней части рис. П2.3. Панель статуса и управления пакета STATGRAPHICS 3.0, занимающая три нижние строки экрана компьютера, изображена на рис. П2.4.

```
                Complete input fields and press F6.
1Help  2Edit  3Savscr 4Prtscr 5   6Go  7Vars  8Cmd 9Device 10Quit
INPUT    10/14/93  12:59 STATGRAPHICS Vers. 3.0      Display  FTAB
```

Рис. П2.4. Пакет STATGRAPHICS. Панель статуса и управления

Панель управления STADIA 6.0 содержит собой систему меню, типичную для Windows-программ, а также линейку пиктограмм. С меню можно работать как с помощью мыши, так и нажатиями функциональных клавиш. А наиболее часто используемые операции могут также быть вызваны с помощью кнопок с пиктограммами. Так, имеются кнопки чтения и записи содержимого активного окна, кнопки работы с буфером обмена (вырезание, копирование, вставка), кнопка выдачи на печать и кнопка изменения шрифта активной страницы. Лаконичность и простота средств управления пакета STADIA 6.0 значительно упрощает освоение и использование пакета.

В пакете STATGRAPHICS первая строка панели статуса (рис. П2.4) указывает возможные действия пользователя или сообщение об ошибке. Например, при загрузке большинства процедур пакета выдается одно и то же сообщение Complete input fields and press F6, предлагая заполнить поля ввода и нажать клавишу [F6]. Во второй строке указано назначение функциональных клавиш. Третья строка панели статуса STATGRAPHICS содержит несколько информационных полей: флажок состояния, текущую дату и время, номер версии, первичное графическое устройство

и имя текущей процедуры или команды. Флажок состояния говорит о том, что в данный момент делает пакет. Так, например, значение флажка INPUT означает, что пакет ждет вашего ввода, значок CALC появляется при выполнении длительных вычислений, значок PROCESS показывает, что система выполняет ваше последнее требование.

Функциональные клавиши. Хотя всю работу с пакетом STADIA можно вести с помощью мыши, для убыстрения работы, как и в DOS-версии пакета, можно использовать функциональные клавиши. Наименования соответствующих функциональных клавиш вынесено в заголовки меню программы (см. рис. П2.3). Так, клавиша F9 вызывает меню статистических методов, а клавиша F1 всегда вызывает на экран раздел гипертекстового справочника, соответствующий текущему режиму работы или выбранной статистической процедуре.

В пакете STATGRAPHICS функциональные клавиши являются контекстно-неориентированными и выполняют основные сервисные процедуры (табл. П2.2). Наша практика показывает, что функциональные клавиши F2 и F9 фактически никогда не используются в работе, а клавиша F5 используется довольно редко. В то же время очень полезны возможности, предоставляемые клавишами F7 и F8. Кроме основных командных клавиш, в пакете STATGRAPHICS доступны сочетания Ctrl F1—Ctrl F10.

Выводы. Простота и удобство интерфейса пакета STADIA 6.0 выделяет его в лучшую сторону по сравнению с Windows-интерфейсами таких статистических пакетов, как STATISTICA, SPSS, MINITAB, S-PLUS и др. За исключением довольно непрозрачной системы заполнения полей ввода данных в статистические процедуры, интерфейс пользователя пакета STATGRAPHICS 3.0 значительно удобнее, чем у многих других известных пакетов общего назначения в среде DOS.

П2.6. Работа с данными

Оба пакета осуществляют хранение информации в специальных форматах в собственных базах данных; однако их возможности различаются. Ввод и редактирование данных в обоих пакетах возможны с помощью встроенных редакторов данных, построенных по принципу электронных таблиц. Основным элементом данных, с которым они оперируют, является переменная (столбец данных). В таблице П2.3 даны основные характеристики переменных в базах данных пакетов.

Пакет STADIA. Работа редактора данных пакета STADIA (рис. П2.3) соответствует основным требованиям, предъявляемым к электронным

Таблица П2.2

Назначение функциональных клавиш пакета STATGRAPHICS

Клавиша	*Надпись*	*ОПИСАНИЕ*
F1	1Help	Вывод подсказки о назначении текущей процедуры и общих требованиях к вводу данных
F2	2Edit	Редактирование текущего экрана как текста или включение режима расстановки меток на текущем графике
F3	3Savscr	Сохранение текущего экрана в файле
F4	4PrtScr	Печать текущего экрана на принтер или в файл
F5	5Opt	Вывод меню опций (для отдельных процедур)
F6	6Go	Исполняет процедуру или операцию
F7	7Vars	Вывод списка переменных, доступных в вашей директории данных. Выбор из этого списка поместит имя переменной в текущее поле ввода
F8	8Cmd	Включает командный режим, который позволяет воспользоваться процедурой или командой внутри другой процедуры
F9	9Device	Определяет выходное устройство, на которое направляется весь последующий вывод (экран, плоттер или метафайл).
F10	10Quit	Выход из текущей процедуры или экрана (=Esc)

Таблица П2.3

Характеристики переменных в базах данных пакетов

Характеристики	*STADIA 6.0*	*STATGRAPHICS 3.0*
Числовые переменные	да	да
Символьные переменные	да	да
Размерность переменных	1	1–9
Число значений в одномерной переменной	до 20000	до 4096
Число переменных в файле	до 500	зависит от вида переменных

таблицам в среде Windows. Данные в электронной таблице могут редактироваться как с клавиатуры, так и путем вставки и удаления из буфера обмена. Последние операции осуществляются стандартным для среды Windows образом. С помощью левой кнопки мыши выделяется блок данных, который затем вырезается или копируется в буфер обмена, после чего содержимое буфера может быть вставлено в нужное место.

Специализированные преобразования данных осуществляются с помощью открывающегося меню преобразования данных (рис. П2.5) па-

нели управления. Оно содержит обширный набор алгебраических, тригонометрических, матричных и других операций. Пункт 2=задаваемая функция этого меню позволяет создавать и запоминать преобразования, часто используемые пользователями.

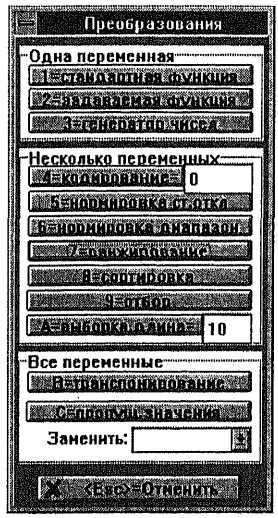


Рис. П2.5. Пакет STADIA. Меню преобразования данных

Данные из электронной таблицы пакета могут быть сразу выведены на график (клавиша F6), на печать (F2), записаны в базу данных (F4) или переданы в статистическую процедуру (F9).

Ввод данных в статистические процедуры в STADIA осуществляется непосредственно только из матрицы данных, находящейся в текущий момент в электронной таблице пакета. Это может иногда приводить к недоразумениям и требовать удаления некоторых лишних переменных из матрицы данных. В документации пакета и в гипертекстовом справочнике указаны требования, которые предъявляет к данным каждая статистическая процедура.

Пакет STATGRAPHICS имеет более широкие возможности по работе с данными: он позволяет также обрабатывать символьные переменные и переменные типа «даты» («мм/дд/гг», например, 11/27/91 для 27 ноября 1991 года) или «месяца» («мм/дд»), которые удобно использовать в ряде графических и статистических процедур (регрессии, анализа временных рядов). Впрочем, это требует дополнительных усилий по формированию структуры электронной таблицы.

Ввод данных в процедуры STATGRAPHICS весьма гибок и удобен. В полях ввода могут быть указаны как непосредственные значения данных

(небольших объемов), так и имена переменных, в которых они находятся, и выражения, использующие широкий набор операторов преобразования данных пакета. Последнее порой требует от пользователя хорошей адаптации к нетрадиционным правилам составления выражений, но позволяет моментально скорректировать данные, не обращаясь к редактору базы данных. Некоторые примеры использования операторов пакета приведены в компьютерных разделах глав 6–8.

