Ф. Мостеллер
Дж. Тьюки

АНАЛИЗ ДАННЫХ И РЕГРЕССИЯ

Ф. Мостеллер
Дж. Тьюки

Том 1

DATA ANALYSIS AND REGRESSION

A second course in statistics

Frederick Mosteller
Harvard University

John W. Tukey
Princeton University
and Bell Telephone Laboratories

Addison-Wesley Publishing Company
Reading, Massachusetts · Menlo Park, California ·
London · Amsterdam · Don Mills, Ontario · Sydney

Ф. Мостеллер, Дж. Тьюки

АНАЛИЗ ДАННЫХ И РЕГРЕССИЯ

Выпуск 1

Перевод с английского Ю. Н. БЛАГОВЕЩЕНСКОГО
Под редакцией и с предисловием Ю. П. АДЛЕРА

Москва «Финансы и статистика» 1982

ББК 22.172
М84

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ
МЕТОДЫ ЗА РУБЕЖОМ

ВЫШЛИ ИЗ ПЕЧАТИ

1. Ли Ц, Джадж Д., Зельнер А. Оценивание параметров марковских моделей по агрегированным временным рядам.
2. Райфа Г., Шлейфер Р. Прикладная теория статистических решений.
3. Клейнен Дж. Статистические методы в имитационном моделировании. Вып. 1.
4. Клейнен Дж. Статистические методы в имитационном моделировании. Вып. 2.
5. Бард Й. Нелинейное оценивание параметров.
6. Болч Б. У., Хуань К. Д. Многомерные статистические методы для экономики.
7. Иберла К. Факторный анализ.
8. Зельнер А. Байесовские методы в эконометрии.
9. Хейс Д. Причинный анализ в статистических исследованиях.
10. Пуарье Д. Эконометрия структурных изменений.
11. Драймз Ф. Распределенные лаги.

ГОТОВЯТСЯ К ПЕЧАТИ

1. Лимер Э. Статистический анализ неэкспериментальных данных. Выбор формы связи.
2. Бикел П., Доксам К. Математическая статистика. Вып. 1 и 2.

М $\frac{0702000000-109}{010(01)-82}$ 39—82

ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

НАУКА И ИСКУССТВО АНАЛИЗА ДАННЫХ

«И если мысль действительно нова,
То надо говорить с азов учиться...»
Е. Винокуров

Перед Вами, читатель, своеобразная книга. Как сказал бы Додо*, очень многие книги — совсем не странные, должны же, хоть иногда, попадаться и такие! (Позвольте напомнить, что именем Додо называл себя английский математик Чарльз Лютвидж Доджсон (1832—1898). Правда, известен он гораздо больше как Льюис Кэрролл — автор знаменитых сказок о непосредственной и доброй Алисе.)

Вам, очевидно, приходилось когда-нибудь что-нибудь обрабатывать статистическими методами, мучиться, терзаться сомнениями, негодовать на бестолковость тех, кто собирал данные, ломать голову над загадками машинных распечаток. Поэтому Вы, без сомнения, сможете оценить многие тонкости этой книги и успешно преодолеете многие «подводные камни», щедро рассыпанные по ее страницам.

Поскольку Вы сделали статистику своей профессией, Вы, очевидно, намерены совершенствоваться в области статистических методов. Дело в том, что чтение этой книги — большой и тяжелый труд, а в наш рациональный век не принято затрачивать труд «просто так», без четкой цели. Эта книга, безусловно, для тех, кто либо стал, либо хочет стать профессионалом. Если перед чтением или в процессе его Вы почувствуете пробелы в своей подготовке, — это не беда: их можно восполнить, обращаясь к другим книгам и своему практическому опыту (краткий перечень отечественных книг, пригодных для этой цели, мы приведем в предисловии ко 2-му выпуску).

Прежде чем приступить к чтению, поговорим о том, что такое «анализ данных». Здесь эти обычные слова звучат как термин, обсудим его. Начнем с «данных». Под этим словом мы обычно понимаем некоторую информацию об окружающем нас мире, безотносительно к тому, как она добыта, но при условиях, что она представляет или сиюминутный, или потенциальный интерес и упорядочена каким-то образом. Примерами данных могут служить результаты переписи населения, сведения о концентрации вредных веществ в водоеме, исходы подбрасывания

*См., например: Льюис Кэрролл. Приключения Алисы в стране чудес. Сквозь зеркало и что там увидела Алиса, или Алиса в Зазеркалье. М., Наука. 1979.

монеты или выходы химической реакции в специально спланированных опытах. «Данные» наступают на нас со всех сторон. Они накапливаются в темпе, значительно опережающем нашу способность их ассимилировать и использовать. Мы их «складируем» впрок, порождая огромные архивы и сложнейшие проблемы хранения, переработки, поиска и использования всего того, что нам удалось «узнать». Значит, с данными надо что-то делать. Но «делать» — это прежде всего означает, насколько возможно, сократить их количество и при этом не потерять слишком много «полезной информации», потенциально в них заложенной.

«Борьба» с данными — традиционная задача математической статистики. И она традиционно осуществлялась с помощью двух взаимно дополняющих принципов: *выборочного метода* и *свертки* информации. Первый из них декларирует отказ от «всей совокупности данных» в пользу специально организованной их части — выборки, а второй заменяет всю выборку несколькими числами — ее характеристиками. В качестве таковых могут быть, например, среднее арифметическое и дисперсия или уравнение регрессии. При получении подобных характеристик данные подвергаются некоторым воздействиям, как говорят, они *обрабатываются*, или исследуются, или анализируются. Поэтому процесс свертки данных называли то обработкой, то исследованием, а то и анализом данных.

Это, однако, обиходное, общепринятое понятие. Дж. Тьюки предложил сохранить термин «анализ данных» только за такими процедурами получения сверток, которые *не допускают* формального алгоритмического подхода. Это предложение и породило новое направление исследований, оказавшееся в центре внимания многих статистиков и специалистов по переработке данных и составляющее содержание книги.

Последние 25—30 лет характеризуются движением, охватившим многие области и направления науки и практики. На знамени этого движения написаны слова «*реальность*» и «*системность*». Классическая наука приучила исследователей к работе с *моделями*, которые существенно огрубляют, упрощают реальные задачи, но зато позволяют получить ответ явный и однозначный. В основе всякой модели лежит система предпосылок, постулатов, аксиом, которые формулируются непосредственно или подразумеваются. В истории науки прослеживается эволюция отношения к предпосылкам от слепой веры в их истинность к требованию их *проверяемости* (хотя бы в принципе). Проверяемость «дослуживалась» даже до «должности» критерия научности модели (которая суть *гипотеза*). Теперь можно констатировать, что сделан следующий шаг. Выяснилось, что от принципиальной проверяемости до конкретной проверки в условиях реальной задачи — дистанция огромного размера, ибо за проверку надо «платить», а плата, как правило, непомерна. Анализ данных предлагает затеять игру с предпосылками: варьировать их и рассматривать последствия такого варьирования. Так, можно сначала смотреть на данные как на числа (детерминированная модель), а потом — как на случайные величины (стохастическая модель) и выбирать такой ответ, который лучше гар-

монирует с требованиями конкретной задачи. У анализа данных нет, конечно, монополии на такой подход. Нечто весьма близкое просматривается уже в принципе дополнительности Бора. Для нас важно лишь, что в анализе данных в принцип возведено именно такое отношение к постулатам.

Займемся теперь системностью. Это, собственно, такая точка зрения, согласно которой о большинстве задач нельзя сказать заранее, что важно, а что второстепенно. Поэтому остается «валить все в кучу», а затем изыскивать какие-то систематические методы поиска «жемчужного зерна». Тогда мир делится на «объект исследования» и то, что его окружает — «окружение», окружающую среду. Какие-то данные извлекаются в ходе жизнедеятельности или специального экспериментирования из объекта — это эндогенная информация, а другие данные возникают вне объекта — это экзогенная информация. Анализ данных изыскивает различные приемы для наиболее полного использования эндогенной информации (что вообще характерно для любых статистических методов), но вместе с тем он постоянно нацелен на максимальное использование информации внешней. Хотя это резко усложняет задачу, зато иногда сулит замечательные находки. Заметим, что системный подход предъявляет исследователям повышенные требования, поскольку он носит принципиально междисциплинарный характер.

Любое исследование имеет начало и конец. Во всяком случае, так было принято считать. Именно анализу данных мы обязаны отказом, от этой «очевидной» точки зрения. Если начала какие-нибудь изыскания, может быть, еще и имеют, то уж концов у них нет и не предвидится. Анализ — способ существования данных. Любые данные, даже те полсотни точек, что Вам принесли вчера для «обсчета», неисчерпаемы. Готовность к постоянному возврату к одним и тем же данным— важная новая психологическая особенность ситуации, характерная для анализа данных. Результат в таком случае выступает как побочный продукт анализа, и, чтобы его извлечь, сам процесс приходится как-то структурировать, разбивать на этапы, шаги. Отсюда — *шаговый* принцип анализа данных. Предлагается различать три основных этапа: пробный, прикидочный и главный. Причем каждый из них обладает своими особенностями, а их логический порядок следования может нарушаться появлением новых идей (в ходе анализа) или экзогенной информацией. Шаговый принцип, сам по себе, тоже не принадлежит всецело анализу данных. Он характерен еще, скажем, для планирования эксперимента, да и для многих других случаев. Его можно рассматривать и как развитие старой математической идеи об итеративных вычислениях. Роль шагового принципа велика прежде всего потому, что с его помощью в непрерывном процессе анализа организуются разрывы, остановы, позволяющие извлекать накопленную информацию и принимать решения, связанные с *управлением* обработкой данных и с их дальнейшим анализом.

Здесь мы сталкиваемся с еще одним любопытным обстоятельством, характерным для работы с любыми сложными системами. Анализ данных, уже в силу определения этого термина, которое мы приводили выше, рисуется в виде некоего «слоеного пирога», в котором «коржи»

формальных операций обильно прослоены «кремом» *неформальных процедур принятия решений*. Мы уже имели повод говорить раньше (см.: Введение в планирование эксперимента. М., Металлургия, 1969; Предпланирование эксперимента. М., Знание, 1980 и др.), что это есть следствие алгоритмической неразрешимости, возникающей в ходе формализации любой реальной задачи. Практически же оно означает, что отрыв анализа данных от текущей содержательной интерпретации и непрерывного квалифицированного управления чреват многими неприятностями и прежде всего риском получить суждения, не полно, не точно, а то и искаженно отражающие всю совокупность данных. Что же касается неформальных решений, то они уже давно стали предметом статистического анализа в рамках теории статистических решений (как байесовского, так и не байесовского типа).

Наука начиналась (и начинается) с определения понятий. Главными признаками научных понятий испокон веков считались точность и однозначность. Им противопоставлялись понятия поэтические, ассоциативные, с присущими им неопределенностью, метафоричностью, раскованностью. Однако время показало, что между понятиями этих двух типов нет непроходимой пропасти. Более того, выяснилось, что в ряде случаев, например при анализе данных, вообще не удается построить последовательную систему представлений, опираясь только на точные и однозначные понятия. Так еще раз в науку «пробралась» *неопределенность*. (Лучше сказать, что она еще раз была осознана, ибо никуда из науки она не исчезала и прежде.) В анализе данных и, следовательно, в этой книге размытые неопределенные понятия используются широко и последовательно. Некоторые из них — авторские неологизмы, что отнюдь не облегчает перевода. Как правило, эти понятия построены так, чтобы одной из их частных узких интерпретаций было известное «точное» понятие статистической теории. Явное введение в науку неопределенных понятий создает новую ситуацию, что хорошо видно, например, по работам Л. Заде (см.: Понятие лингвистической переменной и его применение к принятию приближенных решений. М., Мир, 1976).

Поскольку, как известно, «истина едина» и от нас не зависит, все здание научного исследования было издавна построено для получения единственного однозначного наилучшего ответа. Возникновение вероятностных методов несколько поколебало эту концепцию, так как оказалось, что, задавшись уровнем точности, можно получить множество различных ответов, которые, однако, эквивалентны в смысле точности друг другу. Это, конечно, не означает их эквивалентности в каких-либо иных смыслах, например с точки зрения содержательной интерпретации. Привнесение вероятностных принципов в квантовую механику и возникновение принципа неопределенности Гайзенберга еще более усугубили положение, но за пределами микромира, в обычных «мирских» задачах, старая крепость еще как-то держалась. Теперь же ей ничего не остается, кроме капитуляции. Решающий удар был нанесен в ходе развития регрессионного анализа (и некоторых других методов многомерного статистического анализа), когда выяснилось, что переменные реальных задач обладают «мимикрией»

и склонны выдавать себя не за то, чем они являются. Но тогда понятие эквивалентности ответов приходится распространять на такие решения, которые могут даже не содержать одинаковых переменных! А это уже *многозначность*, множественность ответов. Получать наборы ответов — не слишком сложное дело. Гораздо сложнее научиться ими пользоваться, осуществлять выбор с помощью каких-то новых критериев и какими-то новыми методами. Концепция неоднозначности ответов развивается в многочисленных работах В. В. Налимова.

Ясно, что все перечисленные принципы не реализуешь «голыми руками». Поэтому мы вынуждены завершить их перечисление еще принципом *имитационного* моделирования, опирающегося на современные ЭВМ с развитыми системами внешних устройств и режимом диалога. Системы человек—машина — вот материальная основа анализа данных, инструмент, делающий его принципы практически реализуемыми. Мы уже характеризовали этот инструмент ранее, в предисловии к русскому переводу книги Дж. Клейнена «Статистические методы в имитационном моделировании» (Вып. 1. М., Статистика, 1978), к которому и отсылаем интересующихся подробностями. Среди причин обращения анализа данных к машинному моделированию есть вре́менные и постоянные. Первые связаны с недостаточным развитием современного математического аппарата и могут со временем исчезнуть, вторые же коренятся в сложности проблем изучения реального мира. Таким образом, реализм и системность, непрерывность и шаговость, неопределенность и размытость, неоднозначность предпосылок и ответов, неформализуемость и моделирование — вот те «киты», на которых покоится мир анализа данных. И пусть каждый из принципов не оригинален — вместе они создают неповторимые черты анализа данных.

«Прародина» анализа данных — это статистика. Путь от статистики к анализу данных — это «расшатывание устоев» статистики. На этом пути можно выделить такие этапы: осознание неединственности нормального распределения и создание теории распределений; развитие непараметрической статистики; создание байесовского подхода; теория робастности; аксиоматическое, частотное, субъективное и другие определения вероятности; последовательный анализ; теория малых выборок (отказ от асимптотических оценок); риски, потери и полезности; ошибки первого и второго рода — значимость и мощность; максимум правдоподобия, минимакс; метод «складного ножа»; критерии оптимальности и планы экспериментов. Эти и некоторые другие прорывы и создали предпосылки анализа данных, который, можно думать, стал уже, говоря словами Т. Куна, «новой *парадигмой*», приходящей на смену статистике. Как и положено в науке, традиционная статистика сохраняет свою роль и значение как частный случай анализа данных. И мы уже вправе задумываться над тем, что же придет на смену анализу данных.

Никакой анализ данных был бы невозможен без развития вычислительной техники и вычислительной математики, кибернетики и системного подхода, да и многих других областей современной науки, которые, слившись, привели к его возникновению.

Это случилось в начале 60-х годов нашего века. Условной датой его рождения можно считать публикацию известной статьи Дж. Тьюки «Будущее анализа данных» (Annales of Math. Stat., 1962, 33, p. 1—67).

Чтобы лучше понять книгу, часто, и наш случай именно таков, полезно знать кое-что об ее авторах.

Фредерик Мостеллер родился в 1916 г. в штате Западная Виргиния. Учился в Технологическом институте Карнеги, затем в Принстонском и Чикагском университетах. Работал в промышленности и правительственных учреждениях, преподавал в ряде высших учебных заведений. В последнее время возглавляет статистический отдел в Гарвардском университете. Член многих научных обществ и организаций, в том числе Американской Академии Наук и Искусств, Американской статистической ассоциации, Королевского статистического общества (Великобритания), Биометрического общества, Психометрического общества, Американского общества контроля качества, Центра по изучению общественного мнения и ряда других. Широко известен своими исследованиями в области выборочного метода, анализа общественного мнения, определения авторства (см. гл. 8 настоящей книги), медицинскими приложениями статистики, а также как педагог и теоретик обучения. На русском языке изданы три (насколько известно) работы этого автора: Буш Р., Мостеллер Ф. Стохастические модели обучаемости. М., Физматгиз, 1962; Мостеллер Ф., Рурке Р., Томас Дж. Вероятность. М., Мир., 1969; Мостеллер Ф. 50 занимательных вероятностных задач с решениями. М., Наука, 1971.

Джон Уилдер Тьюки на год старше своего соавтора. Он родился в штате Массачусетс. Учился в университете Брауна, затем в Принстонском университете, Кейсовском технологическом институте, Йельском и Чикагском университетах. О его работе и членстве в различных организациях можно практически полностью повторить сказанное о Мостеллере. В последнее время связан с Принстонским университетом и Белловскими телефонными лабораториями. Проложил новые пути во многих областях исследования. Еще в 30-е годы заинтересовался устойчивостью оценок и вместе с Хубером принял активное участие в создании теории робастных оценок (см.: Andrews D. F., Bickel P. J., Hampel F. R., Huber P. J., Rogers W. H. and Tukey J. W. Robust estimates of location: survey and advances. Princeton University Press, 1972). Общеизвестен его вклад в создание прикладного спектрального анализа: Blackman R. B., Tukey J. W. The measurement of power spectra. New York, 1958 и Tukey J. W. Discussion, emphasizing the connection between analysis of variance and spectrum analysis. — Technometrics, 1961, 3, p. 191—219. Развил (в 1958, 1962 гг.) предложенный Кенуем (1949 г.) метод «складного ножа», играющий важную роль в этой книге: Tukey J. W. Bias and confidence in not-quite large samples. Abstr. in Ann. Math. Stat., 1958, 29, p. 614. Внес вклад в процедуры множественных сравнений в дисперсионном анализе, в анализ остатков и ряд других областей. Уже отмечалась его роль в создании анализа данных. Отметим еще в этой связи фундаментальную трехтомную монографию по ана-

лизу данных: T u k e y J. W. Exploratory data analysis. Addison-Wesley, Reading, Mass., 1970—1971 (ротапринт, 1977 — книги). Ее можно рассматривать как введение в анализ данных, предваряющее настоящую работу*.

Нам известен следующий единственный перевод статьи Тьюки на русский язык: Т ь ю к и Дж. У. Анализ данных, вычисления на ЭВМ и математика. — В кн.: Современные проблемы математики. М., Знание, 1977, с. 41—64.

Таким образом, пути наших авторов издавна и неоднократно пересекались. Большая научная эрудиция, широта взглядов и блестящая профессиональная подготовка обоих авторов, видимо, позволили им преодолеть трудности соавторства, успешно использовать сильные стороны каждого и создать ту книгу, которая сейчас перед Вами.

Эта книга прежде всего учебная. Авторы стремились пройти по «лезвию бритвы» между справедливо осуждаемой и постоянно возрождающейся «поваренной книгой» и интересной лишь адептам строгой теорией. А это компромисс, т. е. положение одинаково неудобное для обеих сторон, но все-таки неизбежное. На каждой странице книги ощущается стремление авторов поделиться с читателем секретами мастерства, теми бесценными крупицами профессиональных знаний, которые даются лишь годами работы с реальными данными и которые упорно сопротивляются проникновению на страницы книг.

Архитектоника книги довольно сложна. Основной текст иллюстрируется примерами и интерпретируется в многочисленных и многоаспектных упражнениях. Благодаря такой структуре возникает значительная связность текста, в котором появившаяся раз тема, как правило, возникает вновь и вновь в различных вариантах. Все это, не создавая формальных препятствий для понимания текста, тем не менее требует от читателя высокой культуры чтения.

Для обозначения «иллюстративных примеров» в тексте авторы используют термин «exhibit», что ассоциируется, скорее, с «выставкой-продажей» или «демонстрацией моделей сезона», чем с нашим контекстом. Как правило, эти примеры ярки и убедительны.

Упражнения играют в книге особую роль. Это и обычное средство тренинга, и вместе с тем документ, ярко свидетельствующий в пользу универсальности статистического метода. Трудно придумать какую-нибудь сферу человеческой деятельности, которая бы не отразилась в упражнениях. Здесь и демография, и химия, и история, и ботаника, и многое другое. Весьма многообразны упражнения и по характеру. Теоретические и экспериментальные, тривиальные и очень трудные, требующие вычислительной техники и рассчитанные только на смекалку.

---

*Когда рукопись этой книги уже была в наборе, вышел в свет русский перевод: Д ж. Т ь ю к и. Анализ результатов наблюдений. Разведочный анализ. Пер. с англ. под ред. В. Ф. Писаренко. М., Мир, 1981, 693 с. Теперь можно рекомендовать начинать чтение именно с него, а затем приступать к изучению данного издания.

Поскольку терминология двух переводов не всегда и не во всем совпадает, читатель может согласовать ее, пользуясь терминологическими указателями, имеющимися в обоих переводах.

Хоть и не принято критиковать то, чего в книге нет, все же скажем, что одно из наиболее слабых мест, как нам кажется, это отсутствие акцента на этап постановки задачи, который мы называем «предпланированием эксперимента», а применительно к этой книге можно было бы назвать «преданализом данных» (скажем, «pre-analysis of data»). Ясно, конечно, что авторы ни на минуту не забывают о важности такой работы, просто они оставляют ее «за бортом» своего рассмотрения. Заметим, еще, что иногда пути, выбираемые авторами при изложении некоторых вопросов, кажутся не самыми прямыми. Впрочем, мы должны сохранить за авторами и право на поиск. Это ведь первая, насколько мы знаем, систематическая монография по анализу данных, выходящая на русском языке.

Поговорим немного о переводе и его принципах, что было делом далеко не обычным. Мало того, что авторы использовали около 100 собственных неологизмов. Там, где их не было, давал себя знать необычный стиль авторов, живой, близкий к разговорной речи, насыщенный идиомами и реалиями. Коллектив, работавший над русским изданием этой книги, ставил перед собой задачу сохранить не только идеи, но и особенности языка авторов. В какой мере это удалось — судить читателю.

Анализ перевода новых терминов, введенных авторами этой книги, потребовал бы специального обширного исследования. Поэтому, пользуясь тем, что все наши интерпретации представлены в «Указателе перевода терминов» в конце 2-го выпуска, ограничимся здесь лишь примерами (примеры терминов, характерных для вып. 2, см. в предисловии к нему). Прежде всего скажем, что иногда нам не хватало смелости или воображения и мы подгоняли новый термин под существующее русское понятие, хотя и сознавали, что это не совсем правомерно (главным образом при этом происходило сужение исходного объема понятия). Например, термин «hinge» мы отождествляли с «квартилем», тогда как это, скорее, «квинтэссенция данных». Или «summary» мы трактовали как «свертку» (иногда «сводку»), тогда как в авторском «numerical summary» нам слышались отголоски «numerology» («числовой магии»), превратившейся в прозаическую «числовую свертку». Понятие «letter value» мы передали как «особая точка» («особое значение») на кривой распределения, т. е. как синоним «процентиля». Здесь подчеркивается важность этой точки для каких-то наших целей. Вместо обычного «transformation» авторы часто используют выражение «re-expression» («перевыражение» что ли), для которого мы сохранили термины «преобразование» и «переформулировка». Разница в оригинале подчеркивает расширение класса преобразований, включающее не только непрерывные, но и дискретные скачкообразные переходы. Найти короткий русский эквивалент не удалось.

Иногда приходилось идти непроторенными путями. Прежде всего укажем на любопытное понятие «stem-and-leaf». Обсудив несколько вариантов, окончательно мы остановились здесь на термине «опора и консоль», сочтя эту механическую аналогию самой удобной. Тогда термин «forget» («забываемое», «пренебрегаемое») пришлось передать словом «мишура» (что висит на консоли, но не весит), ибо это то, что

мы округляем, отбрасываем. Для понятий «indication», «indicator», которым авторы придали новый смысл, мы воспользовались транслитерацией. Специального анализа требует вопрос о том, для чего авторы отказались от общепринятой классификации измерительных шкал Стивенса и построили свою новую классификацию, но для их понятий «amounts and counts», «balances», «counted fractions» и «grades-ordered labels» мы предлагаем соответственно «счетные суммы (итоги)», «счетные разности (балансы)», «счетные доли» и «упорядоченные ярлыки». Явно принеся содержание в жертву краткости, мы перевели интересный термин «PLUS analysis» как «аддитивный анализ». При этом потерялась идея непрерывной достройки модели аддитивными членами. Остается только надеяться на то, что контекст позволяет восполнить эту потерю. Наконец, скажем еще об одном производном от «estimate», а именно о «est mand», что мы несколько неуклюже передали словом «оцениваемое» (то, для чего «оцениватель» (estimator) получает оценку (estimate) в ходе оценивания (estimation).

Русское издание этой книги разбито на два выпуска. Но это все-таки одна книга. И если первый выпуск еще мыслим без второго, то второй без первого нежизнеспособен. (Краткая характеристика второго выпуска дана в предисловии к нему, поэтому здесь мы так мало говорили о регрессии.)

В ходе редактирования перевода нам неоднократно приходилось прибегать к примечаниям, в которых комментировались реалии и сообщались ссылки на отечественные издания по тем или иным вопросам. Все, что мы сочли опечатками, исправлялось без специальных оговорок.

Этим предисловием, безусловно, не ограничивается то, что хотелось бы сказать о предлагаемой книге Ф. Мостеллера и Дж. Тьюки. Надеюсь, читатель согласится с нами в том, что идеи и их авторская интерпретация, заложенные в книге, в свою очередь порождают у читающего новые идеи, ассоциации, исследовательские задачи. В этом, без сомнения, и состоит главная ценность книги.

Ю. Адлер

● ПРЕДИСЛОВИЕ

Две главные темы переплетаются в предлагаемом здесь изложении прикладной статистики — это система общих принципов, необходимых читателю для эффективного анализа данных и совокупность полезных и легкодоступных приемов, обеспечивающих практическую реализацию этих принципов.

После вводного курса статистики студент нуждается в дальнейшей подготовке для борьбы с гораздо более трудными задачами, обычно возникающими при анализе данных, хотя некоторые из этих задач и кажутся простыми. Большинство вводных курсов имеют целью сообщить студентам некоторые понятия, требующиеся при анализе данных (простейшие критерии значимости), и представления об отдельных распределениях вероятности, да иногда еще о регрессионном и дисперсионном анализах. Во вводном курсе преподаватель редко имеет время для обсуждения принципов анализа данных или подходов к их исследованию, оставляя до лучших времен многочисленные жизненно важные «если», «и» и «но» множественного регрессионного анализа. То же можно сказать насчет применения статистических методов для получения эффективных и надежных ответов на основе реальных данных, поэтому мы и обсудим эти вопросы подробно.

Мы предполагаем, что читатель знаком с вводным курсом статистики. (Может помочь и владение другими предметами, однако все, кроме вводного курса, не обязательно.) В зависимости от характера и глубины подготовки, читатель будет продвигаться по главам этой книги более или менее быстро, но большая часть материала, вероятно, будет все же новой при любом уровне подготовки. Мы делаем акцент отнюдь не на математике, хотя некоторые методы неизбежно отягощены формулами. В большинстве случаев математические подробности упрятаны в примеры.

К вычислениям мы обращаемся не часто. В двух-трех таких случаях читатель может пропустить детали и воспользоваться сразу результатами.

КАКУЮ ПОЛЬЗУ МОЖЕТ ПРИНЕСТИ ЭТА КНИГА

Эта книга может научить принципам и подходам, что гораздо важнее, чем овладение конкретными методами. Наши читатели узнают по меньшей мере то, как действуют и как устроены следующие принципы, соглашения и подходы, к которым при желании всегда можно возвращаться и которые можно обдумывать после прочтения тех или иных глав;

● подход к формулировке задач статистического анализа и анализа данных, позволяющий, например, а) правильно понять прямой подход Стьюдента к проблеме статистического вывода и б) прояснить роль концепции неопределенности;

● роль индикаторов (указателей поведения, но не обязательно по заранее выбранной шкале; в отличие от статистических выводов и решений — по фиксированным количественным или качественным шкалам);

● важность формы представления данных и ценность рисунков при попытках добыть информацию, неожиданную для читателя;

● важность переформулировок (вместе с вопросом о том, как это сделать);

● требование тщательно выискивать действительные неопределенности как нетривиальная задача;

● важность итеративного счета, когда мы «крутимся» до тех пор, пока ответ не стабилизируется (однократные вычисления превосходны, если только они обеспечивают наши потребности);

● возможность взаимодействия идей робастности (в частности, робастности эффективности) и устойчивости: что мы делаем и что мы думаем;

● все о регрессии;

● что могут и чего не могут дать нам коэффициенты регрессии;

● каким образом поведение наших данных часто можно использовать в качестве руководства к их анализу (как в регрессии, так и при переформулировке задачи; названы лишь два пути из многих);

● важность рассмотрения остатков и извлечения информации из них;

● идея такого анализа данных и таких вычислений в возможности извлечь пользу из многократных возвратов к началу и выискивания оригинальных подходов, что существенно отнюдь не только для анализа данных.

Тот, кто только приступает к чтению этой книги, может быть, не знает даже значений некоторых из употребленных выше слов; тот же читатель, который проработает все 16 глав, узнает, что же все эти слова означают, и, более того, выработает точку зрения на относительную важность сформулированных сентенций.

Обратимся к самим главам. Из дальнейшего изложения станет ясно, что совсем не обязательно читать их подряд. Первые две главы содержат некоторые практические принципы анализа данных и адресуются начинающим, так же как, скорее всего, и следующие гл. 3 и 4 (а не тому читателю, кто имел повод упражняться в исследованиях по анализу данных, что до 1977 г. для большинства было невозможно). Мы имеем в виду книгу «Exploratory Data Analysis», где эти вопросы изложены более полно, речь идет о ее первом издании 1977 г. (Addison-Wesley Publishing Company, 1977), а не о предварительном труднодоступном варианте. В гл. 4 обосновывается обычная линейная регрессия.

О мнениях вообще, в том числе и мнениях насчет природы и пользы переформулировок, спорить не принято. Мы, однако, будем спорить

там, где уместно (в гл. 5 и 6), поскольку это и просто, и полезно. Правда, с этой дискуссией можно было бы и повременить, даже перенести ее за главы о регрессии. Поэтому мы исключили из гл. 6 многие детали и поместили их в приложение, идущее вслед за гл. 16*.

Главы 7 и 8 — наиболее общие по своим принципам и понятиям, по разнообразию и трудности задач и их технического инструментария. Тезисы о том, что отыскание действительных неопределенностей может оказаться вовсе не легким и что оно все-таки возможно даже для диффузных неупорядоченных ситуаций, где классический математический подход кажется совершенно неприемлемым, можно проиллюстрировать всем, на что мы способны в анализе данных, и здесь есть из чего выбрать. Читатель, усвоивший методы гл. 8, сможет углубиться в анализ данных, так же как и в перепроверку моделей. Для этого годятся универсальные методы — регрессионный и дисперсионный анализ. Глава 8 показывает, как они работают в множественной регрессии. Материал гл. 7 и 8 неявно используется в дальнейшем, поэтому, если их опустить, читатель будет ощущать невосполнимую потерю.

Главы с 9-й по 11-ю образуют необычную подборку из трех методов, требующихся любому статистику или специалисту по анализу данных, но тем не менее не распространенных широко: прямой и гибкий подход к таблицам (сопряженности признаков) с двумя входами (гл. 9), современный взгляд на устойчивые и робастные методы с простейшими приложениями (гл. 10) и достаточно подробное вычисление нормировок (гл. 11). Многие студенты, проштудировав некоторые учебники статистики, так ничего и не узнают о нормировании, и когда они сталкиваются с реальными задачами, это оборачивается потерями. Мы считаем, что материал гл. 11 (9 и 10 тоже) — это квинтэссенция программы обучения статистиков и специалистов по анализу данных. Снова эти главы нет нужды читать перед главами о регрессии, за исключением части гл. 10, которую стоит прочесть перед (или одновременно с) последним разделом гл. 14.

Некоторые темы так взаимосвязаны и переплетены, что образовать из них простую линейно-упорядоченную последовательность по меньшей мере трудно. Опасаемся, что сегодня масса проблем и тем, обозначаемых словом «регрессия», — это именно тот случай. Мы попытались упорядочить их, как могли. Глава 12 излагает азбуку регрессионного анализа. Как правило, такой материал не комментируется. Однако мы считаем, что он важен и достоин внимания при первом знакомстве с регрессией. Затем идет гл. 13 о бедах регрессионных коэффициентов. В гл. 14 исследуются машинные основы обычной работы по подбору линии регрессии и возможности использования робастных и устойчивых моделей наряду с обычными слабыми (неустойчивыми) моделями метода наименьших квадратов.

Глава 14 раскрывает математический подход к пониманию регрессии и для многих читателей может послужить указанием и поддержкой. Не все удалось представить исчерпывающе полно. Для деталь-

*В нашем издании это приложение помещено в конце первого выпуска — после гл. 11. — *Примеч. ред.*

ного исследования надо раскрыть обобщенные соотношения и помнить, что читателю не обязательно участвовать в каждой операции. Тот читатель, который умеет делать выводы из фактов, взятых на веру, сможет извлечь из этой главы больше, не прослеживая за всеми формальными подробностями. Основы содержатся в двух предыдущих главах, особенно в гл. 12.

Мысль о том, что существует одна-единственная регрессия, которую мы могли бы подобрать по заданному множеству данных, часто ложна. Глава 15, которая усиливает (обобщает) результаты гл. 12 и 13, выясняет ложность этой мысли по крайней мере для наших конкретных данных, а гл. 16 рассказывает нам кое-что о том, как можно попробовать «выловить» какие-нибудь дополнительные модели.

Для тех, кто хочет поскорее добраться до глав о регрессии, маршрут проходит через гл. 4, 10 и затем через 12—16 с эпизодическими возвратами к гл. 5, а также к гл. 1 и 2.

НЕКОТОРЫЕ УЗЛОВЫЕ МОМЕНТЫ

Какие-то узловые моменты всегда подразумеваются, но все же удобнее иметь их более или менее полный перечень. Среди прочих мы (в порядке появления) рассмотрим:

- представления типа «опора и консоль»;
- сглаживание текущими медианами;
- лестницу переформулировок для линеаризуемых кривых;
- методы переформулировки задач для анализа и путь оценки их достоинств в частных случаях;
- специальные таблицы, облегчающие ручной счет при переформулировках задач;
- метод прямого оценивания;
- метод «складного ножа» с многочисленными иллюстрациями;
- робастный анализ таблиц с двумя входами;
- робастные и устойчивые меры положения и шкалы;
- прямые и косвенные методы нормировок как индикация или как прием анализа данных;
- методы нормирования с широкими возможностями (с учетом как количественных, так и качественных переменных);
- подход к регрессии с упором на анализ остатков;
- регрессия с ошибками измерений;
- интерпретация регрессионных коэффициентов;
- эффекты «подставных» факторов;
- стандартный метод наименьших квадратов;
- взвешенный метод наименьших квадратов в нескольких разновидностях;
- метод наименьших абсолютных отклонений (модулей);
- выбор среди регрессионных моделей, особенно для шагового подбора;
- выбор новых факторов и оценка старых в множественном регрессионном анализе.

ВЫПОЛНЕНИЕ ЗАДАНИЙ

В конце книги содержатся многочисленные домашние задания и внеаудиторные работы, сгруппированные по главам. Ко многим главам, особенно о регрессии (да и для гл. 10 тоже) их гораздо больше, чем нужно. Преподаватели смогут отбирать из них то, что им нравится, а учащиеся — проверять на них свою решительность. Каждый должен понять, что, как и в реальной практике анализа данных, задачи часто не имеют изящных ответов или простых правильных решений.

Если вы уверены, что это возможно, используйте для громоздких вычислений компьютер и вы обойдетесь без больших усилий. Однако нужно быть уверенным в том, что компьютер действительно выдает именно то, что вам требуется. Далеко не всякая программа из бесчисленных «пакетов» статистических программ для ЭВМ может на самом деле дать нам то, что мы надеемся от нее получить. Еще более коварны отдельные программы, часто написанные теми, кто знает машины и программирование много лучше, чем ловушки, классические и свежевыявленные, статистических алгоритмов. Будьте внимательны, когда пользуетесь такими программами!

КАК ВОЗНИКЛА ЭТА КНИГА

В канун 1963 г. Гарднер Линдзи (Gardner Lindzey) и Элиот Аронсон (Eliot Aronson), которые готовили в то время к печати новое издание «Справочника по социальной психологии» [Handbook of Social Psycology (1968)], предложили авторам этой книги написать новый раздел взамен статьи Мостеллера и Буша (Bush) из первого издания. Лавина идей и издательские планы породили а) гл. 10 переработанного издания Справочника: «Анализ данных — включая статистику» (с. 80—203 во втором томе) и б) осознание того, что мы вольны использовать материал этой статьи как часть самостоятельной книги. Затем имевшимся наброскам была придана более простая, ясная, отшлифованная форма, но собственно эта книга начала складываться лишь тогда, когда цели, связанные с ее написанием, были осознаны и захватили нас. Гл. 1, 2, 7 и 8 близки статье из Справочника, но остальные частично или полностью отражают уже новое понимание того, что:

1) студенты нуждаются в книге, объединяющей исследовательский и рецептурный подходы и ставящей целью, более или менее явно, разбор хода анализа данных с обсуждением именно тех моментов, которые будут проявляться на практике;

2) студенты нуждаются в книге, говорящей правду о регрессии, о ее целях, методах и о том, чего в ней можно достигнуть, а чего —нельзя;

3) студенты нуждаются в материалах, которые подчеркивали бы многие важные моменты и приемы, как правило, остающиеся между строк в руководствах по статистике.

Хотя мы понимаем, что непрерывное развитие знания и новшества делают достижение этих трех целей невозможным, мы все же надеемся, что читатель захочет проверить, насколько это нам удалось.

Кембридж, Массачусетс — *Ф. Мостеллер*
Принстон, Нью-Джерси — *Дж. Тьюки*

Глава 1 ○ НА ПОДСТУПАХ К АНАЛИЗУ ДАННЫХ

ВВЕДЕНИЕ

Каждый, кто приобщается к искусству анализа данных, многократно начинает с имеющихся у него знаний из области статистики и обновляет их, обретая новое видение и переоценивая свой прошлый опыт. Предполагается, что читатель этой книги знаком лишь с простейшими методами и идеями; те же несколько параграфов, что требуют более глубокого знания статистики и анализа данных, отмечены звездочками.

Приложения математики всегда осложнены тем, что сущностью предмета исследования надо овладеть столь же хорошо, как и применяемой математикой. Особенно же трудно говорить о механизмах анализа данных, поскольку этот анализ не ветвится, как дерево, и не поддается, похоже, естественно планомерному изложению. Поэтому мы периодически имеем дело то с новым пониманием, то со старым опытом. И наша цель прежде всего — развитие идей, полезных в анализе данных как для практиков, так и для критиков-теоретиков.

Обычно математические и методологические вопросы начинают обсуждаться с разработки общей теории, из которой извлекаются основные концепции. Затем на их базе создаются конкретные методики, иллюстрируемые в конце концов примерами. Для изложения анализа данных мы считаем более удобным другой порядок, а именно:

а) прежде всего, — что делать? (Какой обработке подвергнуть имеющиеся данные — арифметической или графической?)

б) затем, — почему выбирается именно такой способ обработки? (Каковы основания для выбора? Почему выбран именно этот способ среди многих возможных для данного типа задач? Объяснения, как правило, нужны в терминах конкретных математических моделей.)

в) далее, — в чем проблема? (Что можно сказать о типе моделей? О тех основаниях, что подтверждают разумность нашего выбора, и, наоборот, о тех, что указывают на возможность ошибки?)

г) наконец, — каков механизм наших раздумий обо всем процессе анализа данных? (Что лучше обосновывает пункт (в) — общие теории дедуктивного метода или индуктивные заключения из эксперимента?)

Причем и значительные ревизии на уровне (г) не так важны для практики, как менее впечатляющие изменения на уровне (б).

В этой главе мы затронем несколько тем: о лестнице первичных, вторичных, третичных и т. д. статистик и прямом выводе Стьюдента (Student); о вероятностных свойствах наблюдений и измерений; о по-

требности в неопределенных, размытых, неточно сформулированных концепциях для построения более четких понятий и критериев; о различиях между индикацией, определением и выводом. Ко всему этому еще предстоит возвращаться.

1.1. ЛЕСТНИЦА И ПРЯМОЙ ВЫВОД

До Стьюдента всякий раз анализ данных, отвечающий на вопрос, «Что могло бы быть?», напоминал длинную лестницу от очевидного первого шага до туманных высот. Расчет начинался с первичной статистики, т. е. числа, достаточно хорошо отражающего то основное, что можно сказать о предмете обсуждения на основе данных. Такой первичной статистикой может служить, например, выборочное среднее. Но тогда возникает новый вопрос: «Сколь сильно вычисленное значение может уклониться от действительного?», и рассчитывается вторичная статистика, число, достаточно хорошо характеризующее изменчивость (или, напротив, стабильность, неизменность) первичной статистики. Таким числом могла бы быть оценка стандартного отклонения уже имеющегося выборочного среднего. После этого шага снова во всей красе встает тот же вопрос: «Сколь сильно может уклониться...?», теперь уже для вторичной статистики, которая, как правило, оказывается менее стабильной (относительно самой себя), чем первичная статистика, чью устойчивость она оценивает. В принципе надо было бы переходить к третичной статистике, оценивающей изменчивость вторичной, затем к четвертичной и т. д. все выше и выше по лестнице, которая, — поскольку третичная была худшей характеристикой, чем вторичная, а четвертичная еще хуже третичной, — может рисоваться лишь как нечто все более нечеткое и туманное. Практики обычно останавливаются на первых двух статистиках.

Стьюдент (1908) открыл новое направление, по существу, уже вопросом: «А что, если я имею n наблюдений, взятых случайно из одной действительно нормальной совокупности, о среднем и дисперсии которой я ничего не хочу предполагать?». Пусть $\bar{y}$ — выборочное среднее, μ — генеральное среднее «нормальной совокупности», y_i — измерения, n — объем выборки. Мы будем пока считать $\bar{y}$ и y_i случайными величинами, а не числами, найденными в исследовании*. Следующее *отношение* носит имя Стьюдента:

$$t = \frac{\text{выборочное среднее} - \text{генеральное среднее}}{\sqrt{(\text{выборочная оценка дисперсии распределения})/n}} =$$

$$= \frac{\bar{y} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{\bar{y} - \mu}{\sqrt{\left(\frac{\sum (y_i - \bar{y})^2}{n-1}\right)/n}} =$$

*Авторы хотят подчеркнуть принципиальное различие между двумя подходами: 1) насколько мои данные и их свойства можно объяснить нормальным распределением? 2) что я могу вывести из своих данных, если они действительно реализуют случайный выбор из точно нормальной совокупности? — *Примеч. пер.*

$$= \frac{\text{выборочное среднее} - \text{генеральное среднее}}{\sqrt{\text{выборочная оценка дисперсии числителя}}} =$$

$$= \frac{\bar{y} - \mu}{s_{\bar{y}}} = \frac{\bar{y} - \mu}{\sqrt{\frac{\sum (y_i - \bar{y})^2}{(n-1)\,n}}};$$

его распределение, основанное на теоретическом нормальном распределении, зависит, как было установлено, только от n. Стьюдент рассчитал некоторые числовые характеристики распределения t. Затем, воспользовавшись эмпирической выборкой в 3000 измерений длины пальцев и 3000 измерений роста, которые были собраны как вспомогательные признаки для опознания преступников, он выдвинул предположение о математической форме распределения t. Корректность этого предположения Стьюдента была доказана Р. А. Фишером [Fisher R. A. (1925)], спустя 17 лет.

Этот подход отсекает туманное «восхождение по лестнице» на третьем, а по сути, уже на втором шаге. Для обсуждавшихся целей оценки генерального среднего из всех данных потребовалось лишь: 1) выборочное среднее — первичная статистика; 2) выборочная оценка дисперсии — вторичная статистика; 3) объем выборки — третичная статистика, которую очень легко получить и которая удивительно стабильна, по крайней мере до тех пор, пока эта выборка сравнивается с другими возможными выборками того же объема. Все остальное можно получить за счет допущения о «чистой» нормальности.

Заметим, что в то время как в нашем случае выборочное среднее, выборочная дисперсия и объем выборки служат соответственно первичной, вторичной и третичной статистиками, в иных обстоятельствах они могут играть и иную роль.

Используя предположение о нормальности, эти три числа и любое подходящее (ожидаемое, предполагаемое) значение μ_C для генерального среднего, мы можем рассчитать t по формуле

$$t = \frac{\text{выборочное среднее} - \text{подходящее среднее}}{\sqrt{\text{выборочная оценка дисперсии числителя}}} = \frac{\bar{y} - \mu_C}{s_{\bar{y}}}.$$

Подходящее значение может быть любым числом, в том числе и таким, которое в несколько раз отличается от выборочного среднего при заданной выборке. Его можно положить равным нулю, если мы изучаем групповые различия и считаем всерьез или гипотетически, что мы их не обнаружим. Оно может равняться, скажем, 500, если мы сравниваем группу студентов, например первокурсников какого-нибудь колледжа, с эталоном, представляющим общенациональное «среднее» по стандартизованному тесту Службы педагогического тестирования (Educational Testing Service), который часто имеет среднее 500, а стандартное отклонение 100. Тогда подходящее значение можно взять любым внутри доверительного интервала.

Всякий раз, когда мы подставляем подходящее значение в формулу для t, мы делаем первый шаг к построению критерия значимости. Когда подходящая величина в точности равна среднему (генеральному)

совокупности, из которой взяты y_i, распределение t задается простыми таблицами. Когда же это значение далеко от истинного, t сдвигается по модулю в сторону бóльших чисел. Подобные соображения, примененные к данной и другим ключевым статистикам, требуют несколько бóльшей аккуратности в работе с таблицами критических значений и приводят к понятию об оперативных характеристиках или мощности критериев (см., например [Mosteller F. and Bush R. R. (1954)]), вскрывая всю подноготную критериев значимости и едва ли не все конструкции, используемые на практике для построения доверительных интервалов.

В тридцатые—сороковые годы научились обрывать «лестницу» без каких-либо жестких ограничений. Так, были введены непараметрические, или не зависимые от распределения («свободные от распределения»), методы, что исключило зависимость от нормальности распределения и сделало «5%» действительно 5%, положив таким образом начало совсем новому подходу к старым задачам в «перманентной» революции*.

1.2. РЕАЛЬНЫЙ ВКЛАД СТЬЮДЕНТА

В первые три четверти столетия после своего появления t Стьюдента использовалось на практике особенно часто, что породило множество разнообразных приемов и целую серию усовершенствований. Это в равной мере послужило, как водится, толчком и к усовершенствованию статистической теории.

Ценность работы Стьюдента не в том, что она ведет к большим изменениям в числах, получаемых при анализе данных, так как обычно это совсем не так. Достаточно взглянуть на илл. 1.2.1 (в конце главы), где даны несколько процентных точек, обеспечивающих построение двусторонних $95\% = 19/20 = 38/40$ доверительных пределов (они еще понадобятся и для построения двусторонних $2/3 = 66\ 2/3\%$ доверительных пределов меньшего, чем обычно, уровня, и причины их выбора будут объяснены в параграфе 1.3). Одной из мер эффективности использования t Стьюдента служит отношение длины получаемого доверительного интервала к длине интервала, построенного так, как будто оценка дисперсии и есть дисперсия нормированного нормального закона распределения, что эквивалентно применению t с бесконечным числом степеней свободы. Многие, еще задолго до Стьюдента, использовали отношение, которое теперь носит его имя (если пренебречь мелкими различиями в расчете выборочного стандартного отклонения, обсуждаемыми ниже), а именно

$$\frac{\text{выборочное среднее} - \text{подходящее среднее}}{(\text{выборочное стандартное отклонение})/\sqrt{n}}.$$

Но, не имея таблиц Стьюдента, все они при этом соотносили результат с таблицами нормированного нормального распределения и поль-

*Авторы, очевидно, имеют в виду «революцию», начатую Стьюдентом. — *Примеч. пер.*

зовались самыми разными словесными предостережениями при интерпретации результатов.

Сколь же велико было различие? Из илл. 1.2.1 мы находим, для примера, что если строить 95%-ные доверительные пределы для нормированного нормального распределения, то коэффициент при стандартной ошибке $s_{\bar{y}}$ будет равен 1,96, тогда как для t-распределения с 12 степенями свободы он составит 2,18. Поскольку $2{,}18/1{,}96 \approx 1{,}11$, использование t-распределения добавляет лишь 11% к длине 95%-ного доверительного интервала при 12 степенях свободы[1]. Если взять слегка отличающееся определение $s_{\bar{y}}$, а именно $\sqrt{\Sigma\ (y_i - \bar{y})^2/n^2}$, то (вновь для нормированного нормального распределения) отношение этих длин достигнет величины 1,15. Для умеренных доверительных уровней эффект менее заметен. Так, для двустороннего интервала с 2/3 доверительным уровнем мы можем «опуститься» до 5 степеней свободы при отношении длин доверительных интервалов, основанных соответственно на t-распределении и на нормальном, не превышая 1,1 ($\approx 1{,}07/0{,}97$).

Нас не будет мучить гигантская величина отношения $12{,}7/1{,}96 \approx$ $\approx 6{,}5$ при 95%-ном уровне и 1 степени свободы. Ведь подавляющее большинство исследователей пользуется бóльшими числами степеней свободы. На самом деле замечательное воздействие t-таблиц состоит в подстрекательстве исследователей к работе с выборками больших размеров, дабы избежать таких ужасных отношений, как 6,5; 2,2 и 1,6, возникающих при 95%-ном доверительном интервале для 1,2 и 3 степеней свободы соответственно. Отметим, что для доверительного интервала, равного 2/3 (66 2/3%), соответствующие отношения есть всего лишь 1,8, 1,3 и 1,2.

Ценность работы Стьюдента заключена не в значительности числовых изменений, а в

- осознании того, что если искать подходящие предпосылки, то надо принять во внимание «капризы» малых выборок, причем не только в той задаче, с которой начинал Стьюдент, но и во всех подобных;
- числовой оценке того, сколь мала, как мы уже видели, коррекция границ доверительных интервалов в задаче Стьюдента и сколь велика зависимость этих границ от близости доверительной вероятности к единице;
- создании таблиц, которые можно использовать для определения доверительных интервалов и построения критериев значимости при проверке гипотез, относящихся даже к очень малым выборкам.

С вкладом Стьюдента наряду с положительными моментами сопряжены и некоторые характерные недостатки, в частности:

- стало очень легко пренебрегать оговоркой: «Если сделаны подходящие предположения...»;
- преувеличивается значение «точности» решения Стьюдента для идеализированной задачи;

[1]Мы употребляем символ «$\approx$» вместо слов «приблизительно равны», или «близки по значению», или иных подобных, означающих скорее аппроксимацию, чем точное равенство.

- внимание статистиков-теоретиков отвлекается на развитие «точных» методов в других задачах;
- ослабевает атака «проблем множественности» — трудностей и соблазнов, связанных с многократным использованием разных тестов для одних и тех же данных.

Большое внимание к точности обработки становится еще более удивительным, когда мы заметим, как быстро исчезают малые различия между критическими значениями t Стьюдента и нормальной аппроксимацией (илл. 1.2.2), особенно вблизи наиболее распространенного двустороннего 5%-ного уровня, после умножения (по предложению Бюррау [Burrau Ø. (1943)]) значения t на константу, приводящую дисперсию к единице. (Модификация Бюррау, взвешивая все «за» и «против», освобождает нас от возможных заблуждений по поводу эффектов лишь из-за непостоянства дисперсии t.)

Этот параграф, быть может, слишком растянут, но хотелось обратить внимание читателей и на достоинства работы Стьюдента, и на создаваемые ею трудности, не ограничивая себя в высказываниях о том и о другом.

1.3. РАСПРЕДЕЛЕНИЯ И ИХ НЕДУГИ

Сам Стьюдент всегда помнил, что наблюдения и измерения никогда не бывают распределены по магической колоколообразной кривой, даже если это результаты химических анализов стандартной продукции, выполненные под его собственным наблюдением [Student (1927)]. История статистики и анализа данных — это пестрая смесь здорового скептицизма и наивного оптимизма относительно точных видов распределений наблюдений. Оптимизм такого сорта часто необоснованно раздувается из-за замечательных свойств избранного семейства распределений, — «нормальных» распределений, плотность вероятности которых задается формулой $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2\,(x-\mu)^2/\sigma^2}$ для $-\infty < x < \infty$, где μ и σ — соответственно генеральное среднее и стандартное отклонение, e — основание натуральных логарифмов 2,7182818..., а π — наш старый знакомый 3,1415926... Илл. 1.3.1 дает три примера нормальных распределений с различными сочетаниями μ и σ.

Мы равноправно употребляем прилагательные «нормальное» и «гауссовское» для распределений, которые *точно* соответствуют этой формуле, однако оба термина не вполне удовлетворительны. Слово «нормально» многие неверно истолковывают как «обыкновенно появляющееся», но известно, на практике никогда не бывает распределений, в точности удовлетворяющих этой формуле, — ни для отдельных наблюдений, ни для средних значений, ни для других производных величин, хотя есть как умозрительные, так и фактические основания считать, что многие эмпирические распределения должны хорошо ею аппроксимироваться: иногда вполне успешно, а иногда лишь «на глазок» и не по сути. (Наиболее существенное зачастую замаскировано и не обнаруживается в традиционных гистограммах.)

Связь распределений производных статистик, таких, как выборочное среднее и стьюдентово t, с распределениями индивидуальных значений, как мы еще покажем, обычно слаба. Хотя распределения производных статистик и определяют согласованность наших утверждений о неопределенности, значимости и доверии, реальная практика аппроксимации варьирует от случая к случаю и их вид часто нелегко установить, анализируя отдельную выборку или даже весь объем имеющихся данных, составляющий, быть может, сотни наблюдений.

Говоря о трех распределениях из илл. 1.3.1, мы отмечаем, что все они имеют одну и ту же «форму» кривых. Давайте обсудим само понятие «форма».

Пусть у нас есть так много наблюдений, что их гистограмма «почти совпадает» с точным априорным распределением. Допустим еще, что эта гистограмма построена на миллиметровке так, что рисовавший ее забыл проставить числа вдоль осей. Что же мы можем извлечь из того, что имеем? Что мы потеряли? Без чисел на вертикальной оси мы не можем сказать, сколь велика выборка. Но поскольку нам интересно распределение, а не выборка и выборка велика, можем о них забыть. Без чисел на горизонтальной оси мы не можем сказать даже приблизительно, каковы значения выпавших наблюдений, как распределение растянуто или сжато, каковы его положение и масштаб, если использовать термины из техники. Потерялись именно эти важнейшие характеристики. Что же осталось?

Потеряв положение (сдвиг) и масштаб, объединяя все распределения, отличающиеся только линейными преобразованиями, — мы теряем лишь два числа и соответственно многое остается. Вот все то, что остается, и обозначается обычно словом «*форма*». Даже распределения, принадлежащие к одному и тому же математическому семейству, могут иметь весьма разные формы. Например, семейство бета-функций, хотя это и совершенно особый случай, включает распределения многих форм. На илл. 1.3.2 показан набор кривых плотностей бета-распределений вида $\beta(p) = cp^{a-1}(1-p)^{b-1}$, $0 \leqslant p \leqslant 1$, где c — константа, обращающая всю площадь под кривой в единицу, а параметры (a, b) приведены на графиках.

По гистограмме мы можем определить место центральных 2/3 из всех случаев и найти длину отрезка горизонтальной оси, который их покрывает. Мы можем затем сделать то же самое для центральных 19/20 и сравнить две длины. Если их отношение почти точно равно 2 к 1, то мы подтверждаем одно из свойств нормального распределения. Если же их отношение ощутимо больше, чем 2 к 1, у нас есть основания считать, что двусторонние 5%-ные точки нашего распределения растянуты больше, чем можно ожидать для какого бы то ни было нормального распределения. Если выборка достаточно велика, то это служит веским аргументом в пользу того, что генеральная совокупность не нормальна. В гл. 2 мы покажем, что в случае малых и средних объемов выборок более вескими могут оказаться другие признаки.

Подобные отношения можно получать, сдвигаясь все дальше и дальше в сторону «хвостов» распределения. Если при таком анализе всей

популяции либо достаточно большой выборки эти отношения имеют явную тенденцию быть заметно больше, чем у нормального распределения, то мы понимаем, что «хвосты» распределения *более разбросаны* или более растянуты, чем у нормального. Такое утверждение нельзя сделать для всего размаха распределения, скорее оно касается области «хвостов» в сопоставлении с центральной частью.

Беглый взгляд на илл. 1.2.1 показывает, что t-распределение Стьюдента более разбросано, чем гауссово, конечно, кроме случая с ∞ (бесконечным) числом степеней свободы, который как раз и отвечает кривой Гаусса. Уже упоминавшееся отношение (длины средней части распределения, охватывающей 19/20 к 2/3 его длины), равное примерно 2,02 для гауссовского случая; 2,16 для 12 степеней свободы; 2,4 для 5; 3,4 для 2, резко увеличиваясь, достигает 7,3 для одной степени свободы. Если нам известно, что 2/3 гауссова распределения лежат в некотором интервале, то, взяв интервал с тем же центром, но в семь раз больший, мы знаем, что, за исключением крохотной доли (в одну стомиллиардную), все это гауссово распределение лежит внутри большого интервала. Аналогичное построение для распределения Стьюдента с одной степенью свободы, как можно рассчитать, оставляет за пределами семикратно увеличенного интервала $5\% = 1/20$ распределения. Когда «хвосты» сильно разбросаны, по центральной части распределения можно легко обмануться относительно их поведения. Реальные распределения, как правило, гораздо более разбросаны, чем нормальное.

1.4. КЛАССИЧЕСКИЙ ПРИМЕР: УИЛСОН И ХИЛФЕРТИ АНАЛИЗИРУЮТ ДАННЫЕ ПИРСА

Освещению поднятых проблем должен помочь пример исследования нормальности распределения на необычно обширных данных. В эмпирическом исследовании, предпринятом для проверки аппроксимационных свойств нормального распределения, Пирс [Peirce Ch. S. (1873)] изучал запаздывание между резким звуком короткого удара и реакцией испытуемого, который давал около 500 ответов ежедневно в течение 24 дней.* Хотя Пирс посчитал, что вывод о нормальности

*Взятый авторами пример любопытен. Он принадлежит крупнейшему американскому ученому Чарльзу Сандерсу Пирсу (1839—1914), внесшему большой вклад в логику и топологию, признанному создателю семиотики. Обсуждаемый пример относится к статистическим работам Пирса. Суть эксперимента Пирса кратко излагается здесь по изданию P e i r c e Ch. S. The new elements of mathematics. Vol. 3/1, Hague-Paris, Moiton Publ., Atlantic Hichlands, Humanities Press, 1976, p. 639—676. Хроноскоп (электромеханический прибор для измерения времени с точностью до 0,001 с) запускался телеграфным ключом, синхронизированным с шариком, падавшим на металлическую поверхность. На звук от этого падения реагировал молодой человек, не прошедший никакой предварительной подготовки. Его задача заключалась в том, чтобы, услышав звук, как можно быстрее нажать на ключ другого телеграфного аппарата, который останавливал хроноскоп, измеряя таким образом время запаздывания в тысячных долях секунды. Опыт повторялся 500 раз в день в течение целого месяца. Его результаты и обсуждаются. — *Примеч. ред.*

этого распределения вполне обоснован, дополнительный анализ опубликованных Пирсом материалов, проведенный Уилсоном и Хилферти [Wilson E. B., Hilferti M. M. (1929)], привел их к существенно иным результатам.

Они вычислили по данным Пирса много разных статистик, анализируя каждый день наблюдений отдельно. Мы отобрали некоторые из 23 статистик, описанных Уилсоном и Хилферти, и привели их значения по дням в илл. 1.4.1. Столбец (2) дает оценку разности между точками Q_3 и Q_1, отсекающими 75% и 25% распределения соответственно (определяемой по выборочному распределению наблюдений за один день), деленную на s. В свою очередь s вычислено по тем же данным и взято со специально подобранным множителем. Для нормального распределения разброс этих чисел должен быть приблизительно симметричным относительно единицы. Однако для данных Пирса все 24 числа меньше 1, что чересчур много. Столбец (6) показывает, сколько измерений отклоняется влево, вправо и всего от среднего более чем на 3,1 выборочного стандартного отклонения (3,1s). Для нормального распределения в среднем одно наблюдение из 500 могло бы отклониться от $\bar{x}$ более чем на 3,1s, по «полнаблюдения»на каждое направление, в то время как данные Пирса дают в среднем 5,6 отклонения на 500. Аналогично меры асимметрии и эксцесса почти всегда положительны вместо того, чтобы колебаться около нуля, как это подобает нормальному распределению. К тому же согласно данным столбца (3) слишком много измерений расположено «вблизи» среднего опять-таки в масштабе s по сравнению с нормальным распределением. Так что несоответствие данных Пирса нормальному или гауссовскому закону вполне очевидно.

На это можно было бы резонно возразить, что ни один хороший современный специалист не сочтет время между импульсом и реакцией распределенным нормально, а предположит, что близким к нормальному, может быть, окажется распределение логарифма времени. Взглянув на таблицу илл. 1.4.1, однако, увидим, что и эта модификация неудовлетворительна. Действительно, число наблюдений, смещенных влево более чем на 3,1s, уже теперь в 3 раза превышает среднее для нормальных распределений, а логарифмирование его еще *увеличит*.

Характерные отличия данных Пирса от «нормальных» не обусловлены ни дискретностью, ни большим расхождением в форме. Столбцы (2) и (3), в общем, согласуются с остальными как в «средних», так и в ежедневных колебаниях. Как предсказывает «правило Винзора (Winsor)»*, распределения таких больших выборок для возникающих на

*Винзор предложил эмпирическое правило усреднения, состоящее в том, что заданное число крайних точек с обоих концов упорядоченной выборки не отбрасывается, а отождествляется с крайними из остающихся, после чего находится средняя арифметическая. Тем самым как бы сохраняются степени свободы. Подробности можно найти в статье Тьюки, процитированной в конце этой главы, и в кн.: Смоляк С. А., Титаренко Б. П. Устойчивые методы оценивания. М., Статистика, 1980, с. 155—158. — *Примеч. ред.*

практике распределений, кроме дискретных, вполне «нормальны в своей средней части». Однако то нормальное распределение, которое годится для описания центральной части выборочного, должно иметь разброс лишь примерно в 3/4 от наблюденного значения стандартного отклонения.

Медианное значение в столбце (2) для всех дней, кроме одного, равно 0,756, т. е. оценке дисперсии «нормальной части» распределения вблизи значения $(0{,}756)^2 \approx 0{,}57$. Так что более чем 40% ($\approx$100%—57%) наблюденной дисперсии обусловлены тем, что «хвосты» разбросаны больше, нежели у нормального распределения. Мы вновь видим, что данные Пирса резко отличаются от нормально распределенных.

1.5. ВИДЫ НЕНОРМАЛЬНОСТЕЙ И РОБАСТНОСТЬ

Когда распределение не имеет гауссовской формы, то природа этой негауссовости может быть различной.

1. **Дискретность и нерегулярность.** (1) Обычно измерения или наблюдения не представляются числами от $-\infty$ до $+\infty$, их допустимые значения ограниченны и часто кратны некоторому малому числу (например, число детей и браков в городе не может быть дробным, цены на рынке, как правило, кратны 1/20 рубля*, многие измерения проводятся в конкретных единицах, например в миллиметрах*). (2) Кроме того, как реальные, так и теоретические распределения подвержены и иным нарушениям, скажем, обусловленным предпочтением наблюдателя к определенным числам и даже таким, что вызваны появлением теоретических распределений вроде выборочного распределения рангового коэффициента корреляции в «нулевой» ситуации (когда связываемые нами величины в действительности независимы). Для наглядности последовательные ординаты этого дискретного распределения соединены прямыми ** (илл. 1.5.1).

2. **Резкие различия в форме.** Сюда мы включаем такие различия, которые надежно обнаруживаются уже в выборках объемом, скажем, от 50 до 100, наподобие резкой асимметрии χ^2 (или F) при малом числе степеней свободы. Обрывистые концы равномерного распределения, которое распространяет вероятность равномерно на отрезке, — вот пример такого различия. На илл. 1.5.2 показаны плотности распределений χ^2 с 1 и 3 степенями свободы, на илл. 1.5.3 изображается плотность равномерного распределения, а на илл. 1.5.4 — плотность симметричного треугольного распределения, которое в данном случае представляет собой распределение суммы двух независимых и равномерно распределенных (одинаковых) случайных величин, причем каждая из них задана на отрезке длиной $L/2$ со средним значением суммы μ.

3. **Малые различия в центральной части.** Их трудно отделить от предыдущих нарушений и они редко имеют значение сами по себе.

---

*В оригинале американские единицы измерения. — *Примеч. ред.*

**Аналог плотности. — *Примеч. пер.*

4. **Поведение на «хвостах».** Обнаруживается с трудом, однако часто оно важно, поскольку даже несколько резко выделяющихся значений, далеко отстоящих от основной массы наблюдений, могут, например, изменить выборочное среднее весьма значительно, а выборочную дисперсию — катастрофически.

В обычных при измерениях ситуациях поведение на «хвостах» распределений одновременно и наиболее важно, и наименее регулярно связано с целым. С этими связями большие трудности, поскольку изменения на «хвостах», вызываемые сильнодействующими факторами, увы, появляются лишь в малой доле наблюдений, а значит, они и трудно обнаруживаются в выборке, и легко вызываются мимолетными причинами. Это весьма важно, поскольку далекие «хвосты» распределения индивидуальных значений могут дать весомый вклад во многие статистики.

Рассмотрим случай, где, быть может, из-за действий человека есть один шанс из тысячи огромного отклонения (когда какое-то наблюдение отстоит поразительно далеко от некоторого центрального значения, вроде среднего или медианы). В выборках объема 100 одно такое отклонение должно в среднем приходиться на 10 выборок, или около 100 таких отклонений — на 1000 выборок. Так что 10% всех выборочных средних попадут под их воздействие. Это может сильно изменить 5%-ные точки. Если такой «случайный выстрел» окажется достаточно далеко, то ясно, что он может сильно изменить выборочное среднее. Совершенно иначе, чем выборочное среднее, ведет себя стьюдентово t, хотя и оно есть функция наблюдений. Воздействие одного наблюдения, сильно отличающегося от среднего остальных и от ожидаемого значения, скажется на приближении t к $+1$ или -1. Это обстоятельство, например, при работе с 95%-ными доверительными интервалами, приведет к уменьшению процента риска, обманывая тем самым в надежности; однако можно привести много таких примеров, где, наоборот, «значимое» становится «незначимым». Реально же нам нужны правильные ответы, а не только осторожные формулировки. И часто использование других критериев значимости позволяет избежать таких неприятностей и приближает нас к описанию действительного положения дел.

Вот два желательных свойства, которые суть виды *робастности*, виды потери чувствительности к нарушениям нормальности. Первое — терпимость к ненормальности на «хвостах» распределений — назвали *робастностью* (*устойчивостью*) *к предпосылкам*, примером ее могут служить доверительные интервалы для медианы μ, имеющие 95%-ную вероятность накрыть μ, из какой бы совокупности ни бралась выборка. Мы имеем именно такой вид робастности, когда, например, строим доверительный интервал для медианы по критерию знаков. Второе свойство — высокая эффективность, невзирая на ненормальность «хвостов», — называется *робастностью* (*устойчивостью*) *к эффективности*. Оно иллюстрируется доверительными интервалами для μ, которые проявляют тенденцию к постоянству для любых приблизительно нормальных распределений (например, для нормальных в средней части распределения) и стремятся к наилучшим интервалам, какие можно

было бы построить, если бы действительная форма распределения была известна. Такие процедуры уже есть [Gross A. M. (1976)].

Обратите внимание на то, что большинство аналитических работ по воздействиям «ненормальности» рассматривают лишь то, что происходит, когда поведение на «хвостах» согласуется с результатами наивной экстраполяции поведения центра распределения. Действительное положение вещей, как правило, иное, поскольку многие причины, могущие повлиять на каждое из наблюдений, вроде *грубых* ошибок, действуют лишь изредка. В результате фактические «хвосты» редко согласуются с основной частью распределения и, похоже, плохо состыковываются с ней, когда в очень больших выборках начинает проявляться их реальное поведение.

Важными могут оказаться два случая несогласованного поведения «хвостов». Различие между резко ненормальными и «хвостатыми» распределениями, в общем, не проявляется в выводах, которые весьма похожи и часто равно неудачны. Различие заключается скорее в том, что резкая ненормальность обычно обнаруживается и даже ярко проявляется, а разбросанные «хвосты» часто ускользают и от внимания, и от тщательного исследования.

Рассмотрим смеси нормальных распределений как характерные примеры распределений с растянутыми («разбросанными») «хвостами». Например, можем взять распределение, образованное из двух нормальных распределений, с одинаковыми средними, но разными стандартными отклонениями. Почти все измерения берем из распределения с меньшим стандартным отклонением, но небольшую часть, скажем 1%,— из распределения с втрое большим стандартным отклонением. Такие смешанные распределения иногда называются *загрязненными*, так что это распределение можно назвать 1%-но загрязненным с трехкратным растяжением.

Мы еще должны остановиться на *относительной эффективности*. Грубо говоря, если две оценки одной и той же величины имеют разные дисперсии, то отношение меньшей дисперсии к большей называется относительной эффективностью оценки с большей дисперсией. Для уточнения нашей формулировки распределения двух статистик должны иметь аналогичную форму, а их дисперсии — быть примерно одинаковыми, кратными $1/n$. К счастью, лучшие оценки, как правило, таковы. В этих условиях сравнительная эффективность удовлетворительно оценивает отношение объемов выборок, требуемых для того, чтобы с помощью обеих статистик получить один и тот же результат. Так, например, для больших выборок из нормального распределения выборочная медиана имеет дисперсию, приблизительно равную $(\pi/2)\sigma^2/n$, а выборочное среднее — σ^2/n. Мы скажем, стало быть, что эффективность медианы $2/\pi \approx 0{,}64$. Другими словами, выборка объемом 100, в которой выборочная медиана используется как оценка центра распределения, приблизительно равноценна выборке объемом 64, в которой для тех же целей используется выборочное среднее.

Рассмотрим последний случай, где мы сравним нормальное распределение с 1%-но загрязненым, определенным выше. «Хвосты» нормального распределения этим загрязнением удлиняются так мало, что

нужны тысячи наблюдений для надежного обнаружения самого факта загрязнения. Тем не менее его влияние на статистики от результатов наблюдений может быть вполне ощутимым. При больших выборках можно сравнить две оценки размаха: (1) основанную на стандартном отклонении, s, где суммируются квадраты отклонений; (2) основанную на среднем отклонении, где суммируются абсолютные величины отклонений ($\Sigma \mid y_i - \bar{y} \mid /n$). Для относительной эффективности получим [Tukey J. W. (1960)]:

для нормального — среднее отклонение 88% от s;

для 1%-но загрязненного — среднее отклонение 144% от s.

Отсюда ясно, что и совсем малые различия в форме распределения могут сильно влиять на относительную эффективность и тем самым на сравнительное достоинство разных методов.

1.6. РОЛЬ РАЗМЫТЫХ ПОНЯТИЙ

Для эффективного анализа данных мы вынуждены рассматривать *размытые понятия*, которые допускают определения многими способами. Чтобы разобраться во всем многообразии конкретных понятий, приходится снова и снова возвращаться к понятиям более простым и менее точным. Возьмем простой пример.

Большинство начинающих изучать статистику знают, что стандартное отклонение — одна из полезных характеристик распределения, которую оценивают. Но идея стандартного отклонения представляет серьезную трудность для понимания и использования, а обучающиеся не имеют, как правило, достаточно времени, чтобы разобраться, как и где его применять. Давайте восполним этот пробел.

Первый шаг от размытого к конкретному начинается с грубой качественной идеи о том, что одни распределения более широко растянуты, а другие упакованы более плотно, сконцентрированы, и с понимания того, что было бы хорошо найти для *разброса* какую-нибудь числовую меру. Это ведет нас к следующей последовательности идей:

- разбросы различны;
- какая-нибудь числовая мера может быть полезной;
- одна из числовых мер — стандартное отклонение;
- как мы определим, хороший ли это выбор?

Заметим, что пока мы говорили лишь об идеальном распределении или обо всей совокупности, так что вопросы о выборке или оценивании даже не продуманы. До сих пор мы выясняли, на какие вопросы мы хотим ответить, а не то, как найти на них удовлетворительные ответы.

Вот типичные ответы на последний вопрос «за» стандартное отклонение.

1. Определение стандартного отклонения (для идеального распределения) довольно ясно и широко применяется.

2. Поскольку обычно приходится иметь дело с выборочным (взвешенным или нет) средним либо с массой других традиционных линейных комбинаций наблюдений (скажем, $5 + 2x + 3y$, где x и y — наблюдения), стандартное отклонение (или его квадрат, дисперсия) ока-

зывается чрезвычайно полезной мерой, так как существует известная взаимосвязь между дисперсией определяемой статистики и дисперсией исходного распределения, которая к тому же не зависит от формы распределения. Так, если x и y — наблюдения, a и b — константы, σ_x, σ_y и σ_{ax+by} — стандартные отклонения этих наблюдений, а ρ — коэффициент корреляции между ними, то

$$\sigma^2_{ax+by} = a^2\sigma^2_x + 2ab\rho\sigma_x\sigma_y + b^2\sigma^2_y.$$

3. Для дисперсии между выборкой и совокупностью некоторые известные связи «в среднем» сохраняются (а значит, и для стандартного отклонения, квадратного корня из нее). Так, математическое ожидание или среднее значение s^2 равно σ^2, где s^2 — выборочная дисперсия $s^2 = \Sigma\,(y_i - \bar{y})^2/(n-1)$ и σ^2 — генеральная дисперсия.

4. Если ограничить себя каким-нибудь одним типом распределения (все нормальные, все равномерные, все симметричные треугольные и т. д.), то любые две меры разброса будут различаться лишь на фиксированную константу (множитель), и несущественно, какую именно меру разброса в генеральной совокупности выбрать. (Вспомним из параграфа 1.3, что многие семейства тесно связанных распределений, например все β-распределения, не имеют единой формы.) Следовательно, почему бы не выбрать стандартное отклонение?

С другой стороны, аргументы «против» стандартного отклонения.

● У некоторых распределений стандартные отклонения бесконечны. Действительно, мы видели примеры t-распределения с 1 и 2 степенями свободы (примечание 1 к илл. 1.2.2). Когда такое распределение разбросано в два раза шире, чем другое, их стандартные отклонения не скажут нам об этом, хотя масса других мер разброса, в том числе межквартильный размах, вполне работоспособны.

● Усреднение — весьма шаткий путь для свертки выборочных данных из распределений с бесконечным стандартным отклонением, так что (2) здесь непригодно.

● Даже когда дисперсия генеральной совокупности бесконечна, то выборочные дисперсии все равно конечны, так что и (3) дает в этом случае очень мало, если не совсем ничего.

● Если вид распределения с конечной дисперсией изменить так, чтобы дисперсия сделалась бесконечной, связь стандартного отклонения с другими мерами разброса, которые остаются конечными, резко изменится.

● Любая ничтожная вероятность достаточно сильного отклонения может сделать дисперсию бесконечной.

● Следовательно, в большинстве случаев разница между распределениями с бесконечной дисперсией и с конечной дисперсией, но сильно разбросанными «хвостами» невелика, так что на практике (4) дает нам не много, если вообще дает что-либо.

Мы вовсе не хотим сказать всем этим, что стандартное отклонение — плохой выбор. Во многих случаях он просто идеален для описания генеральной совокупности, зато в других — он весьма далек от идеала.

Подведем итог:

● стандартное отклонение есть только *выбор*;

●чтобы понять, что это лишь выбор, и решить, хорош ли он, нам нужна размытая идея о *мерах разброса.*

Мы снова и снова будем возвращаться к размытым понятиям, таким, как разброс, чтобы облегчать понимание и определение полезности конкретных понятий, таких, как стандартное отклонение, размах и межквартильный размах. Иногда частное понятие возникает первым и тогда неточно формулируемое размытое понятие может помочь в его рассмотрении. Чаще, однако, размытые понятия возникают первыми и ведут к выбору и определению соответствующих частных понятий*.

1.7. ЕЩЕ РАЗМЫТЫЕ ПОНЯТИЯ

Студент, изучающий элементарную статистику, может научиться классифицировать одни экспериментальные результаты как «значимые», а другие — как «незначимые». Затем его или ее можно обучить и тому, что выборочные средние, а также другие статистики, рассчитываемые по выборке, могут порождать «доверительные интервалы» для генеральных средних. Это — конкретный путь реализации другого размытого понятия: «количественные показатели неопределенности».

Здесь ход мысли примерно таков:

●наблюденные числовые «сводки» (такие, как выборочное среднее) для малых или средних выборок не совпадают с теми, что можно было бы рассчитать «точно так же», имея бóльшую выборку или всю генеральную совокупность;

●мы никогда не знаем различий между выборочными оценками и значениями для генеральной совокупности (если бы знали, то могли бы скорректировать), однако мы можем высказывать предположения об их разумной величине;

●эти предполагаемые величины различий будут зависеть от обстоятельств, и очевидно, что одни выборочные «числовые сводки» окажутся более неопределенными, чем другие;

●таким образом, есть основания для создания «количественных показателей неопределенности».

В настоящее время используются различные конкретные подходы к численным показателям неопределенности. Целый класс выборов, часто взаимосвязанных, представляют доверительные интервалы. (Дать 90%-ный доверительный интервал — совсем не то же самое, что 99%-ный; однако в простых случаях информация, даваемая одним, часто в грубом приближении пересчитывается в информацию, даваемую другим.) Кроме доверительных интервалов, есть еще не только критерии значимости, но и много других подходов к количественным показателям неопределенности.

---

*Идея «размытых понятий»—из тех идей, что носятся в воздухе (см.: Налимов В. В. Вероятностная модель языка. 2-е изд. М., Наука, 1979; Заде Л. Понятие лингвистической переменной и его применение к понятию приближенных решений. Пер. с англ., М., Мир, 1976). — *Примеч. ред.*

Мы не можем плодотворно размышлять об этих подходах, не говоря уже о том, чтобы эффективно их использовать, не пользуясь более простым понятием: оценка неопределенности «числовой сводки».

Мы уже ввели выше, не акцентируя это, еще одно важное, но размытое понятие. Термин «числовая сводка» был проиллюстрирован примером выборочного среднего. Мы надеялись, что обычного смысла, вложенного в корни слов «числовая» и «сводка», хватает для их понимания. На деле еще более общее понятие охватывает все те «вещи», что извлекаются из массы данных для того, чтобы выявить по крайней мере одну из особенностей, ей присущих. Таким понятием и может стать «числовая сводка» или «графическая сводка».

Мы назовем любую такую «сводку» *индикацией* (указанием) и будем настоятельно требовать от статистика при встрече с ней учета лишь двух обязательств:

- индикация должна отличаться от анекдота тем, что к участию в ней допущено каждое наблюдение. (Анекдоты обычно учитывают одно наблюдение или малое их число);
- она должна быть выражена таким образом, чтобы по крайней мере некоторые из тех, кто заинтересован в предмете, могли осмыслить ее интерпретацию.

Эти условия важны, однако они не должны пониматься слишком всерьез. Среднее из выборки объемом 15 есть индикация, но и медиана этой выборки тоже, так как каждое наблюдение вносит в нее свой вклад в том смысле, что, хотя зримая связь с медианой и отсутствует, если одно данное наблюдение подходящим образом изменить, то изменится и медиана. Каждое наблюдение влияет на медиану, даже если малые изменения в одном наблюдении (каждое из которых будет действовать на среднее меньше, чем на такое индивидуальное значение, но все же будет действовать) ни на йоту ее не изменят.

И среднее, и медиана принадлежат к частному классу индикаций, которые можно назвать *типичными значениями* или *центрами*, или, на более техническом языке, *мерами положения*. В особенности для «очень приличных» данных, когда наблюдения подчиняются явно выраженному колоколообразному распределению, и среднее, и медиана обычно показывают, где находится центр распределения. (Если же концы сильно разбросаны и достаточно асимметричны, то среднее может приводить к ошибкам.)

Меры разброса, меры связи, все результаты большинства стандартных приемов статистического анализа — суть индикации, точно так же, как и интересные извивы кривых, бóльшие неровности на графиках слева, чем справа, или очевидное расслоение диаграммы рассеяния на компактные группы точек.

В учебниках статистики индикация не выделяется, вместо этого внимание сосредоточивается на том, как выразить неопределенность при заданной индикации. Из-за внушения студентам важности показателей неопределенности, где только возможно (и оправдано), видимо, неизбежно создается впечатление, что сами по себе индикации, неопределенности которых не оцениваются, ничего не стоят. Это, конечно, совсем не так. Мы нуждаемся в оценках неопределенности и

часто, и остро, но нам столь же нужны и индикации с неоцениваемой неопределенностью, особенно, когда такая оценка невозможна или неэкономична.

1.8. ИНДИКАЦИЯ, ПОДСЧЕТ И ВЫВОДЫ

После совершенного Стьюдентом переворота были развиты многие пути формального вывода: сегодня у нас есть выбор между тремя уровнями (или этапами) статистического анализа. Мы можем иметь дело с любым из них или со всеми вместе:

- чистая индикация; в ней, например, мы обращаемся только к первичным статистикам, таким, как средние или процентили, не пытаясь оценивать неопределенность;
- подсчет или расширенное оценивание; здесь вычисляются и исследуются и первичные, и вторичные статистики, например среднее и оценка его стандартного отклонения (причем, хотя ничего и не говорится о выводах, вторичные статистики интересны чаще всего не сами по себе, а лишь как инструмент для оценивания неопределенностей);
- формальный вывод; здесь довольно точная детализация или описание удобной математической модели позволяет нам по меньшей мере увидеть, как связать наши неопределенности в гармоничное целое — через доверительные интервалы, критерии значимости, фидуциальные построения, апостериорные распределения или функции правдоподобия.

В ряде конкретных ситуаций третий уровень анализа для нас закрыт скорее всего по некоторым из следующих причин:

- данные таковы, что некоторым из важных источников изменчивости не удалось проявить свое действие;
- сами источники оказываются явно неслучайными;
- ни один из методов формального вывода не пригоден; одни — не разработаны, другие — разработаны, но основаны на столь жестких предположениях, что их использование становится неразумным или приводит к заблуждениям (или же требует слишком трудоемких расчетов).

Иногда же мы действительно утрачиваем (как, например, во втором из этих случаев) даже самую возможность осмысленного расчета вторичной статистики. Тогда нам, по-видимому, остается лишь уповать на чистую индикацию.

Когда мы соприкасаемся с индикацией в чистом виде, то появляется принципиальное различие между индикациями, которые отмечают нечто определенное, и теми, что дают нечто полезное, но неопределенное, как некий график, который побуждает смотрящего на него воскликнуть: «Взгляните, в поведении кривой есть странность!». Индикации первого вида — это так называемые оценки. Их индикаторы — также «оценивати». И оценки, и оценивати обсуждаются в параграфе 2.4.

Такое различие между «оценивателями» и «неоценивателями» очень похоже на различие между количественными и качественными индикациями, хотя это и не одно и то же. Большие величины χ^2, например,

часто указывают на то, что некоторая гипотеза не проходит, однако же сама по себе величина χ^2 — это число, не имеющее какого-либо значения для реальной ситуации.

РЕЗЮМЕ. АНАЛИЗ ДАННЫХ

Стьюдент в 1908 г. дал альтернативу бесконечной лестнице, в которой оценивание изменчивости каждой статистики вело к следующей статистике, изменчивость которой оценивать еще труднее.

Тремя десятилетиями позже статистики научились обходиться без лестницы, но и без тех жестких предположений, какие делал Стьюдент; для этого ввели «непараметрические», или «свободные от распределения», методы.

Главная ценность работы Стьюдента заключается в том, что он (1) продемонстрировал возможность работы и с малыми выборками, (2) дал числовую оценку того, что же это значит в одном важном случае, (3) построил таблицы для этого случая. Вместе с тем его работа, не по его желанию, очень уж облегчила (1) пренебрежение к «если... сделаны предположения», (2) преувеличение роли «точности» в других задачах и (3) замедление атак на проблему множественности.

Мы обсудили формы распределений, «хвосты» которых длиннее, чем можно было бы предположить по их серединам или «плечам» (мы говорим «растянутые хвосты» или «разбросанные хвосты»). *Фактические данные*, как показали результаты Пирса о запаздывании реакции, *часто имеют растянутые хвосты.*

В обычные отклонения от нормальности мы должны включить: (1) дискретность и нерегулярность, (2) резкие различия формы, (3) малые различия в центральной части, (4) поведение на «хвостах»; если перечислять в порядке увеличения трудности обнаружения, то наиболее важным будет (4), а наименее — (3).

Мы по необходимости коснулись не столько эффективности в гауссовском случае, сколько робастности (устойчивости) к эффективности.

Нечеткие понятия часто определяют ценность и уместность конкретных понятий — это было подробно показано при обсуждении стандартного отклонения.

Прослеживание за ретроспективой развития понятий «значимость» или «доверие» приводит нас к идее «количественных показателей неопределенности», а затем к некой «числовой сводке» как ключевому размытому понятию, которое в свою очередь приводит к концептуальному понятию «индикации».

Три основных уровня анализа данных целесообразно назвать *индикацией*, *подсчетом* и *формальным выводом.*

БИБЛИОГРАФИЯ

Burrau Ø. (1943). Middelfejlen som Usikkerhedsmaal. Mat. Tidskr. B., 1943, 9—16. (См.: Mathematical tables and other aids to computation, 2, 1946, 74-75.)

Cohen J. (1969). Statistical power analysis for the behavioral sciences. New York. Academic Press.

Fisher R. A. (1925). Applications of «Student's» distribution. — Metron, 5 (3), 90—104. (В ссылках на эту работу часто приводится дата 1926, которую привел и сам Фишер в библиографии к книге Statistical methods for research workers (см. русский перевод: Фишер Р. Статистические методы для исследователей. М., Госстатиздат, 1958); журнал, однако, вышел 1 декабря 1925 г.).

Gross A. M. (1976). Confidence interval robustness with long-tailed symmetric distributions. — J. Amer. Statist. Assoc., 71, 409—416.

Kendall M. G., Kendall S. F. H. and Smith B. B. (1938). The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. — Biometrika, 30, 251—273.

Mosteller F. and Bush R. R. (1954). Selected quantitative techniques. B: Lindzey G. (Ed.) Handbook of social psychology. Cambridge, Addison-Wesley, 289—334.

Odeh R. E. and Fox M. (1975). Sample size choice: charts for experiments with linear models. New York, Marcel Dekker, Inc.

Peirce C. S. (1873). Theory of errors of observations. Report of Superintendent of U. S. Coast Survey (for the year ending Nov. 1, 1870). Washington D. C. Government Printing Office. Appendix No. 21, p. 200—224 and Plate No. 27.

«Student» (Gosset W. S.) (1908). The probable error of a mean. — Biometrika, 6, 1—25. См. также в: «Student's» collected papers (edited by Pearson E. S. and Wishart J.), issued by the Biometrika Office, University College, London, 1942. Paper 2, 11—34.

«Student» (1927). Errors of routine analysis. — Biometrika, 19, 151—164. Также в: «Student's» collected papers (edited by Pearson E. S. and Wishart J.), issued by the Biometrika Office, University College, London, 1942, Paper, 14, 135—149.

Tukey J. W. (1960). A survey of sampling from contaminated distributions. B: Olkin I., Ghurye S. G., Hoeffding W., Madow W. G. and Mann H. B. (Eds.) Contributions to probability and statistics. Essays in honor of Harold Hotelling. Stanford, Stanford Univ. Press, 448—485.

Wilson E. B. and Hilferty M. M. (1929). Note on C. S. Peirce's experimental discussion of the law of errors. — Proc. Nat. Acad. Sci., 15 (2), 120—125.

ИЛЛЮСТРАЦИИ

Иллюстрация 1.2.1

Стандартные доверительные значения для *t* Стьюдента

f=степени свободы	1/40=2,5%	1/6=16 2/3%	1/2=50%	5/6=83 1/3%	39/40=97,5%	24/*f* (для интерполяции)
1	—12,71	—1,73	0,00	1,73	12,71	
2	—4,30	—1,26	0,00	1,26	4,30	
3	—3,18	—1,15	0,00	1,15	3,18	
4	—2,78	—1,10	0,00	1,10	2,78	
5	—2,57	—1,07	0,00	1,07	2,57	
6	—2,45	—1,05	0,00	1,05	2,45	4
8	—2,31	—1,03	0,00	1,03	2,31	3
12	—2,18	—1,01	0,00	1,01	2,18	2
24	—2,06	—0,99	0,00	0,99	2,06	1
∞	—1,96	—0,97	0,00	0,97	1,96	0

Примечания:

1. 2,5%- и 97,5%-ные точки вместе дают 95%-ные двусторонние доверительные пределы, или двусторонний 5%-ный критерий значимости.

2. Интерполяция по обратным числам степеней свободы вполне точна. Так, для 48 степеней свободы $24/f = 0{,}5$. Следовательно, соответствующая 97,5%-ная точка, лежащая между 2,06 и 1,96, равна 2,01.

3. Часто, когда предполагается использовать симметричные двусторонние интервалы, удобно соотносить их с распределением $|t|$ абсолютной величины t. Мы пишем $|t|_{0,95}$ для двусторонней 95%-ной точки $|t|$, которая дается в столбце 39/40. Аналогично $|t|_{2/3}$ — для двусторонней 2/3-точки, которая дается в столбце 5/6. Для двусторонних процентных точек t используем обычные обозначения.

Иллюстрация 1.2.2

Стандартные доверительные значения для модификации Бюррау t Стьюдента

$$\frac{f-2}{f}\cdot t = \frac{\bar{y}-\mu}{\sqrt{\Sigma\,(y_i-\bar{y})^2/n\,(n-1)}}\,,\quad \text{где } f=n-1.$$

f = степени свободы	1/40 = 2,5%	1/6 = 16 2/3 %	1/2 = 50%	5/6 = 83 1/3 %	39/40 = 97,5%	24/f (для интерполяции)
3	—1,84	—0,66	0,00	0,66	1,84	
4	—1,96	—0,78	0,00	0,78	1,96	
5	—1,99	—0,83	0,00	0,83	1,[illegible]	
6	—2,00	—0,86	0,00	0,86	2,00	4
8	—2,00	—0,89	0,00	0,89	2,00	3
12	—1,99	—0,92	0,00	0,92	1,99	2
24	—1,98	—0,95	0,00	0,95	1,98	1
∞	—1,96	—0,97	0,00	0,97	1,96	0

Примечания:

1. Дисперсия t равна бесконечности для двух и менее степеней свободы, так что для $f = 1$ и 2 формула Бюррау неприменима.

2. 2,5%- и 97,5%-ные точки вместе дают 95%-ный двусторонний доверительный интервал, или двусторонний 5%-ный критерий значимости.

Иллюстрация 1.3.1.

Три нормальные (гауссовские плотности вероятностей с различными средними и стандартными отклонениями: $\mu_1 = -2$, $\sigma_1 = 1$; $\mu_2 = 0$, $\sigma_2 = 0{,}5$; $\mu_3 = 4$, $\sigma_3 = 2$

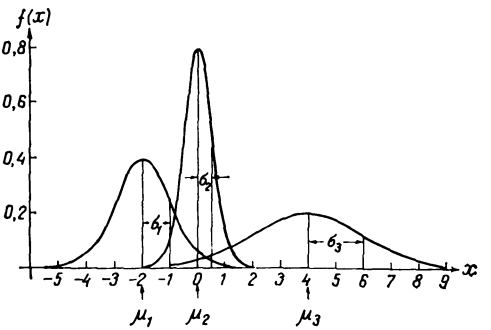


Иллюстрация 1.3.2

Коллекция плотностей β-распределений

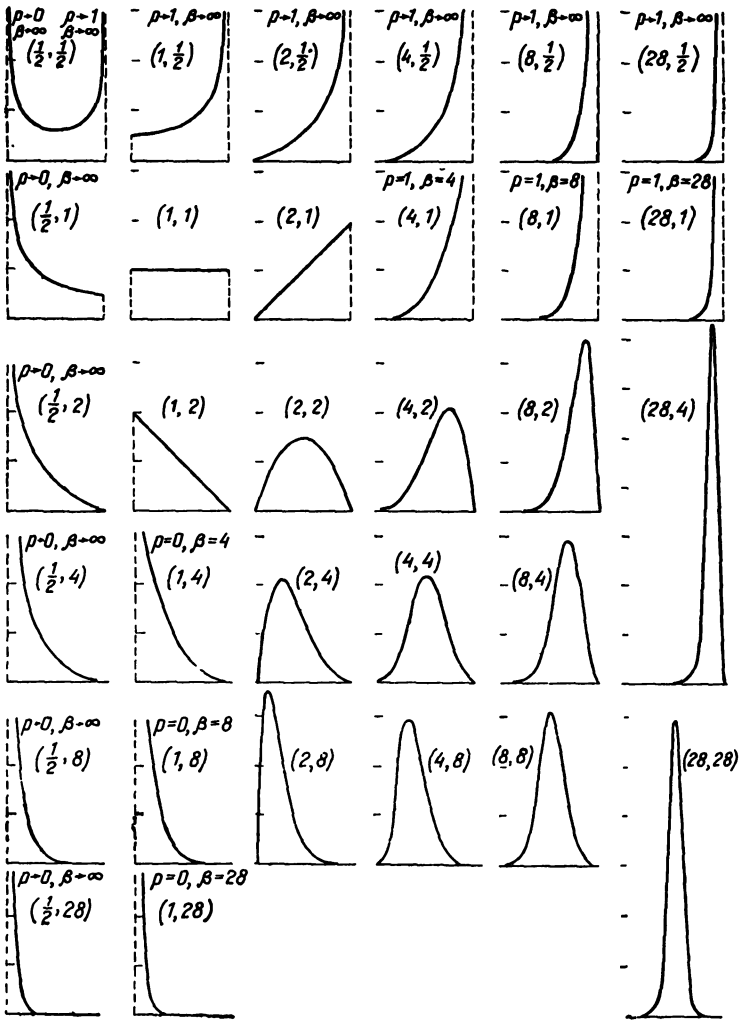


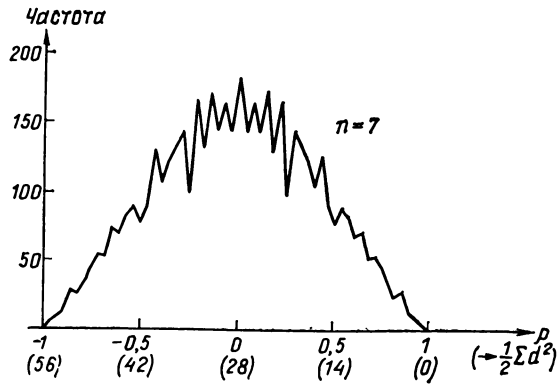
Иллюстрация 1.4.1

Суточные статистики Уилсона и Хилферти для анализа данных Пирса

Дни	(1) $\bar{x} \pm s_{\bar{x}}$ (миллисекунды)	(2) $\frac{Q_3 - Q_1}{2\,(0{,}6745s)}$	(3) Число отклонений от $\bar{x}$ на более чем 0,25 s: наблюдаемое	(3) ожидаемое	(4) Асимметрия	(5) Эксцесс $\beta_2 - 3$	(6) Ошибки, превышающие 3,1 s: отрицат.	положит.	всего
1	475,6±4,2	0,932	110	98	1,18	3,1	1	3	4
2	241,5±2,1	0,842	113	97	0,43	0,9	1	0	1
3	203,1±2,0	0,905	113	97	1,09	3,6	0	7	7
4	205,6±1,8	0,730	134	99	1,82	9,7	1	7	8
5	148,5±1,6	0,912	110	97	0,39	1,3	0	4	4
6	175,6±1,8	0,744	119	97	1,48	6,4	0	6	6
7	186,9±2,2	0,753	132	98	2,96	24,9	0	6	6
8	194,1±1,4	0,840	120	97	0,48	4,1	2	6	8
9	195,8±1,6	0,756	132	98	1,71	13,8	2	4	6
10	215,5±1,3	0,850	120	99	0,84	8,8	2	1	3
11	216,6±1,7	0,782	135	99	1,69	9,8	1	5	6
12	235,6±1,7	0,759	103	78	0,63	4,7	3	5	8
13	244,5±1,2	0,922	101	97	—0,22	2,6	6	1	7
14	236,7±1,8	0,529	192	99	5,74	63,6	2	3	5
15	236,0±1,4	0,662	162	98	1,68	27,9	4	4	8
16	233,2±1,7	0,612	162	98	6,39	90,6	4	2	6
17	265,5±1,7	0,792	123	100	0,21	4,3	3	5	8
18	253,0±1,1	0,959	114	98	0,27	1,8	0	4	4
19	258,7±1,8	0,502	187	99	10,94	143,9	1	3	4
20	255,4±2,0	0,521	179	98	7,71	91,4	0	3	3
21	245,0±1,2	0,790	120	99	0,23	8,2	3	4	7
22	255,6±1,4	0,688	142	99	5,27	68,1	2	4	6
23	251,4±1,6	0,610	158	98	2,73	31,1	0	3	3
24	243,4±1,1	0,730	113	98	—0,02	5,4	3	3	6
	Средние						1,7	3,9	5,6

Иллюстрация 1.5.1

Распределение рангового коэффициента корреляции для $n = 7$



Для получения вероятностей нужно ординаты делить на $7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5040$. Если полусумма квадратов разностей $1/2\Sigma d^2$ не целое число, то вероятность равна нулю.

Пример. Возьмем две согласованные ранжировки $\begin{matrix}1&2&3&4&5&6&7\\3&2&1&4&5&6&7\end{matrix}$; тогда сумма квадратов $\Sigma d^2 = 2^2+0^2+2^2 = 8$ и $1/2\Sigma d^2=4$ [Kendall M. G. and Smith B. B., 1938].

Иллюстрация 1.5.2

Плотности χ^2-распределения с 1 и 3 степенями свободы (ст. своб.) соответственно

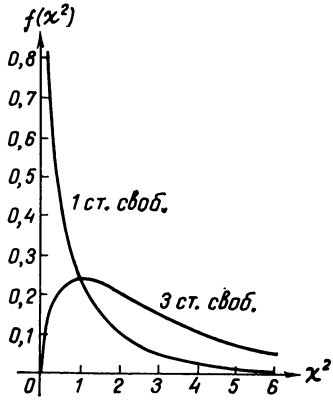


Иллюстрация 1.5.3

Плотность равномерного распределения на отрезке длины L со средним μ

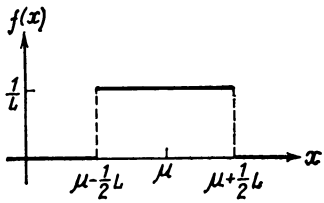
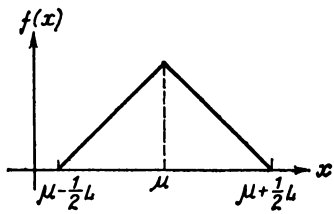


Иллюстрация 1.5.4

Плотность треугольного распределения



Глава 2 ● ИНДИКАЦИЯ И ИНДИКАТОРЫ

Индикация элементарна, важна и пренебрегаема.

Вначале, чтобы излечить от пренебрежения, мы покажем индикацию как нечто ценное, а часто просто самое лучшее из того, что вообще можно сделать. Мы покажем, как разумно можно выбирать индикаторы и сколь осторожно надо при этом действовать, о чем говорят перепроверки. И покажем, что графики очень ценны чуть ли не только из-за их функции индикаторов.

2.1. ЗНАЧЕНИЕ ИНДИКАЦИИ

Один из признаков сведущего в статистике исследователя — непоколебимая убежденность в том, что как бы фактически ни проводились обследования, эксперименты или наблюдения, их условия будут несколько различаться. Это убеждение вместе с соответствующими действиями открывают возможность эффективного использования имеющихся данных. Нет нужды всегда спрашивать в лоб: «А каковы эти различия?», но мы всегда должны подразумевать такие вопросы. Большинство считает неопределенность неудобной; историю анализа данных можно рассматривать как непрерывный поиск определенности для неопределенности. Все те, кто имеют дело с анализом и интерпретацией данных, должны уметь обращаться с неопределенностью там, где это действительно необходимо. Вот карикатура на один из рецептов: ко всякому результату примените критерий значимости и без колебаний отбросьте нулевую гипотезу, если превышен принятый уровень значимости, в противном же случае примите ее. Такой полный уход от действительности с ее неопределенностью, к счастью, редок, однако периодическое обсуждение этой крайности может помочь нам удержать равновесие.

Для исследователя ценность данных в первую очередь состоит в том, что они *указывают*, в чем себя проявляют. Проявление бывает довольно четким и явным или случайным, а чаще всего — смешанным. В ряде сфер деятельности, где данные — всего лишь число голосов или перфокарт, индикации, как правило, не рассматриваются или относятся к второстепенным деталям, хотя исследования в демографии и экологии составляют приятные исключения.

Подсчет голосов и документов представляет собой лишь один предельный случай индикации. Другой, и наиболее существенный, раз-

дел нашей весьма сложной программы анализа данных, включающий многомерный дисперсионный анализ, множественный регрессионный анализ, факторный анализ, латентно-структурный анализ, рассматривает оценку индикации как свою главную задачу.

Слово «индикация» означает размытое понятие, охватывающее широкий диапазон представлений от всех классических описательных статистик (таких, как среднее, медиана, мода, квантиль, стандартное отклонение, корреляция) на одном краю, до любых намеков и догадок, выуженных из данных в процессе раздумий над ними и могущих оказаться информативными для понимающего человека, — на другом. Примерами служат проявления сходства (все кривые выглядят *s*-образно или стреловидно) или общего поведения (частоты, похоже, уменьшаются примерно экспоненциально, или еще — хотя групповые средние меняются сильно, все стандартные отклонения близки по величине к обычно наблюдавшимся в подобных экспериментах). Индикация — это не только отдельные числа (разности, углы наклона и другие характеристики), но также неравенства и тенденции (женщин, по-видимому, умирает больше, чем мужчин; кровяное давление, наверное, растет с возрастом), а также особенности графиков и диаграмм (эти диаграммы рассеяния по форме напоминают бублик).

Чего в индикации *нет*, так это выводов или обсчетов неопределенности; индикация не содержит в себе ни доверительных интервалов, ни критериев значимости, ни апостериорных или фидуциальных распределений, ни даже стандартных ошибок.

Обработка вариации или неопределенности в большинстве дискуссий об анализе данных гипертрофирована, тогда как индикации уделяется совсем мало внимания. Это неравновесие возникает естественно: важные для оценки индикации соображения часто связаны с особенностями конкретных задач и поэтому обсуждаются с трудом. Так как изучением вариации часто пренебрегают из-за рвения начинающих в поисках закономерностей и элементарных проявлений, мы вынуждены фокусировать свое внимание на проблемах неопределенности; но для многих, наоборот, этого психологически достаточно, чтобы искать хоть какую-нибудь определенность. Вот лишь некоторые из множества причин, которые, переплетаясь, привели к появлению объемистой литературы об измерениях вариации и требованиях к ним, а еще к тому, что большинство вводных курсов статистики уделяют главное внимание поведению при заданной индикации. Конечно, все мы хотим адекватной оценки индикаций и их неопределенностей, но стоит ли отказываться от хорошего кекса только из-за того, что мы не можем покрыть его глазурью.