Опишем подробнее порядок работы с редактором базы данных пакета STATGRAPHICS. Его загрузка осуществляется из пункта меню **2. File Operations** пункта **A. Data Management** головного меню пакета. На рис. П2.6 приведен экран, возникающий после загрузки процедуры.

```
                          File Operations
──────────────────────────────────────────────────────────────────────
STATGRAPHICS file name:  EXAMPLES

Operations:  A. Copy        D. Erase       G. Recode      J. Update
             B. Create      E. Join        H. Rename
             C. Edit        F. Print       I. Split

Desired operation:

                          Files on Data Drive
──────────────────────────────────────────────────────────────────────
ANDRNOR1.TXT ANOVA.ASF    ANRNOR.ASF     CARDATA.ASF    CASE1.DAT      DIAMZ.ASF
EMPDATA.DAT  EXPO.ASF     HUBER.ASF      IRONORE.ASF    LOGLIN.ASF     MAK1.CGM
MAKS1.ASF    MAKS2.ASF    MAX1.DAT       MAX2.DAT       MVDATA.ASF     NEW.ASF
NONLIN.ASF   NORM2.ASF    NORM6.ASF      NORM75.ASF     P1_1.TXT       QCDATA.ASF
RANDOM.ASF   SPEEDR.ASF   SPESDR.ASF     STATF1.ASF     STATGHLP.ASF   STATGTXT.ASF
STATR1.ASF   STATX1.ASF   STATX10.ASF    STATX15.ASF    STATX2.ASF     STATX20.ASF
STATX3.ASF   STATX4.ASF   T.ASF          TENZOV.ASF     TSDATA.ASF     TWOSTAGE.ASF
Z.ASF
```

Рис. П2.6. Пакет STATGRAPHICS. Меню процедуры редактора данных

В поле **STATGRAPHICS file name** (имя файла данных STATGRAPHICS) надо указать уже существующее имя (их список приведен ниже в блоке **Files on Data Drive**) или новое. Для создания нового файла данных укажите в поле **Desired operation** (выбор операции) букву **B** и нажмите F6. Для описания переменных файла и их дальнейшего ввода и редактирования в поле **Desired operation** надо ввести букву **C** и нажать клавишу F6. Результат этой операции представлен на рис. П2.7.

Назначение экрана на рис. П2.7 — настройка структуры переменных в файле данных. Заполнение полей **Name** (имя переменной), **Type** (тип переменной) и **Width:** (ширина поля ввода в редакторе) и последующее нажатие клавиши F6 приводит к созданию полей (столбцов) ввода с заданными характеристиками в полноэкранном редакторе базы данных. Для прекращения этого процесса надо нажать Esc, и произойдет выход в редактор базы данных пакета (рис. П2.8). Не касаясь всех тонкостей настройки, укажем возможные типы переменных доступных в пакете. Тип **N** — числовые переменные с плавающей запятой, тип **I** — числовые целочисленные переменные, тип **F** (где F может принимать значения от 1 до 9) — числовые переменные с фиксированным числом десятичных знаков после запятой, тип **D** — числовые переменные, вводимые в формате «мм/дд/гг» (например, 11/27/86 для 27 ноября 1986 года), тип **C** — для символьных переменных.

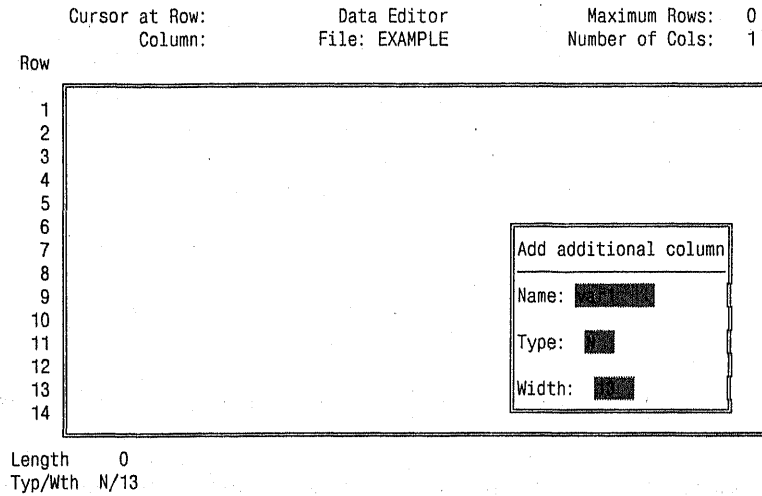


Рис. П2.7. Пакет STATGRAPHICS. Меню настройки структуры переменных в файле данных

На рис. П2.8 приведен вид экрана редактора базы данных пакета с полями для ввода двух действительных переменных с именами var1 и var2. Нажатие клавиши F5 выводит на экран меню преобразования, печати и сохранения переменных.

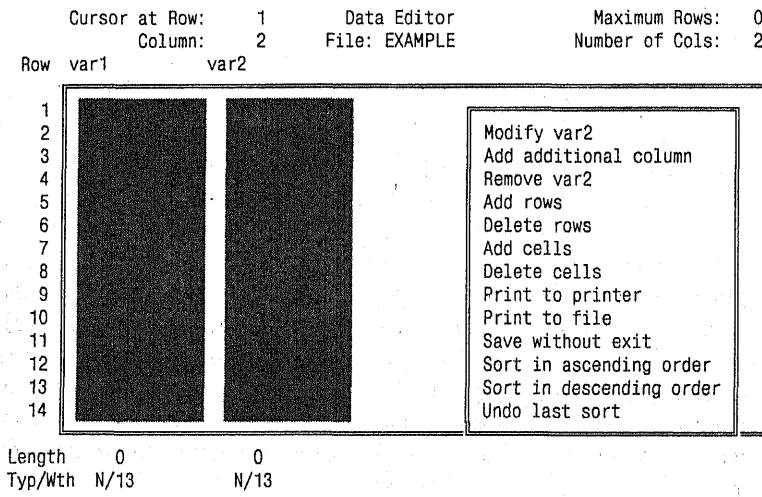


Рис. П2.8. Пакет STATGRAPHICS. Полноэкранный редактор базы данных с вызванным окном опций работы с переменными

Экспорт и импорт данных. Кроме непосредственного ввода данных в блок редактора, STADIA допускает импорт и экспорт файлов данных

формата DBF и ASCII (разделители — пробелы или запятые), что обеспечивает совместимость со многими другими программными средствами хранения и обработки данных. Кроме того пакет поддерживает обмен данными через буфер обмена с любой программой в среде Windows (Excel, Lotus 1-2-3 и т.д.).

STATGRAPHICS поддерживает обмен данными также и в форматах DIF, Lotus и ATLAS*GRAPHICS.

Контроль ввода данных. В пакете STADIA контроль за ошибочным вводом данных осуществляет специальная процедура, позволяющая выделить резко выделяющиеся и пропущенные данные. В STATGRAPHICS этот вопрос частично решается за счет настройки поля ввода электронной таблицы на конкретный формат данных, что особенно удобно для однотипных массивов.

Преобразования данных. Оба пакета содержат широкий набор средств преобразования данных, что весьма важно в практических задачах. Пользователь обнаружит в этих пакетах стандартные функциональные преобразования и преобразования, задаваемые пользователем, разрезание и склеивание переменных, сортировку и ранжирование, генерацию данных с заданными характеристиками и многое другое.

П2.7. Графические возможности

Трудно переоценить важность графических форм представления данных в статистике, где они являются составной частью многих методов анализа и обеспечивают наглядное представление данных и результатов. Пакеты STADIA и STATGRAPHICS не сильно отличаются по разнообразию двумерных и трехмерных типов графического представления данных. Они включают достаточно полный набор двумерных и трехмерных диаграмм рассеяния, функциональных графиков одной или нескольких переменных (с указанием стандартных ошибок и без него), столбиковых и круговых диаграмм, матричных графиков и др. Кроме того, пакеты осуществляют различные типы сглаживания и построение графиков функций распределения и их производных для обширного класса вероятностных семейств. Некоторые возможности пакетов иллюстрируются на рис. П2.9—П2.12.

Графические возможности в пакете STADIA 6.0. Среда Windows предоставляет удобные возможности для создания мощных графических редакторов в своих приложениях. Графический редактор пакета STADIA 6.0 сочетает простоту и удобство с богатыми возможностями.

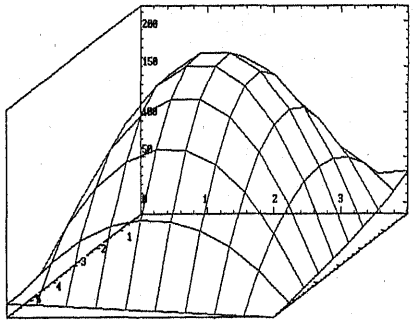


Рис. П2.9. Пакет STADIA.
Трехмерный график поверхности

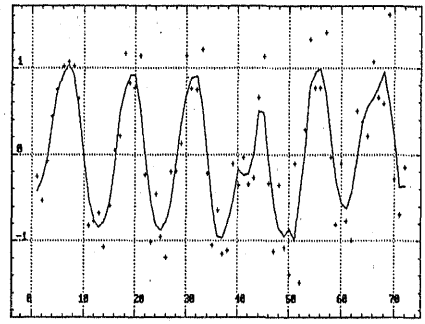


Рис. П2.10. Пакет STADIA.
Сглаживание скользящим средним

Для построения графика исходных данных, загруженных в электронную таблицу, необходимо открыть меню графиков на панели управления (см. рис. П2.3). Кроме того, большинство статистических процедур автоматически строит необходимые по ходу анализа графики. Так, например, процедура регрессионного анализа генерирует графики подобранной зависимости, остатков и прогноза.

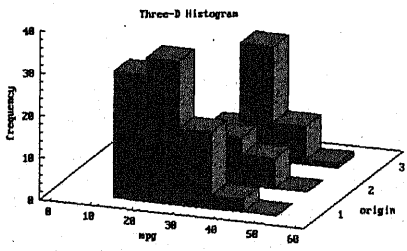


Рис. П2.11. Пакет STATGRAPHICS.
Трехмерная гистограмма частот

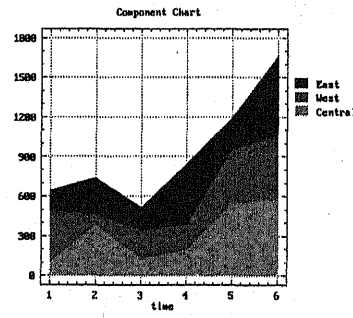


Рис. П2.12. Пакет STATGRAPHICS.
Кумулятивный график компонент ряда

Меню графиков данных (рис. П2.13) предоставляет доступ к различным типам графиков. Дальнейшее редактирование построенного графика позволяет изменить толщину линий, выбрать тип маркера точки и форму самого графика. На график могут быть добавлены названия осей, подрисуночная надпись, легенды. Пакет помещает график в специальное графическое окно и позволяет одновременно работать с восьмью подобными окнами. Любой график может быть сохранен в виде файла в формате BMP. С помощью буфера обмена график может быть также перенесен в любую другую Windows-программу.

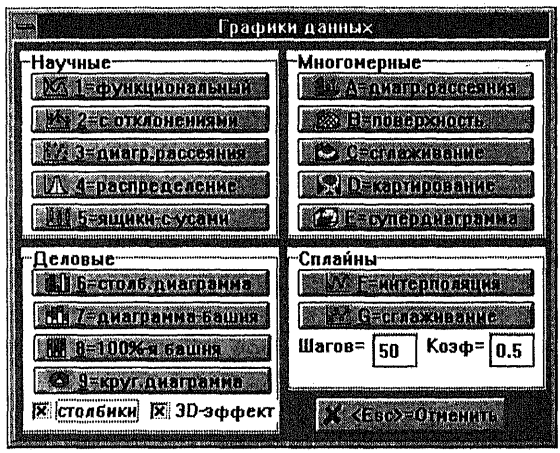


Рис. П2.13. Пакет STADIA. Меню выбора графика

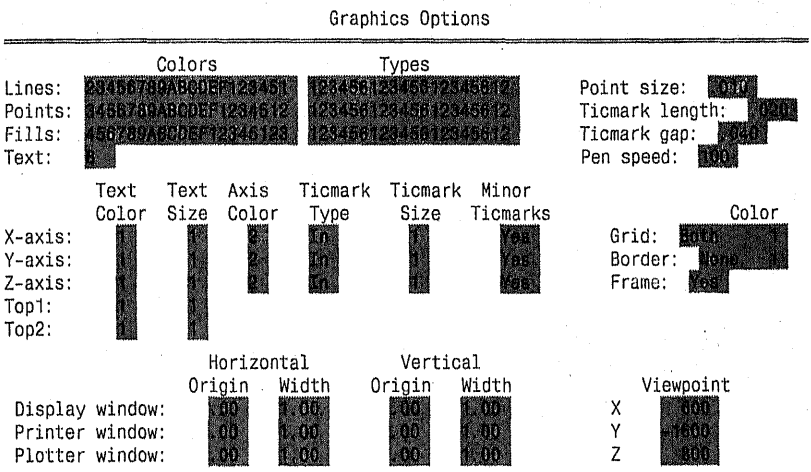


Рис. П2.14. Пакет STATGRAPHICS. Меню настройки графического вывода

Это далеко не полный список графических возможностей пакета. По сравнению с DOS-версией он значительно расширился за счет использования возможностей среды Windows.

Графические возможности пакета STATGRAPHICS. Графический редактор пакета STATGRAPHICS обладает богатыми возможностями настройки графического вывода. Однако их использование, особенно в интерактивном режиме, требует хорошего знания документации пакета. С помощью нескольких меню графический редактор позволяет проводить настройку всех возможных элементов графика, начиная от области определения, размеров, масштаба по каждой из осей, рамки,

заголовка, и кончая типами и цветами линий и точек на графике и детальным оформлением осей координат. На рис. П2.14 приведено одно из меню настройки графического вывода пакета. Значения многочисленных параметров этого меню хранятся в специальном файле и используются по умолчанию.

В режиме интерактивного редактирования графика пакет STATGRAPHICS позволяет расставить метки точек, изменяя при этом как их вид, так и положение, вывести значения координат (или номеров) точек и т.п. Имеются и возможности наложения слайдов ранее построенных графиков и разбиения листа на части для одновременного вывода до 9 различных графических слайдов на лист. В более поздних DOS-версиях пакета порядок графического редактирования существенно упростился.

П2.8. Подготовка отчетов

Важным этапом завершения статистического анализа является подготовка итогового отчета. Он обычно включает в себя, наряду с текстовым описанием задачи, результаты компьютерных расчетов и графические иллюстрации.

Результаты статистического анализа в пакете STADIA 6.0 помещаются в специальное текстовое окно результатов. Его содержимое может быть отредактировано и выведено на печать непосредственно в самом пакете или перенесено с помощью буфера обмена Windows в другой, более привычный тестовой редактор. Аналогично можно поступить с содержимым графических окон пакета.

Пакет STATGRAPHICS 3.0 позволяет осуществить полную подготовку итогового отчета на английском языке. При этом имеется возможность не только скопировать текущий экран в файл, но и предварительно его отредактировать. Процедура генерации отчета позволяет включать сформированные по ходу работы текстовые и графические экраны в итоговый документ. Однако она настолько уступает по простоте и удобству распространенным текстовым процессорам, что Вам вряд ли захочется ей пользоваться без крайней нужды.

П2.9. Документация

Хорошая документация значительно упрощает и убыстряет освоение любого пакета.