2.2. КОГДА ИНДИКАЦИИ ДОСТАТОЧНО?

Часто аналитик прекращает счет после достижения чистой индикации. (Ее дальнейшее обдумывание может создать решающее впечатление об имеющихся при этом неопределенностях.) Давайте взглянем на несколько таких случаев. Пусть, например, в одном взятом наугад

логопедическом классе из 23 школьников обучение ведет один учитель, применяющий определенный метод, причем чтение 5 учащихся значительно улучшилось, а чтение остальных 18 не изменилось. Индикация здесь — это утверждение о том, что данный метод исправляет чтение что-то около 1/4 учащихся. Эта цифра полна неопределенностей, о которых исходные данные ничего не говорят. Во всех случаях мы игнорируем опасность индикации (указания) на то, что этот метод гораздо хуже обычной практики, если другие учителя и методы улучшают чтение 85% учащихся.

Если мы пытаемся оценить неопределенность наших 5/23, то что нам делать с такими вопросами, как:

- из какой популяции случайно выбран этот класс?
- похожа ли эта популяция на те, из которых брались выборки обучавшихся другими методами?
- сколь же сильно личность учителя и его наклонности могут деформировать предписанный метод?

В этом примере хорошая оценка неопределенности вряд ли возможна. Несмотря на это, у нас есть индикация, которая может послужить основой для действия.

Суть дела отнюдь не так проста, как может показаться из этого примера индикации. Для получения из данных чувствительной индикации нередко нужен тонкий и кропотливый анализ, да еще с различными ухищрениями. Возьмем для примера упрощенный вариант задачи, которая бурно дебатировалась научными кругами в начале 60-х годов. Примем, что все требуемые условия сравнения белых с черными в стандартном испытании умственных способностей выполнены безупречно.* (Чтобы показать трудности индикации, нам нет нужды касаться сложного вопроса о точном измерении умственных способностей или биологической проблемы чистоты линий.) Тогда возраст, размер семьи, социально-экономическое положение, теснота жилья, время обучения, вид обучения, уклад домашней жизни и степень ориентации на умственную деятельность могут обеспечить некоторую основу для обсуждения этого вопроса. Но у нас уже есть каверзный вопрос о том, как же воспользоваться этими данными. Очевидно, нужны методы многократной проверки, и мы, по-видимому, сможем удовлетвориться индикацией, если только сумеем применить эти методы эффективно. Техника же индикации часто отстает от описательной статистики, приводимой в начальных курсах количественных методов.

---

*Стандартные испытания умственных способностей, получившие в США и ряде других стран широкое распространение, осуществляются с помощью специальных тестов. Читатель может получить представление об этих тестах, например, по кн.: А й з е н к Г. Ю. Проверьте свои способности! М., Мир, 1972. В последнее время усилилась критика тестов как средства всеобъемлющей характеристики личности.

К чести авторов следует сказать, что в описываемом примере, где было очень легко скатиться на расистские позиции, они остались в рамках подлинно научного подхода. — *Примеч. ред.*

ПРОБЛЕМЫ МНОЖЕСТВЕННОСТИ

Когда мы хотим разделить две группы объектов в зависимости от разницы в изменчивости как минимум одной из большого числа, скажем 100, измеряемых характеристик, имея по одному измерению каждой характеристики для каждого объекта, возникает задача другого типа. Мы можем, если действовать аккуратно, сравнить дисперсии какого-нибудь одного свойства для обеих совокупностей при помощи метода, надежного, скажем, при 95%-ном уровне значимости. Даже если верна нулевая гипотеза о равной изменчивости, все равно среднее число отдельных проверок, констатирующих значимость различия при 5%-ном уровне, как раз равно произведению этих 5% на число всех проверок, а именно $(0{,}05) \times (100) = 5$.

Значит, среднее число отдельных проверок, значимых при 0,05% (1/20 от 1%) равно $(0{,}0005) \times (100)$. Это можно понимать так: не менее одного подобного результата встретится приблизительно в 5% таких ситуаций (причем ситуация означает здесь повторение всего: 100 проверок и каждая на 0,05%-ном уровне). Получается, что если самое большее отношение дисперсий окажется значимым на 5%-ном уровне как максимально возможное из 100, то оно должно быть само по себе значимо на уровне 0,05%!

Такое крайнее значение неизбежно попадает в область, где точный, но неизвестный вид исходного распределения может сильно влиять на вероятность выхода за заданные границы нашего критерия-отношения, все равно будет ли это $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ или что-нибудь более «робастное». Пока еще нет статистических процедур, уверенно преодолевающих эту трудность. Остается лишь такой путь борьбы с этой неопределенностью, как повторное исследование, ограниченное теми немногими характеристиками, которые при первом изучении оказались заметно отличающимися для разных групп. Между тем, даже если из данных можно беспрепятственно извлечь искомые источники вариации, было бы совсем неразумно делать что-либо сверх индикации*.

Пример более «продвинутый», зато в несколько более знакомом направлении — поиск интересных индикаций, которые могут послужить намеками в исследовании дополнительных данных. Цель здесь не в заключениях и не в измерениях ради измерений, а только в охоте за интересными индикациями. Предположим, что мы просмотрели множество свойств, скажем 1000 или даже 10 000 вместо всего лишь 100, и отобрали те значения, которые кажутся интересными. Теперь опасность безоговорочного доверия к крайним процентным точкам даже больше. Здесь мы должны оставить наши вычисления с индикациями и рассматривать результаты скорее как намеки на то, что изучать дальше, нежели как нечто установленное.

По-видимому, ситуация несколько меняется, когда исследование и поиск «намеков» ведутся лишь с частью, скажем 1/3, общего числа

*Обсуждаемая задача близка к задачам статистической классификации, с которыми можно познакомиться, например, по кн.: Айвазян С. А., Бежаева З. И., Староверов О. В. Классификация многомерных наблюдений. М., Статистика, 1974. — *Примеч. ред.*

данных для добывания одного-двух десятков свойств, которые экзаменуются как на свойствах, не участвовавших в отборе, так и на объектах по остальным 2/3 данных. Здесь при переходе от исследования на первой трети к подтверждению на других 2/3 нам ничего не нужно, кроме индикаций; наша задача — выбрать наиболее наглядные из них для последующего эксперимента. (Сравните этот процесс с перепроверкой из параграфа 2.4.)

(Действительно, в очень близкой задаче выбора «наилучшего» среди многих сортов растений (или пород животных), часто разумно сделать первый отбор в условиях, когда известно, что наблюдаемые различия почти наверное недоказуемы [Yates F. (1950)]. Некоторые находят это парадоксальным. Селекционеры могут различать точно, если у них есть много экземпляров и очень мало сортов, понятно, с малой вероятностью того, что среди них есть и очень хороший сорт. Или же они могут изучать много сортов, но различать их сравнительно неточно, поскольку каждый сорт представлен малым числом образцов. Если идти последним путем, то вероятность появления хорошего сорта растет, зато из-за плохой точности снижается вероятность его обнаружения. Попытка достигнуть желанного равновесия и приводит к парадоксу*.)

Вот некоторые из важнейших причин для того, чтобы ограничиться индикацией: а) форма данных (их структура) маскирует главные источники изменчивости; б) у нас нет хорошего способа выяснить, сколь существенными могут быть различия в изменчивости данного вида индикации, различия, которые, как показывает практика, действительно могут встретиться, но их величину нельзя определить в одном-единственном наборе данных; обычно эта трудность возникает из-за многочисленности переменных или исследуемых свойств; в) предварительное изучение очень громоздких данных приводит к выбору нескольких индикаций для дальнейшей проверки.

Каковы бы ни были причина или причины для того, чтобы ограничиться индикацией, их следует зафиксировать.

В качестве последнего примера рассмотрим сравнение двух форм проективного теста типа теста Роршаха**для изучения влияния длины

*Здесь авторы затрагивают очень сложный вопрос о планировании эксперимента в растениеводстве и животноводстве. Желающие углубиться в эту проблему могут для начала обратиться, например, к работам: Доспехов Б. А. Планирование полевого опыта и статистическая обработка его данных. М., Колос, 1972; Снедекор Дж. У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. М., Сельхозиздат, 1961; Уишарт Д., Сендерс Г. Основы методики полевого опыта. М., ИЛ, 1958; Овсянников А. И. Основы опытного дела в животноводстве. М., Колос, 1976. Основоположником этого направления был Р. Фишер. — *Примеч. ред.*

**Пятна Роршаха, или тест Роршаха,—одна из наиболее интересных методик психодиагностики. На карточках-таблицах изображено нечто неопределенное, вроде размытых клякс в тетради плохого ученика. «Что это? На что похоже? Что вам напоминает?» — таков стиль вопросов. Ответы могут быть абсолютно произвольными. Ассоциативные комплексы ответов классифицируются, снабжаются баллами. Ответы могут быть устойчивыми по ассоциативности или нет. По характеру ответов ставится диагноз. Длина цепочки может быть разной, включая от 10 до сотен «рисунков». — *Примеч. пер.*

теста на надежность оценивания, причем одна из форм (короткая) может быть частью второй — длинной. Каждая форма теста допускает 32 разных «балла», характеризующих первичную запись ответа. Пусть, как это имеет место для пятен Роршаха, используется двухбалльная шкала, так что подсчет коэффициентов надежности вполне прост. Чтобы особенности выборки меньше влияли на итоговое сравнение, исследователь испытывает короткую и длинную формы на одних и тех же людях и одновременно. Далее он поступает просто: подсчитывает 32 коэффициента надежности для короткой формы и усредняет их, затем он повторяет эту процедуру для длинной формы и, наконец, рассматривает разность (или, быть может, отношение) этих средних коэффициентов надежности. Такая индикация, как надеется исследователь, разумно отвечает на исходный вопрос о сравнении надежностей длинной и короткой форм теста*.

Что же произойдет, если он попробует перейти от индикации к выводу? Статистик-консультант, видимо, сообщит ему, что все 64 исходных балла коррелированы друг с другом в неизвестной и, скорее всего, различной степени.

Здесь появляются $1/2 \cdot (63) \cdot (64) = 2016$ коэффициентов корреляции. И каждый из них требует еще преобразования (что не просто) в коэффициент корреляции между соответствующими коэффициентами надежности. Только после этого можно найти дисперсии и ковариации для самих показателей надежности и их средних, и, наконец, дисперсию их разности. Если для этого случая применима теория больших выборок, на что можно надеяться, имея 10 000 испытуемых, то можно оценить неопределенность разности. Пойдем ли мы на это, помня о требующихся усилиях?

В описанных условиях большинство исследователей удовлетворилось бы индикацией (в гл. 8 мы введем способ, позволяющий делать некоторые выводы даже в столь тяжелых положениях.)

Иногда результаты, которые в отдельности можно объяснить чистой случайностью, заслуживают внимания, поскольку они усиливают друг друга. Так, когда у нас есть параллельные результаты, полученные из независимых совокупностей данных, входящих в одно исследование, оценки легко взаимоувязываются. Если же параллельные оценки получены из перекрывающихся или взаимосвязанных данных, где, например, на семь вопросов отвечают одни и те же респонденты, результаты согласуются по направленности, но степень их коррелированности неясна. Или же, наконец, какой-нибудь результат современного исследования может оказаться параллельным с результатами других исследований.

Во всех подобных ситуациях может оказаться слишком расточительным отбрасывание результатов только потому, что они не значимы по отдельности или их совместную значимость нельзя удовлетворительно оценить.

---

*Тесты этого вида интенсивно развиваются. Относительно свежую сводку результатов этого развития можно найти в кн.: Соколова Е. Т. Проективные методы исследования личности. М., Изд. МГУ, 1980. — *Примеч. ред.*

2.3. ФИГУРА УМОЛЧАНИЯ

Когда мы изучаем данные ради получения ответа на конкретный вопрос, мы иногда сталкиваемся со столь четкими проявлениями, что ответ оказывается очевидным. Когда суть дела так ясна, и для анализа, и для сообщения обычно достаточен неформальный вывод.

Если данные красноречивы, то читатель или слушатель, как правило, лишь глянув на индикации, и сам «видит» ситуацию. Исследователь в таких случаях обычно говорит: «Никакие статистики не нужны, чтобы увидеть, что...», подразумевая, что простейшими отработанными статистическими приемами владеют все. За подобными утверждениями лежат некоторые представления о статистическом методе, хотя и не выставленные напоказ. Так, например, «Наступление 85 успехов в 100 испытаниях едва ли совместимо с вероятностью успеха 0,1» или «Стандартное отклонение разности явно больше 100, но даже если бы было 50, разности в 10 единиц вряд ли хватило бы для того, чтобы один метод предпочесть другому». Хотя подобные рассмотрения обычны, для «обхода» формального вывода может потребоваться особая изощренность или практический опыт. И многих из нас не удивило бы мнение неофита, что он или она все же не убеждены в необходимости таких длительных демонстраций очевидного.

Если данные еще более очевидны, то читателю или слушателю даются лишь индикации, а о выводах не говорится вовсе.

Такие примеры — отнюдь не об ограничении уровнем индикации в том смысле, как мы употребляем этот термин. Это просто случаи вывода, когда нет необходимости в каких бы то ни было формальных выводах. Неформальный вывод, возможно, выражаемый словами «очевидно» или «посмотрев... мы видим, что» либо совсем без слов, представляет собой одну из необходимых разновидностей вывода*. Так что даже отсутствие слов, сказанных или написанных, — это тоже вывод.

Говоря о ситуациях, допускающих ограничение простой индикацией, мы не собираемся включать сюда приятный случай неформального вывода. Мы касаемся в основном тех случаев индикации, когда не видно ни малейшего подхода к оценкам стабильности, изменчивости или достоверности индикаторов и, следовательно, нет оснований ни для неформального, ни для формального выводов.

2.4. ВЫБОР ИНДИКАТОРОВ

Аналитик часто раздумывает, какой индикатор использовать. Он или она сравнивает чувствительность различных индикаторов к ответам на характерные вопросы. Аналитик, по-видимому, сравнивает легкость их вычисления и часто осведомляется, какой из них дает более, а какой менее устойчивые результаты.

Проведение такого анализа дает гораздо больше, чем чтение о нем в нашей книге. И совсем неудачные индикаторы на самом деле

*Здесь трудно удержаться от ассоциации: в древней Греции геометры заменяли доказательство чертежом с простой подписью «Зри!». — *Примеч. ред.*

участвуют в борьбе за существование, но, как правило, вымирают. Однако, как мы видели в параграфе 1.3, даже малейшее расхождение в форме распределения, которое трудно заметить в выборках огромного объема (порядка нескольких тысяч), может всерьез изменить положение выборочных индикаторов. В указанном примере индикаторы играют роль оценивателей. Так, один из них, почти такой же, как все остальные в идеальных условиях, становится примерно в полтора раза лучше, едва только условия делаются более реальными. Вероятно, подобные сложности и тонкости будут возникать прежде всего там, где выбранный индикатор предназначается для чистой индикации.

Давайте сначала уточним смысл глагола «оценивать», а затем уже обсудим выбор из конкурирующих индикаторов, т. е. оценивателей. Когда индикатор что-либо оценивает? Что он оценивает?

Один из наивных ответов таков: «оценка оценивает параметр». Исторически слово «параметр» означает две совершенно разные вещи:

- численное значение специального символа в особой форме задания семейства распределений; так, семейство нормальных распределений задается параметрами μ и σ^2 либо μ и σ;
- любую численную характеристику распределения вроде медианы, которая частично характеризует совокупность чисел.

Ограничить «оценивание» оценкой специальных коэффициентов семейства распределений — значит потерять многое из того полезного, что заключает в себе это неопределенное понятие. Оно должно быть настолько общим, насколько это возможно. Так, если обычно полезно уметь сравнивать два оценивателя, когда исходные данные подчиняются гауссовскому закону, то нелепо не знать их поведения для других законов.

Следовательно, даже если аналитики точно знают, что именно они оценивают (случай не столь частый, как принято считать), им все равно нужно руководство для выбора оценки с учетом всего, что известно или кажется известным, но с сохранением ориентировки во множестве противоречивых обстоятельств.

Оцениватель — функция наблюдений, особый способ их сопоставления. Он может определяться некоторой арифметической формулой, вроде $\bar{y} = \Sigma x_i/n$, или только словами, как при отыскании выборочной медианы упорядочением и счетом. Мы различаем оцениватель и его значение, оценку, полученную из определенного набора данных. Оцениватель дисперсии $s^2 = \Sigma\,(x_i - \bar{x})^2/\,(n - 1)$ дает оценку 7 для трех наблюдений: 2, 3 и 7. Мы говорим, что s^2 есть оцениватель для σ^2 и называем σ^2 оцениваемым. В этом примере 7 оценивает σ^2.

Иногда оцениватель возникает первым, и тогда мы задаем вопрос: «Что он означает?». Мы гораздо охотнее скажем об *оцениваемом* (о том, что оценивается), что это мишень для оценивателя, чем, что это какой-то параметр. И наша задача — достижение гармонии между оцениваемым и оценивателем. Как же нам ее достигнуть? Насколько точно? И когда для этого существуют подходящие условия?

Такие вопросы пока не привлекали такого внимания исследователей, какого они заслуживают. Хотим мы этого или нет, но централь-

ную роль играет вопрос об объеме выборки. В реальных исследованиях мы пользуемся оценивателями в выборках фиксированного объема. Однако современные пути установления гармонии между оцениваемым и оценивателем, которые зависят от того, как мы можем учесть рост объема выборки, не совсем удовлетворительны.

Теперь при обсуждении этого вопроса разделим оцениватели, используемые для выборок заданного объема, на три класса:

1) оцениватели для сколь угодно больших выборок;

2) оцениватели для больших, но не сколь угодно больших выборок;

3) оцениватели, не представляющие интереса для больших выборок.

Класс 3 трудно обсуждать в нашем контексте. Он никогда не обеспечит искомую гармонию между оценивателем и оцениваемым.

Класс 2 часто встречается в реальной практике. Распространенный пример — выборочный размах. В нашем контексте часто можно получить практические ответы, лишь забывая, что данный оцениватель нельзя использовать для произвольно больших выборок и полагая, что мы имеем дело с классом 1.

Класс 1, пусть нереалистичный, но все же возможный, раз мы действуем так, как будто он есть. Но даже в этой утопической ситуации нельзя установить простых правил для гармонизации оцениваемого и оценивателя. Все, что можно сделать, это перечислить несколько разных случаев, в которых мы, вероятно, смогли бы подобрать оцениватель и оцениваемое вполне удовлетворительно. Вот эти случаи.

● Если среднее значение оценки (результат усреднения по всем выборкам данного объема) не зависит от объема выборки, то это значение — хороший кандидат для искомого оцениваемого, как в случае s^2 и σ^2 при конечной σ^2. Если вдобавок распределения оценок концентрируются около этого оцениваемого по мере роста объема выборки, то большинство сочтет подбор удавшимся. Исключения возникают, например, там, где среднее не совсем подходит для данного распределения, как в случае s^2 при распределении с бесконечной дисперсией.

● Если среднее зависит от объема выборки, но имеет предел, к которому стремится с ростом объема, как s^2 к σ^2 при конечном σ^2, то можно повторить все сказанное выше. Большинство статистиков предпочитают случай, когда зависимость от объема выборки слаба, как для s и σ при нормальном распределении и при n, скажем, не меньше 10.

● Аналогично, если медиана распределения выборочных оценок одна и та же для всех объемов выборок или если ее значение сходится к пределу с увеличением объема, то эти — общее и предельное — значения, видимо, окажутся вполне подходящими оцениваемыми. К тому же, особенно в больших выборках, медиана оцениваемого распределения редко оказывается неразумной характеристикой этого распределения.

● Предел, к которому стремится с увеличением выборки значение какой-нибудь разумной характеристики распределения оценки (см. параграф 1.7), отличной от среднего или медианы, по-видимому, также будет вполне удовлетворительным оцениваемым.

Эти правила расплывчаты и неаккуратны, однако они все же приводят к некоторым подходящим оцениваемым, что многие сочтут достаточным. Наше неопределенное и подозрительное отношение к этой проблеме возникает отчасти из-за малого числа исследований, послуживших основой наших замечаний, а отчасти из-за неудовлетворенности тем, что оцениваемое выбирается в зависимости от свойств совокупности выборок разного объема, не представленных в самом исследовании.

2.5. ОДИН ПРИМЕР ВЫБОРА ИНДИКАТОРА

Предположим, что кто-то может частично повторять наблюдения, которые ведут себя скорее всего как выборка из совокупности с вытянутыми, разбросанными «хвостами». Собрав эти данные, наблюдатель хочет свернуть их с помощью индикаторов положения с той точностью, на какую он только способен. Он или она рассматривает как возможность выборочное среднее всех наблюдений и находит, что сильно вытянутые и разбросанные «хвосты» наделяют его столь большой дисперсией, что оно как индикатор оказывается чересчур расплывчатым. Тогда аналитик переходит к выборочной медиане, поскольку она учитывает около 2/3 информации о параметре положения нормального распределения по сравнению с выборочным средним (действительно, в больших выборках, как отмечалось в 1.4, $2/\pi \approx 2/3$), и весьма вероятно, что ее качество даже улучшится для распределений с более разбросанными «хвостами».

Устойчивость выборочной медианы зависит от плотности распределения вблизи медианы всей совокупности. На илл. 2.5.1 показано распределение, для которого медиана — плохая мера положения.

Пока «длиннохвостое» распределение имеет в середине разумную плотность, наш наблюдатель предпочитает медиану, но не среднее (возможен ли еше лучший выбор?). Кое-кто может подумать, что медиана чересчур вариабельна, что она не похожа на слишком хорошего кандидата для метода «складного ножа» (гл. 8) и может потерять из-за этого существенную часть информации.

Альтернативой к медиане служит усеченное среднее. У выборки «подрезаются» ее разбросанные «хвосты» отбрасыванием некоторой доли измерений с каждого конца выборки. Допустим, например, что мы решили выкинуть по 10% наблюдений слева и справа и найти среднее арифметическое для оставшихся 80%.

Сколь хорошо это для нормального распределения? Удержим в памяти только то, что эффективность медианы — около 2/3 (более точно $2/\pi \approx 63{,}7\%$). Использование «усеченного» среднего возвращает часть остающейся 1/3 информации о генеральном среднем, содержащейся в выборке. Для симметричного усечения эта частичная прибавка меняется, очевидно, от 0% (для 50%-ного усечения каждого «хвоста», так что остается только выборочная медиана) до 100% (для 0%-ного усечения). Исследования, которые мы здесь не приводим, показывают, что эффективность растет быстрее, чем по линейному

закону, поэтому ради осторожности припишем усечению на α с каждой стороны (в долях единицы) значение эффективности

$$2/\pi+(1-2\alpha)\,(1-2/\pi)\approx 2/3+(1-2\alpha)\cdot 1/3;$$

это для нашего примера дает $2/3 + 0{,}8 \cdot (1/3) \approx 0{,}93 = 93\%$ информации о «положении». В выборках из распределений с «хвостами», разбросанными несколько более, чем в гауссовском, усеченные средние уже работают лучше. Усеченные средние вполне подходят для метода «складного ножа», описанного в гл. 8.

Наш наблюдатель выбрал некоторую индикацию. Она определяет «положение», ибо, если мы прибавим ко всем наблюдениям константу, то усеченное среднее изменится на ту же величину и в том же направлении. Работала ли до сих пор какая-нибудь определенная модель распределения? Нет, если не считать того, что должно существовать какое-то теоретическое распределение. Или был ли точно идентифицирован параметр, который надо оценить? Снова нет, хотя конкретный выбор и подразумевает некоторый класс, а именно значения, около которых концентрируются средние из выборок данного объема, усеченных на 10% с каждой стороны.

В нашем примере усеченного среднего с ростом объема выборки ее центральные 80% будут все более и более соответствовать 80% генеральной совокупности. Отсюда естественно выбрать за оцениваемое именно среднее центральных 80% генеральной совокупности.

ФАКУЛЬТАТИВНОЕ ДОПОЛНЕНИЕ

Мы получили ответ и выразили его понятийно. Для тех, кто питает склонность к математике и желал бы видеть результат в более четких формулировках, покажем следующее. Пусть θ — оцениваемый параметр (оцениваемое) и F — интегральная функция распределения. Тогда

$$\theta=\int_{-\infty}^{\infty} y\varphi\,(F\,(y))\,dF\,(y),\ \text{где}$$

$$\varphi\,(u)=\begin{cases}1{,}25=1/0{,}8 & \text{для } 0{,}1\leqslant u\leqslant 0{,}9;\\ 0 & \text{в остальных случаях.}\end{cases}$$

Мы выиграем еще больше, рассмотрев формулу для конечных выборок:

$$\text{Математ. ожидание }\{\text{усеченное на 10\% среднее}\}=\int_{-\infty}^{\infty} y\varphi_n(F\,(y))\cdot dF\,(y),$$

где, если запись $[n/10]$ означает долю (примерно 10% от n) отсекаемых с каждого края наблюдений, то

$\varphi_n\,(u)$ { стремится к нулю как $u^{[n/10]}$ или $(1-u)^{[n/10]}$ вблизи 0 и 1, главным образом концентрируется внутри и вблизи интервала (0,1; 0,9).

Этой формулой можно пользоваться при установлении объема выборки, требуемого для достаточного уменьшения вклада «хвостов» распределения в среднее.

Теперь нам понятно, что же такое оцениваемое и какие неприятдости возникают из-за разбросанных «хвостов» распределения, правна, *для выборки независимых наблюдений.*

2.6. ИНДИКАЦИИ КАЧЕСТВА: ПЕРЕПРОВЕРКИ

Результаты, полученные при помощи некоторых статистических методов, даже если они не всегда используются для предсказания, можно условно считать «предсказывающими» или «прогнозирующими». Множественная регрессия одной переменной по нескольким другим, например, допускает экстраполяцию, что, похоже, вообще есть стандартное свойство, присущее данным.

Практики часто разочаровываются в моделях типа уравнений множественной регрессии, так как «прогноз» удовлетворителен лишь для тех данных, которые участвовали в построении модели. Падение силы предсказания на «свежих» данных действует угнетающе.

Обсудим положение вещей на простом примере множественной регрессии, учитывая, что и наше понимание, и наши выводы с таким же успехом применимы ко множеству других методов. Когда мы говорим о «методике» множественной регрессии, то имеем в виду получение конкретного уравнения, например

$$z = 3{,}4x + 2{,}5y - 5{,}4.$$

Когда же говорим о «модели», то имеем в виду, например, выбор факторов, входящих в уравнение регрессии, и решения в отдельности для каждого фактора о том, использовать ли исходные измерения, их логарифмы, квадратные корни или же другие преобразования. Надо еще решить, стоит ли объединять факторы и комбинировать их в суммы, произведения, отношения или что-нибудь еще в этом роде. В нашем же случае «методика» включает и выбор модели, и поиск численных значений ее коэффициентов.

Когда мы пользуемся данными для получения представлений о некоторой методике, то стремимся ответить на вопрос: «Могу ли я ожидать, что выбранная методика оправдает себя на практике?». Даже после того, как все факторы, входящие в уравнение регрессии, уже выбраны, так что модель задана, коэффициенты все равно выбираются из бесконечного множества комбинаций возможностей таким образом, чтобы результаты подстановки в формулу приближались к данным настолько, насколько это возможно. Проверка действия методики на тех же данных, что породили результаты, почти равноценна «переоценке ценностей», поскольку оптимизация выбора из многих возможных методик волей-неволей настроена на максимальное использование всех и всяких особенностей именно этих, конкретных данных. Иногда шутят: «Оптимизация наживается на случае!». В итоге методика будет работать на этих данных лучше, чем на каких-либо других, которые могут появиться на практике. Кажущаяся степень подгонки модели в среднем будет гораздо лучше, чем действительная.

Никто не знает, как оценивать методику, используя различные части тех данных, по которым велся счет, не опасаясь возможных при этом ошибок. Другими словами, оценка требует определенного вида пере-

проверок. Мы различаем два уровня перепроверок — простую и двойную, причем простая распространена более широко.

● *Простая перепроверка.* Проверка методики на данных, отличных от тех, по которым считались коэффициенты.

● *Двойная перепроверка.* Проверка методики на данных, отличных как от тех, по которым выбиралась модель, так и от тех, что были использованы для расчета коэффициентов.

Второй уровень перепроверки, названный нами двойной перепроверкой по аналогии с тем, что медики называют «дважды слепым» исследованием, возможен лишь при переходе к «свежим» данным. Эти дополнительные данные лучше всего собирать после выбора модели и вычисления коэффициентов. Если добавочный сбор данных невозможен, то хорошие результаты может принести обращение к архивным данным, которые оставались неизвестными пока выбиралась и оптимизировалась модель (как в исследованиях [Macdonald N. J., Ward F. (1963)], [Miller R. G. (1962)], [Mosteller F., Wallace D. L. (1964)]. Для полной обоснованности проверки надо, чтобы данные, оставленные «под замком», были иными, чем те, по которым работала методика. Тогда ожидаемые источники вариации адекватно отражали бы то, что может встретиться в жизни. Например, хорошо было бы брать данные разных лет, разных исследователей или различающихся методик. Несмотря на высокие достоинства двойной перепроверки, мы не всегда можем ее себе позволить.

Достигнем ли мы двойной или будем вынуждены остановиться на простой перепроверке, современные вычислительные машины открывают нам новые горизонты. В классическом подходе с простой перепроверкой наличные данные делились на две (иногда больше) равные части. Одна использовалась для оптимизации, другая — для экзамена. Некоторые энергичные исследователи меняли затем обе части местами и повторяли весь процесс для извлечения из тех же данных добавочной информации. И хотя таким путем удавалось узнать больше, определенные подозрения не переставали возникать из-за неизвестной связи между двумя оценками качества.

Эти подозрения вновь показывают, что настойчивое стремление к статистическому выводу как *само*цели искажает отношение к индикации. Действительно, неизвестные связи между оцениваемыми компонентами должны уничтожать самую возможность использования степени их согласованности для точного установления ценности и *стабильности* комбинированного результата. С другой стороны, если каждая из оценок сама по себе достаточно обоснованна, то взвешенная комбинация двух или более оценок (неважно, сильно или слабо коррелированных), и столь же хороша, и не менее точна.

Тот, кто делит имеющиеся данные пополам и устраивает перепроверки, меняя части местами, использует *все* данные для оценки качества по оптимизации лишь на *половине* всего набора. Если у него к тому же так много данных, что их уменьшение наполовину или увеличение вдвое мало сказывается на качестве оптимизации, — это прекрасно, и ему почти нечего добавить к тому, что уже было сделано. Немногим так везет.

Когда машинный счет был дорог, даже *проведение* перепроверки с переменой мест казалось затруднительным. Сегодня мы смело можем браться за гораздо большее.

Фрэнк Йейтс [Frank Yates (1957)] предложил, а П. Маккарти [McCarthy P. J. (1976)] исследовал недавно многократное деление данных для получения добавочной информации. Предположим, что мы разделили данные на 10 равных частей. Тогда можно взять любые 9 из них, оптимизировать на этих 9/10 и проэкзаменовать на оставшейся 1/10. Если повторить это 10 раз, беря всякий раз новую 1/10 данных, то мы воспользуемся *всеми* данными, чтобы оценить качество, возникшее за счет оптимизации на *9/10* всех данных. Это во многих случаях заметно приближает нас к ответу на вопрос: «Какого качества, пусть приближенно, я могу ожидать от оптимизации по *всем* данным?»

На ЭВМ выполнить десять вычислений немногим более затруднительно, чем традиционные одно или два, поскольку мы повторяем одно и то же, лишь чуть увеличивая время счета и программирования.

В ряде случаев выгодно пойти еще дальше. Предположим, что мы опускаем одно наблюдение и оптимизируем без него, а затем экзаменуем на нем. Повторение такой процедуры с каждым из наблюдений выжмет из данных почти все. Если же мы будем всякий раз проводить полную оптимизацию, то столкнемся уже с серьезными вычислительными трудностями. Иногда удается легко рассчитать либо прямо, либо за счет подходящей аппроксимации, как повлияет на результат оптимизации пропуск определенной очень малой части данных. Тогда можно сравнить скорректированный результат с опущенным наблюдением. Таким образом, мы рассчитываем оптимизацию один раз для всех данных, а затем последовательно повторяем более простой расчет — расчет влияния одного неучтенного наблюдения — и выверяем каждым из них качество результата. Для практики этот подход очень привлекателен.

М. Стоун [Stone M. (1974)] дал обобщенную форму критерия перепроверки, применяемого для выбора и оценки статистического предсказания.

Недостаток всех видов простой перепроверки состоит в том, что очень часто контрольная выборка гораздо больше похожа на рабочую, чем на выборку, типичную для той генеральной совокупности индивидуумов или ситуаций, с которой мы хотим соотнести индикацию. Соответственно результат простой перепроверки слишком часто оказывается слабее (в неизвестной степени), чем это могло бы показаться.

Возможность перепроверки — одно из главных преимуществ многих автоматизированных программ оптимизации. Приближенный, «на глазок», волевым решением сделанный выбор методики зачастую может привести к очень хорошим результатам, и даже лучшим, чем формальная оптимизация, однако тогда нельзя как следует установить суть методики и механизм получения результата и еще нельзя быть вполне уверенным, что контроль производится на «независимых» данных. С другой стороны, на уровне простой перепроверки любая методика, порождаемая программной оптимизацией, легко прослеживается на совокупности данных, по которым считались коэффициенты.

Трудность анализа субъективных суждений снова возникает на уровне двойной перепроверки. Когда мы выбираем модель, вся совокупность использованных данных может быть ясна абсолютно всем. Часто выбор опирается (иногда с большим успехом) на длительный жизненный опыт, отдельный факт или услышанное когда-то предание в равной степени так же, как и на преимущества, даваемые конкретным методом анализа. Опытный исследователь будет твердо настаивать на проведении «дважды слепого» исследования, так как на выбор модели может лечь печать того множества данных, которые уже накопились у него в уме.

Перепроверка, вполне удовлетворительная на немногих последующих выборках, подтвердит предвидение исследователя и убедит нас в применимости его знаний, однако может и слишком успокоить нас насчет будущих применений таких перепроверок. Конечно, мы не решим спора между «плохо» и «хорошо». Исследователь хочет, разумеется, уверенно, насколько возможно, рассматривать весь спектр применений и тем самым быть готовым к другим действиям в других обстоятельствах.

РЕЗЮМЕ. ИНДИКАЦИЯ И ИНДИКАТОРЫ

Указывать (indicate) — значит, «появляться, чтобы проявлять». Иногда индикация, коль скоро мы нуждаемся в этом, составляет основу анализа. Индикация бывает количественной (числовой) и качественной.

Для выхода за уровень индикации нужна оценка ее неопределенности: чем достовернее, тем лучше.

Мы должны были коснуться и проблем множественности, так как совсем не одно и то же — один результат из 1000, отвергнутый на 5%-ном уровне значимости, или один из одного.

В задачах отбора хорошо спланированный эксперимент почти всегда приводит к индикациям, и лишь изредка можно ожидать, что он позволит оценить статистическую значимость результатов.

Зачастую мы сталкиваемся с фигурой умолчания — когда квалифицированному аналитику не нужны формализм и счет, чтобы увидеть, что либо (1) нет никаких сомнений в значимости, либо (2) на значимость нет никакой надежды.

Один из путей определения качества индикаторов: рассмотрение того, сколь хорошо они оценивают то, к чему предназначались.

И хотя мы иногда соотносим с индикатором некую «оценку», при выборе индикатора лучше все же «спросить», что он пытается оценивать, каково его *оценивание*.

Еще одна возможность — выяснить, что происходит с индикатором, когда он используется для выборок все большего объема.

Перепроверка — естественный путь индикации качества всевозможных величин, порожденных данными, — это и оценки, и элементарные индикации, и все прочее в этом роде.

Перепроверка бывает простой (на данных, которые не использовались для расчета числовых параметров модели) и двойной (на данных,

что не использовались, еще и для выбора самой модели, ее формы). Вторая — всегда надежнее.

Основная ценность автоматизированных программ оптимизации, подбора «ближайшей» к данным модели и тому подобных, состоит в том, что они допускают *надежную* перепроверку.

Мы принимаем всерьез индикации или индикаторы, оставшиеся на задворках вывода, но с честным предупреждением: «это только индикация».

Мы подходим к перепроверке со всей возможной осторожностью.

БИБЛИОГРАФИЯ

Macdonald N. J. and Ward F. (1963). The prediction of geomagnetic disturbance indices: 1. The elimination of internally predictable variations. — J. Geophys. Res., 68, 3351—3373.

McCarthy P. J. (1976). The use of balanced half-sample replication in cross-validation studies. — J. Amer, Stat. Assoc., 44, 596—604.

Miller R. G. (1962). Statistical prediction by discriminant analysis. Meteorol. Monogr., 4 (25), 1—54.

Mosteller F. and Wallace D. L. (1964). Inference and disputed authorship: The Federalist. Reading, Mass., Addison-Wesley.

Stone M. (1974). Cross-validatory choice and assessment of statistical predictions. — J. Roy. Stat. Soc., Series B, 36, 111—147.

Yates F. (1950). Recent applications of biometrical methods in genetics: 1. Experimental techniques in plant improvement. — Biometrics, 6, 200—207 (в особенности 204—205).

Yates F. (1957). Частное сообщение (to J. W. Tukey).

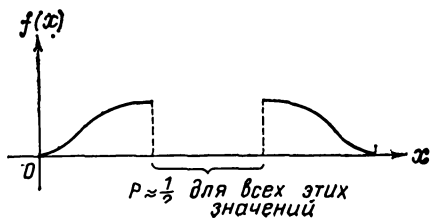
ИЛЛЮСТРАЦИИ

Иллюстрация 2.5.1

Пример распределения, для которого медиана была бы скверным кандидатом в меры положения

Если в обеих частях распределения сосредоточено по 1/2 плотности, то от выборки к выборке медиана будет «скакать» с левой половины на правую и наоборот, создавая значительную вариабельность.

На рисунке P означает вероятность того, что медиана выпадет на соответствующее значение x (из выделенных фигурной скобкой).



Глава 3 ● ПРЕДСТАВЛЕНИЯ И СВЕРТКИ ДЛЯ ОДНОРОДНЫХ ГРУПП ДАННЫХ

3.1. «ОПОРА И КОНСОЛЬ»

Быстрые способы описания однородных групп чисел особенно полезны, когда в них содержится изрядное количество информации о структуре распределения. И прежде всего те из них, что позволяют быстро извлечь нужное, когда, например, мы ищем 15-е значение с каждого края в упорядоченном ряду данных. Все эти нужды может обеспечить *описание типа «опора и консоль»*, хотя для разных целей могут понадобиться разнотипные описания. Такие описания ведут дальше, чем классический счет по гистограммам.

Рассмотрим предварительные данные о добыче полезных ископаемых в США за 1971 г. в стоимостном выражении (по штатам). Алабама дала 291,4 млн. дол. Эту цифру можно разбить на три части следующим образом:

	Опора	Консоль	Мишура	Млн. дол.
Алабама	2	9	14	(291,4)
Аляска	3	3	28	(332,8)
Аризона	9	8	10	(981,0)
Арканзас	2	5	32	(253,2)
Калифорния	19	2	06	(1920,6)

Мы включили данные еще по четырем штатам. Результат по всем пяти штатам представлен в виде «опоры и консоли» на илл. 3.1.1.

Несколько подходов к возможным преобразованиям этих описаний обсуждается в *EDA**. Совсем другие варианты приведены здесь на илл. 3.1.2.

3.2. МЕДИАНЫ, КВАРТИЛИ И ПРОЧИЕ ПРОЦЕНТИЛИ

Уже отмечалось, что середина упорядоченного по возрастанию множества данных называется медианой. Если такой выписанный ряд существует, медиану найти просто. Возьмем, для примера, 25 данных

*Расшифровку *EDA* см. в библиографии к гл. 3, с. 74.— *Примеч. ред.*

(вместе с совпадающими, если они есть), тогда 13-е число с любого конца и есть медиана, которую мы обозначим M. Возьмем теперь всего 24 значения, тогда медиана — $\left(12\frac{1}{2}\right)$-е с любого конца, т. е. середина между 12-м и 13-м значениями. Общее правило:

$$\text{место медианы} = \frac{1}{2}\,(1 + \text{объем выборки}).$$

Место медианы можно установить столь же просто, когда есть представление типа «опора и консоль». Левый столбец на илл. 3.2.1Б показывает (сверху и снизу к середине) накопленное число значений к данному уровню «опоры», включая и его. Значит, общее число данных равно $76 + 5 + 74 = 155$. Медианное значение, 78-е, находится из уровня опоры **18** | 17001 так, как это видно из записи

(76 меньших)
180
180 *M*
181
181
187
(74 больших),

где 78-е значение отмечено буквой M. Мы называем медиану M *особой точкой*. M говорит нам о центре этой группы чисел, не очень зависящем от одного или нескольких выбросов, причем тем меньше, чем меньше их доля в общем числе данных. (Более подробно мы изучим это на примере илл. 14.7.1 в параграфе 14.7 вып. 2).

Другие особые точки. Часто нам полезен не только центр данных, но и значения, отсекающие четверть, восьмую часть и т. д. при движении от меньших к большим или наоборот. Чтобы легче их запомнить, удобно правило последовательного определения, аналогичное правилу для медианы. Оно гласит

$$\text{место следующей особой точки} = \frac{1}{2}\,(1 + \text{место предыдущей}),$$

где «место предыдущей особой точки» берется либо само, если оно целое, либо его целая часть. Если оно было равно $12\frac{1}{2}$, то мы возьмем 12 и место следующей особой точки будет

$$\frac{1}{2}\,(1 + 12) = 6\,\frac{1}{2}.$$

Припишем «местам» последовательных особых точек буквы H (или Q, если вам больше нравится) для квартиля, E для «восьмушки» и далее D, C, B, A, Z, Y, X, W... (в порядке обратного алфавита).

Так, для наших данных о телефонах имеем:

Выборка	Объем 155	Вычисления
M	место 78	$78=\frac{1}{2}(1+155)$
H	» $39\frac{1}{2}$	$39\frac{1}{2}=\frac{1}{2}(1+78)$
E	» 20	$20=\frac{1}{2}(1+39)$
D	» $10\frac{1}{2}$	$10\frac{1}{2}=\frac{1}{2}(1+20)$
И т. д.		

Заметьте, что при расчете места E используется $1+39$, а не $1+39\frac{1}{2}$.

Взглянем теперь внимательнее на 78 (уже найденное), $39\frac{1}{2}$, 20 и $10\frac{1}{2}$-е места. Где лежит $39\frac{1}{2}$-е значение, если отсчитывать от наименьшего (т. е. сверху вниз)? 36-е равно 129, следующие десять значений лежат уже на отметке 13 «опоры»: 130, 130, 131, 131, 133, 134, 134, 135, 136, 137. Но $39\frac{1}{2}$-е лежит посередине между 39-м и 40-м, т. е. между 3-м и 4-м значениями после 36-го, следовательно, 131.

Мы также хотим найти $39\frac{1}{2}$-е значение снизу. Здесь отыщем места после 34-го и большее 40-го и, двигаясь в сторону уменьшения, найдем числа 38, 36, 34, 34, 34, 34, 32, 32, 31, 31, 30, 30, где надо взять середину между 3-м и 4-м, так что $39\frac{1}{2}$-е равно 34 (точнее, числу, лежащему между 340 000 и 349 999).

Совсем просто найти 20-е место. Имеем:

(21-е)	62
(20-е)	63
(19-е)	68
(18-е)	69
(17 больших)	

так что 20-е есть 63.

На илл. 3.2.2 дана простая стандартная форма представления мест особых точек и самих этих точек, что при желании легко проверить.

Даже когда нет ничего, кроме выписанных ниже, в илл. 3.2.2, особых точек, они и сами по себе кое-что говорят о форме распределения наших 155 чисел. Для больших чисел особые точки резко возрастают от E к D, затем к C и далее. А для малых значений соответствующие величины практически не меняются. Ясно, что распределение для выборки из этих 155 чисел весьма асимметрично.

3.3. СЕРЕДИНЫ И РАЗМАХИ

Если мы хотим работать и дальше с нашими особыми точками, то удобными статистиками, основанными на парах таких значений с одинаковыми местами (сверху и снизу), будут их средние (здесь и медиана) и разности. Так, для D (граница $1/16 \approx 6{,}3\%$ всех данных, лежащих либо левее, либо правее ее в упорядоченном ряду) имеем значения

$$87\,\text{x} \quad \text{и} \quad 106^1/_2,$$

откуда находим

$$48\,\text{x} = {}^1/_2\,(87\,\text{x} + 106^1/_2) \quad (\text{«середина»})$$

и

$$76\,\text{x} = 87\,\text{x} - 106^1/_2 \quad (\text{«размах»}).$$

Вычисления размаха говорят в данном случае о «размахе» вокруг медианы. Такие вычисления возможны для любого процентиля (особой точки). Дальнейшие подробности о путях использования «далеких» уровней можно найти в гл. 19, *EDA*.

Когда, как, скажем, в илл. 13.10.2, мы хотим дать размахи как часть представления особых точек, можем записать их примерно так:

		180		
N	155			
M	78			размах
H	39½	34x	131	209
E	20	63x	112	527
D	10½	87x	106½	767

Размахи выписаны в единицах первоисточника, что проще, чем использование обозначения «x» вместо единиц трехзначного числа.

3.4. ВЫБОРКИ ИЗ ВЫБОРОК

Иногда у нас так много данных, что их объем препятствует тщательному анализу. В таких затруднительных обстоятельствах, имеющих свои приятные стороны, выборки из выборок или даже последовательность таких выборок вместе с последовательным анализом могут стать и экономичными, и поучительными. В параграфе 12.8 (вып. 2) это применяется для регрессии, однако идея вполне применима и к другим методикам, начиная с представлений типа «опора и консоль».

3.5. ГРАФИЧЕСКИЙ АНАЛИЗ

Желая предсказать ход зависимости, мы можем обратиться к графикам. В этом параграфе опишем один гибкий подход, который может оказаться полезным, когда у нас довольно много данных. Основная идея — получить совершенно гладкую регрессионную кривую между y («откликом», зависимой переменной) и значениями x (ее «пары», независимой переменной) без сколько-нибудь жестких требований к форме связи.

Когда нам посчастливится обладать очень большим числом «точек», мы можем и отказаться от нанесения всех их на график, а использовать выборки из выборок. А иногда, напротив, более полезными могут оказаться данные, оставшиеся от этих выборок.

От 20 до примерно 400 наборов данных подходящее сокращение дает график с k точками, имеющими наибольшие отклики, и k точками, имеющими наименьшие отклики, причем

$$k = \text{наибольшее из } 10 \text{ и } \sqrt{n},$$

где n — число наборов данных.

При $n > 400$ k будет больше 20, и нам, видимо, захочется его урезать. В таком случае, может быть, стоит начать с наибольшего и наименьшего значений k, а затем выбрать в каждом из них примерно по 10 значений, более или менее равномерно распределенных относительно величины отклика.

В статистическом справочнике «Country and City Data Book» (1962) содержится информация о 88 городках с населением не менее 25 000, в том числе и информация о медианном семейном доходе в этих городках за 1959 г. Илл. 3.5.1 содержит некоторые данные (А) о 10 городках (из 88) с наименьшими и (Б) о 10 городках с наибольшими медианными семейными доходами. На илл. 3.5.2 данные об этих 20 городках нанесены на графики, на всех графиках по оси ординат отложен семейный доход, а по оси абсцисс — 4 показателя из 8, имеющихся в илл. 3.5.1.

График с абсциссой медианного возраста показывает явную зависимость, если не считать нескольких «отбившихся от стада» городков, которые названы поименно. График пользования городским транспортом может показаться неожиданным: выходит, что более богатые пользуются им чаще. Связь на третьем графике относительно слабая, особенно без резко выделившегося городка из Нью-Джерси. На последнем графике зависимости практически нет.

Мы рассмотрели лишь некоторые из 67 показателей, даваемых первоисточником. Но и этого довольно, чтобы увидеть, что, желая выразить медианный семейный доход через другие показатели, мы должны отталкиваться от разнообразия городков.

Давайте начнем с медианного возраста, в зависимости от которого медианный семейный доход возрастает примерно на 300 дол. за каждый год прироста среднего возраста. Коль скоро мы это знаем, мы можем иначе подойти к нашему y, разделив медианный семейный доход на две составляющие:

«модель» ПЛЮС «остаток».

Здесь

«модель» = (300 дол.) · (медианный возраст в годах),

что ее непосредственно задает. Вторая составляющая

«остаток» = (медианный семейный доход) — (300 дол.) · (медианный возраст в годах)

остается неохарактеризованной и представляет для нас сама по себе новую переменную y, требующую дальнейшего изучения.

Было бы заманчивым рассчитать все 88 значений новой y и выбрать новые 10 «наибольших» и 10 «наименьших» городков. Мы с удовольствием делаем это, когда данные находятся в ЭВМ, где все просто. А при ручном счете можно пользоваться данными по одним и тем же городкам для двух или трех шагов анализа и лишь затем пересчитывать.

3.6. ТРЕНДЫ И СКОЛЬЗЯЩИЕ МЕДИАНЫ

При наблюдении пары: x — независимой переменной и y — отклика (зависимой переменной) вид зависимости между ними интересует нас сам по себе, но так же и для последующего анализа «остатков» — отклонений наблюдений от линии регрессии. Иногда значения x образуют правильную решетку (с равным шагом). Однако вполне возможны и другие варианты. Когда нет убедительной точки зрения на форму связи, а линейной мы не ждем, то многие приемы приходят в голову сами собой. Среди них:

1) проведение кривой «от руки» по точкам для получения «подгонки», не раздражающей глаз;

2) разбиение оси x на части, подсчет в этих частях медиан или средних для y и x и проведение кривой поближе к этим точкам;

3) использование скользящих средних или скользящих медиан.

В этом параграфе мы обсудим скользящие (текущие) медианы, а в следующих — другие возможности.

Не для ограничений на метод, а лишь для простоты показа идей будем считать ось абсцисс временем. Предположим, что наблюдения складываются из двух частей — кривой регрессии (того, что мы хотим узнать) и независимой ошибки. Так что отклики (y) в моменты времени t (= 1, 2, 3, ..., T) можно представить, как

$$\text{отклик} = \text{регрессия} + \text{ошибка},$$

или, если $f(t)$ есть значение регрессии в момент t, более формально

$$y(t) = f(t) + \text{ошибка}\ (t).$$

Если ошибка сравнима с вариацией $f(t)$, то возникает искушение оценивать $f(t)$ не по самому $y(t)$, а усреднив предварительно несколько значений в окрестности точки t. Такое усреднение при не слишком сильно закоррелированных между собой ошибках улучшает оценку $f(t)$. Действительно, несмотря на то, что значения $f(t)$ падают вдоль прямой, вовсе не обязательно горизонтальной, среднее арифметическое значений откликов, центрированных относительно t, остается несмещенной оценкой для $f(t)$. Так что мы горим желанием усреднять и приписывать результат моменту t, в окрестности которого велось усреднение. Однако этому препятствует то, что $f(t)$ может быть и нелинейной функцией, а тогда усреднением или сглаживанием, как ни называй, мы портим $f(t)$, регрессию. Беря среднее трех последовательных значений, $y(t-1)$, $y(t)$, $y(t+1)$, мы тем самым будем оценивать

$$\frac{f(t-1)+f(t)+f(t+1)}{3} \text{ вместо } f(t).$$

Число точек для усреднения не стоит брать слишком большим, ведь если взять много, то потеряются именно те характерные черты функции, которые мы и хотим оценить.