Пакет STADIA. Сравнительно небольшая (250 страниц формата А4) документация системы STADIA тщательно продумана и хорошо

оформлена. Разбор назначения каждой процедуры, типовой порядок диалога, демонстрационные примеры, наличие формул по большинству статистических процедур и обсуждения результатов исключают какие-либо неясности при работе с системой.

Пакет STATGRAPHICS. Документация STATGRAPHICS довольно обширна и подробна. Вряд ли Вам удастся далеко продвинуться в освоении пакета без ее внимательного изучения. Особенно это касается вопросов порядка ввода данных во многие процедуры, работы с текстовым и графическим редакторами пакета. Разбор демонстрационных примеров довольно поверхностен со статистической точки зрения и порой остается только гадать, какая величина конкретно вычисляется в тех случаях, когда наряду с точной формулой (таблицей распределения) может использоваться асимптотическое приближение или модификации с поправкой на группировку или непрерывность. В документации нет и каких-либо указаний или ссылок на типы вычисляемых оценок параметров для различных семейств распределения вероятностей. Отсутствие подобной информации делает многие вопросы при использовании пакета неясными даже для специалистов, и может потребоваться даже специальное тестирование, если Вы хотите быть уверены, что вычислили именно то, что хотели.

Начиная с пятой версии пакета, его документация была полностью переработана и большинство из указанных недостатков в ней было устранено.

П2.10. Справочник и экспертная поддержка

Мало кто из современных пользователей при затруднениях в работе с той или иной программой желает рыться в документации. Как правило, небольшие затруднения проще и быстрее разрешаются с помощью встроенного справочника, а к документации пользователи обращаются лишь тогда, когда во встроенном справочнике решение нужной проблемы найти не удалось. Поэтому удобство и информативность встроенного справочника — один из наиболее существенных элементов, обеспечивающих комфортную работу с программным продуктом.

Пакет STADIA. При использовании пакета STADIA практически можно обойтись вообще без чтения документации, так как значительная ее часть присутствует в гипертекстовом справочнике пакета, использующем систему перекрестных ссылок. Гибкий порядок обращения к справочнику позволяет легко получить разъяснение как по сервисным про-

цедурам, так и по статистическим методам. При нажатии клавиши F1 программа выдает фрагмент справочника, соответствующий текущему режиму работы пакета или выполняемой статистической процедуре.

Пакет STADIA содержит встроенную *экспертную систему* по выбору и использованию статистических методов. Для этого используется специальная классификация различных возможных типов статистических данных, и для каждого из этих типов данных указываются возможные статистические методы обработки, их назначение и порядок работы с ними. Это очень полезно для расширения статистического кругозора вообще, и особенно в тех случаях, когда надо использовать альтернативный метод обработки, потому что результаты проведенного анализа являются неудовлетворительными. Мы уже описывали эту экспертную систему в Приложении 1 (см. п. П1.8, рис. П1.3—П1.5).

Пакет STATGRAPHICS. Встроенный справочник пакета STATGRAPHICS 3.0 довольно скуден, контекстно-неориентирован и содержит, в основном, сведения о назначении функциональных клавиш и о порядке ввода данных в статистические процедуры. Эффективность его поэтому крайне низка, и для использования разнообразных возможностей пакета необходимо детально изучить объемную документацию. Современные версии пакета STATGRAPHICS кардинально отличаются от своих ранних аналогов в вопросе оказания помощи при подборе статистической процедуры и толкования ее результатов. Например, в версии пакета STATGRAPHICS в среде Windows предусмотрен уникальный режим StatAdvisor, который по завершении работы статистической процедуры генерирует текстовой отчет, разъясняющий полученные результаты.

П2.11. Технические характеристики

Пакет STADIA 6.0 работает на любом IBM PC-совместимом компьютере с процессором 386 и выше, имеющем не менее 4 Мбайт дополнительной памяти и русифицированную среду Windows версии 3.1 и выше. Все настройки на внешние устройства обеспечиваются непосредственно средой Windows. В отличие от DOS-версии пакета, позволяющей работать непосредственно с гибкого диска, работа с STADIA 6.0 может осуществляться только после инсталляции его на жесткий диск.

Пакет STAGRAPHICS 3.0 может работать практически на любом IBM PC совместимом компьютере и не требует дополнительной памяти. Однако следует иметь ввиду, что прилагаемые к пакету драйверы внешних устройств уже не соответствуют современным принтерам, монито-

рам и другим внешним устройствам. Поэтому у пользователей могут возникнуть проблемы с настройкой пакета.

По современным понятиям, оба пакета очень компактны и занимают не более 3 Мбайт на жестком диске.

П2.12. Ошибки

Количество и разнообразие статистических методов, заложенных в статистические пакеты общего назначения, настолько велики, что в них, как и в любой большой программе, не обходится без ошибок. Нетривиальность многих статистических методов приводит к тому, что методологические и вычислительные ошибки во многих статистических пакетах (особенно сделанных неспециалистами) — не редкость. Ошибки встречаются даже в хороших пакетах, таких как STADIA и STATGRAPHICS.

Например, во время тестирования пакета STATGRAPHICS 3.0 авторами были обнаружены грубые ошибки в процедурах проверки согласия выборочного и теоретического распределения (Distribution Fitting). В них, в частности, не делается различия при вычислении процентных точек для статистик Колмогорова и хи-квадрат для простой и сложной гипотез. Для статистики Колмогорова, например, процентные точки выдаются всегда только для случая простой гипотезы, что приводит к заметному завышению согласия, если нулевая гипотеза сложная. Аналогичная ситуация наблюдается и с критерием согласия хи-квадрат. Процентные точки этой статистики неверны и в случае простой гипотезы, и случае сложной (последнее связано с использованием в статистике критерия неправомерных оценок параметров теоретического распределения). Подробно эти вопросы рассмотрены в главе 10.

Не избежали неточностей и ошибок ранние версии пакета STADIA. Например, в них неправильно считались границы доверительных интервалов в процедурах регрессионного анализа, происходили сбои при построении гистограммы и т.д. Первые варианты пакета STADIA 6.0 также содержали ряд ошибок в организации и работе Windows-интерфейса. Условиями поставки пакета предусмотрена возможность обратиться со всеми замечаниями и претензиями к разработчикам пакета.

Приложение 3

Где приобрести статистические пакеты

В настоящем приложении мы расскажем о том, где можно приобрести описанные в этой книге средства для анализа данных на компьютерах, получить консультации по их использованию, а также по постановкам задач и проведению анализа данных (сведения предоставлены соответствующими фирмами).

П3.1. Универсальные статистические пакеты

Пакет STADIA содержит широкий набор методов анализа данных из всех областей статистики и доступен широкому кругу прикладных специалистов, менеджеров и студентов. Сейчас распространяются версия 5.0 пакета для среды DOS и версия 6.0 для среды Windows. Статистические возможности этих версий эквивалентны, однако в Windows-версии заметно богаче графические возможности пакета. Пакет может поставляться в трех вариантах: study, base и prof, различающихся лишь объемами обрабатываемых массивов и ценой (см. п. П3.3 ниже). Самый дешевый вариант study имеет максимальный объем матрицы данных в 256 чисел в DOS-версии и 400 чисел в Windows-версии, он предназначен главным образом для учебных заведений и задач с небольшими объемами данных. Самые дорогие версии STADIA 5.0 Prof. и STADIA 6.0/w Prof. имеют увеличенный максимальный объем матрицы данных до 12800 и 20000 чисел и расширенные возможности статистических процедур для их обработки по сравнению с базовыми версиями.

Пакет распространяется НПО «Информатика и компьютеры» (Москва).