Такое усреднение широко используется при исследовании временных рядов. Конечно, применяются многие виды усреднения, часто и с неравными весами. В этом параграфе мы сосредоточимся на скользящих медианах по трем смежным значениям с центром в t. В первую очередь попробуем дать представление о том, что происходит при таком сглаживании. В следующем параграфе такое сглаживание работает в реальной задаче. Однако нет особых причин рекомендовать описанный здесь прием сглаживания как трафарет, другие и лучшие методы читатель найдет в *EDA*, гл. 7 и 16, и в [Velleman P. F. (1975)].

Мы хотим остановиться на двух важных моментах: что происходит, когда скользящие медианы применяются к простейшему, идеальному случаю, и что — в условиях более сложной ситуации. Мы надеемся, что сглаживание не привнесет «огромных» потерь и скользящие медианы будут чувствительны к искривлениям и прыжкам уравнения регрессии. В ходе изучения этих неясных вопросов на примерах мы столкнемся также и с некоторыми деталями, которые важно знать.

Скользящие медианы. Наш план таков: рассчитываем скользящие медианы, а затем повторяем тот же расчет для найденных значений до тех пор, пока процесс не стабилизируется. Мы демонстрируем это, беря для расчета по три соседние точки; для любого нечетного числа точек образ действия тот же. (Напомним, что для четных чисел мы берем среднее из двух, ближайших к середине, значений: так, медианой для 7, 2, 10 и 3 будет $1/2\ (3 + 7) = 5$.)

Пример. Расчет скользящих медиан. На илл. 3.6.1 даны 20 наблюдений, следующих друг за другом «во времени»: 1, 2, ..., 20. Эти данные — случайные нормальные отклонения, извлеченные из таблиц случайных чисел. Таким образом, «про себя» нам известно, что их генеральное среднее равно 0, а дисперсия равна 1. Соответственно для процесса, порожденного этими y, для всех t функция $f(t) = 0$. Значит, мы имеем дело с «истинной» регрессией, задаваемой горизонтальной прямой. В реальной ситуации мы очень редко будем все это знать.

Сначала сгладим данные скользящей медианой, а затем применим к тому, что получится, метод наименьших квадратов. После этого мы сравним результат с тем, который получился бы, если бы мы подобрали линию регрессии непосредственно по первоначальным данным, — как мы и должны были бы поступать, будучи уверенными в линейности и нормальности, но не зная углового коэффициента и свободного члена уравнения прямой. Наконец, мы воспользуемся одними и теми же данными дважды, чтобы лучше показать, что происходит, когда мы подбираем функцию регрессии в ситуациях, где она, неизвестная нам, далеко не линейна.

На илл 3.6.1 в первом столбце значения «времени», а во втором — соответствующие значения y. Так как с первым числом надо что-то делать, мы просто-напросто перепишем его в столбец скользящей медианы (третий). Затем берем значения в 1, 2 и 3-й моменты времени и их медиану (для — 0,423; — 0,602; 1,703 медиана равна — 0,423) запи-

сываем в третьем столбце во вторую строку. Таким образом мы постепенно дойдем до 20-й строки, $t = 20$, где опять мы просто-напросто перепишем последнее значение из 2-го столбца в 3-й. Этот процесс повторяем до тех пор, пока столбцы не перестанут меняться.

Запись можно и сократить: достаточно записывать лишь изменяющиеся значения, а не все подряд. Тогда 0,401 в 9-й строке будет записано лишь в столбце y (t), а конечный результат сглаживания надо брать из последней записи в каждой строке.

На илл. 3.6.2 приведены два графика: на правом — точки необработанных исходных данных, а на левом — сглаженные. Как и следовало ожидать, сглаженные данные варьируют меньше, чем исходные. Один из признаков гладкости — число *точек «возврата»* (или локальных экстремумов). Возьмем три последовательных числа. Если среднее из них либо строго больше, чем два соседних, либо строго меньше, то оно называется точкой «возврата». В случайной последовательности длины n ожидаемое число таких точек возврата равно $^2/_3$ $(n - 2)$ в предположении, что все числа в тройке значений y различны. (Среди 6 возможных последовательностей трех разных чисел в 4 случаях имеет место экстремум в середине, так что вероятность точки возврата в последовательности из трех чисел равна $4/6 = 2/3$. А число «— 2» в формуле появляется из-за краевых эффектов.)

В выборке № 1 необработанных исходных данных мы найдем 10 «возвратов» вместо ожидаемых 12. А в сглаженных данных много одинаковых соседних значений, так что разумно, по-видимому, говорить всего о 3 точках возврата, и если интерпретировать их как локальные экстремумы графика, то они соответствуют точкам 4,5; 15, 16 и 17, 18, 19. (Пусть это будет плато — суть не в точной локализации точки экстремума соответствующим t. Важно лишь, имеется ли экстремум.)

На илл. 3.6.3 показаны результаты сглаживания первых 10 случайных выборок объема 20, которые мы взяли подряд из того же первоисточника. Они показывают изменчивость, фактически возникающую при сглаживании скользящими медианами. «Про себя» мы знаем, что должны увидеть как оценку f прямую, идущую вдоль горизонтальной оси. Выборка № 4 — предупреждение для нас: подъем в начале и спад в конце не характерны для настоящей функции, хотя мы ошибочно могли бы «тянуться» к такому предположению, поскольку остальные точки хорошо ложатся на горизонтальную ось.

В некоторых выборках мы видим плоские вершины или впадины длиной в 2 шага (по t). В выборке № 1, например, две начальные точки образуют впадину в 1 шаг; еще одна такая же — между 15-й и 16-й точками. Это обычные признаки скользящих медиан, усредненных по тройкам, хотя они часто причиняют нам беспокойство. Другие методы, уже упоминавшиеся в ссылках (*EDA*, гл. 7), более продвинуты и исключают эти признаки.

Увеличение в начале и падение в конце у выборки № 4 производит сильное впечатление, особенно из-за ровности между 2-й и 18-й точками.

Аналогично бросается в глаза горб в середине выборки № 5, поскольку он нарушает монотонность.

Выборка № 7 имеет хорошо выраженное падение в начале и прогиб в конце, заставляя даже поверить в реальность этих эффектов у процесса, порождающего данные.

Выборки под номерами 2, 6, 10 имеют, похоже, синусоидальный характер.

А теперь вспомним, что все эти результаты предполагалось использовать для того, чтобы оценить точки горизонтальной оси, $f(t) = 0$. Следовательно, вывод таков, что индикацию наклона в сглаженных данных надо рассматривать с осторожностью. Когда мы оглядываемся на необработанные исходные данные илл. 3.6.2, то не замечаем никаких деталей, поскольку очень велик разброс точек. Правда, можно ожидать тренда, ниспадающего слева направо, но три высоких значения в левом краю и два низких, 15-я и 16-я точки, не слишком-то впечатляют, ибо общий разброс велик. Другое дело — сглаженные данные илл 3.6.2, где мы скорее всего обратим особое внимание на горб слева и провал в 15-й и 16-й точках, поскольку в основном разброс этих данных очень мал. Значит, надо проявлять осторожность при интерпретации поведения сглаженных кривых, поскольку ограничения, обусловленные наблюдаемой естественной изменчивостью, не так уж важно исследовать, что мы сейчас показали.

Скользящие медианы дают несколько ломаную линию — часто чуть более ломаную, чем скользящие средние. Но так как несколько исходных данных влияют лишь на малое число значений скользящих медиан (или скользящих средних), прыжки данных вверх и вниз вокруг какого-нибудь значения будут, как мы думаем, выявлены и в скользящих медианах. Правда, они не столь нерегулярны, как сами данные. Когда скользящая медиана (например, по трем значениям) заметно приподнята, то это происходит потому, что по крайней мере два из трех значений достаточно высоки. В таком случае по крайней мере две последовательные скользящие медианы должны быть высокими. Поэтому высокие значения, как правило, «ходят парами» (тройками и т. д.).

Это особенно заметно, когда мы вычисляем скользящие медианы повторно, так как оставшиеся на первом шаге вычислений отдельные экстремумы сглаживаются на следующем шаге. Таким образом, вершины и впадины, остающиеся после сглаживания, имеют тенденцию к образованию разных изломов. Сглаживание, собственно, подавляет высокочастотный шум и оставляет низкочастотный, более выразительный на вид.

Подбор прямых. Давайте теперь посмотрим, что происходит при подборе прямых обычным методом наименьших квадратов (с равными весами) для необработанных и сглаженных данных. Так, для выборки № 1 имеем:

	нецентрированные	центрированные
необработанные данные	$y = -0{,}0588\,x + 0{,}9776$	$= -0{,}0588\,(x - 10{,}5) + 0{,}360;$
сглаженные данные	$y' = -0{,}0834\,x + 1{,}1870$	$= -0{,}0834\,(x - 10{,}5) + 0{,}311.$

На илл. 3.6.4 приведены попарно (первичные, сглаженные) для всех 10 выборок угловые коэффициенты и центрированные свободные чле-

ны прямых. (Мы используем центрирование свободных членов, поскольку это приводит к идеальной ситуации, где для необработанных исходных данных они распределены независимо от угловых коэффициентов.) На илл. 3.6.5 те же данные представлены графически. Теория говорит, что распределение сглаженных результатов должно быть чуть более изменчивым, чем для исходных значений. Совершенно случайно крестики на илл. 3.6.3 (в начале, середине и конце) показывают, как проходит искомая линия по сглаженным данным.

Наконец, на илл 3.6.6 показана диаграмма «опора-консоль» для угловых коэффициентов и центрированных свободных членов прямых, позволяющая непосредственно сравнить маргинальные * распределения; мы вновь видим почти одинаковую изменчивость результатов, хотя теория и говорит нам, что в этой идеальной ситуации работа с исходными данными предпочтительнее (но не для распределений,далеких от гауссовского).

Этим завершается наша демонстрация сглаживания в абсолютно идеальной «нуль-ситуации»: нет тренда и ошибки в каждой точке распределены по одному и тому же нормальному закону. Теперь обратимся к примерам, где кривая регрессии f нелинейна.

3.7. СГЛАЖИВАНИЕ НЕЛИНЕЙНЫХ РЕГРЕССИЙ

Теперь покажем сглаживание нелинейных регрессий, когда случайным разбросом нельзя пренебречь, но он и не слишком велик. На илл. 3.7.1 функция регрессии f_1 проведена по 20 точкам, соединенным пунктиром. Если кривая быстро осциллирует, то трудно ожидать получения информации о колебаниях по достаточно редкой сети значений t. Более того, сглаживая наши данные, мы в некоторой степени сглаживаем и функцию f_1. Сплошной линией на илл. 3.7.1 показана функция f_1^*, сглаженная применением скользящих медиан из трех точек (до стабилизации). Именно эту, сглаженную f_1^* мы хотим видеть скорее, нежели первоначальную функцию, когда сглаживаем данные, в которых f отягощена случайными ошибками. Повторим — сглаживание данных сглаживает и функцию, подлежащую оценке. Фактически сгладились два локальных экстремума в точках 6 и 7, что могло содержать и научную информацию. Но тогда сглаживание подавляет важные эффекты.

На илл 3.7.2 показано, что происходит, когда к f_1 прибавлены несглаженные нормально распределенные ошибки из выборок параграфа 3.6. А затем рассчитаны скользящие медианы для всех 10 выборок.

Для всех выборок, кроме 5, 6, 7 и 8, общий вид сглаженных данных согласуется с видом регрессионной функции f_1^*, представленной на илл. 3.7.1. На четырех отмеченных графиках сильнее сглажен левый, фактически небольшой горб. И все сглаженные выборки имеют явно видимые на правом конце нарушения округлости.

---

*Распределений на пространстве результатов, исходов. — *Примеч. пер.*

Сравнивая графики илл. 3.6.3 с f_1^* на илл. 3.7.1, можно в некоторой степени оценить, что происходит на графиках илл. 3.7.2. Например, начальное падение в выборках с номерами 5, 6, 7 и 8 на илл. 3.6.3 и приводит к уничтожению горба, имеющегося у f_1^*.

3.8. ВЫЯВЛЕНИЕ ЗАКОНОМЕРНОСТЕЙ

Когда мы исследуем отклонения от избранной линии регрессии, остатки, то перед нами открывается много путей для обнаружения закономерностей. Тренды, изгибы, колебания и всплески можно интерпретировать непосредственно. Для исключения резких всплесков возможны преобразования. Иногда простой поворот координатной сетки на 45° помогает выявить сосредоточение «разрешенных» чисел или, наоборот, «исключений».

В параграфах 3.6 и 3.7 обсуждалось применение скользящих медиан к обнаружению взаимосвязей и отклонений для линейных и нелинейных регрессий. Здесь мы рассмотрим другие способы выявления закономерностей, используя процедуру, предлагавшуюся в начале параграфа 3.6, а именно проведение кривой по средним или медианам x и y на отдельных отрезках оси x.

Для демонстрации ряда таких приемов мы изучим эмпирически некоторые числа, связанные с известной математической проблемой, над которой работали многие математики, среди них Л. Эйлер и Г. Харди. Мы изучаем числа, исследовавшиеся Харди [Hardy G. H. (1906)].

Числа Гольдбаха. Гипотеза Гольдбаха представляет собой глубокую математическую проблему, сформулированную более 200 лет назад и еще не решенную. Она утверждает, что каждое четное число, начиная с 6, можно представить как сумму двух простых чисел (т. е. чисел, которые делятся лишь на 1 и на самого себя; при этом 1 простым числом не считается; $6 = 3 + 3$). Но мы не собираемся доказывать ее, нас интересуют *числа Гольдбаха* — так мы назовем для каждого *четного числа E число способов*, которыми его можно представить в виде суммы двух простых чисел*. Например,

$20 = 3 + 17 = 7 + 13$ (число Гольдбаха = 2),

$30 = 7 + 23 = 11 + 19 = 13 + 17$ (число Гольдбаха = 3),

$68 = 7 + 61 = 31 + 37$ (число Гольдбаха = 2).

На илл. 3. 8.1 представлены два набора чисел Гольдбаха: для четных чисел от 2 до 508 и от 9500 до 10 000. Десятки четных чисел даны в столбце T, а единицы (0, 2, 4, 6, 8) — в строках под индексами: $T + 0$, $T + 2$, ..., $T + 8$; соответственно этим четным числам приводятся их числа Гольдбаха.

---

*Христиан Гольдбах (1690—1764) поставил эту проблему в одном из своих писем Л. Эйлеру в 1742 г. В разработке ее, кроме названных авторами, активную роль играли советские математики Л. Г. Шнирельман (1930) и И. М. Виноградов (1937), а также венгр Реньи (1949). Первоначальное знакомство можно получить, например, по работе: Д э в е н п о р т Г. Высшая арифметика. М., Наука, 1965. — *Примеч. ред.*

Чтобы увидеть, какие особенности возникнут, мы на илл. 3.8.2 нанесли числа Гольдбаха в зависимости от половины соответствующего четного числа. Так что на горизонтали проставлены $N = \frac{1}{2}E$.

Первое, что мы видим, — общий тренд вверх и вправо.

Второе — вертикальный разброс чисел при движении вправо расширяется клинообразно. Можно было бы попробовать каким-то образом преобразовать данные, чтобы стабилизировать рассеяние, но мы пока оставим это и исследуем другие закономерности.

Мы видим еще при движении вправо, что на какие-то значения N приходятся особенно низкие числа, а на какие-то другие — особенно высокие. Можно ли здесь обнаружить систему? Для этого было бы хорошо получить регрессионную прямую или кривую, выступающую как исходная. Можно разными способами сгруппировать данные и получить для них вертикальные и горизонтальные средние. Можно, например, взять последовательно группы по 10 значений N от 0 от 9, до 10 до 19, и т. д., а центр по горизонтали — за среднюю точку. В качестве вертикальной координаты можно взять медиану или среднее y, т. е. чисел Гольдбаха. Будем называть такие точки центроидами, хотя в физике этот термин используется лишь для таких точек, координаты которых получаются усреднением по обеим осям *.

Наблюдения, похоже, достаточно хорошо ведут себя в этой области N, так что воспользуемся средним арифметическим из каждой группы. Точки, соответствующие среднему значению N и среднему значению y, на илл. 3.8.2 отмечены крестиками и соединены ломаной линией для грубого представления о ходе регрессии.

Помощь от этой грубой регрессии в том, что она отделяет «высокие» значения от «низких». Пока мы идем вправо до точки $N = 36$, большие значения не возникают, но дальше они появляются и позволяют нащупать в графике закономерность: *каждая третья точка имеет высокое* значение.

Более пристальный взгляд показывает, что действительно все точки вроде 39, 42, 45, 48, 51, 54, 57 и т. д. лежат выше линии, что и обнаруживает закономерность. Похоже, что N, кратные трем, имеют «бо́льшие» числа Гольдбаха, чем остальные, во всяком случае, они лежат выше, чем проведенная линия регрессии.

Теперь, имея эту гипотезу, можем оглянуться и посмотреть, нет ли этой закономерности для меньших N? Констатируем, что 30, 27, 24 и 21 лежат ниже ломаной, проведенной нами.

Можно предположить, что иногда и N, *не* делящиеся на 3, имеют высокие числа Гольдбаха. Они вполне достойны изучения, однако сейчас, добившись некоторого успеха в анализе больших чисел, давайте сосредоточимся на малых. Один из путей исследования — удалить все N, кратные трем, подобрать новую кривую регрессии и тщательно рассмотреть высокие и низкие числа. Это сделано на графике с. 86.

---

*Идею центроида можно рассматривать как обобщение понятия о центре тяжести. — *Примеч. пер.*

На илл 3.8.3 вертикали гораздо более однородны, чем на илл. 3.8.2, и так же возрастает разброс чисел при движении вправо, правда, в меньшей степени. Можем ли мы обнаружить какую-нибудь новую закономерность в оставшихся точках?

Попробуем снова взять высокие и низкие точки относительно кривой регрессии и рассмотреть их. Возьмем последовательные группы по 9 точек оси N. Из каждого такого множества остается по 6 точек, не кратных трем. Их мы и возьмем для построения грубой регрессии так же, как и прежде.

Вначале рассмотрим точки с «высокими» значениями, больше чем на единицу «поднявшимися» над кривой, чтобы выбрать числа N. Находим 17, 32, 50, 56, 65, 71, 77, 80. Это задача из теории чисел, поэтому закономерность может быть связана с делителями N, как мы уже видели, когда обнаружили, что кратность трем, по-видимому, увеличивает число Гольдбаха. Делятся ли эти новые N на что-нибудь еще, кроме 2 и 3, уже рассмотренных нами? Выпишем под каждым N его делители, исключая 2, 3 и N:

17	32	50	56	65	71	77	80
—	—	5	7	5	—	7	5
				13		11	

Заметим, что 5, 7, 11 и 13— простые числа, следующие сразу за тремя. Таким образом, возникает подозрение, что делимость N на небольшие простые числа в среднем увеличивает числа Гольдбаха. Давайте подойдем к этому суждению в духе математического экспериментирования и взглянем на малые числа, те, которые лежат ниже кривой по крайней мере на 1, и посмотрим, что обнаружат их N. Вот они вместе с делителями, отличными от 2 и 3:

34	49	61	64	74	76	79
17	7			37	19	

Мы видим, что среди делителей чисел, лежащих ниже, гораздо меньше малых простых чисел, чем для больших N (нет 5, 11, 13 и лишь однажды 7). Теперь у нас есть общая идея о том, что N, делящиеся по крайней мере 1 раз на малое простое число, дают высокие числа Гольдбаха. Возможно, первая делимость на каждое малое простое число важнее, чем следующая.

Если это действительно так, то, перемножая последовательно разные малые простые числа, мы можем сконструировать такое число, которое *должно* иметь более высокое число Гольдбаха, чем его соседи. Давайте проделаем это.

Пусть $N = 3 \cdot 5 \cdot 7 = 105$, так что соответствующее четное число равно $2N = 210$. Для 210 и его соседей имеем:

200	202	204	206	208	210	212	214	216	218	220
8	9	14	7	7	19	6	8	13	7	9

откуда видно, что мы действительно сконструировали число, которое дает более высокое значение, чем его соседи.

Хотя мы и можем продолжать наслаждаться этим примером, однако все, что мы получили, сводится к тому, что с тщательного анализа остатков от «подсмотренной» линии регрессии можно начинать обнаружение закономерностей. Конечно, мы ничего здесь математически не доказали, да это и не было нашей задачей. Все, что мы хотели,— это подступить к изучению данных эмпирически, путем анализа остатков, и посмотреть, не можем ли мы обнаружить какую-либо закономерность. Пока мы пришли лишь к идее, что при делении четного числа на несколько разных небольших простых чисел мы будем получать в основном более высокие числа Гольдбаха. Мы знаем еще, что кривая регрессии растет при возрастании N, но растет и разброс точек вокруг нее.

Не в наших целях выжимать что-либо сверх этого результата, хотя и можно было бы продолжить анализ остатков в рамках систематического исследования методами анализа данных [Mosteller F. (1972)]. Но мы обязаны сделать из этого примера один важный вывод: довольно простой анализ данных может пролить свет на глубокую проблему. Это хорошо знали математические гении, такие, как Эйлер и Гаусс *.

3.9. ОБ ОСТАТКАХ ВООБЩЕ

Теперь, когда мы на примере повозились с остатками неформально и увидели крупицу того, что может быть получено при их изучении, рассмотрим здесь (и в гл. 16) другие подходы к их анализу. А начнем с общих соображений.

Можно анализировать остатки арифметически и графически. Хороший арифметический анализ вручную утомителен, но легко выполним на ЭВМ, однако мы тем не менее хотели бы использовать, как правило, и графический анализ. Хотя в примере с числами Гольдбаха графики остатков фактически и не строились, а строились лишь графики исходных данных, однако остатки нам были ясны из них. (За подробным арифметическим анализом мы отсылаем читателя к работам [Anscombe F. J., Tukey J. W., (1963)] и [Anscombe at al., 1974]). Вычисление процентилей и выявление «далеко отстоящих» и «выпадающих» значений (см. *EDA*, гл. 5) могут особенно помочь там, где остатки рассчитываются для хорошо подобранной регрессии.

Графическое исследование остатков почти всегда сводится к построению (x, y)-графиков, где y — остатки. Главные вопросы при этом: «Каковы x?» и «Каковы способы изображения?». Вопрос о способе изображения связан с эффективностью наших усилий, так как уменьшение

*Проблема обнаружения закономерностей в массивах данных — одна из центральных в науке. Здесь годятся все научные методы, какие только известны. Понятно, что эта проблематика быстро развивается (см., например: Машинные методы обнаружения закономерностей. Новосибирск. Изд. Ин-та математики СО АН СССР, 1976). С расширением машинной базы усиливается проникновение машинно-эмпирических методов в собственно математику: Нивергельт Ю., Фаррер Дж., Рейнголд Э. Машинный подход к решению математических задач. М., Мир, 1977. Можно ожидать революционизирующего воздействия на поиск закономерностей методов имитации: Клейнен Дж. Статистические методы в имитационном моделировании. Вып. 1, 2. М., Статистика, 1978. — *Примеч. ред.*

затрат на один график вдвое нередко вдохновляет нас на построение более двух графиков, часто весьма работоспособных. Нельзя ожидать, что один-единственный способ будет устраивать нас всегда. Индикации, которые мы выбираем, иногда явно зримы и ясны. С другой стороны, мы хотим обнаружить и нечеткие, тонкие особенности.

Мы достаточно экономно провели исследование в параграфе 3.5, изучая на графиках лишь высокие и низкие значения y и имея порядка дюжины или около того точек на каждом. Это помогло найти самородки там, где этого не ожидали. Что же надо делать, чтобы провести тщательное просеивание золотоносной породы?

Пример. Температура и география. Мы проанализировали остатки для определения зависимости максимальной январской температуры от широты местности. На илл. 3.9.1 представлены такие данные по ряду городов США. Связь достаточно чистая, хотя несколько городов стоят особняком, особенно Джэксонвилл, Сиэтл и Джуно — они выглядят теплее для своей широты.

К этим точкам методом наименьших квадратов была подобрана прямая; ее наклон $b_1 = -1{,}94$. Далее, обозначив температуру через y, широту — x_1, а через $\bar{y}$ и $\bar{x}_1$ — их средние, нашли остатки

$$y - \bar{y} - b_1 (x_1 - \bar{x}_1).$$

Они (илл. 3.9.2) были нанесены на график в зависимости от западной долготы города. Суточное вращение Земли должно, конечно, действовать на температуру на разных долготах усредняюще, однако континентальность местности и влажность (влияющие на температуру) имеют, хотя и запутанную, связь с долготой. Так что долгота в качестве абсциссы полезна. Города, стоявшие особняком, расположились выше других и на графике остатков. Но добавим, что, двигаясь от 100° до 130° долготы, мы наблюдаем общее смещение точек вверх. Что же это за дальние западные города с высокими остатками? Это Лос-Анджелес, Сан-Франциско, Портленд (Орегон), Сиэтл, Спокан. Четыре из пяти расположены у Тихого океана, что говорит о его «утепляющем» действии.

Для последующей индикации возьмем медианы групп по 7 следующих друг за другом (относительно долготы) городов, как медианы долгот, так и медианы остатков. Полученные точки показаны на илл. 3.9.2 кружками. Слабый подъем их около 75° долготы, быть может, обусловлен тем, что есть несколько южных городов, сгруппированных около этой долготы. Хотя линейное воздействие широты и исключено, скопление точек может вызвать сходный эффект. И, конечно, частично ответствен за подъем утепляющий эффект близости океана. Илл. 3.9.3 представляет первичные данные для городов — температуру, широту, долготу, высоту над уровнем моря — и остатки, полученные для зависимости температуры от широты*.

---

*Введением в применение математических и, в частности, статистических методов в географии может служить превосходная книга: Харвей Д. Научное объяснение в географии. М., Прогресс, 1974. — *Примеч. ред.*

3.10. ГРАФИКИ И СГЛАЖИВАНИЕ

Как совместить две крайности: экономичность и полноту исследования? Илл. 3.10.1 дает данные для 88 городков (частично приведенные на илл. 3.5.1) о двух показателях: проценте семей, живущих в отдельных квартирах на 1960 г. (y), и медианном семейном доходе за 1959 г. (x). Мы планируем найти регрессию обеспеченности жильем в зависимости от дохода. Для этого мы разделили данные (по «иксам») на 20 групп по 4 или 5, а иногда в группу включали даже по 1 или по 7 городков. Мы пытались сделать группы более или менее компактными и одновременно старались, чтобы интервалы не были слишком разными. Разделение похоже на то, что производилось с числами Гольдбаха, только здесь мы не выдерживали равных интервалов.

Для каждой группы мы рассчитали медианы и вместо того, чтобы просто соединить их ломаной линией, сглаживаем методом из параграфа 3.6 скользящими медианами по 3 точкам.

Основное впечатление от графика на илл. 3.10.2, где нанесены сглаженные медианы: связь между переменными весьма слаба. Представленная картинка более чувствительна, чем график для 10 высоких и 10 низких значений и чем мог бы быть такой же график для всех 88 точек. При слишком пристальном рассмотрении у нас может возникнуть искушение предположить возрастание семейного дохода от примерно 7200 до 7600 слева направо. Однако опыт сглаживания предостерегает нас от того, чтобы относиться к этому серьезно.

Усилия. Построение графика илл. 3.10.2 тоже требует усилий. Но это, в общем, требует не больше, а, возможно, меньше работы, чем нанесение на график всех 88 точек.

Если бы у нас была колода из 88 перфокарт (по одной на каждый городок), содержащих всю информацию, то было бы легко выполнить такую работу. Для этого надо всего лишь упорядочить карты по «иксам», разбить на группы разумного размера и записать их медианы, что мы и делали. Тогда останется только сгладить и нанести точки на график.

С колодами карт, в которые внесены последовательные остатки, полный графический анализ выполним даже для весьма внушительных объемов данных. (Для нескольких сотен данных желателен в качестве метода изучения и выбор из выборки.)

РЕЗЮМЕ. СВЕРТКИ ДЛЯ ОДНОРОДНЫХ ГРУПП ДАННЫХ И ИХ ПРЕДСТАВЛЕНИЯ

Мы описываем однородные группы чисел двумя-тремя картинками вида «опора-консоль».

Медианы могут естественно дополняться другими особыми точками (процентилями), места которых (в упорядоченном ряду чисел) находятся рекуррентно, по той же схеме, как и место медианы.

Мы даем схему расчета и представления для процентилей, которая включает в себя середины и размахи.

При работе с большими количествами данных придает удобство и гибкость выбор из выборки. Использование такого выбора может

быть решающим: без него можно справиться с анализом лишь за счет огромного труда.

Мы рассмотрели отдельные преимущества и недостатки одной простейшей устойчивой сглаживающей процедуры.

Изучая (x, y) данные, можно поступать по-разному, в том числе строя: (1) график для всех точек (который часто сам по себе недостаточен); (2) графики для 10 или 20 точек с высшими значениями y и с 10—20 низшими (отдельно); (3) прямую и анализируя остатки (возможно, как в (2)); (4) скользящие медианы для y (или, после переупорядочивания, для x); (5) группы соответственно их x-значениям, рассчитывая их медианы и сглаживая эти медианы (скажем, как в (3)) перед построением результирующих графиков; (6) комбинации всего перечисленного.

БИБЛИОГРАФИЯ

Anscombe F. J. and Tukey J. W. (1963). The examination and analysis of residuals. — Technometrics, 5, 141—160.

Anscombe F. J., Bancroft D. R. E. and Glynn J. G. (1974). Tests of residuals in the additive analysis of a two-way table — a suggested computer program. — Technical Report No. 32, November, 1974. Department of Statistics, Yale University.

EDA — Tukey J. W. (1977). Exploratory Data Analysis. Reading, Mass., Addison-Wesley.

Hardy G. H. (1906). Messenger Math., 35, 145.

Mosteller F. (1972). A data-analytic look at Goldbach counts. Statistic Neerlandica, 26, 227—242.

Velleman P. F. (1975). Robust nonlinear data smoothing. Technical Report No. 89, (Series 2), Department of Statistics, Princeton University (AEC).

ИЛЛЮСТРАЦИИ

Иллюстрация 3.1.1

Представление предварительных данных о добыче полезных ископаемых в США по штатам за 1971 г. в виде «опоры и консоли»; единица «консоли» равна 10 млн. дол.

Источник. The World Almanac, 1973, p. 423.

А. Набросок для пяти штатов

«Опора»	«Консоль»
1	
2	95
3	3
4	
5	
6	
7	
8	
9	8
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	2

Б. Все пятьдесят штатов

17	**0**	20229571997066398
21	**1**	1261
8	**2**	95286893
21	**3**	3948
17	**4**	0
16	**5**	82
14	**6**	405
11	**7**	01
	8	
9	**9**	82
7	**(H)**	(192, 555, 104, 118, 114, 680, 127)

Примечания:

1. Нуль на «опоре» и нуль на «консоли» соответствуют 2,241 млн. в Делавэре и 4,299 млн. в Род-Айленде*.

2. В левом столбце — накопленные числа штатов, отсчитанные сверху и снизу к середине, кроме 8,—числа штатов, приходящихся на уровень «опоры» 2.

3. В строке (**Н**) — числа, превышающие 99.

Интерпретация. 9 около 2 — это 29, а 5 около 2 — 25; так что 2|95 — условное обозначение сразу обоих чисел: 29 и 25. Аналогично 8 около 9 — это 98, а 2 и 19 — 192.

Интерпретация. 3948 слева от 3 — это 33, 39, 34, 38, аналогично 95286893 при 2 — это 22, 23, 25, 26, 28, 28, 29, 29.

---

*Поскольку в обоих случаях эти числа меньше единицы «консоли». — *Примеч. ред.*

Иллюстрация 3.2.1

Мощные «опоры» и охват двух порядков в представлении данных о городских телефонах для североамериканских городов с не менее чем 100 тыс. аппаратов в 1972 г.; единица «консоли» — 10 тыс. аппаратов.

Источник. The World Almanac, 1973, p. 419.

А. Не мощные опоры, а опора с переменной толщиной (использовано около половины данных из первых двух столбцов первоисточника)

0	
1*	397169286013746211810748031066131361447810532 14
2	5312173869
3	0142
4	83200
5*	4
6	823
7	4
8	71
9*	6
1**	06, 37,
2	35,
3	
4	85,
5**	

Примечания:

1. «Консоли» в верхней части ствола задаются одной цифрой, а в нижней — двумя. Одна «звездочка» у чисел 1, 5 и 9 на самом деле относится ко всем уровням в верхней части «опоры» и указывает, что мы имеем дело с двузначными числами — однозначной «опорой» и однозначными «консолями». Однако «звездочками» отмечены лишь 1, 5 и 9, чтобы не нагромождать знаки. Аналогично две звездочки в нижней части указывают, что мы имеем дело с трехзначными числами, однозначной «опорой» и двузначными «консолями», а отмечены ими снова лишь 1 и 5.

2. Первая запись 1*|3 читается как 130 тыс. аппаратов.

3. Первая запись в нижней части 1**|06 читается как 1060 тыс. аппаратов.

Б. Мощные «опоры» и все прочее («консоли» всюду задаются одной цифрой)

	0*		
14	**10**	13787253637324	
28	**11**	72416415612651	**От ста до двухсот тысяч**
36	**12**	23953110	
46	**13**	3706140514	
54	**14**	53289349	
60	**15**	357939	1 «консоли» равна 1000
68 |	**16**	11737448	
76 |	**17**	23717516	
5	**18**	17001	
74 |	**19**	131	
71	**2***	31213332003434	
57	**2**	57869888759	**От двухсот до девятисот тысяч**
46	**3***	014204246148	
34	**4**	832000	
28	**5***	4305625	
21	**6**	8239	1 «консоли» равна 10000
17	**7***	41110	
12	**8**	717	
9	**9***	6	
8	**1*x**	03205	**От одного до пяти миллионов**
3	**2*x**	3	
	3*x		
2	**4*x**	8	1 «консоли» равна 100000
1	**5*x**	9	

Примечания:

1. Числа в верхней части — трехзначные, а в нижней — двузначные.
Так, что 19\|131 означает два раза 191 (т. е., конечно, числа от 191000 до 191999) и 193, а 7* \|41110 соответствует 70, три раза по 71 и 74 (опять-таки числу между 740000 и 749999). Наконец, 4* x\| 8 есть 48x (4800000 аппаратов).
2. Две строки уровня 2 ... — первая для цифр, меньших 5, вторая (продолжение первой) — от 5 до 9.
3. Кумулятивный счет данных с двух сторон к середине дан в самой левой колонке. Так, например, есть лишь одно значение 5*x и нет больших, отсюда на «счетчике» — единица. Одно значение и на уровне 4*x, что дает в итоге 2 значения, которые не ниже 4 миллионов. И так же с другой стороны. Заметим еще, что мы не хотим вести счет ни с какого конца далее чем до середины.

Звездочки означают то же, что и в части А, примечание 1 этой иллюстрации. Знак **0*** в вершине опоры оставляет место для пятизначных чисел, пробегающих значения между нулем и единицей «консоли».

Иллюстрация 3.2.2

Структура представления особых точек для данных илл. 3.2.1

N 155		Особые точки (в тыс.)		*N* 155		Особые точки (в тыс.)	
Места		сверху (от максимума)	снизу (от минимума)	Места		сверху (от максимума)	снизу (от минимума)
M	78	180				180	
H	39½	34x	131	*B*	3	23xx	102
E	20	63x	112	*A*	2	48xx	102
D	10½	87x	106½	*Z*	1	535x	101½
C	5½	125x	103	*Y*	1½	59xx	101

Здесь «х» заменяет любую цифру. Так, для 34х, 63х и 87х фактически имеем 340, 639 и 873,5 тыс.соответственно, что можно восстановить по первоисточнику.

Иллюстрация 3.5.1

Данные (А) по 10 городкам, не входящим в общие корпорации городов, с наименьшими и (В) с наибольшими доходами за 1959 г. для семей, живших там в 1960 г.

Источник. Country and City Data Book, 1962, Table 5, p. 468—475.

	Городок и штат	Медианный семейный доход	(212)	(229)	(244)	(246)	(248)	(249)	(256)	(264)
А.	Нью-Гановер, Нью-Джерси	4572	28 8	1,1	42,2	0,4	4,7	78,8	82,2	24,1
	Флоренция-Грехем, Калифорния	4904	25,7	40,2	17,1	16,8	3,9	86,6	41,2	1,3
	Каннаполис, Сев. Каролина	5182	29,0	54,5	20,4	6,1	4,6	96,7	27,1	6,5
	Браунсвилл, Флорида	5306	22,6	35,3	39,1	3,2	4,6	93,5	46,4	26,0
	Уэст Лос-Анджелес, Калифорния	5439	25,1	44,8	26,7	19,7	4,2	79,8	37,1	6,0
	Белл-Гарденс, Калифорния	5567	24,4	26,1	67,6	1,5	3,8	89,8	59,6	30,3
	Хемпфилд, Пенсильвания	5909	29,3	58,4	37,0	5,3	5,2	95,0	24,9	3,0
	Южный Сан-Габриэль, Калифорния	6076	29,3	40,9	36,9	6,3	4,3	90,1	41,5	10,4
	Эссекс, Мэриленд	6160	24,8	46,7	34,5	6,7	5,4	78,5	37,5	12,0
	Мизуен, Массачусетс	6278	34,6	63,1	38,2	6,7	5,3	72,6	19,9	3,6
Б.	Нидхэм, Массачусетс	9282	32,5	56,0	69,3	14,0	6,2	92,3	22,5	11,2
	Тинек, Нью-Джерси	9518	33,0	63,0	62,9	29,4	6,1	80,6	17,9	26,9
	Силъвер-Спрингс, Мэриленд	9540	31,7	48,6	76,2	11,0	5,7	68,9	33,7	40,6
	Гринвич, Коннектикут	9588	35,6	55,8	54,5	9,7	5,8	69,9	20,7	11,6
	Уэст-Хартфорт, Коннектикут	9712	37,4	50,4	72,1	13,5	6,2	79,8	22,1	16,0
	Челтенхэм, Пенсильвания	9985	36,6	57,9	75,0	24,0	6,5	73,5	23,8	41,8
	Маунт Ливан, Пенсильвания	11108	36,9	49,1	62,8	25,5	6,2	81,4	25,4	14,0
	Уэллисли, Массачусетс	11478	31,3	52,4	70,8	15,4	6,7	94,7	23,2	9,8
	Лоуэр Мерион, Пенсильвания	12204	32,6	57,4	69,0	18,8	7,1	78,2	21,2	37,9
	Бизесда, Мэриленд	12357	31,4	36,3	82,6	8,5	6,5	82,3	37,4	41,7

Примечание.

Столбцы 212, ..., 264 следуют нумерации первоисточника и расшифровываются так: (212) — медианный возраст; (229)—% проживающих в типовых домах; (244) — % белого населения; (246) — % пользования общественным транспортом; (248) — медианное число комнат на семью; (249) — % семей, проживающих в отдельных квартирах; (256) — % поменявших местожительство (внутри города) в 1958 — 1960 гг.; (264) — % квартир с кондиционерами.

Иллюстрация 3.5.2

Графики для 10-ПЛЮС-10 городков с высшими и низшими семейными доходами; связь между доходом и четырьмя показателями (по данным илл. 1.5.1)

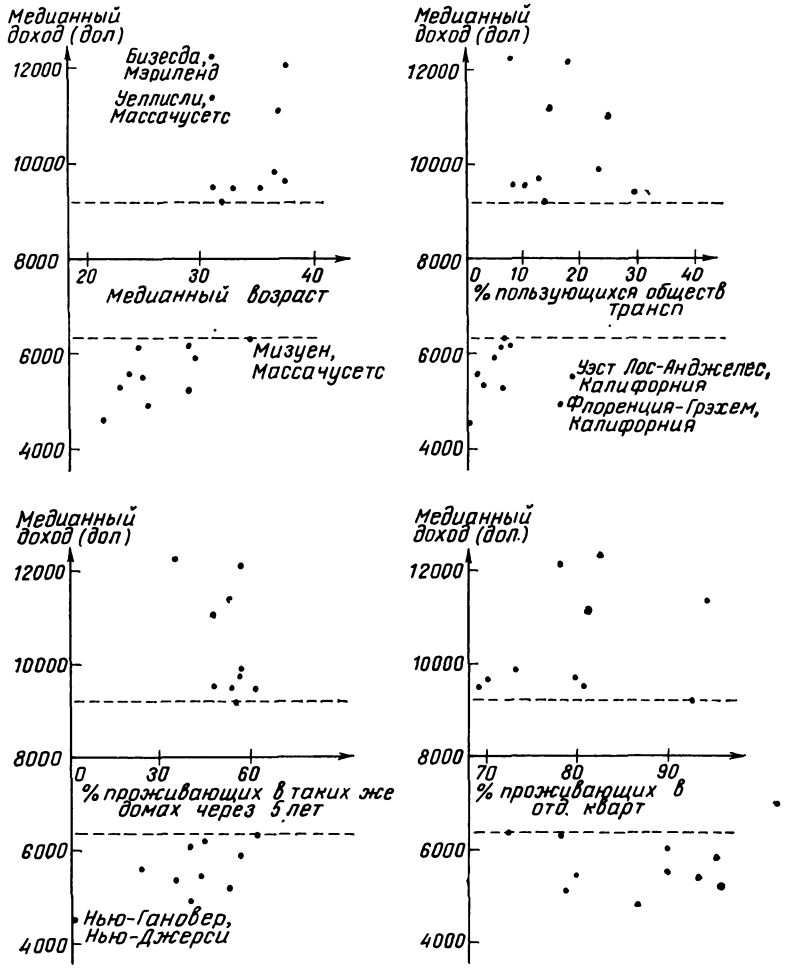


Иллюстрация 3.6.1

Сглаживание множества 20 значений откликов посредством скользящих медиан по трем соседним значениям. Здесь $Y(t)$ определено по выборке № 1 из 10 выборок размера 20 (см. илл. 3.6.3)

t	$Y(t)$ Случайные нормальные отклонения	Скользящие медианы по трем случайным нормальным отклонениям	Скользящие медианы по трем медианам	
			1-я итерация	2-я итерация Y'
1	—0,423	—0,423	—0,423	—0,423
2	—0,602	—0,423	—0,423	—0,423
3	1,703	1,703	1,703	1,703
4	1,887	1,887	1,887	1,887
5	2,049	1,887	1,887	1,887
6	1,127	1,127	1,127	1,127
7	0,651	0,651	0,651	0,651
8	—0,836	0,401	0,401	0,401
9	0,401	0,401	0,401	0,401
10	0,906	0,410	0,410	0,410
11	0,410	0,410	0,410	0,410
12	0,221	0,410	0,410	0,410
13	0,968	0,426	0,426	0,426
14	0,426	0,426	0,426	0,426
15	—0,844	—0,844	—0,844	—0,844
16	—1,290	—0,844	—0,844	—0,844
17	0,657	0,063	0,063	0,063
18	0,063	0,657	0,063	0,063
19	1,283	0,063	0,063	0,063
20	—1,563	—1,563	—1,563	—1,563

Примечание.

Взяты первые 20 чисел из первых 4 столбцов со страницы 154 таблиц нормальных отклонений из книги [Rand Corporation, 1955. A. Million Random Digits with 100000 Normal Deviates. New York: The Free Press].

Иллюстрация 3.6.2

Сравнение необработанных данных выборки № 1 с данными, сглаженными скользящими медианами по трем точкам

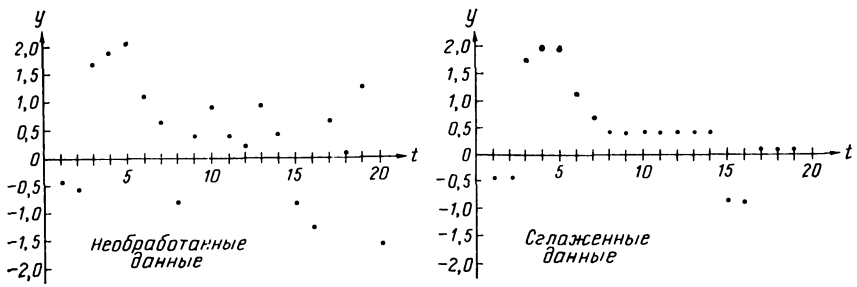


Иллюстрация 3.6.3

Сглаженные графики 10 выборок по двадцать наблюдений нормальны с отклонений в каждой с равным шагом по горизонтали. (Знаком «x» отмечены точки, через которые проходит прямая, проведенная методом наименьших квадратов.

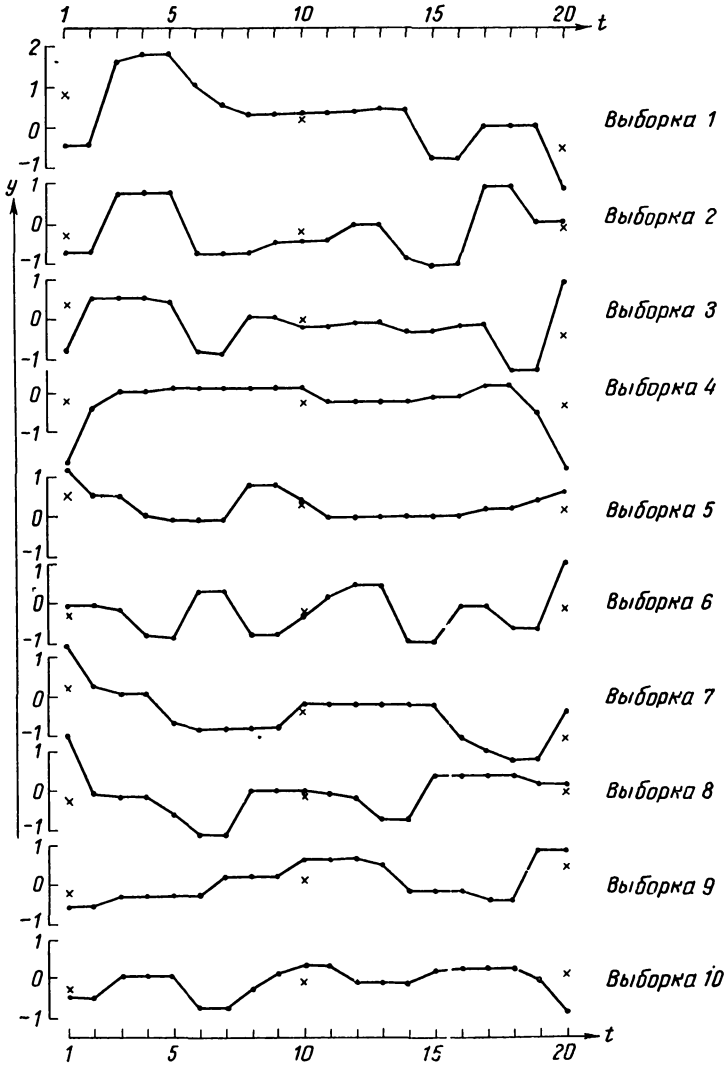


Иллюстрация 3.6.4

Сравнение угловых коэффициентов и свободных членов прямых, подобранных методом наименьших квадратов для первичных и сглаженных данных из илл. 3.6.3

Выборка	Угловые коэффициенты		Центрированные свободные члены (при $t=10{,}5$)	
	первичные	сглаженные	первичные	сглаженные
1	—0,0588	—0,0834	0,360	0,311
2	0,0054	0,0138	—0,148	—0,148
3	—0,0498	—0,0306	—0,188	—0,184
4	—0,0352	—0,0082	—0,125	—0,255
5	—0,0054	—0,0133	0,195	0,373
6	—0,0011	0,0052	—0,136	—0,226
7	—0,0720	—0,0753	—0,590	—0,486
8	0,0373	0,0256	0,120	—0,030
9	0,0505	0,0448	0,140	0,095
10	0,0154	0,0209	—0,059	—0,144
В идеале	0,0	0,0	0,0	0,0

Иллюстрация 3.6.5

Графическое сравнение угловых коэффициентов и свободных членов прямых для первичных и сглаженных данных

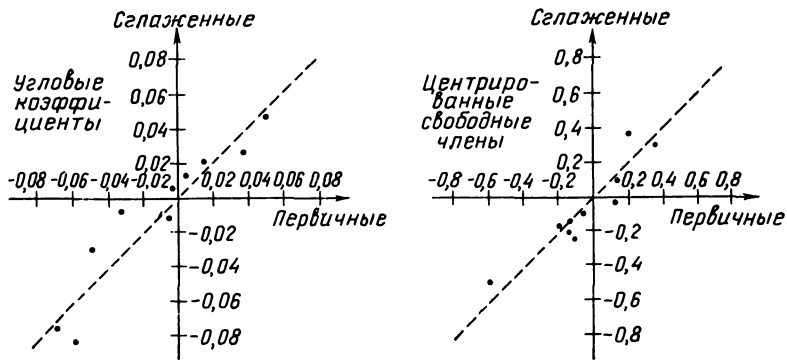


Иллюстрация 3.6.6

«Опора и консоль» для сравнения первичных и сглаженных данных по угловым коэффициентам и центрированным свободным членам как для самих значений, так и для их абсолютных величин

А. Значения

Угловые коэффициенты		
первичные		сглаженные
	0,06	
0	0,05	
	0,04	4
7	0,03	
	0,02	50
5	0,01	3
5	0,00	5
15	—0,00	8
	—0,01	3
	—0,02	
5	—0,03	0
9	—0,04	
8	—0,05	
	—0,06	
2	—0,07	5
	—0,08	3

Центрированные свободные члены		
первичные		сглаженные
	0,5	
	0,4	
6	0,3	17
	0,2	
924	0,1	
	0,0	9
5	—0,0	3
3284	—0,1	448
	—0,2	25
	—0,3	
	—0,4	8
9	—0,5	
	—0,6	

Б. Абсолютные величины (модули)

Угловые коэффициенты		
первичные		сглаженные
	0,08	3
2	0,07	5
	0,06	
80	0,05	
9	0,04	4
57	0,03	0
	0,02	50
5	0,01	33
155	0,00	58

Центрированные свободные члены		
первичные		сглаженные
	0,6	
9	0,5	
	0,4	8
6	0,3	17
	0,2	25
3284924	0,1	448
5	0,0	39

Иллюстрация 3.7.1

Функция регрессии f_1, показанная пунктиром, и ее сглаженный вид f_1^*, причем сглаживание проводилось методом скользящих медиан по трем точкам. Заметим, что $f_1 \neq f_1^*$ лишь при $t = 5$, 6 и 7

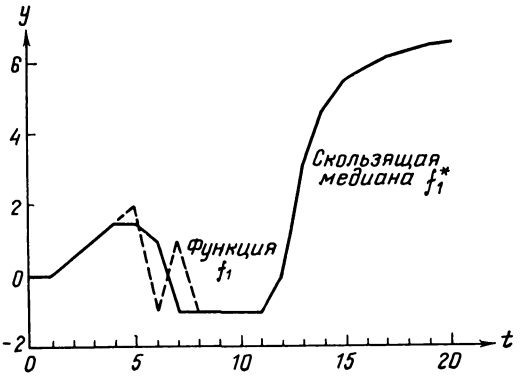
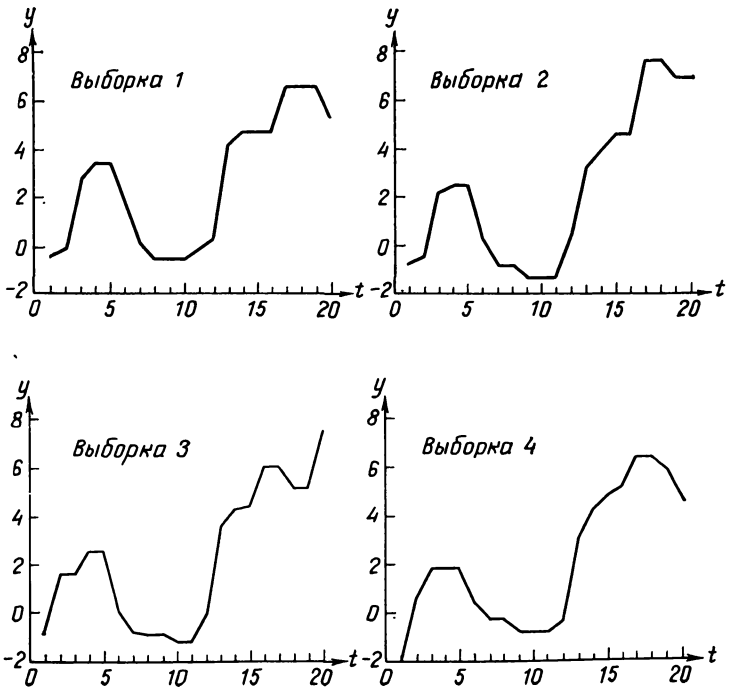
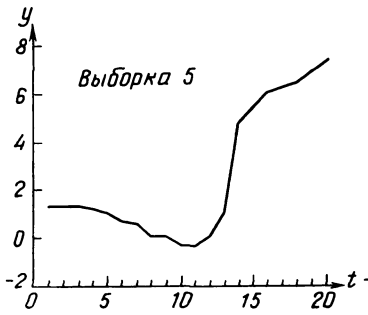
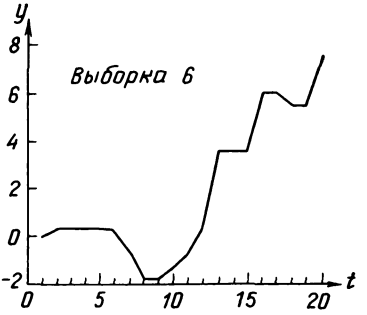


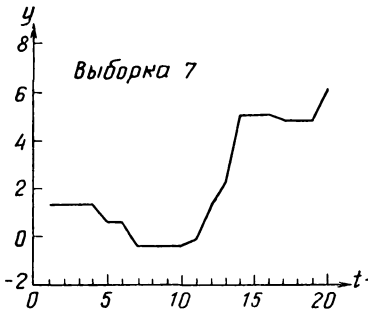
Иллюстрация 3.7.2

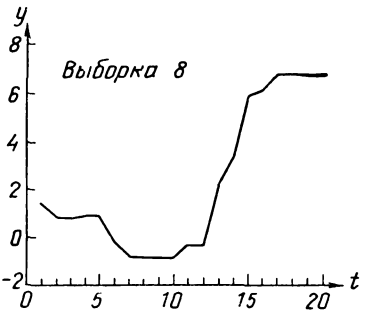
Данные, полученные добавлением к нерегулярной функции f_1 из илл. 3.7.1 несглаженных нормально распределенных ошибок и последующим сглаживанием скользящими медианами по трем соседним точкам

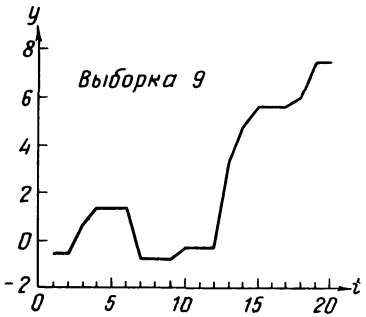



y
8
6
4
2
0
-2
Выборка 5
0
5
10
15
20
t


y
8
6
4
2
0
-2
Выборка 6
0
5
10
15
20
t


y
8
6
4
2
0
-2
Выборка 7
0
5
10
15
20
t


y
8
6
4
2
0
-2
Выборка 8
0
5
10
15
20
t


y
8
6
4
2
0
-2
Выборка 9
0
5
10
15
20
t

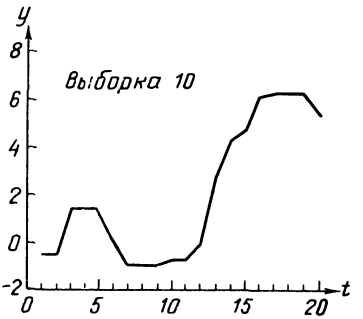

y
8
6
4
2
0
-2
Выборка 10
0
5
10
15
20
t

Иллюстрация 3.8.1

Числа Гольдбаха для двух групп четных чисел E, где $E = T + 0, T + 2, \ldots, T + 8$

Источник. Frederick Mosteller (1972). A dat-aanalytic look at Goldbach counts. — Statistica Neerlandica, 26, No 3, p. 227—242. Воспроизводится с разрешения редакции журнала.

$T/T+$	0	2	4	6	8
0	—	0	0	1	1
10	2	1	2	2	2
20	2	3	3	3	2
30	3	2	4	4	2
40	3	4	3	4	5
50	4	3	5	3	4
60	6	3	5	6	2
70	5	6	5	5	7
80	4	5	8	5	4
90	9	4	5	7	3
100	6	8	5	6	8
110	6	7	10	6	6
120	12	4	5	10	3
130	7	9	6	5	8
140	7	8	11	6	5
150	12	4	8	11	5
160	8	10	5	6	13
170	9	6	11	7	7
180	14	6	8	13	5
190	8	11	7	9	13
200	8	9	14	7	7
210	19	6	8	13	7
220	9	11	7	7	12
230	9	7	15	9	9
240	18	8	9	16	6
250	9	16	9	8	14
260	10	9	16	8	9
270	19	7	11	16	7
280	14	16	8	12	17
290	10	8	19	8	11
300	21	9	10	15	8
310	12	17	9	10	15
320	11	11	20	7	10
330	24	6	11	19	9
340	13	17	10	9	16
350	13	10	20	9	10
360	22	8	14	18	8
370	14	18	10	11	22
380	13	10	19	12	9
390	27	11	11	21	7

$T/T+$	0	2	4	6	8
400	14	17	11	13	20
410	13	11	21	10	11
420	30	11	12	21	9
430	14	19	13	11	21
440	14	13	21	12	13
450	27	12	12	24	9
460	16	28	12	13	24
470	15	13	23	14	11
480	29	11	14	23	9
490	19	22	13	13	23
500	13	15	27	15	14
9500	135	100	209	117	97
9510	253	97	105	211	86
9520	170	199	101	109	188
9530	123	94	232	97	104
9540	257	104	103	202	120
9550	130	199	104	102	193
9560	123	121	191	94	110
9570	295	97	98	245	90
9580	132	194	93	94	199
9590	159	109	221	96	97
9600	261	77	127	197	91
9610	137	190	117	93	234
9620	135	100	194	93	106
9630	264	112	97	212	91
9640	126	191	93	124	202
9650	123	105	192	99	114
9660	324	101	110	194	97
9670	140	220	119	98	191
9680	140	101	197	96	118
9690	284	93	101	193	106
9700	121	254	98	104	184
9710	134	99	192	117	102
9720	254	96	128	195	103
9730	161	193	103	101	196
9740	133	101	235	113	93
9750	286	108	99	194	124
9760	135	185	98	90	219
9770	132	117	191	108	99
9780	260	103	99	230	98
9790	151	205	102	110	200

Продолжение

$T/T+$	0	2	4	6	8
9800	147	118	205	95	101
9810	258	101	118	206	94
9820	135	206	98	100	259
9830	124	98	219	96	91
9840	264	129	104	196	100
9850	130	195	103	130	202
9860	144	104	204	86	99
9870	316	102	104	208	110
9880	156	200	117	101	196
9890	146	103	214	96	118

$T/T+$	0	2	4	6	8
9900	301	98	102	211	94
9910	134	233	100	109	223
9920	141	112	200	122	108
9930	266	105	103	202	103
9940	162	200	113	113	196
9950	126	95	248	98	105
9960	269	113	99	217	120
9970	139	194	93	104	195
9980	136	135	211	103	110
9990	269	102	98	255	99
10000	127				

Иллюстрация 3.8.2

Числа Гольдбаха для четных чисел $2N$, где $N = 1, 2, 3, \ldots$ Ломаная линия проходит через центроиды последовательных групп по 10 точек, $N = 0 \div 9$, $10 \div 19$ и т. д.

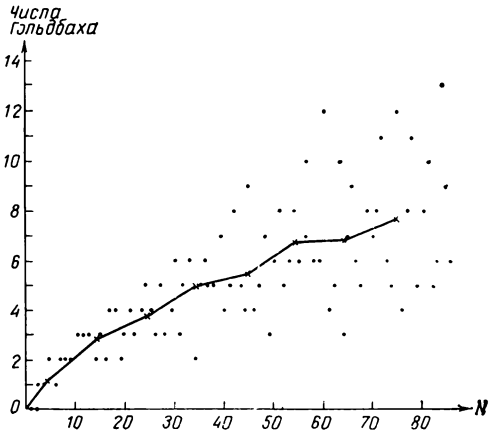


Иллюстрация 3.8.3

Числа Гольдбаха после удаления N, кратных 3. Тренд и увеличение размаха с ростом N остаются

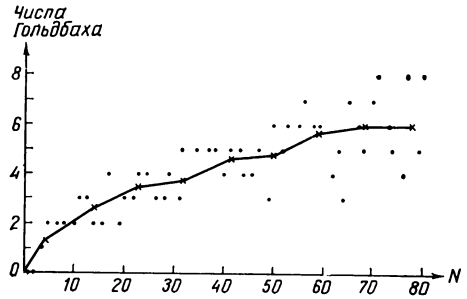


Иллюстрация 3.9.1

Максимум январской температуры в зависимости от географической широты («x» означает, что два наблюдения совпали)

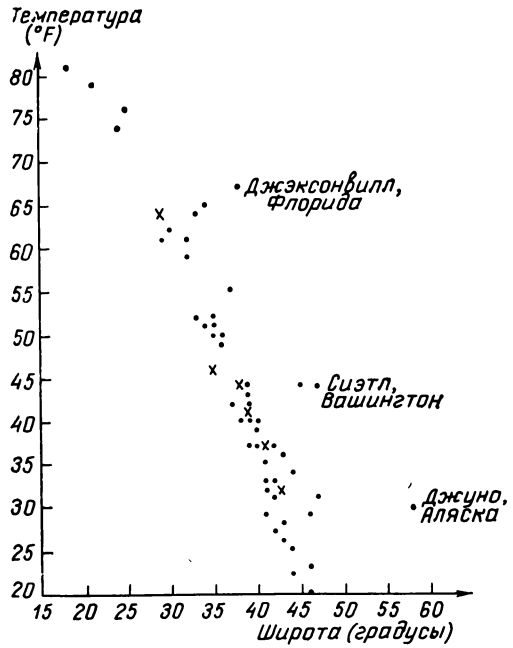


Иллюстрация 3.9.2

Зависимость остатков (от линейной связи максимальной январской температуры с широтой) от долготы

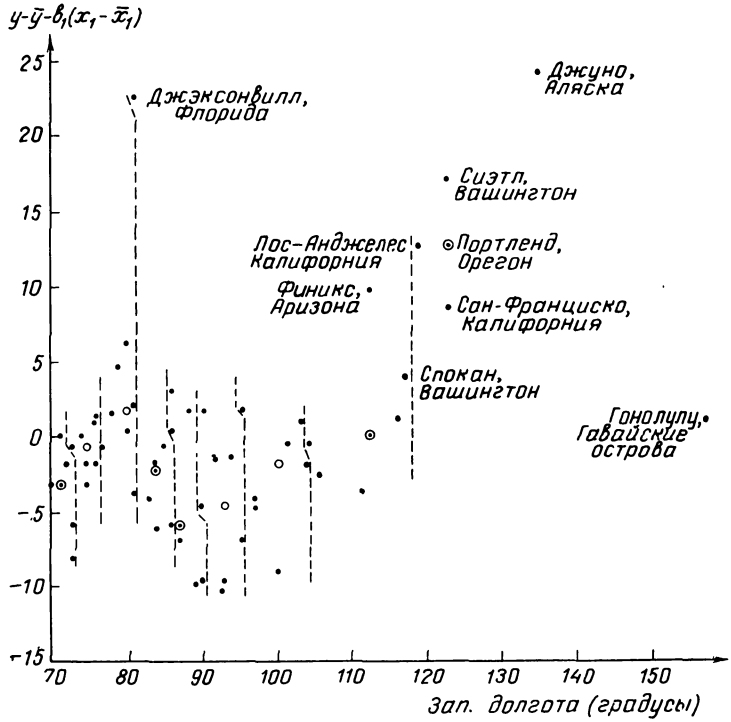


Иллюстрация 3.9.3

Максимум январской температуры по Фаренгейту* за период 1931—1960 гг. для ряда городов США с указанием их широты, долготы и высоты над уровнем моря

Города США	Максимальная январская температура °F	x_1 широта, град.	x_2 долгота, град.	x_3 высота, футы	Остатки $y-\bar{y}-b_1 \cdot (x_1-\bar{x}_1)$
Мобил, Алабама	61	30	88	5	2,0
Монтгомери, Алабама	59	32	86	160	2,9
Джуно, Аляска	30	58	134	50	24,3
Финикс, Аризона	64	33	112	1090	9,8
Литл-Рок, Арканзас	51	34	92	286	—1,2

Продолжение

Города США	Максимальная январская температура °F	x_1, широта, град.	x_2, долгота, град.	x_3, высота, футы	Остатки $y-\bar{y}-b_1 \times (x_1-\bar{x}_1)$
Лос-Анджелес, Калифорния	65	34	118	340	12,8
Сан-Франциско, Калифорния	55	37	112	65	8,6
Денвер, Колумбия	42	39	104	5280	—0,5
Нью-Хейвен, Коннектикут	37	41	72	40	—1,7
Уилмингтон, Делавэр	41	39	75	135	—1,5
Вашингтон, Окр. Колумбия	44	38	77	25	—0,5
Джэксонвилл, Флорида	67	38	81	20	22,5
Ки-Уэст, Флорида	74	24	81	5	2,4
Майами, Флорида	76	25	80	10	6,3
Атланта, Джорджия	52	33	84	1050	—2,2
Гонолулу, Гавайские о-ва	79	21	157	21	1,5
Бойсе, Айдахо	36	43	116	2704	1,2
Чикаго, Иллинойс	33	41	87	595	—5,7
Индианаполис, Индиана	37	39	86	710	—5,5
Де-Мойн, Айова	29	41	93	805	—9,7
Дубьюк, Айова	27	42	90	620	—9,7
Уичито, Канзас	42	37	97	1290	—4,4
Луисвилл, Кентукки	44	38	85	450	—0,5
Новый Орлеан, Луизиана	64	29	90	5	2,1
Портленд, Мэн	32	43	70	25	—2,8
Балтимор, Мэриленд	44	39	76	20	1,5
Бостон, Массачусетс	37	42	71	21	0,3
Детройт, Мичиган	33	42	83	585	—3,7
Су-Сент-Мари, Мичиган	23	46	84	650	—6,0
Мин-Сент-Пол, Миннесота	22	44	93	815	—10,8
Сент-Луис, Миссури	40	38	90	455	—4,5
Хелена, Монтана	29	46	112	4155	0,0
Омаха, Небраска	32	41	95	1040	—6,7
Конкорд, Нью-Гемпшир	32	43	71	290	—2,8
Атлантик-Сити, Нью-Джерси	43	39	74	10	0,5
Альбукерке, Нью-Мексика	46	35	106	4945	—4,3
Олбани, Нью-Йорк	31	42	73	20	—5,7
Нью-Йорк, Нью-Йорк	40	40	73	55	—0,6
Шарлотт, Сев. Каролина	51	35	80	720	0,7
Роли, Сев Каролина	52	35	78	365	1,7
Бисмарк, Сев. Дакота	20	46	100	1674	—9,0
Цинциннати, Огайо	41	39	84	550	—1,5
Кливленд, Огайо	35	41	81	660	—3,7
Оклахома-Сити, Оклахома	46	35	97	1195	—4,3
Портленд, Орегон	44	45	122	77	13,1
Гаррисберг, Пенсильвания	39	40	76	365	—1,6
Филадельфия, Пенсильвания	40	39	75	100	—2,5
Чарлстон, Южн. Каролина	61	32	79	9	4,9
Рапид-Сити, Южн. Дакота	34	44	103	3230	1,2
Нашвилл, Теннесси	49	36	86	450	0,6

Продолжение

Города США	Максимальная январская температура °F	x_1 широта, град.	x_2 долгота, град.	x_3 высота, футы	Остатки $y-\bar{y}-b_1 \cdot (x_1-\bar{x}_1)$
Амарилло, Техас	50	35	101	3685	—0,3
Галвестон, Техас	61	29	94	5	—0,9
Хьюстон, Техас	64	29	95	40	2,1
Солт-Лейк-Сити, Юта	37	40	111	4390	—3,6
Берлингтон, Вермонт	25	44	73	110	—7,8
Норфолк, Виргиния	50	36	76	10	1,6
Сиэтл-Такома, Вашингтон	44	47	122	10	17,0
Спокан, Вашингтон	31	47	117	1890	4,0
Мадисон, Висконсин	26	43	89	860	—8,8
Милуоки, Висконсин	28	43	87	635	—6,8
Шайенн, Вайоминг	37	41	104	6100	—1,7
Сан-Хуан, Пуэрто-Рико	81	18	66	35	—2,3

Источник. The World Almanac and Book of Facts, Newspaper Enterprise Association, New York, 1973; температура — со с. 263, географические сведения — со с. 704—705.

---

* Между температурой по Фаренгейту (°F), принятой в США, и температурой по Цельсию (°C) имеют место следующие соотношения: $°F = {}^5/_9(F°-32)°C$; $°C = ({}^9/_5 C° + 32)°F$. (1 фут $\approx$ 0,3 м.). — *Примеч. ред.*

Иллюстрация 3.10.1
Образование 20 групп и данные о медианном семейном доходе: первоначальные, медианные и сглаженные

x % отдельных квартир	y медианный семейный доход	y'-медиана y	Сглаженные[1] y'	Медиана для x
от 23 до 30	7151, 8380, 6910	7151		25
40	7003	7003	7151	40
от 56 до 59	7538, 8372	7955		58
от 68 до 70	9588, 7451, 9540	9540	7955	69
от 72 до 75	7113, 6908, 6693, 7495, 6278, 9985	7010		74
78	6160, 4572, 6806, 12264	6483	7010	78
от 79 до 80	5434, 6539, 9712, 9518, 8088	8088	7662	79
от 81 до 82	6522, 12357, 7662, 7550, 11108	7662		82
от 83 до 85	7741, 7371, 7003, 8863, 8895, 7413	7577	7577	84
86	8596, 4904, 7475, 6489	7276	7577	86
87	8123, 6338, 7973, 7753	7863	7494	87
88	7474, 8561, 7494, 7276	7494		88
89	5567, 8265, 7260, 6908, 8446,	7260	7420	89

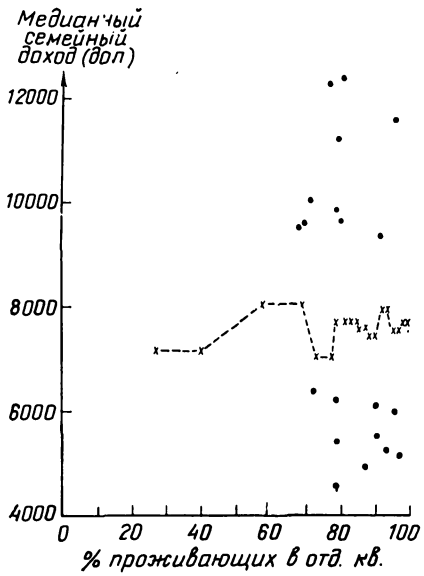
Продолжение

x % отдельных квартир	y медианный семейный доход	y' = медиана y	Сглаженные[1] y'	Медиана для x
от 90 до 91	6613, 6076, 8685, 7597, 7243, 8728	7420		90
от 92 до 93	7869, 7782, 7978, 5306, 9282, 8368	7880		92
94	7189, 11478, 8888, 6922	8039	7880	94
95	8191, 7908, 5909, 7169	7538		95
от 96 до 97	9043, 8336, 5182, 6602, 6615	6615	7538	96
98	8602, 7186, 7365, 8363, 8671, 9236	8482	7656	98
99	7936, 7467, 8998, 6987, 7656	7656		99

[1] Повторными скользящими медианами по трем значениям; показаны только значения, отличающиеся от значений y'.

Иллюстрация 3.10.2

10 максимальных и 10 минимальных среднесемейных доходов из всех данных вообще и данные, сглаженные медианами, из 20 групп с илл. 3.10.1



Глава 4 ● ЛИНЕАРИЗАЦИЯ КРИВЫХ И ГРАФИКИ

ИДЕЯ ЛИНЕАРИЗАЦИИ

Когда связь между переменными хорошо выражена, удобно, чтобы она оказалась линейной. Интерполяция и интерпретация становятся легкими, да и анализ остатков от такой регрессии много проще. В этой главе мы предлагаем некоторые пути к спрямлению кривых.

Относительно эмпирически установленной связи между переменными нельзя надеяться, что ее спрямление в области эксперимента будет спрямлять ее и далеко за пределами этой области. Конечно, нам может и повезти, но, как правило, нужна хоть какая-нибудь теория или априорная информация, чтобы на это рассчитывать.

В данной главе мы будем учиться преобразовывать одну или обе переменные для спрямления зависимости. Если y и x — это переменные, то рассматриваются преобразования y или x (или и y, и x). Основные инструменты при этом — последовательность преобразований и руководящие правила для определения этой последовательности.

Изучение технических приемов упростится, если сначала мы сконцентрируем внимание на спрямлении детерминированных функций, а затем распространим результаты на диаграммы рассеяния и другие представления эмпирических данных.

4.1. ПОСЛЕДОВАТЕЛЬНОСТЬ ПРЕОБРАЗОВАНИЙ

Так как нам нужна какая-то система преобразования, прежде всего рассмотрим степенные преобразования. Пусть, для начала, степень p имеет значения

$$-3,\ -2,\ -1,\ -\frac{1}{2},\ \#,\ \frac{1}{2},\ 1,\ 2,\ 3$$

(о знаке $\#$ — несколько позднее). Рассмотрим только положительные значения переменной, которую для удобства обозначим t.

Во-первых, хотелось бы иметь множество преобразований, каждое из которых монотонно изменяется в одном и том же направлении. Функция t^p возрастает вместе с t при $p > 0$ и убывает при $p < 0$. Чтобы сделать все эти функции возрастающими, воспользуемся при отрицательном p выражением $-t^p$.

Во-вторых, какую форму имеют эти кривые? Когда $p > 1$, они выпуклы вниз (вогнуты) $\smile$, когда $p = 1$, кривая спрямляется, а при $p < 1$ все кривые выпуклы вверх $\frown$. На илл. 4.1.1 представлены эти кривые, причем изменены шкалы (как на илл. 5.5.1). Они несколько растянуты за счет добавления констант для прояснения характера y.

Что же мы выберем для #? Значение $p = 0$ дает константу, которая бесполезна. Поэтому мы берем $\log t$. (Эти степени t можно ведь рассматривать как возникающие из выражения $\int t^{p-1}\,dt$. Тогда $p = 0$ приводит к $\log t$.) Можно придумать и что-нибудь еще и прийти, возможно, к иному результату. Кривая $\log t$ хорошо подходит, и у нас нет желания отбросить ее еще и потому, что логарифмическим преобразованием пользуются наиболее часто. Так что в качестве # возьмем $\log t$, а не t^0.

Выбирая преобразование, мы будем перебирать разные степени в поисках той, что спрямляет лучше. Пробуждению нашей интуиции должен помочь разбор одного примера, в котором вся информация находится в нашем распоряжении. Затем мы сформулируем ряд правил.

4.2. ПРЕОБРАЗОВАНИЕ $y = x^2$

Для нашего условного примера возьмем функцию

$$y = x^2, \quad x \geqslant 0.$$

Ее график дан на илл. 4.1.2. Отметим, что она вогнута и растет с ростом x.

1. *Каким преобразованием y можно спрямить эту кривую?* Посмотрим, что произойдет, если заменить y на y^2 или $y^{1/2}$. Достаточно рассмотреть два интервала: от 0 до 1 и от 1 до 4, поскольку при $t = 1$ все t^p равны единице.

Преобразование y^2. Если y заменить на y^2, то все точки $0 < x < 1$ станут ниже, чем были ранее, так как возведение в квадрат числа между нулем и единицей уменьшает его. Возведение в квадрат y при $x > 1$ увеличивает y. В итоге мы получим еще более изогнутую кривую, чем раньше.

Три точки, две прямые. Проведем тот же анализ количественно. Можно рассчитать угол наклона хорды между $x = 0$ и $x = 1$ и то же между $x = 1$ и $x = 4$ как для исходной, так и для преобразованной кривой. В идеале углы наклона должны совпасть в обоих интервалах:

исходные углы наклона	$0 < x < 1$	$1/1 = 1$	$1 < x < 4$	$15/3 = 5$;
новые углы наклона	$0 < x < 1$	$1/1 = 1$	$1 < x < 4$	$205/3 = 85$.

Отношение исходных наклонов (5 к 1) равно 5, а новых (85 к1) равно 85, значит, переход к y^2 не сближает углы, а наоборот. Так что мы пошли не в том направлении.

Попробуем теперь использовать $y^{1/2} = \sqrt{y}$. Точки в интервале $0 < x < 1$ *приподнимаются* при извлечении корня, действительно, извлечение корня или возведение в *любую* положительную, меньшую единицы степень *увеличивает* результат, например,

$$\sqrt{0{,}01} = 0{,}1; \quad \sqrt[3]{0{,}001} = 0{,}1.$$

А в интервале $1 < x < 4$ результаты снижаются. Итак, мы видим подъем слева и спад справа (от 1), что и должно быть при движении

в верном направлении. Таким образом, при замене y на $y^* = \sqrt{y}$ мы получаем связь

$$y^* = \sqrt{x^2} \quad \text{или} \quad y^* = x, \; x > 0,$$

а это — уравнение *прямой*, проходящей через начало координат.

Из этого примера мы извлекаем такой урок: для вогнутых монотонно возрастающих кривых в поисках подходящих преобразований y надо сдвигаться по шкале порядков в отрицательном направлении. Но мы ничего не узнали о том, как далеко надо идти. Здесь сама идея возведения в квадрат и извлечения квадратного корня возникла из-за того, что формула была известна.

Перейдем ко второму уроку, вернувшись к первоначальной кривой $y = x^2$.

2. *Как преобразовать x для линеаризации этой кривой?* Вновь наше знание функциональной формы допускает две замены: либо $x^* = x^2$, либо $x^* = \sqrt{x}$. Если мы заменим x на $x^* = x^2$, то точками графика будут (x^2, y) или (x^2, x^2), раз $y = x^2$, т. е. они лягут на прямую. Будем действовать последовательно. Мы неявно допускали, что в нашем расположении все степени p, но лишь немногие из них нужны. Предположим теперь, что нам разрешены лишь порядки — 3, —2, —1, # 1, 2, 3. Что мы взяли бы для y?

3. *Проба* $\log y$. Так как надо идти вниз по порядкам для y, перейдем от y к $\log y$ и посмотрим, что получится. Возьмем логарифм по основанию e, $\log_e$. Имеем $\log 0 = -\infty$, $\log 1 = 0$, $\log 4 = 1{,}39$:

исходные углы наклона	$0 < x < 1$,	1;	$1 < x < 4$,	5;
новые углы наклона	$0 < x < 1$,	∞;	$1 < x < 4$,	0,46.

Логарифм увеличивает угол наклона в левом интервале и уменьшает — в правом; оба изменения — в нужном направлении, но чересчур сильны.

Сдвиг. Мы могли бы избежать здесь бесконечностей, прибавив константу к y до логарифмирования. Выясним, какую же константу надо добавить, чтобы углы наклона двух хорд сравнялись. Мы хотим, чтобы

$$\frac{\log(1+c) - \log c}{1} = \frac{\log(16+c) - \log(1+c)}{3},$$

$$\log \frac{1+c}{c} = \frac{1}{3} \log \left(\frac{16+c}{1+c} \right).$$

Проверка показывает, что хорошим приближением будет $c = 0{,}95$. Мы могли бы, как обычно, округлить c до 1, но пусть лучше останется неокругленным. Итак мы пришли к преобразованию

$$y \text{ к } y^* = \log(y + 0{,}95).$$

Табулированные значения (с точностью до сотых) для $x = 0, 1, 2, 3, 4$ таковы:

x	0	1	2	3	4
y^*	—0,05	0,67	1,60	2,30	2,83

График представлен на илл. 4.2.2. С одной стороны, кривая не стала прямой, но, с другой стороны, она стала много прямее, чем исходная, и вполне пригодна для работы. К этой кривой можно подобрать прямую, отличающуюся от нее (по ординате) не более чем на 0,1, для всех $0 < x < 4$.

Отсюда делаем вывод, что совершенство не всегда необходимо. Обычно вполне годится и достаточна грубая сетка для p. Как правило, мы берем $p = 1/2$, $p = -1/2$, иногда $p = 1/3$ или же еще какие-нибудь удобные степени. Сколь дробные градации стоит проверять, зависит от материала.

Хотя мы и не доказали это, но все-таки вогнутость или выпуклость вместе с характером монотонности ведут к набору правил для подбора преобразования. В следующем параграфе мы дадим эти правила без дальнейших обсуждений.

4.3. ПРАВИЛО ВЫПУКЛОСТИ

Основное правило: двигайся по порядкам (ступенькам) в ту сторону, куда указывает выпуклость. Для x и для y выпуклость определяется отдельно. На илл. 4.3.1 дана памятка, как пользоваться порядками степеней при поиске преобразования. Дуги представляют четыре типа изменения кривых, четыре типа выпуклости *.

Возьмем новый пример. Данные для него представлены на илл. 4.3.2, они же нанесены на график илл. 4.3.3.

Пример 1. Воспользуйтесь заданным множеством преобразований для линеаризации кривой из илл. 4.3.3, меняя y.

Решение. Относительно y кривая выпукло возрастает, и надо двигаться вверх от $p = 1$. Выберем три точки и рассчитаем пары углов наклона. Пусть $x = 0{,}1;\ 3;\ 9$. Исходные углы наклона хорд:

$$0{,}1 < x < 3,\ \frac{2{,}88 - 0{,}93}{3 - 0{,}1} = 0{,}67;\ 3 < x < 9,\ \frac{4{,}16 - 2{,}88}{9 - 3} = 0{,}21.$$

Их отношение 0,67/0,21 = 3,2.

Перейдем теперь от y к y^2:

x	0,1	3	9
$y^* = y^2$	0,86	8,29	17,31

Получим углы наклона:

$$0{,}1 < x < 3,\ \frac{8{,}29 - 0{,}86}{3 - 0{,}1} = 2{,}56;\ 3 < x < 9,\ \frac{17{,}31 - 8{,}29}{9 - 3} = 1{,}50.$$

Их отношение 2,56/1,50 = 1,7.

---

*Если отсчитывать от «зенита» по часовой стрелке, то на илл. 4.3.1 представлены: выпукло убывающие, вогнуто возрастающие, вогнуто убывающие и выпукло возрастающие кривые. — *Примеч. пер.*

Отношение уменьшилось, но мы сдвинулись недостаточно сильно. Попробуем y^3:

x	0,1	3	9
$y^{**}=y^3$	0,80	23,89	72,0

$$0{,}1<x<3,\ \frac{23{,}89-0{,}80}{3-0{,}1}=7{,}97;\ 3<x<9,\ \frac{72{,}0-23{,}9}{9-3}=8{,}01.$$

Отношение 7,97/8,01 = 0,995.

Оно в пределах ошибки округления равно 1,00, и, следовательно, замена y на y^3 спрямит нашу кривую.

Поскольку «про себя» мы знаем, что исходные данные были округленными значениями функции $y=2\sqrt[3]{x}$, можно видеть, что наш поиск привел к практически точному преобразованию. Несмотря на это, будет поучительно посмотреть, что произошло бы, если бы вместо y мы попытались бы преобразовать x.

Пример 2. Линеаризовать кривую илл. 4.3.3, меняя x.

Решение. Кривая выпукла влево (см. илл. 4.3.1), так что мы будем двигаться влево от $p=1$ к порядку типа $\log x$ или $-1/x$.

Испробуем эти выражения для старых точек:

y	0,93	2,88	4,16
x	0,1	3	9
$x^*=\log x$	−1	0,48	0,95
$x^{**}=-1/x$	−10	−0,33	−0,11

Как показывают приведенные ниже расчеты, отношения постоянно растут.

	Левый интервал	Правый интервал	Отношение и его логарифмы	
x	$\frac{2{,}88-0{,}93}{3-0{,}1}=0{,}67$	$\frac{4{,}16-2{,}88}{9-3}=0{,}21$	0,31	−0,51
x^*	$\frac{2{,}88-0{,}93}{0{,}48-(-1)}=1{,}32$	$\frac{4{,}16-2{,}88}{0{,}95-0{,}48}=2{,}72$	2,06	0,31
x^{**}	$\frac{2{,}88-0{,}93}{-0{,}33-(-10)}=0{,}20$	$\frac{4{,}16-2{,}88}{-0{,}11-(-0{,}33)}=5{,}82$	29,01	1,46

Уже взяв логарифм, мы ушли слишком далеко.

Если нанести на график логарифмы отношений как функции от p в трех точках*, то можно проинтерполировать зависимость, чтобы найти нужную для спрямления степень преобразования. В результате получим 0,38, или примерно 1/3 (точный ответ).

*Знаку # на «оси порядков» p будет соответствовать начало координат. — *Примеч. пер.*

4.4. БОЛЕЕ СЛОЖНЫЕ КРИВЫЕ

Если кривая имеет S-образный вид, то маловероятно, что мы смогли бы избавиться от искривленности способами, описанными выше. Можно условно разбить ее на две части в точке перегиба $\int$ и спрямлять каждую отдельно. Аналогично можно обрабатывать и более сложные кривые. А иногда этот подход можно улучшить.

В отдельных случаях у нас есть путеводная нить теории, от которой мы можем ожидать пользы. Если бы мы, например, знали, что число простых чисел y среди всех целых, меньших x, равно приблизительно $x/\log_e x$, то построение графика для наблюдаемого числа y в зависимости от $x/\log_e x$, по-видимому, дало бы желаемую линейность.

4.5. ДИАГРАММЫ РАССЕЯНИЯ

Когда данные не столь гладки, как те, с которыми мы имели дело, приходится заменять значения в узкой области некоторыми средними — медианой или средним арифметическим как для y, так и для x, а затем работать с этими точками, как прежде. Причем медиана имеет некоторые преимущества, поскольку для преобразованных значений она совпадает с преобразованной медианой исходных значений, тогда как среднее этим свойством не обладает. Естественно, что среднее арифметическое привлекает своей аддитивностью, но в нелинейных ситуациях это не то, что нужно.

РЕЗЮМЕ. ЛИНЕАРИЗАЦИЯ КРИВЫХ

Линеаризация связи между y и x в пределах имеющихся данных не обязательно обеспечивает спрямление вне этих пределов.

«Линеаризация» сначала была атакована методом трех хорошо подобранных точек.

Мы используем отношение углов наклона двух отрезков, соединяющих среднюю точку с левой и правой соответственно, пытаясь сделать это отношение поближе к 1,0, пробуя преобразования разного порядка и, где это помогает, интерполируя. Если облако точек (диаграмма рассеяния) оказывается слишком неопределенным для эффективного сглаживания, мы делим (x, y)-точки на группы по x-значениям, затем находим x-медиану и y-медиану для каждой группы и работаем с ними.

Простейшие преобразования упорядочиваются по значениям t^p при разных $p \neq 0$ и $\log_e t$, занимающим место $t^0 = 1$ (которое всегда равно 1 и потому бесполезно).

Если наша кривая или облако точек все еще искривлены, мы сдвигаем порядок преобразования в направлении, которое подсказывает выпуклость кривой. Это правило можно применять с переменным успехом для преобразования как x, так и y.

ИЛЛЮСТРАЦИИ

Иллюстрация 4.1.1

Формы кривых $z = t^p$ для $p = -3, -2, -1, \#, 1, 2, 3$

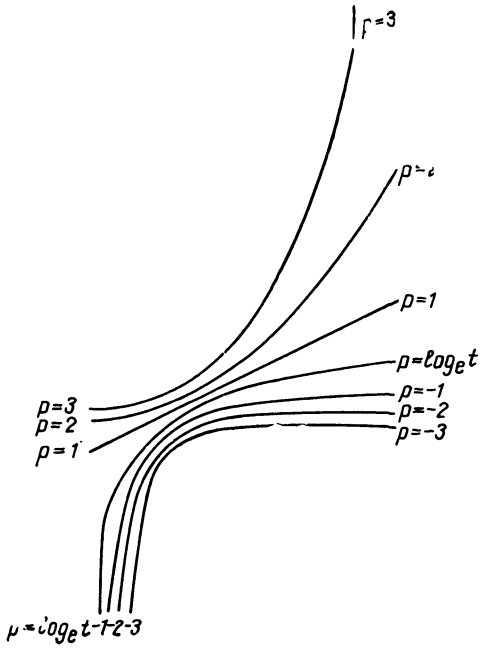


Иллюстрация 4.2.1

График $y = x^2$, направленный вверх; y возрастает вместе с x

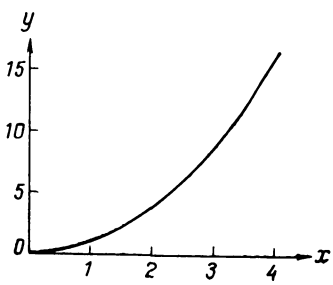


Иллюстрация 4.2.2

График функции $y^* = \log (x^2 + 0,95)$

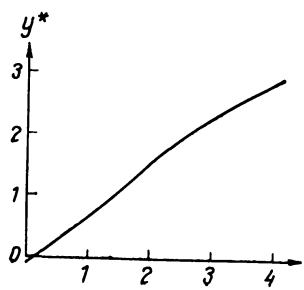


Иллюстрация 4.3.1

Памятка для выбора порядка преобразования. Стрелками отмечены направления выпуклостей для четырех типов кривых (по каждому переменному отдельно)

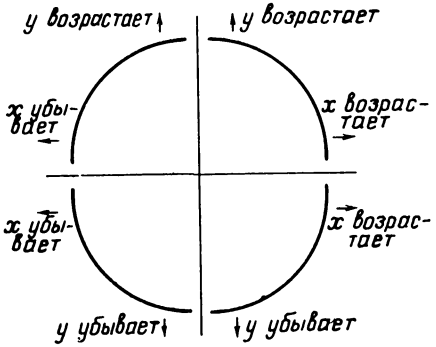


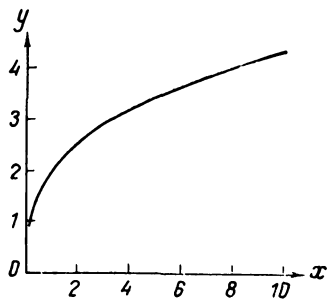
Иллюстрация 4.3.2

Данные по x и y для примера 1 параграфа 4.3

x	y	x	y
0,1	0,93	1	2,00
0,3	1,34	3	2,88
0,5	1,59	5	3,42
0,7	1,78	7	3,83
0,9	1,93	9	4,16

Иллюстрация 4.3.3

График для данных примера 1



Глава 5 ● ПРАКТИКА ПРЕОБРАЗОВАНИЙ

Числовые данные записываются или сообщаются преимущественно в форме, отвечающей удобству или привычке, а не сообразно требованиям анализа. Из-за этого нам часто приходится преобразовывать данные прежде, чем приступать к их анализу. В этой главе предлагаются преобразования, допускающие экономную реализацию, давая тем самым некоторую путеводную нить на первых шагах анализа.

Здесь не обсуждаются цели преобразований. Некоторые из них уже появлялись (в частности, в гл. 4) и некоторые появятся позже (например, в гл. 9 и 12). Каковы бы ни были цели, общие законы воздействия тех или иных преобразований на тот или иной набор чисел остаются одни и те же. Они-то и являются предметом настоящей главы.

Некоторые читатели могут при первом чтении ограничиться лишь начальными параграфами (5.1 и 5.2 — обязательны), а к остальным обращаться по мере надобности.

5.1. РАЗНОВИДНОСТИ ЧИСЛОВЫХ ДАННЫХ

Данные поступают к нам в самом разном виде, однако бóльшая часть их укладывается всего в несколько широких классов. В помощь кустарному методу — отыскивать известные аналоги к неизвестной ситуации — мы дадим эти широкие классы для ориентации в большинстве данных, какие нам хотелось бы анализировать.

Вот эти классы:

● счетные суммы («итоги»), где ни одно из слагаемых не может быть отрицательным;

● счетные разности («балансы»), которые могут иметь любой знак и быть как угодно велики и в положительную, и в отрицательную сторону;

● счетные доли (отношения счетных сумм) вроде того, как во фразе «Я видел 37 сорок и у 4 были желтые клювы, значит 4/37 принадлежат к разновидности желтоклювых»;

● ранги, где либо 1-й — наибольший, 2-й — следующий за ним ... либо 1-й — наименьший, 2-й — следующий за ним;

● упорядоченные ярлыки вроде A, B, C, D, E, F, или *, **, ***, ****.

(Заметим, что все ярлыки упорядочены. Некоторые данные поступают в виде «имен» (скажем, малиновка, черный дрозд, воробей и т. д.), и они не переводятся в числа, хотя сосчитать их *число* вполне возможно, так что в преобразовании «имен» нет нужды. Номера на футболках у

футболистов в действительности те же «имена». Просто-напросто само приписывание чисел еще не создает числовых данных).

Если приходится иметь дело с числовыми данными, не подходящими ни под один класс, то мы пытаемся найти аналогичные правила, например:

● если данные в одном направлении меняются неограниченно, а в другом имеют определенную границу, скажем A, то естественным аналогом будет класс «счетных сумм». Правило таково: берем $y - A$, если A — нижняя граница, и $A - y$, если A — верхняя граница, и обращаемся с ними так, как будто это счетные суммы;

● если же данные ограничены с двух сторон, скажем $B \leqslant y \leqslant C$, естественным аналогом будут счетные доли. Правило таково: к этому классу относим величины $(y - B)/(C - B)$.

Преобразования счетных сумм. Если y — счетная сумма, то наиболее часто используются преобразования

$$\log(y+c) \text{ или } (y+c)^p,$$

где c и p — константы, причем c часто равно нулю. Выбор c, не равного нулю, позволяет избежать трудности логарифмирования там, где $y = 0$, и помогает таким же образом для степеней $p < 0$, например для $p = -1$. Мы называем такое c «сдвигом» и говорим, когда $c \neq 0$, о «логарифмировании со сдвигом» или о «возведении в степень со сдвигом».

Наиболее обычные степени: $p = 1/2$, $p = -1$, $p = 1/3$ (в порядке частоты использования; см. параграф 4.1 и последовательность преобразований из илл. 4.1.1 гл. 4; $p = 1$ — это особый случай).

Если отношение максимального значения y к минимальному велико, то мы обычно начинаем с анализа $\log y$.

Преобразования разностей. «Балансы» сами по себе преобразуются редко. Чаще за счет погружения в суть данных удается установить, что «баланс» y есть разность

$$y = z - u,$$

где z и u — величины класса «счетные суммы». Затем мы изучаем новые числа, y^* — часто как «балансы» в форме

$$y^* = (\text{преобразованное } z) \; (\text{преобразованное } u).$$

Так как знание только значения y не задает y^*, то эти числа означают нечто большее, чем просто преобразование y.

Если V — сумма, то $\log V$ — «баланс», так как логарифмы чисел между нулем и единицей отрицательны и не ограничены в обоих направлениях. Таким образом, один из возможных способов преобразования «баланса» y дается функцией e^{cy}. Так как $e^{cy} = (e^c)^y$, то, взяв логарифм по основанию $b = e^c$, можно записать

$$\log_{e^c} e^{cy} = y,$$

и, таким образом, e^{cy} суть сумма, а его логарифм (по основанию $b = e^c$) и есть наш «баланс» y. Некоторые относительно редкие преобразования счетных разностей построены по этому образцу.

Преобразования долей. Для анализа счетных долей мы часто сравниваем доли в одном классе, скажем A, с долями в другом, скажем не A, симметричном во всем, *кроме* знака. Для иллюстрации одного метода воспользуемся нашим примером с 4 и 37 сороками:

$$\left(\text{преобразование } \frac{4}{37}\right) - \left(\text{то же самое преобразование } \frac{33}{37}\right).$$

Такое преобразование можно назвать «свертыванием». Некоторые свертывания основаны на преобразованиях, уже описанных выше для класса сумм. В более сложных случаях обычно полезны свертывания свертываний. Однако здесь мы о них почти ничего не скажем. Те, кто уже знаком с такими вещами, могут припомнить такие термины, как пробит-анализ, нормализующие преобразования, нормированные нормальные (или гауссовы) отклонения, прямые и обратные тригонометрические преобразования арксинуса и др. Логистическое преобразование, принадлежащее к «сложному» случаю, есть просто-напросто свертывание логарифма, так как

$$\log\left(\frac{p}{1-p}\right) = \log p - \log(1-p) = (\text{преобразование } p) - (\text{то же самое преобразование } (1-p)).$$

Преобразования рангов. Как мы убедимся позже, свертывание логарифмов со сдвигом служит одним из удобных и эффективных преобразований рангов.

Преобразования ярлыков. Наиболее полезный подход здесь, видимо, состоит в том, чтобы задаться некоторым стандартным распределением и поставить вопрос: «Если мы разделим это распределение соответственно наблюденным долям, то на что придутся центры тяжести отдельных частей?» Подробнее — в свое время.

Примечание. К каждому из названных классов мы еще вернемся в этой главе.

5.2. БЫСТРОЕ ЛОГАРИФМИРОВАНИЕ

Мы живем в век быстродействующей вычислительной техники, поэтому многие будут удивлены, что сплошь и рядом ручным способом многое можно сделать быстрее, чем с использованием сложных программ. Однако этой возможностью часто пренебрегают, особенно при исследовании данных, для которых многомерные анализы и итерации — правило. Так, если только не быть асом программирования и не иметь солидного запаса программ и машинного времени, то разумный объем рациональной ручной работы может оказаться выгоднее программирования на ЭВМ какого-то единственного избранного алгоритма. Поэтому, даже при большой любви к вычислениям на ЭВМ, мы все же высоко ценим ручную обработку и даем рекомендации для ее ускорения.

Если у нас есть настольный калькулятор, делающий больше операций, чем $+$, $-$, $\times$, $\div$, то надо лишь нажать кнопку, чтобы ввести число под знак логарифма. Более того, мы даже можем выбирать при расчете число десятичных знаков. Обычно мы удерживаем их больше, чем нуж-

но, замедляя вычисления. Хороший практический совет — начинать работу с двумя десятичными знаками, что, как правило, на один больше, чем нам надо, ну а если окажется, что этого мало, что ж, вернемся и пересчитаем заново. Если наш калькулятор требует двух «нажатий» кнопок для $\log_{10}$ и одного для $\log_e = \ln$, то давайте брать $\log_e$ с двумя десятичными знаками.

Когда нет настольной машинки, делу могут очень помочь таблицы. Илл. 5.2.1 дает удобную форму таблицы, взятой из *EDA*, гл. 3 (к которой мы и отсылаем вас за дальнейшими деталями, хотя таблица говорит сама за себя).

Использование илл. 5.2.1 для быстрого логарифмирования по основанию 10. Сначала находим характеристику, затем мантиссу, определив место нашего числа между двумя числами в первом столбце и взяв число во втором столбце.

Мы назвали таблицу «сканирующей», так как в ней x разбиваются путем сканирования значений $f(x)$ с заданным шагом на интервалы с равными значениями $f(x)$ в каждом из них (с точностью до принятого числа десятичных знаков). Эта таблица, следовательно, существенно отличается от обычных, где $x = y$ соответствует значение $f(x)$, а для отсутствующих в таблице x нужна интерполяция.

Пример. Так, число 82,1 имеет характеристику 1 (из треугольника в илл. 5.2.1Б, левая сторона). А теперь рассмотрим его как четырехзначное, 8210, и прочитаем в последнем столбце илл. 5.2.1А значение 0,91; таким образом, log 82,1 = 1,91. Если бы у нас было 82,23, то по правилу округления мы выбрали бы *четный* вариант, т. е. 0,92.

Логарифмы со сдвигом. Для

$$\log(y+c),\ c > 0$$

мы прибавляем c к y, где y — «итог», и обращаемся к илл. 5.2.1. Для целочисленных y возникают некоторые дополнительные удобства. Здесь возможны три случая — с малым, средним и большим c. Если c — малое, им можно пренебречь при $y > 1$; если c — среднее, берем долю, скажем, 1/6; если c — большое, когда оно достаточно велико, его стоит округлить до целого.

Для малых и больших c (кроме совсем малых y в первом случае) мы должны только упомянуть удобную таблицу, которая всегда эффективна при отыскании логарифмов целых чисел. Следующая дополнительная таблица $\log(y + c)$ показывает, почему это так:

Краткая таблица $\log(y+c)$, $c \geqslant 0$

	$y=0$	$y=1$	$y=2$	$y=3$	$y=4$
$(c=0)$	$(-\infty)$	(0,00)	(0,30)	(0,48)	(0,60)
$c=0{,}01$	−2,00	0,00	0,30	0,48	0,60
$c=0{,}03$	−1,52	0,01	0,31	0,48	0,61
$c=0{,}1$	−1,00	0,04	0,32	0,49	0,61
$(c=0{,}25)$	(−0,60)	(0,10)	(0,35)	(0,51)	(0,63)

Ясно, что нам не слишком нужна эта таблица, когда $y \geqslant 1$, а $c = 0{,}01$ или 0,03, когда $y \geqslant 2$, а $c = 0{,}1 \div 0{,}25$. Изменение c вряд ли сильно изменит столбцы соответствующих y.

Для средних c удобно ввести какое-нибудь одно значение. Есть некоторые основания выбрать $c = 1/6$. Но более важно, что 1/6 работает, видимо, так же, как и классические 1/10 или 1/4. Так, илл. 5.2.2 дает некоторые значения $\log(y + 1/6)$ с точностью до двух десятичных знаков. Когда $y \geqslant 10$, $\log(y + 1/6)$ настолько приближается к $\log y$, что мы вполне можем пользоваться таблицей илл. 5.2.1.

Те читатели, которые этого хотят, могут теперь перейти к гл. 7, возвращаясь к пропущенному материалу по мере надобности.

5.3. БЫСТРОЕ ВЫЧИСЛЕНИЕ КВАДРАТНЫХ КОРНЕЙ И ОБРАТНЫХ ВЕЛИЧИН

После логарифмов чаще всего мы пользуемся корнями (квадратными) и обратными величинами. Ручной калькулятор опять-таки был бы здесь лучше всего. Однако его может не быть и тогда обращаемся к таблице.

При извлечении корней мы должны быть несколько внимательнее, так как $\sqrt{20} = 4{,}47$ — это отнюдь не $\sqrt{2} = 1{,}41$ или $\sqrt{200} = 14{,}1$. Подробности даны и разъяснены в примерах илл. 5.2.3А.

Примеры. Числа 124,2 и 1242 разбиваются на два двузначных периода слева от десятичного знака. Каждый период «порождает» одну значащую цифру корня; они обозначены здесь буквами a, b.

Образцом для 124,2 служит 1 24,2; изолированное однозначное число слева устанавливает место исходного числа в первом столбце илл. 5.2.3В. Так как 1242 имеет два знака в левом крайнем периоде (12 42), то его место в третьем столбце таблицы из илл. 5.2.3 В.

Число 0,00654 разбивается вправо от «запятой» 0,00 65 4 и имеет, следовательно, две цифры в первом (слева направо) ненулевом периоде, так что его место в 5-м столбце таблицы. Каждый 00-период (после запятой) дает один 0 в ответ: значит, корень из 0,00 00 00 65 4 даст 0,00080.

Сжатая таблица для корней, показанная на илл. 5.2.3В, взята из *EDA* гл. 3, где ее применение обсуждается более подробно.