Пакет STATGRAPHICS — универсальный, многопрофильный пакет с хорошо методически продуманным меню-ориентированным интерфейсом пользователя. Ранние версии пакета, по-видимому, являются самыми распространенными в России из западных статистических пакетов. Краткая информация о новых версиях пакета, начиная с четвертой, опубликована в [34]. В настоящее время распространяются три модификации седьмой версии этого пакета. Первые две из них функционируют

в среде DOS. STATGRAPHICS v.7.0 работает на всех типах процессоров, начиная с 8086. STATGRAPHICS Plus v.7.0 требует не ниже 386 процессора. В этих версиях, в отличие от более ранних, предусмотрена поддержка русского алфавита. Версия STATGRAPHICS Plus for Windows функционирует в среде Windows и поставляется в виде базового модуля Base System и дополнительных специализированных модулей контроля качества, планирования эксперимента, анализа временных рядов и многомерных методов анализа. Для версий STATGRAPHICS Plus имеются также сетевые аналоги.

По сравнению с разобранной в книге третьей версией пакета следует прежде всего отметить значительное улучшение качества документации пакета, работы графических процедур и значительно возросшие объемы обрабатываемых массивов.

В России пакет распространяется Санкт-Петербургским СП «ИнфоСтрой», являющимся официальным дилером фирмы Manugistics Inc. (ранее STSC Inc.), которая владеет всеми правами на распространение пакета STATGRAPHICS.

Пакет SPSS — универсальный статистический пакет фирмы SPSS Inc., одного из крупнейших производителей и распространителей статистического программного обеспечения в мире. Версии системы SPSS написаны для самых популярных платформ — MS DOS, Windows, OS/2, Macintosh, UNIX, IBM/370 MVS и др. Все они совместимы между собой по принципу работы, командному языку и форматам файлов. Версия SPSS для Windows продолжает сохранять позиции лидирующего статистического пакета в мире. Сейчас распространяется версия SPSS 7.0 for Windows 95, которая предлагает удобные возможности управления данными, широкий спектр статистических функций, интегрированных графиков и отчетов. Исследовательские задачи выполняются в едином режиме, начиная с ввода исходных данных и кончая получением отчета о результатах анализа.

Пакет также предлагает:

- справочную систему, ориентированную на конкретные задачи, а также глоссарий по научным терминам;
- набор средств для быстрого доступа к ресурсам программы;
- возможность обработки неограниченного количества переменных и наблюдений;
- усовершенствованный доступ к данным и новый 32-разрядный код, обеспечивающий более быструю обработку;
- обмен с другими приложениями посредством DDE, OLE, ODBC.

SPSS является модульной программой. Базовая система предоставляет пользователям возможности для преобразования данных, функции работы с файлами, описательную статистику, дисперсионный анализ (ANOVA), корреляцию, линейную регрессию, средства построения графиков и подготовки отчетов. Дополнительные модули пакета включают: анализ и конструирование таблиц (Tables), анализ временных рядов (Trends), анализ категорий (Categories), методы углубленного и расширенного статистического анализа (Prof. Stats и Adv. Stats) и др.

Документация к системе SPSS признана лучшей документацией для систем подобного типа и может использоваться в качестве подробного и доступного учебника по прикладной статистике. Часть документации пакета переведена на русский язык.

Пакет распространяет эксклюзивный авторизованный дистрибьютор SPSS Inc. в России «Статистические системы и Сервис» (Москва).

Пакет SYSTAT — универсальный статистический пакет фирмы SPSS, Inc. (США), которая поглотила бывшего разработчика этого пакета SYSTAT, Inc. Пакет содержит богатый набор статистических методов и отличается прекрасными графическими возможностями. Имеются версии пакета для DOS и Windows. Пакет распространяется Центром СТАТ-ДИАЛОГ (Москва) (поставки по предварительным заказам).

Пакет STATISTICA/w — универсальный статистический пакет фирмы StatSoft, Inc. (США). Пакет был создан в начале 90-х годов сразу для среды Windows, опередив Windows-версии других статистических пакетов. В пакете нашли отражение многие последние достижения теоретической и прикладной статистики. Однако у специалистов пока не сформировалось единого мнения по поводу этого пакета, поскольку многие принятые в пакете подходы имеют свои преимущества и недостатки. Часть документации пакета переведена на русский язык.

Пакет распространяется фирмой «СофтЛайн» (Москва).

П3.2. Специализированные пакеты

Пакет Эвриста предназначен для анализа временных рядов и включает широкий набор возможностей анализа регрессионных зависимостей, трендов, фазового пространства, гармонического, спектрального, факторного и кросскорреляционного анализа, построения моделей интервенций, сезонных и несезонных ARMA-моделей, передаточных функций, а также многочисленные методы прогнозирования временных рядов. Этот пакет особенно популярен как мощное средство анализа и прогноза финансовых рынков.

Документация к пакету одновременно является популярным учебником по курсу прикладного анализа временных рядов, читаемого на ряде факультетов МГУ им. М.В.Ломоносова. Пакет распространяется Центром Статистических Исследований (Москва).

Пакет Мезозавр предназначен для анализа временных рядов и содержит методы сглаживания, выделения сезонных колебаний, спектрального анализа, частотной фильтрации, а также линейные и нелинейные модели тренда, авторегрессионные модели, модели Бокса-Дженкинса и т.д. Распространяется Центром СТАТ-ДИАЛОГ (Москва).

Пакет КЛАСС-МАСТЕР предназначен для кластерного анализа количественных, качественных и логических (вида «да–нет») данных. Пакет может строить логическое описание найденного разбиения на классы. Распространяется Центром СТАТ-ДИАЛОГ (Москва).

Пакет САНИ предназначен для анализа и визуализации разнотипных данных, в том числе данных нечисловой природы. Позволяет представлять данные в удобном для восприятия виде, строить группировки, выявлять аномальные наблюдения, проверять гипотезы о независимости и т.д. Распространяется Центром СТАТ-ДИАЛОГ (Москва).

Пакет СТАТИСТИК-КОНСУЛЬТАНТ предназначен для анализа широкого класса регрессионных и факторных задач и включает в себя ряд уникальных и довольно эффективных процедур анализа временных рядов. Первая версия этого пакета поддерживала очень ограниченный набор методов статистического анализа. Этот недостаток был учтен при создании второй версии этого пакета, которая находится сейчас в распространении. Следует отметить хороший встроенный гипертекстовой справочник пакета по статистическим методам анализа и наличие экспертной поддержки при подборе модели описывающей данные.

Пакет разработан фирмой «Тандем» совместно с Карельским отделением Академии наук и распространяется фирмой «Тандем» и ее дилерами.

Пакет ОЛИМП:СтатЭксперт предназначен для анализа и прогнозирования финансовых, экономических, инженерных и научных данных. Базовая версия программы ориентирована на анализ экономических временных рядов. В профессиональной версии представлены также методы факторного, кластерного, частотного и структурного анализа и анализа нечисловой информации. Программа работает под Windows и использует в качестве интерфейса пользователя Excel. Пакет разработан и распространяется департаментом информационных технологий ТОО «Росэкспертиза».

Пакет Forecast Expert предназначен для анализа временных рядов, построения сезонных и несезонных ARIMA-моделей и осуществления прогноза на их базе. Пакет также позволяет учитывать зависимость рассматриваемого временного ряда от другого ряда и строить прогноз с учетом этой зависимости. Главной особенностью этой узкоспециализированной программы является автоматический подбор и тестирование модели, не требующий от пользователя знаний в области математической статистики. Пакет функционирует в среде Windows 3.1 и выше.

Пакет разработан фирмой «Про-Инвест Консалтинг» (Москва) и является продолжением технологической цепочки программ (Project Expert, Project Questionnaire & Risk, Invest Expert) для экономики и финансов. Распространяется фирмой-разработчиком, а также ее представителями более чем в 40 городах России и СНГ.

П3.3. Цены и телефоны фирм

В приведенной таблице указаны цены на описанные выше статистические пакеты и телефоны фирм-распространителей.