Обратные величины. Когда мы работаем с обратными величинами, зачастую удобнее использовать

$$-\frac{D}{y} \text{ или } C-\frac{D}{y},$$

где C и D — положительные константы, в таком случае результат растет с ростом y. Илл. 5.3.1 также из *EDA* гл. 3 дает компактную таблицу значений — $1000/y$.

Корни и обратные величины со сдвигом. Опять, имея дело с целочисленными «итогами», мы подбираем «сдвиги» подкоренного выражения или знаменателя. И опять, если «сдвиги» пренебрежимо малы или целочисленны, можем воспользоваться удобными таблицами квадратных корней и обратных величин. Что же касается небольших значе-

ний c, мы можем увидеть из табличек, приводимых здесь, что стоит беспокоиться лишь при очень малых c и «малых» y. Вот эти таблицы:

Краткая таблица квадратных корней из $y + c$

	$y=0$	$y=1$	$y=2$	$y=3$	$y=4$	$y=5$
$(c=0)$	(0,00)	(1,00)	(1,41)	(1,73)	(2,00)	(2,24)
$c=0{,}01$	0,10	1,00	1,42	1,73	2,00	2,24
$c=0{,}03$	0,17	1,01	1,42	1,74	2,01	2,24
$c=0{,}1$	0,32	1,05	1,45	1,76	2,02	2,26
$(c=0{,}3)$	(0,55)	(1,14)	(1,52)	(1,82)	(2,07)	(2,30)

Краткая таблица величин, обратных к $y + c$

	$y=0$	$y=1$	$y=2$	$y=3$	$y=4$	$y=5$
$(c=0)$	$(-\infty)$	(—1000)	(—500)	(—333)	(—250)	(—200)
$c=0{,}01$	—100000	—990	—498	—332	—249	—200
$c=0{,}03$	—33333	—971	—493	—330	—248	—199
$c=0{,}1$	—10000	—909	—476	—323	—244	—196
$(c=0{,}3)$	(—3333)	(—769)	(—435)	(—303)	(—233)	(—189)

Ситуация с $c = 1/6$ представлена в илл. 5.3.2 и 5.3.3. Опять-таки можно использовать гораздо бо́льшие таблицы (иногда с изменениями на 1 в последнем знаке) для бо́льших y.

5.4. БЫСТРОЕ ПРЕОБРАЗОВАНИЕ ДОЛЕЙ, ПРОЦЕНТОВ И ТОМУ ПОДОБНЫХ ВЕЛИЧИН

Мы предлагаем свертывать наши преобразования дробей так, чтобы 50% всегда приходились на начало отсчета (0,00). Как далеко может завести это соглашение? Величина свернутости здесь $p - (1 - p) = 2p - 1$. Можно ли представить 51% как 0,02, а 48% — 0,04 при всех преобразованиях, какие мы только захотим провести? Как показывает илл. 5.4.1, это вполне возможно с точностью до второго знака, пока не дойдем до 38% или соответственно 62%, дальше преобразования расходятся, сначала медленно, а затем все быстрее.

Теперь нам остается определить «свертывание корнем» как

$$\sqrt{2}\,(\sqrt{f} - \sqrt{1-f}) \text{ вместо } \sqrt{f} - \sqrt{1-f},$$

а «свертывание логарифмом» — как

$$\frac{1}{2}(\log_e f - \log_e (1-f)) = 1{,}1513\,(\log_{10} f - \log_{10} (1-f)),$$

вместо

$$\log_e f - \log_e (1-f) = \log_e \frac{f}{1-f}$$

или

$$\log_{10} f - \log_{10} (1-f) = \log_{10} \frac{f}{1-f}.$$

Для больших чисел порядка тысячи двух десятичных знаков вполне достаточно. Для гораздо больших чисел, видимо, имеет смысл брать больше знаков.

Сдвиг долей. Преобразования долей с целочисленными числителем и знаменателем и со сдвигом (пусть часть суммы, числитель, есть x, а вся сумма, знаменатель, есть n и c — сдвиг) мы начнем с такого варианта:

$$\frac{(\text{часть суммы}) + (\text{сдвиг})}{(\text{вся сумма}) + (\text{сдвиг}) + (\text{сдвиг})} = \frac{x+c}{n+2c},$$

так что эта доля совпадает с

$$1 - \frac{(\text{недостающая часть суммы}) + (\text{сдвиг})}{(\text{вся сумма}) + (\text{сдвиг}) + (\text{сдвиг})} = 1 - \frac{n-x+c}{n+2c} = \frac{x+c}{n+2c}.$$

Тогда в исходном выражении для свертывания логарифмом для $c = 1/6$

$$\log (\text{часть суммы}) - \log (\text{недостающая часть суммы})$$

превращается в

$$\log (\text{часть суммы} + 1/6) - \log (\text{недостающая часть суммы} + 1/6).$$

Заметим, что сдвигу подверглись $p = x/n$ и $(1 - p) = (n - x)/n$, так что в $\log n$ добавляются два сдвига. Точно так же при «логарифмировании со сдвигом» члены $\log (n + 2c)$ прибавлялись к нулю. Следовательно, мы можем воспользоваться илл. 5.2.2 для вычисления $\log (y + 1/6)$, где y — целое. Если нам выгодно выбрать 50% в качестве начала, то появляется множитель 1,1513, даже если мы берем доли со сдвигом. Мы можем забыть об этом множителе, когда расхождения малы.

5.5. СОВМЕСТИМОСТЬ СТЕПЕНЕЙ И ЛОГАРИФМОВ

Проанализировав совместимость в окрестности 50% преобразований долей, мы задаемся вопросом о совместимости степеней и логарифмов для преобразования «итогов». Все, что мы хотим, — это подобрать такой параметр A для y, чтобы сделать преобразования совместимыми. Проделав это однажды, можем брать любое из шести выражений, например следующих (см. илл. 5.5.1):

$$(p=2), \qquad \frac{A}{2}\left(\frac{y}{A}\right)^2 + \frac{A}{2};$$

$$(p=1), \qquad y = A\left(\frac{y}{A}\right);$$

$$(p=1/2), \qquad 2A\left(\frac{\sqrt{y}}{\sqrt{A}}\right)-A,$$

$$(p=\text{псевдонуль}), \qquad A\log_e\frac{y}{A}+A,$$

$$(p=-1), \qquad -A\left(\frac{y}{A}\right)^{-1}+2A,$$

$$(p=-2), \qquad -\frac{A}{2}\left(\frac{y}{A}\right)^{-2}+\frac{3}{2}A$$

или в общем виде

$$\frac{A}{p}\left(\frac{y}{A}\right)^p+\left(1-\frac{1}{p}\right)A$$

для всех $p \neq 0$ как преобразование y, которое совместимо с y в окрестности $y = A$.

Для удобства можно взять $A = 300$, снимая результаты с илл. 5.5.1. Значения в зависимости от p гладко меняются слава направо, от столбца к столбцу. Четвертый столбец илл. 5.5.1 ясно показывает, что совместимость степеней с y при $y = 300$ согласует логарифм, а вовсе не y^p при $p = 0$, которое ни с чем не сообразно.

В илл. 5.5.2 затабулированы эти же преобразования, но более подробно, с двумя знаками после запятой, в ближайшей окрестности $y = 300$, чтобы показать степень совместимости их около этой точки.

5.6. ПРЕОБРАЗОВАНИЯ ЯРЛЫКОВ

Допустим, что у нас есть объекты, классифицируемые как A, B, C, D или E, и что частота их появлений такова, как на илл. 5.6.1. Для каждого уровня мы можем рассчитать накопленные доли, как в илл. 5.6.1:

- долю p индивидуумов в предшествующих уровнях;
- долю P индивидуумов в предшествующих уровнях плюс данный уровень,

и затем воспользоваться либо таблицей, либо формулой из илл. 5.6.2 для вычислений соответствующих значений $\varphi(\cdot)$. Тогда значения этих функций станут пробными преобразованиями каждого уровня:

$$\frac{\varphi(P)-\varphi(p)}{P-p},$$

где P и p — две рассчитываемые доли.

В илл. 5.6.1Б и 5.6.1В даны подобные вычисления для гипотетических групп «хороших» и «плохих» студентов, классифицированных, как мы предполагаем, без связи с преобразуемыми уровнями. Было бы хорошо, если бы преобразования для этих двух групп были согласованы численно между собой (и с объединенной группой). Согласие здесь было бы превосходным, если бы преобразования этих двух групп отличались лишь на аддитивную константу. Нас также не должен беспокоить и постоянный множитель, если преобразование рассогласуется

лишь им. Отсюда возникает вопрос, будет ли предложенное преобразование одной группы приближенно линейной функцией преобразования другой?

Илл. 5.6.3 показывает график такой зависимости. Результат — не прямая линия, но и не сильно изогнутая или зазубренная. Соответственно мы считаем разумным каждое из преобразований, но от избытка осторожности предполагаем их усреднить, а именно:

A	B	C	D	E
—4,48	—2,54	—0,08	2,80	5,20

Действительно, они расположены вполне равномерно, так что можно ожидать удачи и от замены на —2, —1, 0, 1, 2. Это хорошо видно, если прибавить (—0,09) и поделить затем на 2,6, что дает

A	B	C	D	E
—1,75	—1,01	0,00	1,05	1,97

5.7. ПРЕОБРАЗОВАНИЯ РАНГОВ

При преобразовании рангов удобно обращаться с ними так, как будто это доли. Если наблюдение стоит на 5-м месте, а всего их 37, то можно считать, что оно делит все наблюдения на две части, а именно:

- 4 в одном классе и 33 в другом, если секущее значение приближается к 5-му с одной стороны;
- 5 в одном классе и 32 в другом, если секущее значение приближается к 5-му с другой стороны.

Эти два «долевых» случая приводят к свертыванию логарифмов со сдвигом

$$\log (4+1/6)-\log (33+1/6)$$

и

$$\log (5+1/6)-\log (32+1/6),$$

так что естественно связанное с рангом 5 выражение дается усреднением соответствующих аргументов:

$$\log (4\,1/2+1/6)-\log (32\,1/2+1/6)$$

или эквивалентно

$$\log (5-1/3)-\log (33-1/3),$$

где ранг 5 при отсчете с одного конца будет рангом 33 при отсчете с другого.

В общем виде имеем

$$\log (i-1/3)-\log (n+1-i-1/3)=\log \frac{i-1/3}{n+1-i-1/3}$$

для ранга i (от заданного конца). Расчеты облегчаются при использовании таблицы илл. 5.7.1, дающей значения $\log (i - 1/3)$. Очевидно, что можно забыть о (— 1/3) при i, превышающих 30, если нам хватит двух знаков.

5.8. ПЕРВАЯ ПОМОЩЬ ПРИ ПРЕОБРАЗОВАНИЯХ

Не всегда легко выбрать верное преобразование для имеющихся данных. Чтобы улучшить дело, было бы полезно знать (1) о чувствительности имеющихся данных к относительно слабым индикациям, (2) об опыте работы с другими наборами данных или (3) полагаться на знание объекта исследования. Но и этого может быть мало. Либо мы окажемся не подготовленными к тяжелому выбору преобразования, либо для надежного выбора будет слишком мало информации. И тогда возникает нужда в простых правилах, чтобы оказать «первую помощь» и дать преобразование, которое почти всегда не плохо, а часто и весьма хорошо.

Вот четыре правила, которые вполне эффективно обеспечат большинство наших нужд:

1. Берите логарифмы «итогов» (а если есть нули или бесконечности, то, может быть, их придется рассмотреть отдельно, см. следующий параграф).

2. Берите логистическое преобразование или свертывание логарифмом для долей и процентов; используйте его с точностью до множителя

$$\log\left(\frac{p}{1-p}\right)$$

(нули и единицы требуют специальной обработки). Если на данные наложены ограничения, так что $A \leqslant x \leqslant B$, где $A \neq 0$ и $B \neq 1$, то используйте

$$\log\left(\frac{x-A}{B-x}\right).$$

3. Преобразуйте i-й ранг из n по формуле

$$\log\left(\frac{i-1/3}{n-i+2/3}\right)=\log\left(\frac{3i-1}{(3n+1)-(3i-1)}\right).$$

4. «Балансы» оставляйте как есть.

Эти правила — не окончательный ответ (так же, как первая помощь пострадавшему не заменяет врача), но они — попытка доброго начала.

Дополнительная помощь. Что, если первой помощи недостаточно и негде взять хоть какой-нибудь руководящий принцип? В таком случае мы можем обратиться к «дополнительной помощи» в следующих направлениях.

Если мы начинали с «итогов», где первая помощь рекомендует логарифмы, то, может быть, придется уничтожить сделанное и вернуться от

$$x^* = \log x \text{ к } x = \operatorname{antilog} x^*.$$

Если мы готовы это сделать с x^*, то точно так же должны быть готовы поступить с cx^* вместо x^*, получая

$$x^{**} = \operatorname{antilog} cx^*.$$

что из-за равенства

$$\log(x^c) = c \log x = cx^*$$

даст нам

$$x^{**} = x^c$$

с некоторым показателем степени c.

Если x было долей или процентом и первая помощь рекомендовала

$$x^* = \log \frac{x}{1-x},$$

то обратный перерасчет cx^* дает

$$x^{**} = \frac{\left(\frac{x}{(1-x)}\right)^c}{1+\left(\frac{x}{(1-x)}\right)^c} = \frac{x^c}{x^c + (1+x)^c},$$

а для использования такого выражения у нас мало опыта. Поэтому, может быть, лучше начать с

$$x^* = \log \frac{x}{1-x} = \log x - \log (1-x)$$

и, уничтожив каждый логарифм в отдельности, перейти к

$$x^{**} = x^c - (1-x)^c,$$

что полезно, например, при $c = 1/2$.

Для рангов мы можем делать похожие вещи — опять же здесь мало опыта. Если мы рассматриваем

$$\log\left(\frac{i-1/3}{n-i+2/3}\right) = \mathrm{Gol}\left(\frac{i-1/3}{n+1/3}\right),$$

где

$$\mathrm{Gol}\left(\frac{i-1/3}{n+i/3}\right) = \log (p) - \log (1-p)$$

— функция, обратная к кумулятивному логистическому распределению, то почему бы нам не рассмотреть и

$$\mathrm{Aug}\left(\frac{i-1/3}{n+i/3}\right),$$

где Aug означает функцию, обратную к кумулятивной функции нормального (гауссовского) распределения. (Результат весьма близок к тому, что известно как «нормализованные баллы» (или ранги).)

С «балансами» можно обращаться так, как будто это ранги: взять n значений и упорядочить, затем приписать им ранги и производить те же преобразования рангов:

$$\log \frac{3i-1}{3(n+1)-(3i+1)} = \log \frac{i-1/3}{n-i+1-1/3} = \log (i-1/3) - \log (n-i+1-1/3),$$

что обсуждались в параграфе 5.7.

Ни один из этих путей не гарантирует хорошей организации данных, однако не тот, так другой часто оказывается удачным. Ну а если ни

один из них нас все-таки не устроит, попробуем отыскать сумму двух или более разных преобразований нашего x.

Упорядоченные значения. Случается, что мы имеем дело с переменными, значения которых не числа, но имеют определенный порядок, так же как уровни A, B, C, ... когда они идут в порядке «поступления». Чтобы перевести такие переменные в «регрессионные», надо придать их уровням числовые значения. Каждому уровню можно поставить в соответствие процент, предписанный ему куском логистического распределения. Тогда каждый уровень получит числовое значение, соответствующее центру тяжести (CG) этой части логистического распределения.

При расчетах на ЭВМ мы считаем CG отсекаемого числами $p = A$ и $p = B$ куска логистического распределения по формуле

$$\frac{B \log B + (1-B) \log (1-B) - A \log A - (1-A) \log (1-A)}{B-A},$$

или, что то же самое, можно провести анализ, как в параграфе 5 6 с таблицей из илл. 5.6.2.

Для ручного счета часто вполне годится формула

$$\frac{1}{6}\left(\log \frac{A}{1-A} + 4 \log \frac{(A+B)/2}{1-(A-B)/2} + \log \frac{B}{1-B}\right),$$

которая проще, поскольку нуждается лишь в таблице значений

$$\log\left(\frac{A}{1-A}\right).$$

Так можно поступать и со всеми данными, и с любой их частью, имеющей особенности в поведении. В последнем случае мы прежде всего смотрим на результаты для частей, ведущих себя аналогично (см. пример из илл. 5.6.1 и его обсуждение в тексте), а затем комбинируем их друг с другом.

5.9. ЧТО ДЕЛАТЬ С НУЛЯМИ И БЕСКОНЕЧНОСТЯМИ?

Как нам поступать, когда правила (1) и (2) из параграфа 5.8 предписывают брать логарифм от нуля? Ответ до некоторой степени зависит от того, что мы собираемся делать с переменной, и от того, сколь много у нас нулей.

Если среди y попадется лишь несколько нулей (или процент нулей мал), а задача состоит в подборе устойчивой кривой, так что любому заметно отклоняющемуся значению придается нулевой вес, то можно считать (условно), что

$$\log 0 = L,$$

где L берется существенно *меньшим любого из чисел* (но большим по модулю). Это позволяет приписать малому числу нулей нулевые веса вне зависимости от того, как они связаны с предшествующей подгонкой. Хотя иногда такой подход вполне приемлем при малом числе нулей, он не годится, когда нулей становится порядочно.

Другое простое решение — «сдвиг» логарифмов* с помощью

$$\log(x_j + c) \text{ или } \log(y + c)$$

вместо $\log x_j$ или $\log y$. Когда имеют дело с целочисленными «итогами», распространенными значениями для c служат 1,00 и 0,25, хотя изредка берутся числа и бóльшие, и меньшие. Аналитик, желающий работать всегда с одной-единственной поправкой (включая ранги и т. п.), может избрать для сдвига значение 1/6, а не 1/3, как мы уже пояснили. Когда мы имеем дело с произвольными «итогами», которые включают нули, то выбор $\log 0 = L$ вполне может «сработать», однако у нас нет большого опыта и твердого представления о том, что именно надо делать.

Хотя со сдвигами все понятно, мы не знаем, лучше ли обходиться для целочисленных «итогов» сдвигами или же лучше брать вместо $\log 0$ какое-нибудь подходящее отрицательное число (так как $\log 1 = 0$, любые отрицательные величины идут в нужном направлении). Если мы хотим взять такое значение «из головы», то должны сознавать, что мы предпочтем, например, при таких альтернативах:

$$-\log 4, \text{ или } -\log 6, \text{ или } -\log 8.$$

(Заметим, что $-\log 6$ есть логарифм среднего аргументов y от $-\log 4$ и $-\log 8$). Если мы хотим поменьше исказить суть, то можем разрешить величине, заменяющей $\log 0$, зависеть от частоты нулей. Бохидар, Грубер и Тьюки (Bohidar, Gruber, Tukey) изучали отдельные аспекты такого выбора и нашли, что разумно использовать $(\log p)/(1 - p)$, где p — доля нулей (если все элементы суть нули, то p рассчитывается так, как будто один элемент — не нуль). В разных частях таблицы могут быть и разные оценки для p.

Бесконечности. Иногда «итоги» становятся бесконечными либо на самом деле, либо с точки зрения практики. Время, потребное для того, чтобы крыса пробежала по лабиринту, может быть бесконечным, если она не бежит. Время, потребное ученику, чтобы стать мастером, может быть бесконечным, если он бросит учебу раньше, чем выучится. Время, за которое рыба погибнет из-за загрязнения воды, может быть практически бесконечным, если она жива в течение длительного периода, а эксперимент пора заканчивать, тем более, что большинство рыб уже давно погибло.

Одно из легких «лекарств» для таких бесконечностей — не логарифмическое преобразование результатов, а рассмотрение обратных величин. Величина, обратная времени для крыс, не желающих бегать, в точности равна нулю. Здесь фактически мы анализируем уже не время, а скорость.

Наши скорости — это «итоги», избавленные от бесконечностей, но зато теперь появились нули. Если мы хотим испробовать правила первой помощи, то мы можем совершать все те процедуры, что были описаны ранее в этом параграфе, в том числе и взять преобразование log (скорость ПЛЮС сдвиг), что целесообразно.

---

*Точнее, «сдвиг» данных от нуля вправо *перед* логарифмированием. — *Примеч. пер.*

Нули и бесконечности одновременно. А что, если у нас есть и нули, и бесконечности? Тогда мы вначале прибавим сдвиг, а затем возьмем обратные величины, что даст

$$\frac{1}{\text{итог ПЛЮС сдвиг}}.$$

Если же пожелаем перейти к логарифмам, то снова прибавим сдвиг, быть может, на другую величину, со «звездочкой», т. е.

$$\log\left(\frac{1}{\text{итог ПЛЮС сдвиг}}\ \text{ПЛЮС сдвиг}^*\right).$$

Знаки. Во избежание отрицательных значений некоторые используют положительные обратные величины, и если времена либо итоги, как правило или всегда, больше единицы, а сдвиг невелик, они берут

$$-\log\left(\frac{1}{\text{итог ПЛЮС сдвиг}}\right),$$

так как иначе большинство логарифмов будет отрицательным. Если данные измеряют «время до...», то анализируются скорости («быстроты») как обычные обратные величины и «медленности» — как логарифмы от обратных величин со сдвигом [1].

Другие предпочитают сохранять при преобразовании направленность. Если мы обратимся к «медленности», то при таком подходе надо брать отрицательные обратные величины

$$-\frac{1}{\text{итог}}$$

и всегда со знаком «минус» выражения

$$-\log\left(\frac{1}{\text{итог ПЛЮС сдвиг}}\right) \text{ или } -\log\left(\frac{1}{\text{итог ПЛЮС сдвиг}}+\ \text{сдвиг}^*\right),$$

к какому бы знаку (плюс или минус) это ни приводило.

Выбор знака не влияет на результат в большинстве конкретных вычислений, но может влиять на чувствительность к результатам промежуточных шагов или к намекам, появляющимся в процессе счета.

Доли, ранги и им подобные величины. Если мы исходим из долей и берем затем

$$x^* = \log\left(\frac{p}{1-p}\right)$$

как первую помощь, то тотчас же сталкиваемся с проблемой нулей — бесконечностей при $p = 0$ и $p = 1$. Однако возможна и запись

$$x^* = \log\frac{p}{1-p} = \log\frac{k}{n-k},$$

где k — число наблюдений из n, что, естественно, приводит к

$$x^* = \log\frac{k+\text{сдвиг}}{n-k+\text{сдвиг}},$$

где значения сдвига берутся теми же, то и для «целочисленных итогов».

[1] Здесь в оригинале игра слов. — *Примеч. пер.*

РЕЗЮМЕ. ПРЕОБРАЗОВАНИЯ

Мы различаем следующие классы переменных: (1) счетные суммы («итоги»), (2) счетные разности («балансы»), (3) счетные доли, (4) ранги, (5) ярлыки.

Самые распространенные преобразования y для (1) — это $\log(y + c)$ и $(y + c)^p$ с разными значениями для p и c.

Как правило, мы не преобразуем «балансы», находя, что разумнее преобразовывать две (или более) величины, разность между которыми и представляет собой баланс (в параграфе 9.6 мы встретим контрпример, где преобразование e^{cy} выглядит естественней).

Доли неплохо обрабатывать в терминах «свертывания» часто не самих величин, а логарифмов (включая логистическое преобразование) или корней исходных данных.

Сканирующие таблицы позволяют преобразовывать данные быстро и просто.

Наши сканирующие таблицы дают для логарифмов, корней и обратных величин два десятичных знака.

Согласованием в окрестности 50 % можно достигнуть такой степени свернутости как логарифмов, так и корней, что они дадут совпадения для интервала от 38 до 62% и будут достаточно близки для интервала от 25 до 75%.

Естественный подход к преобразованию ярлыков направлен на интервал, покрываемый данным уровнем (% до него, % в нем, % после него). Тогда уровень задается центром тяжести этого интервала либо для всех данных, либо для их хорошо отделимой части.

Илл. 5.6.2 в этой главе помогает определять центры тяжести всех или части данных для любых уровней.

Информацию от отдельных частей мы комбинируем, чтобы получить простое и полезное в широком интервале преобразование.

Ранги можно преобразовывать в терминах «свертывания логарифмов» со сдвигами и числителя, и знаменателя соответствующей дроби на — 1/3.

Мы обучаемся «первой помощи незнакомым данным» с помощью (1) логарифмирования «итогов», (2) свертывания дробей, процентов или других долей с логарифмическим преобразованием внутри «свертывания», (3) обработки рангов так, как уже объяснялось, (4) оставляя «балансы» в покое.

В первую помощь входит и выбор «сдвигов», когда это желательно или необходимо. Правила первой помощи для преобразования данных развиваются в правилах дополнительной помощи, которые заметно более гибки.

Идея «сдвигов» распространяется и на борьбу одновременно с «нулями» и «бесконечностями». Мы «сдвигаем» корни и логарифмы, взяв $c > 0$, в частности, $c = 1/6$. Можно «сдвигать» доли, прибавляя по 1/6 к порождающим их «итогам». Мы понимаем, как совместить любые два простых преобразования одновременно вблизи какого-нибудь заданного A. (Это обеспечивает вышеупомянутое простое линейное преобразование, совмещающее преобразования с общим c вблизи $(A—c)$. Читателю надо научиться совмещать преобразования с разными c.)

БИБЛИОГРАФИЯ

Bohidar N. R., Gruber D. G. and Tukey J. W. Efficacy estimates for parasite-count data, including zero counts. Submitted to Experimental Parasitology. (Направлено в печать.)

EDA — Tukey J. W. (1977). Exploratory Data Analysis. Reading, Mass., Addison-Wesley.

ИЛЛЮСТРАЦИИ

Иллюстрация 5.2.1

«Сканирующая» таблица десятичных логарифмов с двумя знаками

А. Основная сканирующая таблица для мантисс

Сканирующее значение	log	Сканирующее значение	log	Сканирующее значение	log	Сканирующее значение	log	Сканирующее значение	log
9886		**1567**		**2484**		**3936**		**6238**	
	0,00		0,20		0,40		0,60		0,80
1012		**1603**		**2540**		**4027**		**6382**	
	0,01		0,21		0,41		0,61		0,81
1036		**1641**		**2601**		**4121**		**6532**	
	0,02		0,22		0,42		0,62		0,82
1059		**1678**		**2660**		**4216**		**6683**	
	0,03		0,23		0,43		0,63		0,83
1084		**1718**		**2723**		**4316**		**6840**	
	0,04		0,24		0,44		0,64		0,84
1109		**1757**		**2786**		**4415**		**6998**	
	0,05		0,25		0,45		0,65		0,85
1136		**1799**		**2852**		**4519**		**7162**	
	0,06		0,26		0,46		0,66		0,86
1161		**1840**		**2917**		**4623**		**7328**	
	0,07		0,27		0,47		0,67		0,87
1189		**1884**		**2986**		**4732**		**7499**	
	0,08		0,28		0,48		0,68		0,88
1216		**1927**		**3054**		**4841**		**7673**	
	0,09		0,29		0,49		0,69		0,89
1245		**1973**		**3127**		**4955**		**7853**	
	0,10		0,30		0,50		0,70		0,90
1273		**2018**		**3198**		**5069**		**8035**	
	0,11		0,31		0,51		0,71		0,91
1302		**2066**		**3274**		**5187**		**8223**	
	0,12		0,32		0,52		0,72		0,92
1333		**2113**		**3349**		**5308**		**8413**	
	0,13		0,33		0,53		0,73		0,93
1365		**2163**		**3428**		**5433**		**8610**	
	0,14		0,34		0,54		0,74		0,94
1396		**2213**		**3507**		**5559**		**8810**	
	0,15		0,35		0,55		0,75		0,95
1429		**2265**		**3590**		**5689**		**9016**	
	0,16		0,36		0,56		0,76		0,96
1462		**2317**		**3672**		**5821**		**9225**	
	0,17		0,37		0,57		0,77		0,97
1495		**2372**		**3759**		**5957**		**9441**	
	0,18		0,38		0,58		0,78		0,98
1531		**2426**		**3845**		**6095**		**9660**	
	0,19		0,39		0,59		0,79		0,99
1567		**2484**		**3936**		**6238**		**9886**	

В спорных случаях берется «четное» число; так, для 1462 берем 0,16, а для 1495 берем 0,18.

Б. Таблица характеристик (с границами по степени десяти)

1		1
	+0	−1
10		0,1
	+1	−2
100		0,01
	+2	−3
1000		0,001
	+3	−4
10000		0,0001
	+4	−5
100000		0,00001
	+5	−6
1000000		0,000001

В. Примеры

	Число	Б		А		Логарифм числа
log	137,2	2	+	0,14	=	2,14
log	0,03694	−2	+	0,57	=	−1,43
log	0,896	−1	+	0,95	=	−0,05
log	174,321	+5	+	0,24	=	5,24

Иллюстрация 5.2.2

Таблица значений $\log_{10}(y + 1/6)$

Десятки \ Единица	0	1	2	3	4	5	6	7	8	9
00	−0,78	0,07	0,34	0,50	0,62	0,71	0,79	0,86	0,91	0,96
10	1,01	1,05	1,09	1,12	1,15	1,18	1,21	1,23	1,26	1,28
20	1,30	1,33	1,35	1,36	1,38	1,40	1,42	1,43	1,45	1,46

Пример

На пересечении строки 10 и столбца 2 имеем значение $y = 10 + 2 = 12$. Выход этой ячейки есть $\log 12\frac{1}{6} = 1{,}09$.

Иллюстрация 5.2.3

Сканирующая таблица для извлечения корня (квадратного)

А. Примеры

Считая от запятой, разбиваем число на пары цифр (периоды); так, 124,2 представляется как 1 24 2, а 1242 представляется как 12 42. Аналогично 0,00654 представляется как 00 65 4 (*или* 65 4).

Число	Периоды	Из табл. Б	Из табл. В	Число
124,2	1 24 2	ab,	112	11,2
1242	12 42	ab,	35	35,0
0,00654	00 65 4	0,0x	80	0,080

Б. Сканирующая таблица определения места запятой (входы — «жирные» числа, результат — между ними)

1			**1**
	a,	$0{,}x$	
1 00			**0,01**
	ab,	$0{,}0x$	
1 00 00			**0,00 01**
	abc,	$0{,}00x$	
1 00 00 00			**0,00 00 01**
	$abcd$,	$0{,}000x$	
1 00 00 00 00			**0,00 00 00 01**

В. Основная сканирующая таблица (вход и выход, как в табл. Б)

Сканирующее значение	Корень	Сканирующее значение	Корень	Сканирующее значение	Корень	Сканирующее значение	Корень	Сканирующее значение	Корень
98 01		**2 49 64**		**5 66**		**15 60**		**35 40**	
	100		160		240		40		60
1 02 01		**2 62 44**		**5 90**		**16 40**		**37 21**	
	102		164		246		41		62
1 06 09		**2 75 56**		**6 20**		**17 22**		**39 69**	
	104		168		252		42		64
1 10 25		**2 89 00**		**6 50**		**18 06**		**42 25**	
	106		172		258		43		66
1 14 49		**3 02 76**		**6 81**		**18 92**		**44 89**	
	108		176		264		44		68
1 18 81		**3 16 84**		**7 13**		**19 80**		**47 61**	
	110		180		270		45		70
1 23 21		**3 31 24**		**7 45**		**20 70**		**50 41**	
	112		184		276		46		72
1 27 69		**3 45 96**		**7 78**		**21 62**		**53 29**	
	114		188		282		47		74
1 32 25		**3 61 00**		**8 12**		**22 56**		**56 25**	
	116		192		288		48		76
1 36 89		**3 76 36**		**8 47**		**23 52**		**59 29**	
	118		196		294		49		78
1 41 61		**3 92 04**		**8 82**		**24 50**		**62 41**	
	120		200		30		50		80
1 48 84		**4 08 04**		**9 30**		**25 50**		**65 61**	
	124		204		31		51		82
1 58 76		**4 24 36**		**9 92**		**26 52**		**68 89**	
	128		208		32		52		84
1 69 00		**4 41 00**		**10 56**		**27 56**		**72 25**	
	132		212		33		53		86
1 79 56		**4 57 96**		**11 22**		**28 62**		**75 69**	
	136		216		34		54		88
1 90 44		**4 75 24**		**11 90**		**29 70**		**79 21**	
	140		220		35		55		90
2 01 64		**4 92 84**		**12 60**		**30 80**		**82 81**	
	144		224		36		56		92
2 13 16		**5 10 76**		**13 32**		**31 92**		**86 49**	
	148		228		37		57		94
2 25 00		**5 29 00**		**14 06**		**33 06**		**90 25**	
	152		232		38		58		96
2 37 16		**5 47 56**		**14 82**		**34 22**		**94 09**	
	156		236		39		59		98
2 49 64		**5 66 44**		**15 60**		**35 40**		**98 01**	

Иллюстрация 5.3.1

Сканирующая таблица для (отрицательных) обратных величин (преобразование: —1000/число)

А. Таблица — указатель места запятой (на с.119).

Б. Основная таблица: цифры обратных величин (отрицательных)

Сканирующее значение	Число	Сканирующее значение	Число	Сканирующее значение	Число	Сканирующее значение	Число	Сканирующее значение	Число
990	—100	**1639**	—60	**2469**	—40	**4115**	—240	**617**	—160
1010	—98	**1681**	—59	**2532**	—39	**4202**	—236	**633**	—156
1031	—96	**1709**	—58	**2597**	—38	**4274**	—232	**649**	—152
1053	—94	**1739**	—57	**2667**	—37	**4348**	—228	**667**	—148
1075	—92	**1770**	—56	**2740**	—36	**4425**	—224	**685**	—144
1099	—90	**1802**	—55	**2817**	—35	**4505**	—220	**704**	—140
1124	—88	**1835**	—54	**2899**	—34	**4587**	—216	**725**	—136
1149	—86	**1869**	—53	**2985**	—33	**4673**	—212	**746**	—132
1176	—84	**1905**	—52	**3077**	—32	**4762**	—208	**769**	—128
1205	—82	**1942**	—51	**3175**	—31	**4854**	—204	**794**	—124
1235	—80	**1980**	—50	**3279**	—30	**4950**	—200	**820**	—120
1266	—78	**2020**	—49	**3367**	—294	**505**	—196	**840**	—118
1299	—76	**2062**	—48	**3436**	—288	**515**	—192	**855**	—116
1333	—74	**2105**	—47	**3509**	—282	**526**	—188	**870**	—114
1370	—72	**2151**	—46	**3584**	—276	**538**	—184	**885**	—112
1408	—70	**2198**	—45	**3663**	—270	**549**	—180	**901**	—110
1449	—68	**2247**	—44	**3745**	—264	**562**	—176	**917**	—108
1493	—66	**2299**	—43	**3831**	—258	**575**	—172	**935**	—106
1538	—64	**2353**	—42	**3922**	—252	**588**	—168	**952**	—104
1587	—62	**2410**	—41	**4016**	—246	**602**	—164	**971**	—102
1639		**2469**		**4115**		**617**		**990**	

А.

Сканирующее значение	Указатель	Указатель	Сканирующее значение
1000			1000
	0,x	a,	
10 000			100
	0,0x	ab,	
100 000			10
	0,00x	abc,	
1 000 000			1,0
	0,000x	abcd,	
10 000 000			0,1
	0,0000x	abcde,	
100 000 000			0,01

Примеры

Числа	A	B	—1000/число
124,2	a,	—80	—8,0
0,04739	abcde	—212	—212**
1242,0	0, x	—80	—0,80

Иллюстрация 5.3.2

Табличка значений $\sqrt{y + 1/6}$

Десятки	Единицы									
	0	1	2	3	4	5	6	7	8	9
00	0,41	1,08	1,47	1,78	2,04	2,27	2,48	2,68	2,86	3,03
10	3,19	3,34	3,49	3,63	3,76	3,89	4,02	4,14	4,26	4,38
20	4,49	4,60	4,71	4,81	4,92	5,02	5,12	5,21	5,31	5,40

Для $30 \leqslant y \leqslant 288$ к $\sqrt{y}$ добавляется 0,01; для $y > 288$ надо брать $\sqrt{y}$ в табл. илл. 5.2.3

Иллюстрация 5.3.3

Табличка значений $-1000/(y + 1/6)$

Десятки	Единицы									
	0	1	2	3	4	5	6	7	8	9
00	—6000	—857	—462	—376	—240	—194	—162	—140	—122	—109
10	—98,4	—89,6	—82,2	—75,9	—70,6	—65,9	—61,9	—58,3	—55,0	—52,2
20	—49,6	—47,2	—45,1	—43,2	—41,4	—39,7	—38,2	—36,8	—35,5	—34,3
30	—33,1	—32,1	—31,1	—30,2						

Для $34 \leqslant y \leqslant 56$ значение $-1000/y$ уменьшается на 0,1; для $y \geqslant 57$ надо брать $-1000/y$ из табл. илл. 5.3.1.

Иллюстрация 5.4.1

Преобразования долей (%), отсчитываемых от 0,5 (или 50%)

«Свернутости», преобразованные («свернутые») корни и логарифмы — возможные альтернативные выражения для счетных дробей (знак ответа берите из шапки таблицы над столбцом процентов).

А. Основная таблица

+	«Свернутость»	«Свернутый» корень	«Свернутый» логарифм	−
50%	использовать	0,00	использовать	**50%**
51	⟶	0,02	⟵	**49**
52		0,04		**48**
53		0,06		**47**
54		0,08		**46**
55%	использовать	0,10	использовать	**45%**
56	⟶	0,12	⟵	**44**
57		0,14		**43**
58		0,16		**42**
59		0,18		**41**
60%	использовать	0,20	использовать	**40%**
61	⟶	0,22	⟵	**39**
62		0,24		**38**
63	0,26	0,26	0,27	**37**
64	0,28	0,28	0,29	**36**
65%	0,30	0,30	0,31	**35%**
66	0,32	0,32	0,33	**34**
67	0,34	0,35	0,35	**33**
68	0,36	0,37	0,38	**32**
69	0,38	0,39	0,40	**31**
70%	0,40	0,41	0,42	**30%**
71	0,42	0,43	0,45	**29**
72	0,44	0,45	0,47	**28**
73	0,46	0,47	0,50	**27**
74	0,48	0,50	0,52	**26**
75%	0,50	0,52	0,55	**25%**
76	0,52	0,54	0,58	**24**
77	0,54	0,56	0,60	**23**
78	0,56	0,59	0,63	**22**
79	0,58	0,61	0,66	**21**
80%	0,60	0,63	0,69	**20%**
81	0,62	0,66	0,73	**19**
82	0,64	0,68	0,76	**18**
83	0,66	0,71	0,79	**17**
84	0,68	0,73	0,83	**16**
85%	0,70	0,76	0,87	**15%**
86	0,72	0,78	0,91	**14**
87	0,74	0,81	0,95	**13**
88	0,76	0,84	1,00	**12**
89	0,78	0,87	1,05	**11**
90,0%	0,80	0,89	1,10	**10,0%**
90,5	0,81	0,91	1,13	**9,5**
91	0,82	0,92	1,16	**9**
91,5	0,83	0,94	1,19	**8,5**
92	0,84	0,96	1,22	**8**

Продолжение

+	«Свернутость»	«Свернутый» корень	«Свернутый» логарифм	−
92,5%	0,85	0,97	1,26	7,5%
93	0,86	0,99	1,29	7
93,5	0,87	1,01	1,33	6,5
94	0,88	1,02	1,38	6
94,5	0,89	1,04	1,42	5,5
95,0%	0,90	1,06	1,47	5,0%
95,5	0,91	1,08	1,53	4,5
96	0,92	1,10	1,59	4
96,5	0,93	1,12	1,66	3,5
97	0,94	1,15	1,74	3
97,2%	0,94	1,16	1,77	2,8%
97,4	0,95	1,17	1,81	2,6
97,6	0,95	1,18	1,85	2,4
97,8	0,96	1,19	1,90	2,2
98,0	0,96	1,20	1,95	2,0
98,2%	0,96	1,21	2,00	1,8%
98,4	0,97	1,22	2,06	1,6
98,6	0,97	1,24	2,13	1,4
98,8	0,98	1,25	2,21	1,2
99,0	0,98	1,27	2,30	1,0
99,2%	0,98	1,28	2,41	0,8%
99,4	0,99	1,30	2,55	0,6
99,6	0,99	1,32	2,76	0,4
99,8	1,00	1,35	3,11	0,2
100,0%	1,00	1,41	∞	0,0

Б. Дополнительная таблица для «свернутых» логарифмов от значений, меньших 1% и больших 99%

+	«Свернутый» логарифм	−	+	«Свернутый» логарифм	−
99,0%	2,30	1,0%	99,80%	3,11	0,20%
99,1	2,35	1,9	99,82	3,16	0,18
99,2	2,41	1,8	99,84	3,22	0,16
99,3	2,48	1,7	99,86	3,28	0,14
99,4	2,55	1,6	99,88	3,36	0,12
99,50	2,65	1,50	99,90	3,45	0,10
99,52	2,67	1,48	99,91	3,51	0,09
99,54	2,69	1,46	99,92	3,57	0,08
99,56	2,71	1,44	99,93	3,63	0,07
99,58	2,73	1,42	99,94	3,71	0,06
99,60	2,76	1,40	99,95	3,80	0,05
99,62	2,78	1,38	99,96	3,91	0,04
99,64	2,81	1,36	99,97	4,06	0,03
99,66	2,84	1,34	99,98	4,26	0,02
99,68	2,87	1,32	99,99	4,61	0,01

Продолжение

+	«Свернутый» логарифм	−	+	«Свернутый» логарифм	−
99,70	2,90	**1,30**	Примеры:		
99,72	2,94	**1,28**			
99,74	2,97	**1,26**	99,29% дает 2,47		
99,76	3,01	**1,24**	0,37% дает —2,80		
99,78	3,06	**1,22**			

Иллюстрация 5.5.1

Некоторые значения преобразованных логарифмов и степеней, совместимых со значением $y = A = 300$ и его окрестностью

Значения p

2	1	1/2	(Псевдонуль)	−1	−2
$\frac{y^2}{2A}+\frac{A}{2}$	y	$\sqrt{4Ay}-A$	$A\log_e\left(\frac{y}{A}\right)+A$	$-\frac{A^2}{y}+2A$	$-\frac{A^3}{2y^2}+\frac{3A}{2}$
150	**0**	—300	—∞	—∞	—∞
151	**25**	—127	—445	—3000	—21150
154	**50**	—55	—238	—1200	—4950
167	**100**	46	—30	—300	—900
217	**200**	190	178	150	112
254	**250**	248	245	240	234
281	**280**	280	279	279	278
290	**290**	290	290	290	289
299	**299**	299	299	299	299
300	**300**	300	300	300	300
301	**301**	301	301	301	301
321	**320**	320	319	319	318
354	**350**	348	346	343	340
417	**400**	393	386	375	366
567	**500**	475	453	420	396
1817	**1000**	795	661	510	436
6817	**2000**	1249	869	555	447
∞	∞	∞	∞	600	450
$p=2$	1	$\frac{1}{2}$	[псевдонуль]	—1	—2

Иллюстрация 5.5.2

Подробное табулирование преобразований из илл. 5.5.1 в окрестности значения $y = A = 300$, где преобразования хорошо совмещаются

$\frac{y^2}{2A}+\frac{A}{2}$	y	$\sqrt{4Ay}-A$	$A\log_e\left(\frac{y}{A}\right)+A$	$-\frac{A^2}{y}+2A$	$-\frac{A^3}{2y^2}+\frac{3A}{2}$
280,67	**280**	279,66	279,30	278,57	277,81
285,38	**285**	284,81	284,61	284,21	283,80
290,17	**290**	289,92	289,83	289,66	289,48

$\frac{y^2}{2A}+\frac{A}{2}$	y	$\sqrt{4Ay}-A$	$A\log_e\left(\frac{y}{A}\right)+A$	$-\frac{A^2}{y}+2A$	$-\frac{A^3}{2y^2}+\frac{3A}{2}$
292,11	**292**	291,95	291,89	291,78	291,67
294,06	**294**	293,97	293,94	293,88	293,82
296,03	**296**	295,99	295,97	295,95	295,92
298,01	**298**	298,00	297,99	297,99	297,98
299,00	**299**	299,00	299,00	299,00	299,00
300,00	**300**	300,00	300,00	300,00	300,00
301,00	**301**	301,00	301,00	301,00	301,00
302,01	**302**	302,00	301,99	301,99	301,98
304,03	**304**	303,99	303,97	303,95	303,92
306,06	**306**	305,97	305,94	305,88	305,82
308,11	**308**	307,95	307,90	307,79	307,69
310,17	**310**	309,92	309,84	309,68	309,52
315,38	**315**	314,82	314,64	314,29	313,95
320,67	**320**	319,68	319,36	318,75	318,16
$p=2$	1	1/2	[псевдонуль]	—1	—2

Иллюстрация 5.6.1

Пример с баллами

А. Данные по всем студентам

Уровень	Число	p = доля до данного уровня	P = доля до следующего уровня	$\varphi(p)$	$\varphi(P)$	$\frac{\varphi(P)-\varphi(p)}{P-p}$
A	127	0,0000	0,0304	0,0000	—0,1361	—4,48
B	497	0,0304	0,1496	—0,1361	—0,4220	—2,40
C	3243	0,1496	0,9269	—0,4220	—0,2616	0,21
D	231	0,9269	0,9823	—0,2616	—0,0889	3,12
E	74	0,9823	1,0000	—0,0889	0,0000	5,02
(Всего)	(4172)					

Замечания: 0,0304 = 127/4172; 0,1496 = (127 + 497)/4172 и т. д.; значения $\varphi(p)$ и $\varphi(P)$ берутся из илл. 5.6.2.

Б. Гипотетически хорошие студенты

A	64	0,0000	0,0792	0,0000	—0,2768	—3,49
B	127	0,0792	0,2364	—0,2768	—0,5469	—1,72
C	560	0,2364	0,9295	—0,5469	—0,2549	0,42
D	54	0,9295	0,9963	—0,2549	—0,0244	3,45
E	3	0,9963	1,0000	—0,0244	0,0000	6,59
	(808)					

В. Гипотетически плохие студенты

A	12	0,0000	0,0114	0,0000	—0,0623	—5,46
B	53	0,0114	0,0617	—0,0623	—0,2316	—3,37
C	821	0,0617	0,8406	—0,2316	—0,4387	—0,27
D	107	0,8406	0,9421	—0,4387	—0,2212	2,14
E	61	0,9421	1,0000	—0,2212	0,0000	3,82
	(1054)					

Иллюстрация 5.6.2

Значения (и формула) для φ (*p*) и φ (*P*)

А. Значения для некоторых *p* или *P*.

q (последняя цифра) =

p или *P*	0	1	2	3	4	5	6	7	8	9
0,000*q*	0,0000	—0,0010	—0,0019	—0,0027	—0,0035	—0,0043	—0,0051	—0,0058	—0,0065	—0,0072
0,00*q*	0,0000	—0,0079	—0,0144	—0,0204	—0,0261	—0,0315	—0,0367	—0,0417	—0,0466	—0,0514
0,0*q*	0,0000	—0,0560	—0,0980	—0,1347	—0,1679	—0,1985	—0,2270	—0,2536	—0,2788	—0,3025
0,1*q*	—0,3251	—0,3465	—0,3669	—0,3864	—0,4050	—0,4227	—0,4397	—0,4559	—0,4714	—0,4862
0,2*q*	—0,5004	—0,5140	—0,5269	—0,5393	—0,5511	—0,5623	—0,5731	—0,5833	—0,5930	—0,6022
0,3*q*	—0,6109	—0,6191	—0,6269	—0,6342	—0,6410	—0,6474	—0,6534	—0,6590	—0,6641	—0,6687
0,4*q*	—0,6730	—0,6769	—0,6803	—0,6833	—0,6859	—0,6881	—0,6899	—0,6913	—0,6923	—0,6929
0,5*q*	—0,6931	—0,6929	—0,6923	—0,6913	—0,6899	—0,6881	—0,6859	—0,6833	—0,6803	—0,6769
0,6*q*	—0,6730	—0,6687	—0,6641	—0,6590	—0,6534	—0,6474	—0,6410	—0,6342	—0,6269	—0,6191
0,7*q*	—0,6109	—0,6022	—0,5930	—0,5833	—0,5731	—0,5623	—0,5511	—0,5393	—0,5269	—0,5140
0,8*q*	—0,5004	—0,4862	—0,4714	—0,4559	—0,4397	—0,4227	—0,4050	—0,3864	—0,3669	—0,3465
0,9*q*	—0,3251	—0,3025	—0,2788	—0,2536	—0,2270	—0,1985	—0,1679	—0,1347	—0,0980	—0,0560
0,99*q*	—0,0560	—0,0514	—0,0466	—0,0417	—0,0367	—0,0315	—0,0261	—0,0204	—0,0144	—0,0079
0,999*q*	—0,0079	—0,0072	—0,0065	—0,0058	—0,0051	—0,0043	—0,0035	—0,0027	—0,0019	—0,0010

Б. Формула $\varphi(p) = p \log_e p + (1 - p) \log_e (1 - p)$.

Иллюстрация 5.6.3

Два предположительно взаимосвязанных преобразования

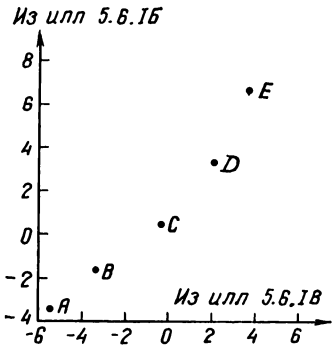


Иллюстрация 5.7.1

Значения $\log (i - 1/3)$ **при** $i = 1, \ldots, 99$ **с двумя десятичными знаками**

q (последняя цифра) =

i	0	1	2	3	4	5	6	7	8	9
$0q$	—	—0,18	0,22	0,43	0,56	0,67	0,75	0,82	0,88	0,94
$1q$	0,99	1,03	1,07	1,10	1,14	1,17	1,19	1,22	1,25	1,27
$2q$	1,29	1,32	1,34	1,36	1,37	1,39	1,41	1,43	1,44	1,46
$3q$	1,47	1,49	1,50	1,51	1,53	1,54	1,55	1,56	1,58	1,59
$4q$	1,60	1,61	1,62	1,63	1,64	1,65	1,66	1,67	1,68	1,69
$5q$	1,70	1,70	1,71	1,72	1,73	1,74	1,75	1,75	1,76	1,77
$6q$	1,78	1,78	1,79	1,80	1,80	1,81	1,82	1,82	1,83	1,84
$7q$	1 84	1,85	1,86	1,86	1,87	1,87	1,88	1,88	1,89	1,90
$8q$	1,90	1,91	1,91	1,92	1,92	1,93	1,93	1,94	1,94	1,95
$9q$	1,95	1,96	1,96	1,97	1,97	1,98	1,98	1,99	1,99	1,99

При $i \geqslant 30$ можно пользоваться и значениями $\log i$.

Глава 6 ● НУЖНЫ ЛИ НАМ ПРЕОБРАЗОВАНИЯ?

Из опыта работы с разными классами данных мы можем знать, что определенные преобразования исходных x полезны. Или у нас могут быть какие-то иные причины для веры, что преобразования надо делать. И все же мы можем предпочесть бездействие либо из-за непомерных усилий, либо из-за трудностей интерпретации и аргументации анализа преобразованных данных, либо просто из-за ничтожности выигрыша. Однако тогда предметом нашей заботы становятся возможные потери от *не*преобразования.

Хотя мы и уделили немало внимания приемам преобразований (в гл. 5), иногда затраты на них все же не окупаются, как, например, при выборе для y между описаниями

$$bx \text{ и } b^* \log x$$

или между вариантами

$$a+bx \text{ и } a^*+b^* \log x,$$

когда $1 \leqslant x \leqslant 1{,}001$.

Это отражает характерные ситуации выбора между исходным, в примере выше,— x, и преобразованным — $\log x$ носителями (информации). Ограничимся теперь такими x, которые всюду положительны. В этой главе принимается гипотеза, что преобразованный носитель работает лучше, чем исходный, а вопрос ставится так: *достаточно* ли хорош преобразованный носитель, чтобы оправдать затраты на преобразование? Качество работы оценивается в терминах остатков. Худшим подбором считается такой, который приводит к слишком большим остаткам.