Таблица П3.1

НАЗВАНИЕ	Назначение пакета	Цена	Тел./факс
SPSS для Windows	универсальный	4290	(095)125-09-68
STADIA 5.0 study (DOS)	универсальный	50/40	(095)939-53-06
STADIA 5.0 base (DOS)	универсальный	100/80	(095)939-53-06
STADIA 5.0 prof (DOS)	универсальный	150/120	(095)939-53-06
STADIA 6.0 study (Win)	универсальный	200	(095)939-53-06
STADIA 6.0 base (Win)	универсальный	500/400	(095)939-53-06
STADIA 6.0 prof (Win)	универсальный	800/640	(095)939-53-06
STATGRAPHICS 7.0 (DOS)	универсальный	998	(812)312-26-73
STATGRAPHICS Plus 7.0	универсальный	1695	(812)312-26-73
STATGRAPHICS Plus/Win	универсальный	1699	(812)312-26-73
STATISTICA 5.1/w	универсальный	1756/1580	(095)916-88-38
SYSTAT 5.0	универсальный	около 1100	(095)125-21-31
Forecast Expert	временные ряды	550/300	(095)216-64-26
Мезозавр (DOS/Win)	временные ряды	520	(095)125-21-31
Олимп:СтатЭксперт Std	временные ряды	485/340	(095)955-01-03
Олимп:СтатЭксперт Prof	временные ряды	895/627	(095)955-01-03
Эвриста 2.21 (DOS)	временные ряды	520/260	(095)939-17-96
Эвриста 3.0 (Win)	временные ряды	700/350	(095)939-17-96
Знак (SIGN)	регр., факт. анализ	120	(095)939-53-06
Класс-Мастер (DOS/Win)	многомерный анализ	320	(095)125-21-31
САНИ (DOS/Win)	нечисловые данные	320	(095)125-21-31
Статистик-консультант	регр., факт. анализ	395	(814)6-79-92, (095)939-53-06

Замечания. 1. Цены приведены в долларах США, оплата в рублях по курсу.

2. Если в графе «Цена» указаны две суммы (например, 200/100), то первая означает обычную цену, а вторая — цену со скидкой (для учебных заведений и учреждений РАН).

3. Многие фирмы предоставляют специальные скидки для учебных заведений, хотя и не объявляют это в своих прейскурантах. О наличии и величине таких скидок можно узнать, обратившись в фирму или ее представительство.

4. У пакета SPSS можно приобрести отдельные модули: Base (980 дол.), Tables (550 дол.), Prof. Stats. (550 дол.), Adv. Stats. (550 дол.), Categories (550 дол.), Trends (550 дол.), Exact Test (740 дол.), CHAID (740 дол.), LISREL 7 (740 дол.).

5. У пакета STATGRAPHICS Plus для Windows указана цена за основной набор статистических модулей (можно приобрести также дополнительные модули контроля качества, планирования эксперимента, анализа временных рядов и многомерных методов анализа).

П3.4. Консультации и обучение

Консультации по вопросам использования пакетов, указанных в табл. П3.1, можно получить у фирм-распространителей этих пакетов (телефоны фирм приведены выше).

Кроме того, отдел статистических исследований НПО «Информатика и компьютеры» производит консультации по постановкам задач, планированию исследований, подбору методов анализа и программного обеспечения для анализа данных в бизнесе, банковском деле, маркетинге, торговле и других областях, а также осуществляет индивидуальное и групповое обучение основам прикладной статистики и анализу данных на компьютере. Возможно проведение курса на базе заказчика и/или формирование программы курса под задачи заказчика. Выполняются расчеты на компьютере по данным заказчика. Адрес: 119899, Москва, Мичуринский просп., 1, к. 119. Тел. (095)939-53-06, факс (095)939-52-95.

Аналогичные услуги в С.-Петербурге оказывает Научно-консультационный центр «Тренд» Санкт-Петербургского государственного университета. Адрес: 198904, Санкт-Петербург, Старый Петергоф, Библиотечная пл. 2, НИИ математики и механики, НКЦ «Тренд». Телефон: (812) 428-4282, Факс: (812) 428-7039, E-mail: com@ trend.niimm.spb.su или trend@ stat2.math.lgu.spb.su

Литература

1. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Классификация и снижение размерности. Справочное издание под ред. Айвазяна С.А. — М.: Финансы и статистика, 1989. — 607 с.

2. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное издание под ред. Айвазяна С.А. — М.: Финансы и статистика, 1983. — 471 с.

3. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Исследование зависимостей. Справочное издание под ред. Айвазяна С.А. — М.: Финансы и статистика, 1985. — 471 с.

4. *Андерсен Т.* Введение в многомерный статистический анализ. — М.: Физматгиз, 1963. — 500 с.

5. *Андерсен Т.* Статистический анализ временных рядов. — М.: Мир, 1976. — 756 с.

6. *Аптон Г.* Анализ таблиц сопряженности. — М.: Финансы и статистика, 1982. — 144 с.

7. *Аренс Х., Лейтер Ю.* Многомерный дисперсионный анализ — М.: Финансы и статистика, 1985. — 230 с.

8. *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ. — М.: Мир, 1982. — 488 с.

9. *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.

10. *Бард Й.* Нелинейное оценивание параметров. — М.: Финансы и статистика, 1979. — 349 с.

11. *Бендат Дж., Пирсол А.* Прикладной анализ случайных данных. — М.: Мир, 1989. — 540 с.

12. *Бендат Дж., Пирсол А.* Применение корреляционного и спектрального анализа. — М.: Мир, 1979. — 311 с.

13. *Бернулли Я.* О законе больших чисел. Под общей ред. Ю.В.Прохорова. — М.: Наука, 1986. — 176 с.

14. *Бикел П., Доксум К.* Математическая статистика. — М.: Финансы и статистика, 1983. Вып. 1 — 280 с.; Вып. 2 — 254 с.

15. *Бокс Дж., Дженкинс Г.* Анализ временных рядов. Прогноз и управление. М.: Мир, 1974. Вып. 1 — 288 с.; Вып. 2 — 197 с.

16. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. — М.: Наука, 1983. — 416 с.

17. *Боровков А.А.* Теория вероятностей, 2-е изд., доп. — М.: Наука, 1986. — 431 с.

18. *Бриллинджер Д.* Временные ряды. — М.: Мир, 1980. — 536 с.

19. *Векслер Л.С.* Статистический анализ на персональном компьютере//МИР ПК, № 2, 1992, с. 89–97.

20. *Вучков И., Бояджиева Л., Солаков Е.* Прикладной линейный регрессионный анализ. — М.: Финансы и статистика, 1987. — 239 с.

21. *Гаек Я., Шидак З.* Теория ранговых критериев. — М.: Наука, 1971. — 376 с.

22. *Гитис Э.И., Пискулов Е.А.* Аналого-цифровые преобразователи. — М.: Энергоиздат, 1981. — 360 с.

23. *Гнеденко Б.В.* Курс теории вероятностей. — М.: Физматгиз, 1988. — 406 с.

24. *Гоноровский И.С.* Радиотехнические цепи и сигналы. — М.: Сов. Радио, 1977. — 608 с.

25. ГОСТ 23554.2–81. Система управления качеством продукции. Экспертные методы оценки качества промышленной продукции. Обработка значений экспертных оценок качества продукции. — М.: Изд. Стандартов 1982. — 66 с.

26. *Готтсданкер Р.* Основы психологического эксперимента. — М.: МГУ, 1982. — 463 с.

27. *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). С.-Петербург, 1992. — 104 с.

28. *Демиденко Е.З.* Линейная и нелинейная регрессия. — М.: Финансы и статистика, 1981. — 302 с.

29. *Дженкинс Г., Ваттс Д.* Спектральный анализ и его приложения. — М.: Мир. Вып. 1, 1971. — 316 с.; Вып. 2, 1972. — 288 с.

30. *Джонсон Н., Лион Ф.* Статистика и планирование эксперимента в технике и науке. — М.: Мир. Т. 1, 1980, — 610 с., Т. 2, 1981, — 520 с.

31. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ: В 2-х книгах, Кн. 1. — М.: Финансы и статистика, 1986. — 366 с., Кн. 2. — М.: Финансы и статистика, 1987. — 351 с.