Для изучения этого вопроса мы хотим использовать как можно меньше информации о данных, чтобы упростить себе задачу. Мы предлагаем метод анализа, требующий вначале информации по трем пунктам:

число точек в данных (число опытов);

теснота связи между исходным и преобразованным носителями;

теснота связи между исходным носителем и откликом.

Для описания двух мер близости мы возьмем коэффициент корреляции. Полезно напомнить, что

$$\frac{\text{дисперсия (остатков)}}{\text{дисперсия (отклика)}} = 1 - r^2,$$

где r^2 — коэффициент парной корреляции между откликом и носителем в условиях линейной регрессии.

Хоть и разумно предположить, что у нас есть по крайней мере графики зависимости y от x — исходного носителя, и можно, хотя бы на глаз, оценить тесноту связи между ними, но мы никак не можем связать ни y, ни x исходный с x спрямленным, преобразованным носителем. Можно, вероятно, избежать затрат на преобразование тогда, когда есть какой-нибудь простой способ оценки близости между $x_{\text{исх}}$ и $x_{\text{спр}}$. Большую пользу для оценки степени спрямления имеет отношение максимального значения исходного x к его минимальному значению $x_{\text{исх}}$. Грубо говоря, если это отношение значительно больше 1, то скорее всего мы будем преобразовывать x, а если оно того же порядка, что 1, то можно и не беспокоиться.

Здесь мы приведем лишь общие соображения о том, как выглядит ответ, когда соответствующий преобразованный носитель есть логарифм исходного носителя, оставляя для приложения (после гл. 11) большинство деталей этого случая и все, что относится к преобразованиям квадратного корня или (отрицательной) обратной величины.

Общие соображения о том, когда преобразование $\log x$ целесообразно

Если наибольшее значение x вдвое превышает наименьшее, обычно преобразование нужно при высокой (скажем $>0{,}9$) корреляции между x и y (откликом). (Если мы очень осторожны, то будем преобразовывать и при меньших корреляциях, а если не боимся риска, то не пошевелимся и при корреляциях, меньших 0,95.)

Если отношение максимального x к минимальному меньше двух, то мы ничего не делаем при корреляции между откликом и $x_{\text{исх}}$, меньшей 0,95.

Если большее значение в 20 или более раз превосходит меньшее, то, видимо, почти всегда предпочтительнее преобразовывать, конечно, когда есть надежда на какую-нибудь полезную связь между исходным x и y.

В сомнительных ситуациях можно либо читать приложение, либо перебирать преобразования, следя, хорошо ли они работают, либо делать и то, и другое.

Глава 7 ● ОХОТА ЗА ИСТОЧНИКАМИ НЕОПРЕДЕЛЕННОСТИ

Чтобы выйти за пределы индикаций, надо приспособиться к их неопределенностям. Хотя точность оценки есть число, более существенна ее достоверность, поскольку нам легко обмануться из-за переменных, отсутствующих или неузнанных в анализе.

Мы соотносим вклад неопределенности с двумя источниками: один позволяет судить исходя из имеющихся данных о внутренней неопределенности, другой, обусловленный причинами, не обнаруживаемыми нашими данными, — о дополнительной неопределенности. Так что внутренняя и дополнительная неопределенности — два размытых понятия, введенных, чтобы помочь нам в понимании неопределенности, изменчивости и устойчивости. Забвение того, что есть два источника неопределенности, может повести к ее серьезной недооценке и как следствие к слишком большому оптимизму насчет стабильности индикации. Во избежание таких «ловушек» надо найти подходящие формулы для ошибок, определяемых данными, и помнить об источниках, которые незримо участвуют в процессе сбора данных.

Хорошо спланированная программа наблюдений или экспериментов, как правило, уменьшает влияние всех видов вариации на неопределенность результатов. План особенно ценен, когда есть уверенность, что все основные источники вариации включены в исследование. Зачастую это осуществляется «расширением базы» сфокусированного на узкую задачу исследования, что сводит практически всю действительную вариабельность к внутренней неопределенности, существенно снижая вклад дополнительной неопределенности

Коль скоро удачный план осуществлен, возникает желание получить разумную оценку внутренней неопределенности по имеющимся данным, т. е. провести *прямое оценивание* неопределенности. Кроме того, нам еще надо разумно оценить величину эффектов, которые нельзя извлечь из данных. Отдельными источниками их могут быть: систематическая ошибка измерения (пример — тенденция исключать маленьких детей из переписи); несоответствие между произведенной выборкой и исследуемой популяцией (например, при опросе общественного мнения выборка, как правило, не совпадает с совокупностью голосующих и вдобавок мнение до официальной баллотировки не обязательно совпадает с мнением в момент голосования); эффект «первого взгляда», когда при повторных наблюдениях одного и того же объекта наблюдатель оказывается под впечатлением первого наблюдения.

Мы начнем обсуждение внутренней неопределенности с иллюстрации того, как классическая формула

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

может сослужить и плохую службу. Далее мы остановимся на том, что естественная или искусственная рандомизация некоторых возможных разбиений служит практической основой для оценивания внутренней неопределенности. Различные уровни (варианты) группировки открывают пути к прямому оцениванию изменчивости (разброса), и мы предлагаем некоторые принципы выбора подходящей группировки. После разбора примера мы обсудим несколько более серьезных трудностей, связанных с прямой оценкой неопределенности, и наметим подходы к их преодолению.

Вернувшись к дополнительной неопределенности и подчеркнув ее важность, мы рассмотрим ее оценку и обсудим ее взаимодействие с внутренней неопределенностью.

7.1. КАК $\sigma/\sqrt{n}$ МОЖЕТ ОБМАНУТЬ

И математические, и нематематические введения в статистическую обработку данных стараются подчеркнуть, что стандартное отклонение (квадратичная ошибка) выборочного среднего равно $\sigma/\sqrt{n}$, где σ — стандартное отклонение генеральной совокупности, а n — объем выборки. Это утверждение исключительно важно и именно оно входит в число основных положений теории выборок. Основывается это, однако, как и положено вводным курсам, на слишком упрощенной точке зрения о сути происходящего. Потом дисперсионный анализ, возможно, приведет к идее о разделении источников вариации, но мы вновь и вновь будем обращать внимание на непреложность того, что сами данные говорят непосредственно лишь о своей собственной вариабельности. Исследование Пирсом времени реакции (см. параграф 1.4) снова дает нам пример, интересный и сам по себе, и методологически.

Из илл. 1.4.1 выберем следующие значения: .

Дни	(1)	2	3	4	5	от 6-го до 24-го
Наблюденные средние	(475,6)	241,5	203,1	205,6	148,5	от 175,6 до 265,5
$s/\sqrt{n}$	(4,2)	2,1	2,0	1,8	1,6	от 1,1 до 2,2

Исключая результаты первого дня, которые явно отличаются от остальных, мы видим, что $s/\sqrt{n}$ меняется от 1,1 до 2,2. Если значения $s/\sqrt{n}$ измеряют стандартные отклонения наблюденных средних,

то дисперсия разности между средними пар соседних дней должна быть равна:

$$(1,1)^2+(1,1)^2=2,42 \text{ для наименьшего разброса}$$

и

$$(2,2)^2+(2,2)^2=9,68 \text{ для наибольшего.}$$

Эти пределы соответствуют стандартным отклонениям разностей между средними для соседних дней в пределах от $\sqrt{2,42}\approx 1,6$ до $\sqrt{9,68}\approx 3,1$. Если эти стандартные отклонения, основанные на $\sigma/\sqrt{n}$, действительно были бы таковы, то львиная доля разностей между днями была бы меньше, чем два стандартных отклонения, или меньше, чем 3,2 и 6,2, и почти все разности были бы меньше, чем три стандартных отклонения, или меньше, чем 4,8 и 9,3. В действительности же разности для соседних дней, кроме первого (см. илл. 1.4.1), равны

$$\begin{array}{cccccccccccc} -38, & +2, & -57, & +27, & +11, & +7, & +2, & +20, & +1, & +19, \\ +9, & -8, & -1, & -3, & +32, & -12, & +6, & -3, & -10, & +11, & -4, & -8 \end{array}$$

вопреки всем и всяким рассчитанным пределам.

На языке дисперсионного анализа данные Пирса обнаруживают значительный междневный разброс. Можно сказать, что они «бесконтрольны» (по определению Шьюхарта (Shewhart)): вариация внутри дня не объясняет должным образом вариацию между днями. Не эффект ли это «новобранца», наблюдающего впервые в жизни? Ведь он продолжает вздрагивать даже после 20 дней работы.

Колебание в этих данных напоминает историю с проблемой «уравнения наблюдателя» в астрономии. Была надежда, что систематические ошибки каждого наблюдателя можно сначала стабилизировать, а затем учесть, улучшая тем самым точность. К несчастью, попытки такого сорта постоянно терпели неудачи, причины которых могут подсказать рассмотренные данные. Так что причуды каждодневного наблюдения требуют учета, по крайней мере путем введения нового источника вариации — междневного. (А как насчет межчасного или межнедельного? Об этом можно лишь догадываться, так как эти конкретные данные представлены для целых дней и охватывают лишь чуть больше трех недель). Сильное отличие первого дня от всех последующих — вполне обычное явление для многих видов данных, от которого в ответственных экспериментах важно в какой-то мере избавиться за счет предварительной работы и тренировок.

Уилсон и Хилферти (см. параграф 1.4) выяснили, что данные Пирса иллюстрируют «тезис о том, что если мы хотим оценить разброс некоторой статистики, то надо иметь много выборок, и что верность такой формулы, как $\sigma/\sqrt{n}$, на практике научно не подкрепляется даже для оценивания надежности среднего».

Даже имея дело с такой элементарной статистикой, как арифметическое среднее, часто гибельно использовать прямую оценку его внутренней неопределенности как единственно возможную. Получение достоверной меры неопределенности — это отнюдь не просто выискивание формул.

7.2. ЕЩЕ ОДИН ПРИМЕР, ГОВОРЯЩИЙ О НЕОБХОДИМОСТИ ПРЯМОЙ ОЦЕНКИ ВАРИАЦИИ

Давайте от задачи измерения перейдем теперь к задаче счета в предположении биномиального распределения. Мы стремимся автоматически находить разброс данных при биномиальном законе, когда стандартное отклонение наблюдаемой доли «успехов» есть $\sigma = \sqrt{pq/n}$, где n — объем выборки, а $p = 1 - q$ — доля «успехов» в совокупности. И снова мысль о стандартном отклонении «микрокосма» полезна во многих отношениях, но в то же время чревата недооценкой действительной вариации. Возьмем массовое производство как источник примеров величин, у которых индивидуальные различия смешаны с особенностями поведения станков.

Если принять, что среди тысяч произведенных штучных деталей наблюдаемая доля брака равна p и что один рабочий на одном станке в одну смену производит 1000 деталей, то утверждать, что средняя доля брака за долгое время лежит между $p - 3\sqrt{pq/1000}$ и $p + 3\sqrt{pq/1000}$, было бы крайне рискованным делом. Доля брака, вероятно, зависит от многих вещей, таких, как день недели (понедельник — день тяжелый), рабочий, станок, смена, мастер, контролер, и от многих других, перечислить которые здесь нет никакой возможности. Осознание этого букета источников неоднородности привело Уолтера Шьюхарта в 1931 г. к разработке методов контроля качества с пределами $p \pm 3\sqrt{pq/n}$ как идеалом, к которому удается приблизиться только ценой огромных и изощренных инженерных усилий.*. То, из-за чего не удается получить стабильности в массовом производстве со всеми его контрольными и измерительными чудесами, в других областях хоронит надежды на готовые формулы. Вера в их существование может породить фантастические гарантии и досадные неожиданности.

У относительно прямой оценки изменчивости нет альтернативы. В этой оценке, когда мы изучаем различия в больших группах наблюдений, различия более или менее представляют все то множество источников вариации, с которым мы вынуждены считаться. Так, в промышленной ситуации мы можем обнаружить массу брака при некоторых сочетаниях рабочих, смен, станков, дней недели, которые дают представление о вариации реального промышленного производства. (Как мы покажем в параграфе 7.7, обычно нельзя ожидать, чтобы данные содержали информацию обо всех источниках вариации.) В хорошо организованных исследованиях многообразие источников вариации вынуждает нас на прямую оценку.

Но если даже и нет такого множества источников, все равно отсутствие подходящих формул, связывающих микроразличия с макроразл-

*Профессор У. Шьюхарт был одним из основателей статистического контроля качества в промышленности. Первоначальное представление об этой области исследования можно получить, например, по работам: Хэнсен Б. Л. Контроль качества. Теория и применение. М., Прогресс, 1968; Адлер Ю. П. Управление качеством: статистический подход. М., Знание, 1979. — *Примеч. ред.*

бросом, часто требует прямой оценки. Даже ценою сверхупрощения (вроде постулата о независимости, отбрасывания многих заведомо существующих источников вариации и условия нормальности распределений) не всегда удается вывести соответствующую формулу, поскольку ее вывод или поиск может оказаться непомерно громоздким, если, конечно, вообще возможным. Требование всесторонности и полноты анализа тоже толкает нас на путь прямой оценки.

7.3. ВЫБОР КОМПОНЕНТОВ ОШИБКИ

Большая совокупность данных предоставляет значительную свободу для измерения ее внутренней неопределенности. Так, при изучении вариантов контрольных работ можно в качестве объектов взять всех девятиклассников, обучающихся в школах города. Естественно, возникает иерархия:

а) *ученик*; б) *класс*; в) *школа*; г) *город*.

Будем считать, раз уж мы выбираем, что ученики, учившиеся в этом году в девятом классе некоторой школы, образуют случайную выборку из тех, кто «мог бы» сюда попасть с учетом, например, социально-экономического, этнического и криминалистического фона ареала, из которого эта школа черпает своих учеников. Если мы поступим так для каждой школы в городе и сочтем их (город и школы) заданными, то получим подходящую частную модель, пригодную для оценки внутренней неопределенности. В этой модели мы могли бы обратиться к различию между школьниками внутри класса как к основе для мер стабильности.

В той мере, в какой мы соотносим результаты с конкретной школой в конкретном городе, такая оценка может вполне нас устроить. Тем более, когда расширенные, более всеохватывающие оценки невозможны (как если бы только одна школа в городе имела девятый класс), наша оценка, быть может, лучшее из того, что мы вообще могли бы сделать.

Если мы работаем в мегаполисе с его уравниловкой, где какие-нибудь частные распределения, скажем социально-экономического или этнического происхождения, не играют особой роли, то выгоднее, с точки зрения добычи полезной информации, выбрать различия между школами за основу для оценивания внутренней неопределенности. Поступая так, мы тем самым практически превращаем изучаемые *школы* в случайную выборку из большой совокупности школ *.

Таким способом мы вводим в оценку по крайней мере часть межрайонных различий нашего огромного города. Если есть региональные различия и «наш» город принадлежит к исключительному региону, то мы не получим верного представления о межрайонных различиях для всех регионов, пользуясь нашей оценкой нестабильности. Если же, напротив, социально-экономические группы играют преобладающую роль и доли разных групп во всех городах примерно одинаковы, то мы получим сверхпредставительную межрайонную изменчивость. Наша оцен-

*Читатель, интересующийся существом подобных задач, может обратиться, например, к кн.: Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М., Прогресс, 1976. — *Примеч. ред.*

ка может оказаться недостаточной в разных смыслах, но с данными по одному-единственному городу ничего лучшего, видимо, сделать нельзя.

Хотя нетрудно выписать формулы, основанные на предположениях других типов, оценка вариабельности на практике, кажется, останется универсальной основой обработки таких объектов, как ученики, классы, школы, города и даже сборники контрольных работ, используемые при обучении, как если бы именно они составляли случайную выборку. Это важно понять и для использования некоторых широко применяемых (см. гл. 8) общих приемов оценивания вариабельности, и для уяснения того, что практический выбор обычно происходит между «будем считать это случайной выборкой» и «забудем об этом» («заметем под ковер»).

Мы ведем в весьма прагматическом духе обсуждение того, что ориентированный на дисперсионный анализ читатель назовет просто «выбором компонентов ошибки».

Можно доказывать, что было бы хорошо ограничиться вычислениями индикаций нестабильности для «как будто бы случайных» данных лишь теми случаями, когда эти «как будто бы» были действительно рандомизированы. Авторы, стоящие на такой позиции, часто приводят искусственно заниженные оценки нестабильности из-за исключения источников изменчивости, представленных в выборке, хотя, возможно, не очень «случайно» или полно. Следовательно, мы поощряем обработку эффектов как рандомизированных при самых разных обстоятельствах, где рандомизация по меньшей мере сомнительна.

Мы оказали бы читателю плохую услугу, если бы сохранили у него впечатление, что наиболее широкая оценка внутренней неопределенности всегда наилучшая. Наши цели могут требовать и более узких оценок. Если мы исследуем, например, *только* три больших города в некотором штате, то в той мере, в какой принимаемые решения проводятся одинаково во всех больших городах одного штата, подходящий средневзвешенный результат по этим трем городам будет естественным показателем. В этом показателе проявляются только неопределенности отдельных исследований; межгородские различия не вносят дополнительной неопределенности, поскольку учтены все города, которые могли повлиять на показатель. (Оставаясь в пределах США, укажем на Бостон, Нью-Орлеан и Сиэтл как на контрпример, по всей вероятности *.)

Кроме того, в большинстве случаев огромные различия при столь малом числе сравниваемых объектов делают любые оценки, на них основанные, совершенно неустойчивыми. Если бы сравнивались только, скажем, два города, то, видимо, было бы лучше получить раздельные оценки внутренней неопределенности для каждого города, а средние результаты принимать во внимание лишь для «выборки» *из* этих двух

*Эти города образуют наибольший треугольник, который вписывается в территорию США без Аляски и островов и имеет вершины, совпадающие с крупными городами. Бостон — в Новой Англии, на побережье Атлантического океана, Нью-Орлеан — на юге, на берегу Мексиканского залива, а Сиэтл — на северо-востоке, в заливе Тихого океана. В каком смысле авторы говорят здесь о контрпримере, сказать трудно. Возможно, что эта выборка хорошо представляет не только себя, но и все крупные города США. — *Примеч. ред.*

городов, но настаивая на дальнейших шагах по изучению межгородской вариации.

Руководствуясь этим планом, иные читатели смогут комбинировать суждения и наблюденные межгородские разности по-своему, так чтобы оценить соответствующее межгородское влияние на неопределенность. Здесь нет какого-то единственного «правильного» ответа. Разным целям могут лучше служить и различные оценки нестабильности*.

7.4. НЕКОТОРЫЕ ПОДРОБНОСТИ ВЫБОРА КОМПОНЕНТОВ ОШИБКИ

Перечисляя группы (городов, лет и т. п.), различия между которыми порождают компоненты ошибки, и оценивая каждый компонент в отдельности, мы, вероятно, обратимся к t Стьюдента, чтобы получить границы для эффектов внутренней изменчивости.

Обычно мы ограничены в выборе числа используемых групп, но иногда наш выбор свободнее. Если условия выбора компонентов ошибки позволяют иметь 100 эквивалентных групп, то мы могли бы разделить эти 100 групп случайным образом на 20 новых групп (классов) по пять групп в каждой или на 5 классов по 20 групп в каждом. Экономит ли уменьшение числа классов вычисления? Во что это обойдется с других точек зрения?

Сколько же групп надо брать? Вообще говоря, чем больше, тем лучше. Однако обычно давление экономики и потока данных говорят: поменьше. Посмотрите на илл. 7.4.1, где двусторонние 5%-ные точки для t проливают на это некоторый свет. Так, заметим, что, взяв три степени свободы, мы не добираем около 89% от того, что дает бесконечное число степеней свободы для 5%-ных точек, измеряя в шкале t. А при 10 степенях свободы — лишь около 10%. (Фактически потери в терминах дисперсий, что, возможно, более подходит к случаю, хорошо согласуются с квадратом отношения входов t-таблицы). Следовательно, число групп от 4 до 10 практически достаточно. Мы часто берем 10 из соображений удобства.

Если бы было всегда ясно, какие данные должны составлять «правильные» группы, и если бы всегда было легко рассчитать результат для каждой такой группы, да еще если бы такие результаты всегда эффективно использовались, то и проблема внутренней неопределенности тотчас отпала бы.

7.5. ВОЗМОЖНОСТЬ ПОЛУЧЕНИЯ ПРЯМЫХ ОЦЕНОК

Когда разбиение данных на группы не происходит автоматически, часто помогает сознательное конструирование равноценных подвыборок. Так, например, во время выборочного обследования из списка объектов мы можем построить любые систематические подвыборки (так, чтобы каждая подвыборка содержала каждый m-й объект), выбирая

*Рассуждения на близкие темы читатель найдет в, к сожалению, независимо написанной популярной статье: Адлер Ю., Грановский Ю. Опыт, опыт, повторись. — Химия и жизнь, 1978, № 10. — *Примеч. ред.*

первые объекты случайно среди m начальных объектов списка. Для оценки генерального среднего μ подсчитаем сначала средние $\bar{y}_i$ для каждой подвыборки. Каждое из k таких средних считается в предположении независимости измерений. Среднее их $\bar{\bar{y}} = \Sigma \bar{y}_i/k$ служит оценкой для μ. Оценка дисперсии $\bar{\bar{y}}$, равная $s^2_{\bar{\bar{y}}} = \Sigma (\bar{y}_i - \bar{\bar{y}})^2/k\ (k-1)$, позволяет найти доверительные границы для μ:

$$\bar{\bar{y}} \pm |t_{k-1}| s_{\bar{\bar{y}}}, \qquad \text{(а)}$$

где $|t_{k-1}|$ берется из таблиц для t Стьюдента при $(k-1)$ степенях свободы для какого-нибудь заданного двустороннего уровня значимости. В границах, заданных формулой (а), мы использовали прямую оценку изменчивости $\bar{y}_i$.

Не думайте, что $s_{\bar{\bar{y}}}$ эквивалентно $s/\sqrt{n}$, обсужденному в примере с данными Пирса (параграф 7.1). Правда, мы используем обозначение дисперсии для меры изменчивости, но оценка этой изменчивости посредством $s_{\bar{\bar{y}}}$ включает различия между хорошо соразмеренными частями данных, вроде «дней» у Пирса, тогда как $s/\sqrt{n}$ у него получено из расхождений между исходными наблюдениями.

В том же духе можно было бы извлечь несколько разно расслоенных выборок, когда каждая повторная выборка извлекается из всей популяции методом, устанавливающим определенный тип расслоения. В этом случае для оценки генерального среднего получаются разные $\bar{y}_i$, и взвешенные оценки генерального среднего можно строить для каждой из таких систем выборок. Тогда расчет доверительных границ для генерального среднего μ можно будет вести по той же формуле (а). Использование равноценных (часто называемых «взаимопроникающими») подвыборок имеет то преимущество, что они могут с легкостью учитываться в изменчивости, представляющей, скажем, интервьюеров и контролеров, чью службу надо относить к разным выборкам.

Точно так же можно получать многие оценки. Рассмотрим пример, где оценивается угол наклона (регрессионные коэффициенты в данном случае — линейные комбинации наблюденных y с множителями, зависящими от x), составляющий часть экспериментальных данных Джонсона и Цзао [Johnson P. O. and Tsao F. (1944)], [Johnson P. O. (1949)].

Эксперимент Джонсона — Цзао. Объектом эксперимента служил круглый диск, который давит вниз с одним из следующих семи усилий: 100, 150, 200, 250, 300, 350 и 400 г. Гидравлическая система с клапаном обеспечивает нарастание нагрузки с одной из следующих четырех скоростей: 100, 200, 300, 400 г/мин. Испытуемый объявлял «готово», когда он замечал увеличение тяжести. Наблюдение (т. е. отчетливо фиксируемое изменение) бралось на основе субъективных отчетов об изменении нагрузки. Испытуемые предварительно обучались работе на аппаратуре и методике опыта. Джонсон и Цзао испытывали 8 вариантов: 4 с открытыми глазами и 4 вслепую, 4 с мужчинами, 4 с женщинами. Весь эксперимент повторялся дважды. Каждый опыт повторялся 5 раз, причем все 28 вариантов (4 скорости и 7 начальных «весов») рандоми-

зировались вместе с повторами (для каждого из 8 испытуемых в каждом цикле)*.

Графическая проверка изменчивости в функции уровня показывает, что логарифмы наблюдений более пригодны для анализа, чем исходные данные. Результаты подсказывают, что при фиксированной скорости зависимость от начальных весов не слишком сильна. Давайте рассмотрим это подробнее.

Вычислим угловой коэффициент регрессионной прямой для четырех «зрячих» испытуемых и при скорости 300 г/мин между десятичными логарифмами наблюдений и начальными весами, закодированными 1, 2, ..., 7 соответственно их семи возрастающим уровням. На илл. 7.5.1 приведены исходные данные и их логарифмы. Коэффициенты при x (в логарифмическом масштабе на каждые 50 г привеса) для четырех испытуемых равны: 0,0029; 0,0154;—0,0064; 0,0021. Усредненное значение есть $\bar{b} = 0{,}0035$ и $s_{\bar{b}} = 0{,}0045$. Следовательно, по таблицам t при трех степенях свободы 95%-ные доверительные границы для $\bar{b}$ составляют интервал от — 0,0108 до 0,0178.

Для всего диапазона изменения от 100 г до 400 г. (6 = (7—1) в кодовых единицах) мы должны умножить коэффициент и границы на 6, что в нашем примере дает $6\,\bar{b}=0{,}0210$ и 95%-ный интервал: —0,065 ÷ 0,107. Для оцениваемого изменения в исходных наблюдениях это соответствует лишь 5 % примерно (1,05 антилогарифм от 0,0210), с 95%-ным доверительным интервалом от — 14% до 28% (множитель, характеризующий изменение, меняется от 0,86 до 1,28). С точки зрения разброса данных из других источников есть много целей, для которых разброс из-за начальных весов не только не отличается значимо от нуля, но и вообще не очень важен.

Заметим, что этот результат, полученный прямым оцениванием внутренней неопределенности, точно такой же, как и следующий из непосредственно полного дисперсионного анализа, в котором взаимодействие «коэффициент × испытуемый» рассматривается как компонент ошибки для коэффициента. Может быть, величайшие обобщающие возможности дисперсионного анализа состоят в том, что он обеспечивает прямую оценку внутренней неопределенности, а не в его обычных целях.

Можно распространить простую прямую оценку внутренней неопределенности на результаты с более сложной зависимостью от данных, чем линейные комбинации с постоянными весами. Иногда этого вполне достаточно, но в иных случаях возникают трудности, требующие более серьезного рассмотрения.

---

*Авторы примера использовали рандомизированный дробный факторный план. К сожалению, оригинал их работы 1944 г. не удалось отыскать в библиотеках СССР, что препятствует более полному комментированию этой задачи. — *Примеч. ред.*

7.6. ТРУДНОСТИ, СВЯЗАННЫЕ С ПРЯМЫМИ ОЦЕНКАМИ

С вопросом о дополнительной внешней изменчивости, которому посвящен следующий параграф, непосредственно связаны только что упомянутые основные трудности прямых оценок изменчивости:

а) может статься, подсчет разумного результата сделать не удастся, ввиду столь малочисленных данных, что из них и групп-то как следует не выберешь;

б) даже если эти результаты окажутся вполне разумными, они все равно будут столь сильно смещенными, что об их использовании не может быть речи. Ни одна из этих трудностей не возникает в примере из предыдущего параграфа, поскольку там все результаты были линейными комбинациями наблюдений с постоянными весами. Так, анализируй мы те же самые данные Джонсона — Цзао без разделения четырех испытуемых, нашим первым делом должно было бы быть определение арифметических средних по испытуемым, и результирующий коэффициент регрессии стал бы средним арифметическим коэффициентов для каждого из четырех индивидуумов. Подобные результаты возможны и для некоторых типов взаимно проникающих подвыборок из параграфа 7,5. До тех пор пока имеет место линейность с постоянными весами, — все просто.

В гл. 8 мы выясним, как обращаться с более сложными случаями, когда появляются трудности обоих видов.

Наряду с этими трудностями есть и более известная проблема, возникающая при наличии двух или более независимых друг от друга мер изменчивости, каждая из которых будет вносить свой компонент в ошибку. Допустим, мы проводили исследование реакции учеников на международные новости в 20 школах из самых разных областей страны ежегодно в течение 10 лет. Различия между школами, обусловленные региональными различиями, по-видимому, неизбежны и, конечно, внесут вклад как важный вид изменчивости. Но нельзя пренебречь и изменениями от года к году воздействия международных новостей.

К счастью, воздействие каждого из этих главных источников неопределенности можно оценить по самим данным, однако всегда открытым остается вопрос о том, не обусловлена ли неопределенность просто недостаточной случайностью выборки. Мы должны быть готовы к вопросу о том, как одновременно работать с двумя компонентами ошибки. Подробно объяснить, что надо делать, труднее, чем просто сделать. (Делать это легче в терминах дисперсионного анализа, а не общих средних, но существует стандартный перевод. Формулы есть в [Scheffé H. (1959)]. Рабочий пример, иллюстрирующий счет, можно найти в [Cooper B. E. (1969)]. Несколько более сложный пример в [Anderson R. L. (1947)].)

Сказать, что у нас есть 200 групп, каждая из которых представляет одну из школ в один год наблюдения, так что нужно лишь воспользоваться вариацией результатов из этих групп, чтобы оценить общую изменчивость результатов, явно недостаточно.

7.7. ДОПОЛНИТЕЛЬНАЯ НЕОПРЕДЕЛЕННОСТЬ И ЕЕ ВЗАИМООТНОШЕНИЯ С ВНУТРЕННЕЙ НЕОПРЕДЕЛЕННОСТЬЮ

Мы хотим быть во всеоружии перед воздействием систематических ошибок и источников вариации, исключаемых из оценки внутренней неопределенности. Если наблюдения ограничены одним городом, то изменений от города к городу попросту нет, и они не могут внести вклад в оценку внутренней неопределенности. (А если обследуются всего два города, то, как мы уже видели, может оказаться разумным отнести вариацию от города к городу к дополнительной неопределенности). Подобные утверждения можно сделать для годов, областей и многих других признаков, определяющих данные.

Помимо вариаций, которые можно оценить лишь из данных большего объема, могут быть и такие, что заключены в способах сбора данных. Используемые инструменты,— будь то письменные психологические тесты, социологические анкеты или ртутные термометры, — часто страдают неточностью калибровки и бывают чувствительны не к тем факторам, для которых проводятся измерения.

Рассмотрим в качестве простого примера работу специалиста по анализу рынка, который планирует включить в анкету для своих респондентов несколько дополнительных вопросов, выходящих за пределы обычной анкеты. При пробном обследовании он обнаруживает, что интервьюеры без большого труда получают от респондентов ответы на дополнительные вопросы. Действительно, в пробном эксперименте из 50 опрошенных лишь двое не ответили на анкету и дополнительные вопросы одновременно. Чего же теперь наш специалист может ожидать при полном обследовании? Ему может помочь, например, знание о том, что в таких обследованиях примерно на 15% анкет не отвечают даже при персональном опросе. Так что его 96% превращаются по крайней мере в 85%. Далее, он должен учесть, что цифра 96% может оказаться завышенной, причем многое зависит от того, как будет поощряться участие респондентов в этом эксперименте. По крайней мере потеря еще 15% не должна быть неожиданной для исследователя.

(Если важна доля ответов, то можно запланировать пробное обследование со случайно выбранными респондентами и отмечать различные побудительные мотивы, чтобы понять, какую долю ответов они приносят).

Такими источниками дополнительной неопределенности нельзя пренебрегать. То, что эта неопределенность часто оценивается при не полной рандомизации, хотя иногда и смягчаемой данными из других источников, не оправдывает мнения, что ее не существует. Нет никаких оснований, чтобы ее (что обычно) недооценивать (даже когда физики оценивают наиболее фундаментальные константы (см. [Du Mond J.W.M. and Cohen E. R. (1958)]), ничего не остается, как только разрешить не совсем случайные выборки).

Что же нам все-таки со всем этим делать? Рассматривать как нечто совершенно отделенное от внутренней неопределенности? Или как нечто с ней взаимодействующее? Авторам хотелось бы научиться оценивать дополнительную неопределенность и систематические ошибки сов-

местно с внутренними неопределенностями. Исследователь может считать столь же важным сообщение внутренней неопределенности, как и общей. Когда же общая оценка представляет собой нечто такое, что вне нашей власти, то лучше всего, наверное, сообщать оценки отдельных компонентов.

Как же нам получить общую оценку практически?

Легко раскрыть выборочное стандартное отклонение — оценку корня из квадратичной ошибки. Возведем в квадрат стандартное отклонение, добавим квадрат смещения и извлечем квадратный корень. Иногда это сводит натуральный объем выборки до эффективного.

Сравнительно легко объединить дополнительные неопределенности с результатами, выраженными через доверительные интервалы. Пусть мы хотим добавить неопределенность типа «формочки для печенья»: «нечто между — 4% и + 1%». Если доверительный интервал уже определен границами от 62% до 70%, то надо лишь раздвинуть его к границам до 58% и 71%. Аналогично, если мы хотим построить критерий значимости, то можно взять в качестве нулевой гипотезы сначала левую, а затем правую границу. Если же мы захотим добавить неопределенность типа «распределенности», такую, что «систематическая ошибка примерно следует нормальному закону со средним — 1% и стандартным отклонением 3%», то нам придется сочетать ее с нашей моделью и проводить расчеты более искусно. Так, если доверительный интервал для стандартной ошибки в 2% уже построен и, как в предыдущем примере, задан границами 62% и 70%, то надо просто сложить дисперсии $2^2 + 3^2 = 13$ и получить общую оценку в 3,6. При этом новым центром интервала будет 68% — 1% = 67%. А окончательные доверительные границы равны* 67% ± 7,2%, или от 59, 8% до 74,2%.

Если дополнительная неопределенность оценивалась по не слишком случайной выборке и вполне сравнима с изменчивостью, ожидаемой при повторениях всего исследования, то ясно, что не столь уж важно, каковы наши внутренние «правдоподобные рассуждения» на ее счет. Это и есть основы описания *и* общей, *и* внутренней неопределенности. Ведь читатели и слушатели вправе знать, к чему могут повести их собственные решения. Желая сообщить общую неопределенность, мы должны поступать именно так, и наша способность к правдоподобным рассуждениям должна быть достаточно хороша, чтобы мы смогли все это сделать.

Хотя подобные детали анализа и требуют искусного обращения, исследователи, авторы и их консультанты-статистики всегда обязаны серьезно относиться к выбору, чтобы всякий раз вместе с оценкой внутренней неопределенности оценивать и дополнительную. Результаты могут проявиться в форме «словесных предупреждений», а не обязательно в виде чисел. Слова — часто допустимый минимум, но каждый из нас должен стараться сделать лучше то, что возможно. Наш долг — оценивать дополнительную и общую неопределенности как для наших чи-

---

*Видимо, описка: центр будет 66% — 1% = 65% и соответствующими границами доверительного интервала числа 65% ± 7,2%, или от 57,8% до 72,2%. — *Примеч. пер.*

тателей, так и для тех исследователей, которые будут работать после нас, даже если мы можем прямо оценить внутреннюю неопределенность.

Несмотря на то что хорошие методы оценивания дополнительной неопределенности кажутся глубоко связанными с сутью самого анализа, широкое обсуждение внутренней сущности явления со специалистами может помочь статистикам найти новые методы с более широким спектром применения. Это направление нуждается в разработке.

В заключение подчеркнем еще раз, что по очень разным соображениям оценка дополнительной неопределенности требуется и в строго контролируемых лабораторных работах, и в тех крупномасштабных исследованиях, в которых хотелось бы «поиграть» методами, прежде чем они будут завершены.

РЕЗЮМЕ. ОХОТА ЗА ИСТОЧНИКАМИ НЕОПРЕДЕЛЕННОСТЕЙ

Все результаты включают два вида неопределенностей: внутреннюю, которая допускает оценки по данным, и внешнюю, которая таким образом не оценивается.

Величина $\sigma/\sqrt{n}$ может и не оценивать стандартной ошибки даже тогда, когда n наблюдений с дисперсией, заведомо равной σ^2, вносят одинаковый вклад в среднее арифметическое; то же самое можно сказать и о $\sqrt{pq/n}$.

Большой объем хорошо сформированных данных позволяет рассмотреть несколько мер внутренней неопределенности, основанных на сравнении более или менее тесно связанных частей данных. Часто эти меры существенно различны. Выбор одной из них (1) может оказаться существенной проблемой и (2) может основываться на разумной целесообразности.

Можно воспользоваться и взаимно проникающими подвыборками (иногда определяя их постфактум) как основой для оценки внутренней неопределенности.

Когда сравнивается мало объектов и, следовательно, есть мало степеней свободы, мы несем убытки в виде роста изменчивости при оценке внутренней неопределенности.

Основные трудности оценивания неопределенности с помощью равноценных подвыборок (возможное отсутствие определения, вероятное смещение) связаны с малыми подвыборками.

Оценивание внутренней неопределенности, когда результаты допускают классификацию двумя или более способами, — это нетривиальная задача. В частности, обычно недостаточно просто сложить неопределенности, полученные в разных классификациях, или объединить классификации и обрабатывать их как одну.

Дополнительную (внешнюю) неопределенность приходится оценивать при известной неслучайности выборки, а объединение внутренней и внешней неопределенностей в одной число требует осторожности и обоснованности.

Мы часто нуждаемся в оценке дополнительной неопределенности как для экспериментальных данных, так и для результатов наблюдений.

БИБЛИОГРАФИЯ

Anderson R. L. (1947). Use of variance components in the analysis of hog prices in two markets. J. Amer. Stat. Assoc., 42, 612—634.

Cooper B. E. (1969). Statistics for Experimentalists. Elmsford, N. Y., Pergamon Press, Inc., 167.

DuMond J. W. M. and Cohen E. R. (1958). Fundamental constants of atomic physics. B: Condon E. U. and Odishaw H. (Eds.) Handbook of physics. New York, McGraw-Hill (LC: 57—6387), 7—143; 7—173.

Johnson P. O. (1949). Statistical methods in research. New York, Prentice-Hall, 299.

Johnson P. O. and Tsao F. (1944). Factorial design in the determination of differential limen values. Psychometrika, 9, 107—144.

Scheffé H. (1959). The Analysis of Variance. New York, John Wiley and Sons, p. 247. Русский перевод: Шеффе Г. Дисперсионный анализ. Пер. с англ. М., Физматгиз, 1963.

Shewhart W. A. (1931). Economic Control of Quality of Manufactured Product. New York, Van Nostrand, 361.

ИЛЛЮСТРАЦИИ

Иллюстрация 7.4.1

Двусторонние 5%-ные точки для t Стьюдента, $|t_k|_{0,95}$

Степени свободы	$[t]_{0,95}$	Степени свободы	$[t]_{0,95}$
1	12,71	6	2,45
2	4,30	7	2,36
3	3,18	8	2,31
4	2,78	9	2,26
5	2,57	10	2,23
		∞	1,96

Иллюстрация 7.5.1

Эксперимент Джонсона—Цзао

Кодированный вес	(1) Мужчины	(2) Мужчины	(3) Женщины	(4) Женщины
Исходные данные: «зрячие», скорость 300 г/мин Привес в граммах до реакции «есть!»				
1	15,8	35,0	27,2	12,2
2	18,6	39,3	41,1	9,6
3	12,2	47,8	32,2	11,7
4	12,8	38,2	21,3	12,4
5	16,5	57,7	33,7	11,9
6	15,8	39,7	28,2	12,8
7	17,0	44,8	29,6	10,5

Преобразование — десятичные логарифмы исходных данных:

x	$y=\log_{10}$ (исходные данные)			
1	1,20	1,54	1,43	1,09
2	1,27	1,59	1,61	0,98
3	1,09	1,68	1,51	1,07
4	1,11	1,58	1,33	1,09
5	1,22	1,76	1,53	1,08
6	1,20	1,60	1,45	1,11
7	1,23	1,65	1,47	1,02
$\Sigma x=28$	$\Sigma y=8,32$	11,40	10,33	7,44
Σxy	33,36	46,03	41,14	29,82
$\Sigma x^2=140$	$(\Sigma x)^2=784$		$n=7$	

Глава 8 МЕТОД ПРЯМОГО ОЦЕНИВАНИЯ

8.1. «СКЛАДНОЙ НОЖ»

Стоит встретиться более сложной статистике, чем взвешенное среднее, и мы тотчас сталкиваемся с трудностями оценки ее устойчивости, даже если объем выборки довольно велик. Так, например, при построении регрессионной зависимости с k факторами требуется не менее чем $k + 1$ опыт, да и не многие удовлетворились бы результатом при столь малом числе точек. Следовательно, если каждая группа нуждается в значительном объеме данных, то число групп, требуемых для прямой оценки изменчивости стандартным методом из параграфа 7.5, может быть лишь весьма ограниченным. Более того, многие статистики имеют смещенные оценки, если они получены по малым выборкам. Как правило, главный член этого смещения пропорционален $1/n$, где n — объем выборки. Значит, среднее, подсчитанное по нескольким подвыборкам, может оказаться более смещенным, чем простое среднее по всем данным, по крайней мере, когда первичные выборки малы. «Складной нож» — это широко распространенный метод, предназначенный для преодоления или смягчения названных проблем.

Понятие «складной нож» относится к универсальному методу, призванному заменить частные методики, которые не всегда пригодны, подобно бойскаутскому ножу, годящемуся на все случаи жизни. «Складной нож» открывает пути к построению разумных доверительных границ в сложных случаях. Основная идея заключается в оценке эффекта каждой из групп, на которые были разбиты данные, но не самого по себе, что мы делали в параграфе 7.5, а скорее через тот результат, который получится, если данную группу *выкинуть*.

Иллюстрации облегчают понимание. При отыскании невзвешенного среднего пяти чисел любое число можно легко представить как взвешенную разность двух средних, среднего всех пяти значений и среднего тех четырех, которые останутся после отбрасывания искомого. Действительно, пусть, например, нам надо выразить число 7 в выборке чисел 3, 5, 7, 10, 15. Имеем

$$5\left(\frac{3+5+7+10+15)}{5}\right)-4\left(\frac{3+5+10+15}{4}\right)=7.$$

Этот результат не только элементарно доказывается, но и имеет тривиальные следствия. Для усреднения с равными весами следствия действительно тривиальны. Но едва только появляется более

сложная статистика, чем среднее, как оказывается, что «те же самые» вычисления с «кусками» данных не сводятся к тем же результатам, что и для всех данных. Взамен этого мы получаем нечто гораздо более полезное. В частности, как мы увидим позднее, такими сложными статистиками могут быть даже уравнения регрессии, а не только простые числа.

Вот два главных момента в методе «складного ножа»: сначала мы проделываем желаемые вычисления для всех данных, затем данные разбиваем на группы и повторяем вычисления, последовательно отбрасывая каждую группу для несколько уменьшившихся данных.

Пусть теперь $y_{(j)}$ — результат, полученный после отбрасывания j-й подгруппы для сложной статистики, т. е. полученный для объединения (k — 1) подгруппы. Пусть далее $y_{\text{общ}}$ соответствует аналогичному результату для всей исходной выборки. Введем *псевдозначения* следующим образом:

$$y*_j = k y_{\text{общ}} - (k-1)\, y_{(j)}, \quad j = 1, 2, \ldots, k. \tag{1}$$

Теперь эти псевдозначения играют ту же роль, что и исходные данные в параграфе 7.5 при подсчете результатов отдельно для каждой группы. Отметим, что, как и в примере со средним из пяти чисел, при вычислении средних с равными весами $y*_j = y_j$, где y_j есть результат отдельной j-й группы. Соответственно для простых средних «складной нож» сводится к методу из параграфа 7.5, на что мы и надеялись.

Для $y*_j$ в конце концов требуется примерно такая же точность, как и для y_j. Значит, умножения на k и (k — 1), которые могут быть велики, как правило, следует вести с бóльшим числом десятичных знаков, чем требовалось бы, если бы ими пользовались непосредственно. Следовательно, значения $y*_j$ особенно чувствительны к ошибкам счета и округлению в $y_{\text{общ}}$ и $y_{(j)}$, хотя их чувствительность к изменчивости данных обычна, а пожалуй, и меньше, чем y_j, которые они заменяют.

Ключевая идея состоит в том, что для широкого класса задач псевдозначения должны использоваться при построении доверительных границ с помощью t-критерия так, как если бы они были результатами, предназначенными для вычисления какой-нибудь сложной статистики в каждой из k независимых групп данных. Слова «как если бы» здесь главные. Критерий Стьюдента тем не менее во многих случаях прекрасно работает, если отклонения от $y*_j$ действительно независимы.

Среднее «складного ножа», которое есть наилучшая оценка, а также и оценка дисперсии s^2* даются выражениями

$$y* = \frac{1}{k}\,(y*_1 + \ldots + y*_k);$$

$$s^2 = \frac{\Sigma y^2*_j - \frac{1}{k}\,(\Sigma y*_j)^2}{k-1};$$

$$s^2* = s^2/k.$$

(Когда $y_{(j)}$ округляются или как-то квантуются после или во время вычислений, можно по предложению Тьюки (неопубликованная

работа) пользоваться консервативной тактикой увеличения s_*^2 на $k^2\tau^2$, где τ^2 часто можно рассматривать как дисперсию равномерного распределения на интервале округления или квантования. Так, если мы округляем $y_{(j)}$ до третьего знака, соответствующее смещение будет иметь равномерное распределение в диапазоне 0,001, т. е. от $x - 0{,}0005$ до $x + 0{,}0005$, где x — округленное значение. Используя обычную формулу $L^2/12$ для дисперсии прямоугольного распределения на интервале длины L, мы получим $\tau^2 = (0{,}001)^2/12 = 0{,}00000008$. Если только k невелико или s^2* очень мало, то можно не обращать внимания на член $k^2\tau^2$ при столь мизерном τ^2. Сравнить $k^2\tau^2$ с s^2* стоит не дорого.)

Корректировка числа степеней свободы

Трудности метода «складного ножа», похоже, чаще всего возникают по двум причинам:

- слишком разбросанные «хвосты»;
- дискретность получаемых значений.

Для слишком разбросанных «хвостов» вообще не получается хороших результатов ни в какой ситуации и ни для какого метода анализа. Если только мы готовы соответствующим образом преобразовать то, что мы оцениваем, то могут помочь преодолению этой трудности методы, обсуждаемые в гл. 10. Но если мы настаиваем на более классических оценках, вроде среднего арифметического, то нас ничто не спасет.

Что-то надо делать и с дискретностью. Давайте рассмотрим естественный, но весьма крайний случай. Пусть мы отыскиваем медиану «складного ножа» для выборки с четным числом элементов, скажем $k = 2m$. Если мы отбросим любое наблюдение в верхней половине этих $2m$ значений, медианой станет m-е значение снизу. А если отбросить любое из нижних значений, то в медиану превратится $(m + 1)$ значение в исходной выборке.

Значит, $y_{(j)}$ примет только два разных значения, повторенных по m раз, и то же самое относится к псевдозначениям. Тогда стабильность s^2* будет не выше, чем у квадрата разности двух значений. Но это оставляет нам лишь одну степень свободы для t-критерия при таких s.

Можно ли повысить стабильность $s*$? Да. Метод «складного ножа» для групп, обсуждаемый в параграфе 8.3, даст нам более двух различных псевдозначений. Это, должно быть, очень ценно, поскольку только успешное применение всех элементов исходной выборки сохраняет степени свободы, чего практически нельзя было сделать в нашем примере с медианой.

Было бы, конечно, полезно эмпирическое правило для определения уменьшенного числа степеней свободы. Могут пригодиться следующие простые правила:

а) сосчитайте число различных псевдозначений, отнимите единицу и пользуйтесь результатом как числом степеней свободы.

Это правило применимо только тогда, когда одинаковость псевдозначений обусловлена способом вычислений, как в примере с медианой или размахом, а не в случаях, когда она обусловлена природой исходных наблюдений или вытекает из арифметики. В частности,

б) если малые изменения исходных наблюдений — вроде того, что вместо фактических 0 и 1 фигурируют — 0,001, + 0,002, 0,997 и 1,004 — приводят к двум разным наборам псевдозначений, то *не* стоит считать их «одинаковыми», применяя правило (а). *Пример*: процент «успехов» в биномиальных наблюдениях,

в) если сохранение большего числа десятичных знаков приводит к двум разным наборам псевдозначений, то *не* стоит считать их «одинаковыми», применяя правило (а).

Дополнение к 8.1. Сочетания и преобразования

Результаты анализа данных не всегда просто числа. Но когда мы имеем дело с некоторыми числами, скажем y, z, v и w, можем взять просто k групп и применить «складной нож» отдельно для каждой из них. Имея согласованные множества псевдозначений $(y*_j,\ z*_j,\ v*_j,\ w*_j)$ для каждого j, мы легко найдем множество k замещающих значений для любого сочетания или функции этих результатов, образующееся, например, из значений

$$(y*_j+z*_j)/(v*_j+w*_j),$$

которые как-то оценивают соответствующие величины

$$(y+z)/(v+w).$$

Аналогично $\log y*_j$ что-то говорит об оценке $\log y$. Этот подход распространяется на сочетания, которые зависят от дополнительной переменной или переменных, как в выражениях

$$ye^{-zt} \text{ или } y\cdot x_1+z\cdot x_2+v\cdot x_3+w\cdot x_4,$$

где мы можем построить такие замещающие функции:

$$y*_j e^{-z*_j t} \text{ и } y*_j\cdot x_1+z*_j\cdot x_2+v*_j\cdot x_3+w*_j\cdot x_4.$$

При обдумывании вопросов о таких сочетаниях и преобразованиях нам следует помнить, что порядок операций — главный момент. Так, например, почти всегда

$$(\log y)*_j \neq \log (y*_j),$$

хотя оба эти выражения скорее всего будут подобны. По определению, мы должны иметь

$$(\log y)_{(j)}=\log(y_{(j)}).$$

Увы, это равенство нас не спасает, поскольку

$$(\log y)*_j = k\cdot(\log y)_{\text{общ}}-(k-1)\,(\log y)_{(j)} = k\cdot\log(y_{\text{общ}})-(k-1)\log(y_{(j)}) =$$
$$= k\cdot\log(y_{\text{общ}})-(k-1)\log\left(\frac{k\cdot y_{\text{общ}}-y*_j}{k-1}\right),$$

что не равно $\log (y*_j)$. Так, например, при $k=2$, $y_{\text{общ}}=4$ и $y*_j=3$ подстановка дает $2\log 4-\log 5=\log 3{,}2$, что отнюдь не равно $\log 3$. Два следствия из этого заслуживают внимания:

● может статься, что «складной нож» с одними выражениями будет работать лучше, чем с другими (как с $\log y$ или y^2 вместо y);

● если мы имеем дело с *линейными* комбинациями данных, вроде

$$y - 3z + 2v, \text{ или } y + zt + vt^2 + wt^3,$$

где t — дополнительная переменная, или еще

$$y \cdot x_1 + z \cdot x_2 + v \cdot x_3 + w \cdot x_4,$$

где x_1, x_2, x_3 и x_4 — дополнительные переменные (регрессионные переменные — факторы), то порядок действий *не* существен, т. е., используя «складной нож» отдельно для y, z, v, w, получим те же самые части, вполне достаточные для построения всех сочетаний. Следовательно, когда возможно много наборов x, вычисление коэффициентов методом «складного ножа» становится экономичным.