32. *Дэниел К.* Применение статистики в промышленном эксперименте. — М.: Мир 1979. — 299 с.

33. *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», С.-Петербург, 1996 — с. 193 – 196

34. *Дюк В.А., Мирошников А.И.* Эволюция STATGRAPHICS//МИР ПК, № 12, 1995.

35. *Енюков И.С.* Методы, алгоритмы, программы многомерного статистического анализа. — М.: Финансы и статистика, 1986.

36. *Кендэлл М., Стьюарт А.* Статистические выводы и связи. — М.: Наука, 1973. — 899 с.

37. *Кендэлл М., Стьюарт А.* Теория распределений. — М.: Наука, 1966.

38. *Кендэлл М., Стюарт А.* Многомерный статистический анализ и временные ряды. — М.: Наука, 1976. — 736 с.

39. *Кендэлл М.* Временные ряды. — М.: Финансы и статистика, 1981. — 199 с.

40. *Кендэлл М.* Ранговые корреляции. — М.: Статистика, 1975. — 212 с.

41. *Кокрен У.* Методы выборочного исследования. — М.: Статистика, 1976. — 440 с.

42. *Кокс Д.Р., Оукс Д.* Анализ данных типа времени жизни. — М.: Финансы и статистика, 1988. — 192 с.

43. *Колмогоров А.Н.* Об одном новом подтверждении законов Менделя//ДАН СССР, 1940, т. 27, № 1, стр. 38–42.

44. *Крамер Г.* Математические методы статистики. — М.: Мир, 1975. — 648 с.

45. *Крылов В.Ю.* Геометрическое представление данных в психологических исследованиях. — М.: Наука, 1990. — 117 с.

46. *Кулаичев А.П.* Пакеты для анализа данных//МИР ПК, № 1, 1995.

47. *Левин Б.Р.* Теоретические основы статистической радиотехники. В трех кн. — М.: Сов. Радио, 1975.

48. *Леман Э.* Проверка статистических гипотез. — М.: Наука, 1964. — 498 с.

49. *Леман Э.* Теория точечного оценивания. — М.: Наука, 1991. — 448 с.

50. *Ликеш И., Ляга И.* Основные таблицы математической статистики. — М.: Финансы и статистика, 1985. — 356 с.

51. *Литтл Р.Дж., Рубин Д.Б.* Статистический анализ данных с пропусками. — М.: Финансы и статистика, 1991. — 336 с.

52. *Лукашин Ю.П.* Адаптивные методы краткосрочного прогнозирования. — М.: Статистика, 1979. — 254 с.

53. *Макаров А.А.* STADIA против STATGRAPHICS, или кто ваш «лоцман» в море статистических данных//МИР ПК, № 3, 1992, с. 58—66.

54. *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.

55. *Макаров А.А.* Статистические пакеты в обучении математической и прикладной статистике//Тезисы доклада на международной конференции «Статистическое образование в современном мире: идеи, ориентации, технологии». С.-Петербург, 1996. — с. 193—196.

56. *Макино Т., Охаси М., Доке Х., Макино К.* Контроль качества с помощью персональных компьютеров. — М.: Машиностроение, 1991.

57. *Мардиа К., Земроч П.* Таблицы F-распределений. — М.: Наука, 1984. — 255 с.

58. *Марпл-мл. С.Л.* Цифровой спектральный анализ и его приложения. — М.: Мир, 1990. — 584 с.

59. *Мэйндоналд Дж.* Вычислительные алгоритмы в прикладной статистике. — М.: Финансы и статистика, 1988. — 350 с.

60. *Мюллер П., Нойман П., Шторм Р.* Таблицы по математической статистике. — М.: Финансы и статистика, 1982.— 278 с.

61. *Оуэн Д.Б.* Сборник статистических таблиц. Изд. 2-е, испр. — М.: ВЦ АН СССР, 1973. — 586 с.

62. *Поллард Дж.* Справочник по вычислительным методам статистики. — М.: Финансы и статистика, 1982. — 344 с.

63. *Рао С.Р.* Линейные статистические методы и их применение. — М.: Наука, 1968. — 548 с.

64. *Рунион Р.* Справочник по непараметрической статистике. Современный подход. — М.: Финансы и статистика, 1982. — 198 с.

65. *Семенов Н.А.* Программы регрессионного анализа и прогнозирования временных рядов. Пакеты ПАРИС и МАВР. — М.: Финансы и статистика, 1990. — 111 с.

66. *Смирнов Н.В., Дунин-Барковский И.В.* Курс теории вероятностей и математической статистики для технических приложений. Изд. 2-е, испр. и доп. — М.: Наука, 1965. — 511 с.

67. *Смоляк С.А., Титаренко Б.П.* Устойчивые методы оценивания. — М.: Статистика, 1980. — 206 с.

68. Справочник по прикладной статистике. В 2-х т., под ред. Э.Ллойда, У.Ледермана, Ю.Н.Тюрина — М.: Финансы и статистика, 1989, 1990.

69. Справочник по специальным функциям с формулами, графиками и таблицами/Под ред. М.А.Абрамовица, И.Стиган. — М.: Наука, 1979. — 830 с.

70. Статистические методы для ЭВМ/Под ред. К.Эйнслейна, Э.Рэлстона, Г.С.Уолфа — М.: Наука, 1986. — 459 с.

71. Статистические методы повышения качества/Под ред. Хитоси Куме. — М.: Финансы и статистика, 1991.

72. *Стреляу Я.* Роль темперамента в психическом развитии. — М.: Прогресс, 1982. — 231 с.

73. *Теннант-Смит Дж.* Бейсик для статистиков. — М.: Мир, 1988. — 207 с.

74. *Тутубалин В.Н.* Границы применимости (вероятностно-статистические методы и их возможности). — М.: Знание, 1977. — 61 с.

75. *Тутубалин В.Н.* Теория вероятностей и случайных процессов — М.: Изд-во МГУ, 1992. — 400 с.

76. *Тьюки Дж.* Анализ результатов наблюдений. Разведочный анализ. — М.: Мир, 1981. — 693 с.

77. *Тюрин Ю.Н., Симонова Г.И.* Знаковый анализ линейных моделей //Обозрение прикладной и промышленной математики, т. 1, вып. 2, 1994. — с. 214—278.

78. *Урбах В.Ю.* Математическая статистика для биологов и медиков. — М.: Изд-во АН СССР, 1963. — 323 с.

79. Факторный, дискриминантный и кластерный анализ. — М.: Финансы и статистика, 1989. — 215 с.

80. *Феллер В.* Введение в теорию вероятностей и ее приложения. — М.: Мир. Т. 1, 1964, — 498 с., Т. 2, 1967, — 752 с.

81. *Фигурнов В.Э.* IBM PC для пользователя. Краткий курс. — М.: Инфра-М, 1997. — 480 с.

82. *Флейс Дж.* Статистические методы для изучения таблиц долей и пропорций. — М.: Финансы и статистика, 1989. — 319 с.

83. *Хальд А.* Математическая статистика с техническими приложениями. — М.: Изд-во Иностранной литературы, 1956. — 664 с.

84. *Хампель Ф., Рончетти Э., Рауссей П., Штаэль В.* Робастность в статистике. Подход на основе функций влияния. — М.: Мир, 1989. — 512 с.

85. *Хан Г., Шапиро С.* Статистические модели в инженерных задачах. — М.: Статистика, 1980. — 444 с.

86. *Хартман Г.* Современный факторный анализ. — М.: Статистика, 1972.

87. *Хастингс Н., Пикок Дж.* Справочник по статистическим распределениям. М., Статистика, 1980. — 95 с.

88. *Хенинен А.Я., Павлов Ю.Л.* Статистик-Консультант, или Еще один довод в пользу неизбежного//МИР ПК, № 6, 1994.

89. *Хеттсманпергер Т.* Статистические выводы, основанные на рангах. — М.: Финансы и статистика, 1987. — 334 с.

90. *Хикс Ч.* Основные принципы планирования эксперимента. — М.: Мир, 1967. — 406 с.

91. *Холлендер М., Вулф Д.* Непараметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.

92. *Хьюбер П.* Робастность в статистике. — М.: Мир, 1984. — 304 с.

93. *Шиндовский Э., Шюрц О.* Статистические методы контроля производства. М. Госкомстандарт, 1969. — 542 с.

94. *Ширяев А.Н.* Вероятность. — М.: Наука, 1980. — 574 с.

95. *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.

96. *Яглом А.М.* Корреляционная теория стационарных случайных функций (с примерами из метеорологии). — Гидрометеоиздат, 1981. — 280 с.

97. *Box G.E.P., Cox D.R.* An analysis of transformations, J. Roy. Stat. Soc., 1964, B 26, p. 211–243

98. *Chatfield C.* The Analysis of Time Series: an Introduction, 4th ed. — Chapmam and Hall, 1989. — 242 p.

99. *Elliott A.C., Gray Y.L.* Directory of Statistical Microcomputer Software. — N.Y.: Basel, 1986.

100. *Everit B.* A Handbook of Statistical Analyses using S-PLUS. Chapman & Hall, 1994. — 143 p.

101. *Granger C.W.J., Newbold P.* Forecasting Economic Time Series, 2nd ed. — Academic Press, Inc., 1986. — 338 p.

102. *Hartley H.O.* Testing of homogencity of a set of variances. — Biometrika, 1940, 31, pp. 249–255.

103. *Mosteller F., Tukey J.W.* Data Analysis and Regression: A Second Course in Statistics. Reading, MA: Addison-Wesley, 1977.

104. *Neter J., Wasserman W., Whitmore G.A.* Applied Statistics, Allyn and Bacon, Inc., 1988. — 1006 p.

105. *Sen P.K.* Nonparametric simultaneous inference for some MANOVA models/Handbook of Statistics. — v.1. Holland, 1980.

106. Software Digest Rating Report. 1991, v. 8, № 5.

107. *Spector P.* An introduction to S and S-PLUS. Duxbury Press, 1994. — 286 p.

108. *Venables M.N., Ripley B.D.* Modern Applied Statistics with S-PLUS. Springer-Verlag, 1994 — 462 p.

109. *Woodwant W.A, Elliott A.C., Gray Y.L., Mattlock D.C.* Directory of Statistical Microcomputer Software. — N.Y.: Basel, 1988.

Краткий путеводитель по списку литературы

Для удобства читателей мы помещаем краткие пояснения к списку литературы.

Справочники по прикладной статистике — [68], [62].

Стандартные учебники теории вероятностей и статистики — (строгий, аксиоматический подход) — [23], [17], [94], [44].
Учебники и пособия по статистике, рассчитанные на прикладных специалистов — [14], [66], [30], [68], [85], [8], [2], [3], [1].
Вероятностные распределения — [16], [37], [50], [68], [87].
Таблицы распределений — [16], [50], [60], [61], [69].
Случайный выбор — [41].
Непараметрические методы статистики — [91], [64], [89], [21].
Компьютерные алгоритмы статистики — [59], [8], [73], [70].
Разведочный анализ данных — [76].
Анализ данных с пропусками — [51].
Регрессионный анализ — [31], [28], [20], [92], [103].
Дисперсионный анализ — [31], [32], [30], [90], [1].
Планирование эксперимента — [30], [90], [32].
Таблицы сопряженности — [6], [82], [36].
Меры связи признаков — [40], [64], [78].
Анализ временных рядов — [12], [15], [18], [29], [39], [68], [38], [96], [9], [101], [98], [5].
Многомерные методы — [4], [68], [79], [35], [38], [1].
Факторный анализ — [8], [79], [86], [38], [1].
Дискриминантный анализ — [8], [79], [1].
Кластерный анализ — [35], [68], [79], [1].
Многомерное шкалирование — [45], [68], [70].
Методы контроля качества — [56], [71].

Замечания. Из книг, приведенных в списке литературы, выделим двухтомный «Справочник по прикладной статистике» [68], 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.

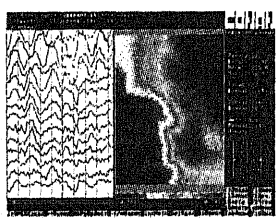
Кроме того, мы хотели бы отметить трехтомник «Прикладная статистика» С.А.Айвазяна и его соавторов В.М.Бухштабера, И.С.Енюкова, Л.Д.Мешалкина [2], [3], [1]. Это справочное издание вобрало в себя многолетний опыт работы как его авторов, так и всей школы прикладной статистики в СССР. Издание отражает широкий 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.

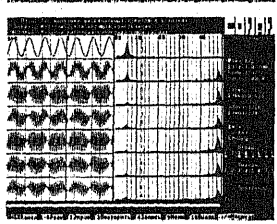
Из книг, носящих обзорный, справочный характер в конкретных областях прикладной статистики, обратим внимание на следующие: в области вероятностных распределений — [16], [87]; в области регрессионного и дисперсионного анализа — [31]; в области непараметрических методов — [91]. Те, кого интересуют вопросы разработки компьютерных алгоритмов статистики, могут найти полезную информацию в [59], [73]. Более строгое аксиоматическое изложение основ теории вероятностей и статистики содержится в [23], [17], [94], [44], [14].

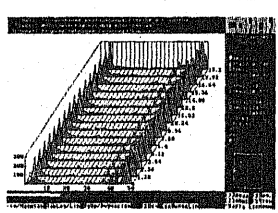
КОМПЛЕКСНЫЙ АНАЛИЗАТОР СИГНАЛОВ

Выполнение технических и физических измерений: акустика: вибродиагностика связь, радиотехника, аэро/гидродинамика, механика конструкций и материалов и пр.

на любом компьютере!







Возможности CONANt:

- спектральные характеристики: АЧХ, ФЧХ, ПФ, корреляции, огибающая, сигнал/шум, кепстр, октавные спектры, импульсная х-ка и др.
- усреднение сигналов и ЧХ, сглаживание, фильтрация, установки окон и эпох
- статистический анализ, построение зависимостей, любые преобразования и вычисления
- прецизионные средства просмотра, маштабирования и анализа записей
- мониторинг и анализ в реальном времени
- гибкое управление внешней аппаратурой
- поканальная калибровка и тарировка
- 1-32 независимых канала от 100КГц до 100Мгц
- экспорт/импорт DBF- и текстовых файлов
- 2/3-мерная цветная графика
- экранный справочник и методическая книга

Цена от $250 до $500
(АЦП: $250–900)

CONAN работает в ведущих научных и технических центрах страны!

InCo
НПО «Информатика и компьютеры»
117602, Москва, а/я 365
☎ (095) 437-36-95

КОМПЛЕКСНАЯ электрофизиологическая лаборатория на одном компьютере?

ДА!
Это теперь наконец-то возможно!!!

Цена от $700 до $2200
(аппаратура $2200-$5000) это в 3-5 раз доступнее аналогов!

CONAN — самое выгодное решение:

- ✓ **полностью заменяет** энцефалограф, ВП-анализатор, кардиограф, реограф, миограф и полиграф, **но значительно превосходит** их по своим возможностям

Возможности CONAN:

- ✓ полный комплект методов регистрации и анализа ЭЭГ и ВП;
- ✓ многопациентный и индивидуальный ЭКГ-монитор-анализатор;
- ✓ специальные средства полного анализа полиграфии (РЭГ, КГР, ЭМГ, ЭОГ, дыхание, и др.);
- ✓ управление записью, периферией, экспресс-анализом и стимуляцией;
- ✓ настройка на любые методики;
- ✓ Всесторонняя клиническая проверка!

InCo

НПО «Информатика и компьютеры»
117602, Москва, а/я 365
☎ (095) 437-36-95

Оглавление