Мы мало знаем насчет того, какие выражения способствуют улучшению поведения «складного ножа». Разве что можем засвидетельствовать следующее:

● весьма желательно *избегать* ситуаций, где выборочное распределение величин, прошедших через метод «складного ножа», имеет резкую границу или где возможные значения оценок ограничены интервалом либо полупрямой. Так, например, если оценки в «складном ноже» — вероятности, то в конечных результатах вряд ли будет много значений меньших, чем нуль, или больших, чем единица. Одним из возможных подходов в этом случае мог бы быть переход от исходных данных к логитам $\log [p/(1-p)]$, а затем обратное преобразование (антилогиты) конечных результатов. Это позволит сохранить числа в границах;

● весьма желательно *избегать* выборочных распределений с разбросанными «хвостами» (даже одним «хвостом»);

● по-видимому, стоит *избегать* явно асимметричных выборочных распределений.

Короче говоря, можно подвергать воздействию метода «складного ножа» некоторые числа, несущие информацию о любых сочетаниях исходных данных. Наши выводы вообще будут несколько отличны от тех, которые получились бы при прямом применении «складного ножа» к выбранным сочетаниям. Это дает возможность выбора, что в свою очередь иногда сулит улучшение наших выводов. Если же мы работаем с линейными комбинациями, как, скажем, в случае оценки коэффициентов уравнения множественной регрессии, то никаких различий не возникает.

Теперь было бы полезно перечитать два последних параграфа гл. 2, заменяя «перепроверку» методом «складного ножа».

8.2. ПРИМЕРЫ ДЛЯ ЭЛЕМЕНТОВ ВЫБОРКИ

Рассмотрим четыре примера. Первый из них — напоминание об идеях и их применение в случае статистического вывода о стандартном отклонении для асимметричного распределения с разбросанными «хвостами». Второй — более простой и ясный — для нестандартной обработки, когда общая теория «не работает». С другой стороны, нет задачи, для которой бы «складной нож» дал наилучший результат, но как быть без него? В третьем примере мы сталкиваемся с небольшим об-

следованием, где решено брать группы по несколько элементов в каждой. Наконец, четвертый — более подробное обсуждение сложной задачи с многими откликами (многокритериальной задачи). Он не однажды демонстрирует мощность метода в ситуации, где в противном случае нам пришлось бы отыскивать самые сильные индикации, на какие мы только способны.

Пример (первый из четырех). Доверительные границы стандартного отклонения. Выборка из некоторого распределения дала следующие 11 значений:

0,1; 0,1; 0,1; 0,4; 0,5; 1,0; 1,1; 1,3; 1,9; 1,9; 4,7.

Нет никаких оснований предполагать, что распределение нормально, и, более того, можно, видимо, думать, что это не так. Попробуем построить доверительные границы для стандартного отклонения σ.

Первое решение. Поскольку данных мало, давайте рассматривать каждое измерение как группу единичного объема. Обозначим измерения $x_1, x_2, \ldots, x_{11}$. Поскольку объем каждой группы равен единице, стандартное отклонение в группе считать не надо. Найдем сначала стандартное отклонение для всех измерений вместе, т. е. для объединения всех групп:

$$y_{\text{общ}} = \sqrt{\Sigma (x_i - \bar{x})^2/10}\,.$$

Оно приведено в шапке илл. 8.2.1. Затем выкинем j-е измерение, т. е. j-ю группу, и найдем

$$y_{(j)} = \sqrt{\Sigma (x_i - \bar{x}_{(j)})^2/9}\,,$$

где $\bar{x}_{(j)}$ — среднее из 10 значений, оставшихся после отбрасывания j-го, а сумма для $y_{(j)}$ подсчитана без $(x_j - \bar{x}_{(j)})^2$. Найдем эти y для каждого из 11 измерений (групп). Их значения сведены в третьем столбце илл. 8.2.1. Теперь можно найти и псевдозначения

$$y^*_j = 11 y_{\text{общ}} - 10 y_{(j)}$$

(см. четвертый столбец таблицы илл. 8.2.1).

Оценкой для σ как раз и будет y^* — среднее наших псевдозначений, которые оказываются близкими к 1,49. Детали видны из илл. 8.2.1, где представлены также 2/3-доверительные интервалы для σ — от 0,85 до 2,13 и 95%-ные доверительные интервалы — от 0,10 до 2,88. А так как «про себя» мы знаем, что эти данные извлечены из экспоненциального распределения со средним и стандартным отклонениями, равными каждое единице, то нам не слишком нужны все эти границы.

Второе решение. Мы вовсе не обязаны применять метод «складного ножа» непосредственно к s, как мы только что сделали. Можно было бы, например, работать с $\log s$. Обдумав эту возможность, мы намерены реализовать ее тут же, поскольку, как известно, выборочное распределение $\log s$ ведет себя лучше, чем само s. В частности, распределение логарифма s обычно ближе к симметричному и имеет менее разбросанные хвосты, чем выборочное распределение s. Конечно, логарифм может оказаться более смещенным, но несмещенность — не главное требо-

вание, особенно в свете способности метода «складного ножа» к снижению смещения.

Особенности и результаты приведены на илл. 8.2.2. Воспользуемся обозначениями Y для логарифмов, оставляя y для данных илл. 8.2.1. Теперь 2/3-доверительные интервалы для σ меняются от 0,94 до 3,29, а 95%-ные интервалы — соответственно от 0,44 до 6,93.

Неудовлетворительные решения. Давайте для сравнения рассмотрим некоторые подходы к обработке тех же данных, основанные на гипотезе о нормальном распределении. Общая s^2 равна 1,805. А вот фрагмент таблицы значений хи-квадрат для 10 степеней свободы:

	2,5%	1/6	1/2	5/6	97,5%
Значения	3,25	5,78	9,34	14,15	20,48

Следовательно, обычные симметричные доверительные границы для σ^2 равны

$$\frac{1{,}805}{20{,}48/10}=0{,}88 \text{ и } \frac{1{,}805}{3{,}25/10}=5{,}55,$$

в то время как 2/3-границы — соответственно 1,28 и 3,12. Если перейти к σ, получим 2/3-границы от 1,13 до 1,77, а 95%-ные — от 0,94 до 2,36. Такие границы, видимо, слишком оптимистически узки.

Точно так же, переходя к размаху $w = 4{,}7 - 0{,}1 = 4{,}6$ и имея следующие процентные точки для w/σ в выборке объема 11 из нормального распределения:

	2,5%	1/6	1/2	5/6	97,5%
Значения	1,78	2,41	3,12	3,93	4,86

найдем границы для σ от 0,95, до 2,58 при 95%-ном уровне и от 1,17 до 1,91 для уровня 2/3.

Комментарии. Сравним четыре множества решений:

Источник	2/3-границы	95 %-ные границы
«складной нож» для s	$0{,}85 \leqslant \sigma \leqslant 2{,}13$	$0{,}10 \leqslant \sigma \leqslant 2{,}88$
«складной нож» для $\log s$	$0{,}94 \leqslant \sigma \leqslant 3{,}29$	$0{,}44 \leqslant \sigma \leqslant 6{,}93$
s^2/σ^2	$1{,}13 \leqslant \sigma \leqslant 1{,}77$	$0{,}94 \leqslant \sigma \leqslant 2{,}36$
w/σ	$1{,}17 \leqslant \sigma \leqslant 1{,}91$	$0{,}95 \leqslant \sigma \leqslant 2{,}58$

Сравнение двух множеств, связанных со «складным ножом», демонстрирует явную аналогию между границами для уровней 2/3 и 95%. Правда, надо принять во внимание тот факт, что «складной нож» для $\log s$ не может привести к отрицательной нижней границе σ, тогда как для s — вполне может. В этом смысле лучше работать с ло-

гарифмами. Тем не менее в данном случае и само s сработало прекрасно. Есть и другие возможности вроде «складного ножа» для $\sqrt{s}$ или $\sqrt[3]{s^2}$.

Для экспоненциального распределения, откуда фактически извлекалась наша выборка, $\operatorname{var} s \approx 2\,(\sigma^2/n)$, тогда как для нормального — $\operatorname{var} s \approx \frac{1}{2}(\sigma^2/n)$. Отношения этих величин — меры относительной изменчивости стандартных отклонений для выборок из этих различных распределений. Значит, в выборке экспоненциального вида изменчивость компонентов дисперсии может оказаться в четыре раза выше, чем для нормальной выборки. Тогда можно ожидать и множителя порядка $\sqrt{4} = 2$ в ширине доверительных интервалов. Следовательно, прямое применение нормальной теории просто не может дать сколько-нибудь приемлемых доверительных границ, что и заставляет нас отнести два последних решения к неудовлетворительным.

Пример (второй из четырех). Оценивание 10%-ных точек объединения популяций. Каждому объекту некоторой генеральной совокупности соответствует множество измерений. Для каждого из 11 объектов выборки сделано по 5 измерений. На илл. 8.2.3 показаны гипотетические измерения, упорядоченные по убыванию в каждом объекте. Требуется оценить верхнюю 10%-ную точку по всей совокупности измерений.

Решение. Для оценки 10%-ной точки можно избрать различные пути. Один из подходов: объединить группы, принять, что расстояние между любой парой соседних измерений пропорционально $1/(n+1)$ общего распределения и проинтерполировать на 10%. Так, 5 измерений делят совокупность на 6 частей, на каждую из которых приходится в среднем 1/6 вероятности. Правда, имея всего 5 измерений, трудно оценить верхнюю 10%-ную точку, поскольку эта оценка придется, конечно, на ту часть интервала, которая лежит за самым большим наблюдением и часто неопределенно длинна. Здесь же — другой случай, где объединение есть просто соединение всех частей, но этого достаточно.

Счет можно упростить, выбирая, как на илл. 8.2.4, немногие самые большие наблюдения. Давайте сперва оценим 10%-ную точку по всем 55 измерениям. Здесь 55 точек образуют 56 интервалов, а мы желаем найти среди них интервал $56/10 = 5{,}6$, начиная сверху. На илл. 8.2.4 положим $y_{\text{общ}} = 0{,}6$ между 5-м и 6-м измерениями. Тогда

$$5{,}172 - 5{,}137 = 0{,}035; \quad 0{,}6 \cdot 0{,}035 = 0{,}021; \quad 5{,}172 - 0{,}021 = 5{,}151 = y_{\text{общ}}.$$

Когда мы отбрасываем первый объект, можно найти $y_{(1)}$, основываясь на 50 точках, или 51 интервале. Значит, мы можем дойти до $51/10 = 5{,}1$ интервала. При отбрасывании первого объекта из илл. 8.2.4 пропадают два измерения, 6,880 и 5,172, поскольку они принадлежат этой группе. Интерполируя между 5-м и 6-м из оставшихся измерений, найдем

$$4{,}756 - 4{,}660 = 0{,}096; \quad 0{,}1\,(0{,}096) = 0{,}0096 \approx 0{,}010$$

и

$$y_{(1)} = 4{,}756 - 0{,}010 = 4{,}746.$$

Илл. 8.2.5 показывает все результаты: $y_{(j)}$, $y*_j$, $y*$, $s*$, а также 95%-ные доверительные границы, причем приходится пользоваться лишь двумя степенями свободы, так как получилось только три разных значения. Обратите внимание, что оценка $y* = 5{,}874$ заметно больше, чем прямой результат $y_{\text{общ}} = 5{,}151$. В данной конкретной задаче мы «по секрету» знаем, что 10%-ная точка равна 5,773, ибо совокупность, связанная с объектами, построена из смеси трех нормальных распределений с $\mu = 2$, 3 и 4 (с вероятностью 1/3 каждое) и независимыми $\sigma = 1,2,3$ с теми же вероятностями. Стало быть, мы имеем 9 разных равновероятных нормальных распределений. Мы выбирали одно из них с возвращением для каждого из 11 объектов.

Наш пример из тех, для которых метод «складного ножа» подходит не слишком хорошо. Для него гораздо лучше несколько итераций с обычными порядковыми статистиками. Повторение значения 4,981 на илл. 8.2.5 — это симптом трудности. Вообще говоря, вариации максимумов и минимумов и размахов существенно зависят от точной формы лежащего в основе данных распределения. Соответственно вполне возможно, что нельзя построить устойчивых и корректных доверительных границ для их значений. Но даже и в таких ситуациях «складной нож» часто лучший метод из всех возможных. Грубая идея о неопределенности все же лучше, чем ничего.

8.3. «СКЛАДНОЙ НОЖ» ДЛЯ ГРУПП ДАННЫХ: ОЦЕНИВАНИЕ ДОЛИ В ВЫБОРОЧНОМ ОБСЛЕДОВАНИИ

На практике обычно мы делим наши данные не столь дробно, сравнивая группы, образованные более чем из одного объекта или опыта.

Пример (третий из четырех). Оценка доли. Поясняя технику оценивания долей, Кокрэн [Cochran W. G. (1953, p. 113); (1963, p. 156)]* приводит данные о размерах городов (число жителей) в 1920 и 1930 гг. для случайной выборки объема 49, извлеченной из совокупности 196 больших городов США. На илл. 8.3.1 воспроизведены эти данные с итогами по семеркам и в целом. Когда считаешь подобный пример вручную, пользуясь только таблицами, такое разбиение на семерки не затруднительно, поскольку не вносит никакой дополнительной работы, зато короткие суммы гораздо легче проверять, чем (если бы пришлось) сумму всех 49 значений.

Формула для оценки доли (прироста) во всей совокупности в 1930 г. такова:

$$\frac{(\text{вся выборка} = 1930)}{(\text{вся выборка} = 1920)} \times (\text{вся совокупность} = 1920),$$

так что для логарифма оценки всей совокупности — 1930 будем иметь

$$\log(\text{вся выборка} = 1930) - \log(\text{вся выборка} = 1920) + \log(\text{вся совокупность} = 1920).$$

*В русском переводе (см. библиографию к гл. 8) эти данные приводятся на с. 174. — *Примеч. ред.*

Следовательно, будет вполне естественно работать и при «складном ноже» с величиной

$$z = \log(\text{вся выборка} = 1930) - \log(\text{вся выборка} = 1920),$$

ибо такой выбор минимизирует число умножений и делений.

Дальнейшие вычисления демонстрирует илл. 8.3.2, где в столбце «Все» числа 5054 и 6262 взяты непосредственно с предыдущей илл. 8.3.1, а в столбце «$i = 1$» числа 4303 $=$ 5054 — 751 и 5347 $=$ 6262— 915 есть результаты отбрасывания первых 7 городов; все остальные столбцы — аналогично. Пятизначные логарифмы обычно дают более чем достаточную точность, так что псевдозначения z удобно округлить до трех знаков. Для удобства ручного счета из каждого значения $z*_i$ мы вычли произвольное центральное значение 0,001, после чего умножили результаты на 1000. Эти условные (кодированные) значения и использовались для отыскания значений

$$95\% \text{ границ} = \text{среднее} \pm \text{допуск}.$$

Все результаты сведены на илл. 8.3.3. Итоговая оценка, равная 28300, примерно на 100 ниже оценки, полученной без «складного ножа». (Поскольку действительное значение итога 1930 г. равно 29351, автоматическая корректировка смещения, эффективная только для средних, в данном случае не помогает.) Границы этой оценки обычно оказываются несколько шире, чем если бы мы рассматривали каждый город как отдельную группу; действительно,

$$|t_6|_{0,95} = 2{,}447; \quad |t_{47}|_{0,95} = 2{,}012.$$

Стандартная ошибка здесь получается равной $\pm 0{,}0125$ условных единиц, или примерно $\pm$ 830 исходных (antilog 0,0125 $\approx$ 1,0292; 0,0292 $\cdot$ 28300 $\approx$ 830). Это значение, имеющее всего 6 степеней свободы, вполне согласуется с результатом Кокрэна — 604 [Cochran W. G. (1953, p. 119); (1963, p. 163)].

Теперь, как мы обещали в параграфе 2.2, рассмотрим метод, позволяющий исследователю, сравнивающему два вида проективного теста, сделать нечто большее, чем просто индикация.

Эксперимент должен включать некоторое множество объектов, мы же хотим обобщить результат на широкий класс подобных объектов. Нам надо только разделить эти объекты на подходящее число групп и воспользоваться «складным ножом» для всех вычислений, включая оценки и доверительные границы для разностей усредненных коэффициентов надежности, соответствующих двум разновидностям теста. Хотя это и потребует довольно большого счета, приложение «складного ножа» здесь стандартно.

Умение работать с группами так, как будто они простые объекты,— ключевое достижение метода «складного ножа», и это не только способ сведения счета к разумным границам, но, что более важно, это предпосылка для борьбы с гораздо большими ошибками, чем раньше. В частности, это подход к планированию выборки для соответствующей оценки стабильности любого результата обследования. Если объекты извлекаются группами (кластерами), то в ошибку входит как составляющая вариация между группами, и мы должны быть уверены, что

каждый кусок данных состоит из одного или нескольких целых кластеров. А если объекты отбираются с расслоением и объемы слоев известны, то вариацию между слоями надо исключить из ошибки. Это условие, которое иногда может обеспечить выбор кусков (и всегда важно при выборе приемов счета).

8.4. БОЛЕЕ СЛОЖНЫЙ ПРИМЕР

Обратимся теперь к примеру «складного ножа», где мы не сможем предложить никакой приемлемой альтернативы. Аналогия между «отбрасывай по одному» как способом перепроверки из конца параграфа 2.6 и методом «складного ножа», несомненно, бросится в глаза читателю. Пример, к которому мы теперь приступаем, включает как перепроверку, так и полный анализ устойчивости. Следовательно, мы не должны удивляться тому, что мы применим и «отбрасывание по одному» (т. е. «складной нож») при оценивании устойчивости, и «отбрасывание по одному» при перепроверке. Действительно, когда мы доберемся до обсуждения, то увидим, что наш пример порождает вопросы, при решении которых методы «отбрасывания по одному» вполне естественны не только при «двукратном», но и при «трехкратном» применении.

Пример (четвертый из четырех). Дискриминация. Этот пример относится к проблеме авторства. Александр Гамильтон (Alexander Hamilton) и Джеймс Мэдисон (James Madison) некоторое время писали по одним и тем же политическим вопросам; сходны были и их биографии. Один из подходов к установлению авторства в отношении некоторых их работ основан на анализе частоты, с которой каждый из них употреблял высокочастотные слова. Весьма подробное и успешное исследование этого вопроса осуществлено Мостеллером и Уоллисом [Mosteller F., Wallase D. L. (1964, 1963)], чем мы и воспользуемся. Но поскольку исследователи часто убеждаются сами в существовании хронической главной трудности задач — много факторов, мало данных, мы представили здесь, для обнажения методологии «складного ножа» при оценке изменчивости, небольшое новое исследование этого вопроса.

Мы попытаемся дискриминировать некоторые тексты Гамильтона и Мэдисона по пяти словам, которыми они пользовались наиболее часто. Тогда мы увидим, хорошо ли работает метод. Преимущество выбора именно 5 наиболее часто употребляемых слов с точки зрения данного примера вовсе не в том, что этот выбор основан на какой-то априорной оценке того, хорошо или плохо разделяют отобранные слова тексты наших авторов, это преимущество простоты, но отнюдь не обязательно простоты дискриминации. Решение взять 5 слов — произвольное решение: хотелось бы, чтобы пример достаточно полно воспроизводил действительность, но довольно скромное место, которое мы (и читатель) можем этому уделить, заставляет нас мчаться во весь опор. Еще одно преимущество: поскольку в зависимости от большего или меньшего числа случаев появления частоты варьируют столь слабо, что для использования таких высокочастотных слов они должны

вести себя хорошо, возникает уверенность в эффективности всех стандартных методов обработки данных*.

Мы отобрали по 11 статей каждого автора, принадлежность которых известна; они отбирались в основном из статей в «The Federalist». Были взяты именно эти 22 статьи, поскольку среди примерно 100 известных статей они были ближе всего к объему в 2500 слов, варьируя в пределах от примерно 2200 до 2800. В некотором смысле было бы лучше выбрать их случайно. Для удобства применения «складного ножа» каждая статья Гамильтона была случайным образом сопоставлена с одной из статей Мэдисона. Возможно, что было бы разумнее упорядочить пары, но мы поступили иначе. Число $k = 11$ выбрали отчасти потому, что оно больше круглого числа 10, а нам часто надо умножать или делить на $k - 1$. Кроме того, 10 лишь вдвое больше, чем 5 — число факторов, включенных в анализ, а одна из наших целей — показать изменчивость, которую можно ожидать в исследовании со слабо дискриминирующими факторами и весьма умеренным объемом данных, пригодных для определения метода различения. В данном случае мы работаем только со статьями, авторство которых установлено, тогда как подход имеет целью показать, конечно, средства различения «неизвестных» статей.

Попарное сопоставление сделали экономно, иначе нам пришлось бы удалять по одной статье и проделывать вычисления 22 раза. У этого метода нет никаких особых достоинств и он может вводить в заблуждение, если не принимать его как способ экономии.

Стандартная процедура, которая и здесь кажется наиболее разумной, основана на линейной дискриминантной (классифицирующей) функции. Когда есть два класса объектов, задаваемых факторами x_1, x_2, ..., x_k, линейная дискриминантная функция — это некая линейная функция

$$\hat{y} = A\,(b_1 x_1 + b_2 x_2 + \ldots + b_k x_k) + B.$$

Коэффициенты b_i при x выбираются таким образом, чтобы разделение наблюдаемых выборочных значений из двух классов объектов было настолько глубоким, насколько это возможно, принимая во внимание внутреннюю изменчивость двух классов. А числа A и B — просто масштабные коэффициенты, выбираемые для удобства исследователя или вычислений.

Когда это удобно, как в данном случае, можно положить, что у каждой статьи Гамильтона значение y равно 1, а Мэдисона — соответственно нулю. Искомая дискриминантная функция подбирается как уравнение множественной регрессии со значениями отклика (который надо предсказывать) 0 и 1, в зависимости от того, кто автор — Мэдисон или Гамильтон. В соответствии с этим свободные коэффициен-

*Частотный анализ в лингвистике получил широкое распространение. Его основоположником можно считать Эдгара А. По, автора знаменитого рассказа «Золотой жук». За дальнейшей информацией читатель может обратиться, скажем, к работам: Арапов М. В., Херц М. М. Математические методы в исторической лингвистике. М., Наука, 1974; Вероятностное прогнозирование в речи. М., Наука, 1974. — *Примеч. ред.*

ты A и B автоматически принимают такие значения, чтобы прогноз отклика $\hat{y}$ по дискриминантной функции в среднем давал 1 для статей Гамильтона в том подмножестве, для которого эта функция построена, тогда как среднее $\hat{y}$ у Мэдисона должно быть 0. Положим, хоть и произвольно, но вполне естественно, что дискриминантный балл выше 0,5 указывает на Гамильтона, а ниже 0,5, — понятно, на Мэдисона. Это решение облегчит нам жизнь и вовсе не обязательно находить оптимальную точку деления шкалы — мы можем обойтись и ослабленным разделением.

На илл. 8.4.1 представлены частоты на 1000 слов для следующих 5 наиболее частых слов: *and* (и) — (x_1), *in* (в) — (x_2), *of* (указатель принадлежности) — (x_3), *the* (определенный артикль) — (x_4) и *to* (к) — (x_5) во всех 22 статьях. Статьи пронумерованы так же, как и в работах Мостеллера и Уоллеса [1964, p. 12—14, 269—270]. На этой иллюстрации представлены еще суммы квадратов отклонений и суммы их попарных произведений для обоих авторов (отклонения взяты относительно средних). Объединенные суммы квадратов и парных произведений использованы для получения коэффициентов дискриминантной функции $D_{\text{общ}}$. Суммы столбцов x_i для Гамильтона вычитаются из соответствующих сумм для Мэдисона и эти средние разности тоже представлены на илл. 8.4.1. Давайте сначала изучим изменчивость дискриминантной функции в терминах изменчивостей ее коэффициентов. В первой части илл. 8.4.2 приведены коэффициенты при 5 факторах и постоянный член, подсчитанные для полной выборки из 22 статей. А во второй части — при последовательном отбрасывании каждой группы (в данном случае — каждой пары) статей и счете по оставшимся 20 публикациям. Здесь полная дискриминантная функция строится методом «складного ножа» от столбца к столбцу, т. е. счет коэффициентов идет так:

11 (общий коэффициент) — 10 (коэффициент без j-й пары).

Например, псевдокоэффициент при x_3, когда отброшена четвертая пара, получается из выражения (где оставлено больше знаков, как и в фактических вычислениях, чем в илл. 8.4.2):

$$11\,(0{,}0526442) - 10\,(0{,}0563169) = 0{,}015917.$$

Этот результат можно найти во второй части илл. 8.4.2 на четвертом месте в третьем столбце. Результат усечен на 5-м знаке. Эти 11 новых дискриминантных функций вместе с 12-й, полученной их усреднением, и есть псевдодискриминанты и дискриминант «складного ножа» соответственно. Отметим, что обобщенные функции получаются методом «складного ножа» либо для самих значений функций, либо для коэффициентов, как угодно, поскольку связь между ними линейна.

Давайте рассмотрим дискриминантную функцию «складного ножа». $D*$ — стандартные ошибки ее коэффициентов, найденные по формуле параграфа 8.1:

$D^* = -3{,}0141 - 0{,}0195x_1 + 0{,}0301x_2 + 0{,}0547x_3 - 0{,}0167x_4 + 0{,}0420x_5$										
Стандартные ошибки («складной нож» для s_{b_i})	0,0193		0,0149		0,0105		0,00645		0,0181	
Критическое отношение $\lvert b_i \rvert / s_{b_i}$	1,0		2,0		5,2		2,6		2,3	

Эти результаты показывают, что только третий коэффициент в самом деле отличается от 0; он связан с частотой употребления слова «*of*», однако здесь мы не будем далее развивать этот тезис.

Если подставить в дискриминантные функции $D_{\text{общ}}$ и D^* данные для всех 22 статей Гамильтона и Мэдисона, то получается результаты, представленные на илл. 8.4.3, где они произвольно ограничены тремя знаками. Прежде всего бросается в глаза, что результаты для $D_{\text{общ}}$ и D^* очень близки. Кроме того, если принять значение 0,5 как границу, т. е. если относить значения дискриминантной функции, превышающие это число, к статьям Гамильтона, а меньшие — Мэдисона, то видно, что и $D_{\text{общ}}$, и D^* правильно классифицируют все статьи.

Удивительно ли это? Нисколько. Мы ведь использовали для классификации те же самые работы, по которым велась дискриминация. Нужна перепроверка.

8.5. ПЕРЕПРОВЕРКА ПРИМЕРА

Пока что «складной нож» дал нам четкую оценку изменчивости коэффициентов. Давайте теперь попробуем получить столь же качественную оценку способности к окончательной дискриминации текстов Гамильтона и Мэдисона. Это новая и важная задача.

Если бы дискриминация удалась, то даже без всякой оценки устойчивости перепроверка становится необходимой. Конечно, имея 11 пар и 5 факторов, не разделить данные пополам. Но, как указано в параграфе 2.6, мы могли бы по очереди выкидывать каждую пару и проводить дискриминацию на оставшихся десяти. А чтобы это сделать, как раз и нужны дискриминанты, уже вычисленные раньше в другой связи. Подставляя в каждую $D_{(i)}$ частоты для соответствующих статей наших авторов H_i и M_i, мы найдем те значения, которые стоят в левой части илл. 8.5.1. Теперь появилась одна ошибка в классификации: статья M_3 со значением 0,510 явно относится к Гамильтону.

Фактически поведение этих дискриминантов, основанных на десяти наблюдениях над пятью факторами, неправдоподобно хорошо. Индикатор качества разделения подгрупп статей Гамильтона и Мэдисона, задаваемый разностью средних, равен 0,975—0,015 = 0,960. Это из ряда вон выходящее значение, однако мы пока что не знаем его

стабильности, оцениваемой уже знакомым путем. (Если разделить 22 статьи на две равные группы *случайным образом*, то среднее индикатора качества разделения в таком случае должно быть равным нулю, ибо знаки разностей будут чередоваться случайно.)

Разброс внутри каждой группы относительно ее выборочного среднего отнюдь не мал; выборочное стандартное отклонение колеблется от 0,28 до 0,31 для обоих авторов. Между тем нам не только не хватает индикации стабильности результата, но у нас вообще нет убежденности в том, что результат заслуживает доверия. Например, все данные, входящие в величину D_3 для H_3, входят и в величину D_7 для H_7. Значит, мы не застрахованы от каких-нибудь корреляций между значениями, которые сделают средние квадраты разностей совсем не такими, как суммы соответствующих дисперсий.

Взяв среднеквадратичные отклонения для статей Гамильтона относительно 1, а для Мэдисона — относительно 0 (вместо того, чтобы считать их относительно наблюденных средних), мы получим индикации, указывающие цену комбинированной оценки разброса и сжатия, но не устраняющие наше затруднение, поскольку их образуют *суммы* членов, в каждой из которых участвуют лишь единичные значения. (В данном примере значения почти те же самые.) Эти меры, впрочем, вполне законны, но все-таки они ничего не говорят о стабильности.

Отбросив только одну пару, мы окажемся перед выбором:

1) оценивать ли стабильность дискриминантной функции в целом (не оценивая качества ее работы), как это делалось в параграфе 8.4;

2) либо оценивать качество перепроверки (не оценивая стабильность самой оценки), как это сделано в данном параграфе.

8.6. ДВА АНАЛОГИЧНЫХ ИСПОЛЬЗОВАНИЯ ПРИНЦИПА «ОТБРАСЫВАЙ ПО ОДНОМУ»

Если надо оценить и качество работы, и стабильность, то приходится комбинировать перепроверку и «складной нож». А это значит пропускать по одной паре для каждой цели. Отсюда нам нужно искать дискриминантные функции, основанные на множествах в $9 = 11 - 2$ пар статей. Обозначим, произвольно, дискриминант, полученный без пары i и пары j, через

$$D_{(i)\,(j)} = D_{(i)\,(j)} = D_{(j)\,(i)}.$$

То, к чему мы теперь приступаем, с обеих точек зрения (и «складного ножа», и перепроверки) сводится к исключению одной пары и повторению всего анализа заново, но теперь уже с 10 (вместо 11) парами статей.

Когда исключается для перепроверки i-я пара, «складной нож» приводит к псевдодискриминантам

$$D*_j(i) = 10D_{\text{общ}}(i) - 9D_{(j)}(i) = 10D_{(i)} - 9D_{(i)(j)}.$$

Для каждого i получается десять $D*_j(i)$ и одно $D*_i$, которое есть среднее $D*_j(i)$. Причем в образовании одной из этих дискриминантных функций участвуют статьи H_i и M_i. Значит, когда мы пользуемся эти-

ми дискриминантами для H_i и M_i, получаются $2 \cdot (10 + 1) = 22$ отличные перепроверки.

Все результаты таких перепроверок сведены на илл. 8.6.1 с тремя десятичными знаками. (Для большинства целей двух знаков должно хватить.) Начнем с первой группы из илл. 8.6.1. Первое число 1,250 получается, если применить $D*_2(1)$ к H_1. Отметим, что по текстам Гамильтона все результаты для $D*_{11}(1)$, вплоть до десятого 0,535, превышают 0,5. Вместе с тем среди значений по текстам Мэдисона отмечено одно, тоже превышающее 0,5, что свидетельствует об ошибке псевдодискриминантов. Следовательно, для первой пары статей ошибка частоты для псевдодискриминантов равна 5%. А для всех пар эта ошибка составляет около 16%.

Для каждой статьи мы можем теперь узнать, как выносятся решения в пользу Гамильтона или Мэдисона. Давайте представим себе бесконечное множество статей этих авторов, причем все они отличаются от i-й пары, на которой могла быть основана дискриминантная функция $D*(i)$, и рассмотрим те 10 статей, которые фактически использовались, как выборка из этой бесконечной совокупности. В результате применения генерального значения $D*(i)$ к H_i и M_i должны были бы получиться μ_{Hi} и μ_{Mi} — истинные значения для этих статей. А существуют значения, для получения которых и придуман «складной нож». Такие значения, обозначим их $\bar{y}_{Hi}$ и $\bar{y}_{Mi}$ при i-й паре, вместе с их изменчивостью, оцениваемой из индивидуальных псевдозначений по s^2*, могут служить для построения доверительных границ μ_{Hi} и μ_{Mi} при любом уровне значимости. И если только эти границы не перекроют числа 0,5, мы сможем ясно судить (при заданном уровне, конечно) о принадлежности статей Гамильтону или Мэдисону. Понятно, мы хотим теперь знать, сколько стандартных ошибок каждого $\bar{y}_{Hi}$ или $\bar{y}_{Mi}$ не включают в себя числа 0,5.

На илл. 8.6.2 обобщены результаты для всех 22 статей и, кроме того, указаны значения s^2*.

Заметим, что по 8 статей обоих авторов отстоят от 0,5 на 2 стандартные ошибки, а одна у Мэдисона — на 1,7, правда, остальные 5 имеют лишь относительно узкие пределы.

Теперь мы, наконец, получили оценки перепроверки вместе с оценками их стабильности. И результаты, хотя и гораздо более полезные, чем прежде, отнюдь не блестящи. Оценив стабильность, мы поднялись на вторую ступеньку лестницы, но, поскольку у нас нет оценки стабильности этой стабильности, мы еще не взошли на третью. Учитывая столь малое число рассмотренных статей, не стоит удивляться и тому, что мы не смогли добыть сколько-нибудь приемлемой информации о форме распределения μ_{Hi} (или μ_{Mi}).

8.7. РАССЕЯНИЕ СРЕДНИХ μ

Но мы можем научиться еще кое-чему. Если бы у нас было бесконечно много статей для построения дискриминантов, то потеря одной пары не оказала бы никакого влияния ни на дискриминанты, ни на их коэффициенты. Следовательно, должно быть значение 1 для μ_{Hi}

и 0 для μ_{Mi}, так как эти ограничения существовали до потери одной пары. Тогда на основе величины типа

$$\Sigma(\bar{y}_{Hi}-1)^2=1{,}011$$

мы могли бы построить приближенные оценки компонент дисперсии стандартного отклонения для распределения (относительно i) величины μ_{Hi}; причем это значение соответствовало бы i-й статье Гамильтона, если бы выборка для дискриминантной функции была бесконечной.

Усредненное по статьям и дискриминантам выражение $(\bar{y}_{Hi}-1)^2$ будет равно $\sigma_\mu^2+\sigma^2$, где σ_μ^2 есть дисперсия μ_{Hi}, а σ^2 — изменчивость, связанная с любой дискриминантной функцией, полученной лишь по конечной выборке пар статей. Мы оцениваем σ^2 через s^2*. А генеральное среднее значение для $(\bar{y}_{Hi}-1)^2$ оцениваем непосредственно, объединяя результаты для обоих авторов, чтобы получить затем σ_μ^2 вычитанием.

Отыщем среднюю суммы квадратов отклонений (относительно 1 или 0 в зависимости от того, Гамильтон это или Мэдисон), равную $^1/_2(1{,}011+0{,}865)=0{,}938$. Чтобы получить оценку генерального среднего для $(\bar{y}_{Hi}-1)^2$, разделим предыдущий результат на 11 (а вовсе не на 10, что было бы уместно, если бы мы оценивали среднее $\bar{y}_{Hi}$ по тем же данным) и найдем 0,0853, подтверждая тем самым возможность смещенного примерно на 0,30 предсказания для стандартных отклонений (в параграфе 8.5).

Для 22 значений s^2* мы имеем сумму 0,8364 и среднее 0,0380, которое служит оценкой для σ^2. Вычитание дает 0,0853—0,0380 = 0,0473 в качестве оценки для σ_μ^2. Отсюда оценка σ_μ будет около 0,22. Как для Мэдисона, так и для Гамильтона оценки для μ будут чуть более чем на два стандартных отклонения отстоять от границы 0,5. Значит, согласно гауссовской теории, да и для многих других распределений около 2,5% статей каждого автора могут классифицироваться ошибочно, если, конечно, брать только 5 слов, но зато при бесконечном объеме выборки для дискриминантной функции. (Наша оценка σ_μ далеко не идеальна, в первую очередь, может быть, из-за числа степеней свободы.)

Теперь интересно сравнить средние разности и стандартные отклонения для разных вариантов дискриминантов в подгруппах статей Гамильтона и Мэдисона. Вот эти оценки:

	Разделение групповых средних	Стандартное отклонение распределения	Отношение
Дискриминация по 10 статьям	0,96	0,29	3,3
Дискриминация по ∞ статьям	1,00	0,22	4,5

И снова мы имеем здесь дело с индикациями, стабильность которых неизвестна. Улучшение за счет дискриминации по гораздо большему

числу пар рассматриваемых статей будет значительно меньшим, чем можно допустить, хотя прирост сам по себе и значителен.

Взятые порознь и вместе эти результаты рисуют картину сильных и слабых сторон дискриминантной функции на практике. Чтобы практически воспользоваться ею, нужны ЭВМ и повторное применение принципов «отбрасывай по одному» для «складного ножа» и для перепроверки.

Давайте опять подчеркнем, что главный результат здесь — это беглая демонстрация «складного ножа» в сложной задаче с умеренным объемом данных, а вовсе не исчерпывающий анализ всей проблемы «Federalist» с ее узловыми вопросами о выборе слов.

8.8. ДАЛЬНЕЙШЕЕ ОБСУЖДЕНИЕ ПРИМЕРА

Может быть, полезно исследовать вопрос об устойчивости оценки σ_μ. Правда, при попытке оценить ее устойчивость мы столкнемся с огромным объемом счета. Снова потребуется «складной нож». Теперь мы должны будем отбросить еще одну пару, сосчитать $(11 \cdot 10 \times \times 9)/6 = 165$ дискриминантных функций — по одной на каждое множество из 8 пар статей — и пересчитать 11 новых оценок «складного ножа» для σ_μ. Объединяя их в новую итоговую оценку «складного ножа» вместе с уже существующей, мы и получим то, что ищем: оценку σ_μ *и* оценку стабильности этой оценки. Какой другой метод позволяет хотя бы грубо приближенно оценить устойчивость оценки σ_μ?

В качестве следующего шага прислушаемся к человеку, который читает фразу (в конце параграфа 8.4), намекающую на то, что всю работу делает частота одного лишь слова «*of*», и спрашивает: «А каковы, собственно, достоинства и недостатки пятифакторной дискриминантной функции против прямого использования частоты слова «*of*»?»

Чтобы ответить на подобный вопрос, требуется уже не обычная, а двойная перепроверка (см. параграф 2.6). Нам следовало бы изучить ответ на этот вопрос на иных данных, чем те, которые его породили.

Но если это почему-либо невозможно и нам приходится снова и снова обращаться все к тем же 22 статьям, то нетрудно сделать обычную перепроверку с оценкой устойчивости. Для этого надо только вернуться к каждому набору по 9 пар и найти 2 дискриминантные функции, одну для всех 5 слов (уже готовую) и одну только для слова «*of*» (что совсем просто). Переходя от пары к паре, мы сможем в конце концов для каждой статьи подсчитать:

1) результаты первой дискриминации;
2) результаты второй дискриминации;
3) разность между ними как индикатор того, какая лучше.

Тогда можно будет применить «складной нож» к величинам такого типа, как в (3), найденным для каждой статьи, для оценки улучшения и для оценки устойчивости этой оценки. Если учесть, что наш исходный вопрос фактически в том, как сравнить два способа дискриминации при условии, что они оба опираются на очень большую совокупность статей, то ответ получится подходящим и полезным.

Дальнейшие подробности о «складном ноже» можно найти в работах: [Quenouille M. H. (1956)], [Tukey J. W. (1958)], [Durbin J. (1959)], [Mickey M. R. (1959)], [Kendall M. G. and Stuart A. (1961)], [Brillinger D. R. (1964)], [Miller R. G. (1964)], [Robson D. S. and Whitlock J. H. (1964)], [Jones H. L. (1965)], [Gray H. L. and Adams J. E. (1972)], [Gray H. L., Watkins T. A. and Adams J. E. (1972)], [Egman R. K., Meyers C. E. and Bendel R. (1973)], [Frawley W. H. (1974)], [Jones H. L. (1974)], [Miller R. G. (1974a)], [Miller R. G. (1974b)], обзор с библиографией 45 названий, [Wainer H. and Thissen D. (1975)]. Есть уже и обобщения «складного ножа» (см. [Gray H. L. and Shucany W. R. (1972)]).

РЕЗЮМЕ. «СКЛАДНОЙ НОЖ»

Метод «складного ножа» довольно хорошо приспособлен для решения многих задач, но следует помнить, что какой-нибудь частный метод может сработать лучше в том или ином конкретном случае анализа.

Мы пользуемся этим методом и непосредственно для оценивания, и при оценке дисперсии.

Мы поняли, на какие части естественно и разумно делить данные.

Мы умеем вычислять псевдозначения (т. е. значения «складного ножа») и s^2* и идем дальше до (доверительных) границ соответствующих индикаций.

Вот главные трудности работы с элементарным методом «складного ножа»: (1) наличие у распределений оцениваемых статистик сильно разбросанных «хвостов» и (или) (2) наличие слишком малого числа различных псевдозначений. (Иногда методы гл. 10 позволяют бороться с трудностью (1). Изменением объема выборки или ее структуры часто преодолевают трудность (2).)

Там, где «складной нож» используется для сравнения различных значений, целиком полученных параллельными путями, часто можно считать, что метод применен ко всей функции или ко всей ситуации целиком (а не что имеет место класс «равноправных» наборов чисел).

Желательно *избегать* (в «складном ноже») выборочных распределений с (1) обрубленными краями и (2) с одним или несколькими разбросанными «хвостами». Было бы также желательно избегать и резкой асимметрии (часто преодолеваемой преобразованиями).

Мы можем использовать наш метод для получения оценок неопределенностей, связанных с работой дискриминантной функции, а также и ее коэффициентов, в частности, комбинируя принцип «отбрасывай по одному» для «складного ножа» и для перепроверки. (То же самое можно сделать для многих методов статистического анализа, а не только для дискриминантных функций.)

Приступая к третьему этапу «отбрасывания по одному», стоит добавить еще одну процедуру «складного ножа» к двум уже упомянутым, а именно оценку неопределенности нашей оценки качества работы дискриминантной функции. (Это общие требования метода «отбрасывания по одному».)

Мы показали технику «складного ножа» на нескольких небольших и одном сложном большом примерах.

БИБЛИОГРАФИЯ

Brillinger D. R. (1964). The asymptotic behavior of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates. — Rev. Int. Statist. Inst., 32, 202—206.

Cochran W. G. (1953). Sampling techniques. New York, Wiley (2nd ed., 1963). Русский перевод со второго издания: Кокрен У. *Методы выборочного исследования*. М., *Статистика*, 1976.

Durbin J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. — Biometrika, 46, 477—480.

Egman R. K., Meyers C. E. and Bendel R. (1973). New methods for test selection and reliability assessment using stepwise multiple regression and jackknifing. — Educ. Psychol. Meas., 33, 883—894.

Frawley W. H. (1974). 364: Using the jackknife in testing dose responses in proportions near zero or one — revisited. — Biometrics, 30, 539—545.

Gray H. L. and Adams J. E. (1972). Jackknifing stochastic processes. — Texas J. Sci., 23, 559.

Gray H. L. and Shucany W. R. (1972). The generalized jackknife statistic. New York, Marcel Dekker, Inc.

Gray H. L., Watkins T. A. and Adams J. E. (1972). Jackknife statistic, its extensions, and its relation to e_n-transformations. — Ann. Math. Stat., 43, 1—30.

Jones H. L. (1965). The jackknife method. B: Proc. IBM Scientific Computing Symposium on Statistics, October 21—23, 1963. White Plains, New York, IBM Data Processing Division, 185—201.

Jones H. L. (1974). Jackknife estimation of functions of stratum means. — Biometrika, 61, 343—348.

Kendall M. G. and Stuart A. (1961). The advanced theory of statistics. Vol. 2, Inference and relationship. London, Charles Griffin & Company, 5—7. Русский перевод: Кендалл М., Стьюарт А. Статистические выводы и связи. М., Наука, 1973.

Mickey M. R. (1959). Some finite population unbiased ratio and regression estimators. — J. Amer. Statist. Assoc., 54, 594—612.

Miller R. G., Jr. (1964). A trustworthy jasckknife. — Ann. math. Statist., 35, 1594—1605.

Miller R. G., Jr. (1974a). A unbalanced jackknife. — Ann. of Stat., 2, 880—891.

Miller R. G., Jr. (1974b). The jackknife — a review. — Biometrika, 61, 1—15 (45 references).

Mosteller F. and Wallace D. L. (1963). Inference in an authorship problem. — J. Amer. Statist. Assoc., 58, 275—309.

Mosteller F. and Wallace D. L. (1964). Inference and disputed authorship: The Federalist. Reading, Mass., Addison-Wesley.

Quenouille M. H. (1956). Notes on bias in estimation. — Biometrika, 43, 353—360.

Robson D. S. and Whitlock J. H. (1964). Estimation of a truncation point. — Biometrika, 51, 33—39.

Tukey J. W. (1958). Bias and confidence in not-quite large samples. Abstract в: Ann. math. Statist., 29, 614.

Tukey J. W. (unpublished). Data analysis and behavioral science. Princeton University and Bell Telephone Laboratories.

Wainer H. and Thissen D. (1975). When jackknifing fails (or does it?). — Psychometrika, 40, 113—114.

ИЛЛЮСТРАЦИИ

Иллюстрация 8.2.1

«Складной нож» для выборочного стандартного отклонения ($y_{общ}$ = выборочному стандартному отклонению = 1,34347)

j	x_j	Стандартное отклонение без $x_j = y_{(j)}$	$y^*_j = 11\, y_{общ} - 10\, y_{(j)}$
1	0,1	*1,36382*	1,1400
2	0,1	*1,36382*	1,1400
3	0,1	*1,36382*	1,1400
4	0,4	*1,38888*	0,8894
5	0,5	*1,39539*	0,8243
6	1,0	*1,41457*	0,6325
7	1,1	*1,41578*	0,6204
8	1,3	*1,41563*	0,6219
9	1,9	*1,39427*	0,8355
10	1,9	*1,39427*	0,8355
11	4,7	*0,70742*	7,7040
	$\Sigma = 13,1$		$y^* = 1,4894$

$s^* = 0,6244$.

Двусторонние доверительные границы для σ:

2/3: $1,4894 \pm |t_{10}|_{2/3}\, s^* = 1,4894 \pm 1,02$ (0,6244), или от 0,85 до 2,13.

95%: $1,4894 \pm |t_{10}|_{0,95} s^* = 1,4894 \pm 2,23$ (0,6244), или от 0,10 до 2,88.

Иллюстрация 8.2.2

«Складной нож» для $\log_{10} s$ данных из илл. 8.2.1 ($Y_{общ} = \log_{10} y_{общ} = 0,12823$)

j	$Y_{(j)} = \log_{10} y_j$	$Y^*_j = 11 \log_{10} y_{общ} - 10 \log_{10} y_{(j)}$
1	*0,13476*	0,06293
2	*0,13476*	0,06293
3	*0,13476*	0,06293
4	*0,14266*	—0,01607
5	*0,14470*	—0,03647
6	*0,15062*	—0,09567
7	*0,15100*	—0,09947
8	*0,15095*	—0,09897
9	*0,14435*	—0,03297
10	*0,14435*	—0,03297
11	*—0,15032*	2,91373
		$Y^* = 0,24454 =$ Среднее

$s^{*2} = 0,071605$; $s^* = 0,2676$

2/3-границы для $\log_{10}$ σ: 0,2445 ± 1,02 (0,2676), или интервал от —0,028 до 0,517.

95%-ные границы для $\log_{10}$ σ: 0,2445 ± 2,23 (0,2676), или интервал от —0,352 до 0,841.

2/3-доверительный интервал для σ: от 0,94 до 3,29.

95%-ный доверительный интервал для σ: от 0,44 до 6,93.

Иллюстрация 8.2.3
Измерения 11 объектов

Объекты

1	2	3	4	5	6	7	8	9	10	11
6,880	4,660	6,950	4,756	4,411	4,257	2,642	12,541	8,404	3,262	3,286
5,172	4,522	3,948	3,792	4,357	3,572	2,276	4,081	5,137	2,874	2,858
3,598	3,403	3,062	2,458	3,571	1,809	2,007	3,853	3,172	2,120	2,787
3,034	3,211	2,906	0,412	2,983	1,801	1,922	0,364	1,432	1,456	2,752
0,628	0,070	0,482	—0,458	1,825	1,480	1,588	—2,945	—1,415	**0,780**	2,047

Иллюстрация 8.2.4
Восемь наибольших значений, упорядоченных и соотнесенных со своими объектами

Ранг	Величина	Объект	Ранг	Величина	Объект
1-й	12,541	8	5-й	5,172	1
2-й	8,404	9	6-й	5,137	9
3-й	6,950	3	7-й	4,756	4
4-й	6,880	1	8-й	4,660	2

Иллюстрация 8.2.5.

Значения, полученные методом «складного ножа»

$y_{общ}$	5,151	Псевдозначения
$y_{(1)}$	4,746	$y_{*1}=9,201$ (=11(5,151)—10(4,746))
$y_{(2)}$	5,168	$y_{*2}=4,981$
$y_{(3)}$	5,099	$y_{*3}=5,671$
$y_{(4)}$	5,168	$y_{*4}=4,981$
$y_{(5)}$	5,168	$y_{*5}=4,981$
$y_{(6)}$	5,168	$y_{*6}=4,981$
$y_{(7)}$	5,168	$y_{*7}=4,981$
$y_{(8)}$	5,099	$y_{*8}=5,671$
$y_{(9)}$	4,746	$y_{*9}=9,201$
$y_{(10)}$	5,168	$y_{*10}=4,981$
$y_{(11)}$	5,168	$y_{*11}=4,981$
Итог	55,866	Итог = 64,611
тог/11	5,0787	Итог/11 = 5,874 = y_*

Проверка: $y_* = 11\ (5,151) - 10\ (5,0787) = 5,874.$
$s^2 = 2,78024.$ $s_*^2 = 2,78024/11 = 0,25275.$ $s_* = 0,503.$
$y_* \pm |t_2|_{0,96} s_* = 5,874 \pm 4,30\ (0,503)$, или 95%-ный доверительный интервал от 3,71 до 8,04.

Иллюстрация 8.3.1

Совокупность 49 больших городов в 1920 (x) и 1930 (y) годах (тыс. жителей)

1-я 7		2-я 7		3-я 7		4-я 7		5-я 7		6-я 7		7-я 7		Сводка	
x	y	x	y	x	y	x	y	x	y	x	y	x	y	x	y
76	80	120	115	60	57	44	58	38	52	71	79	36	46	(751)	(915)
138	143	61	69	46	65	77	89	136	139	256	288	161	232	(977)	(1122)
67	67	387	459	2	50	64	63	116	130	43	61	74	93	(965)	(1243)
29	50	93	104	507	634	64	77	46	53	25	57	45	53	(385)	(553)
381	464	172	183	179	260	56	142	243	291	94	85	36	54	(696)	(881)
23	48	78	106	121	113	40	60	87	105	43	50	50	58	(830)	(937)
37	63	66	86	50	64	40	64	30	111	298	317	48	75	(450)	(611)
Итог751	915	977	1122	965	1243	385	553	696	881	830	937	450	611	5054	6262

Иллюстрация 8.3.2

Подробности применения «складного ножа» для оценки доли (прироста) по данным илл. 8.3.1

	Все	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$
$x_{(i)}$ (выборка = 1920)	5054	4303	4077	4089	4669	4358	4224	4604
$\log x_{(i)}$	3,70364	3,63377	3,61034	3,61162	3,66922	3,63929	3,62572	3,66314
$y_{(i)}$ (выборка = 1930)	6262	5347	5140	5019	5709	5381	5325	5651
$\log y_{(i)}$	3,79671	3,72811	3,71096	3,70062	3,75656	3,73086	3,72632	3,75213
$z_{(i)} = \log [y_{(i)}/x_{(i)}]$	*0,09307*	*0,09434*	*0,10062*	*0,08900*	*0,08734*	*0,09157*	*0,10060*	*0,08899*
$z^*_i = 7\, z_{\text{общ}} - 6\, z_{(i)}$	—	0,08545	0,04777	0,11749	0,12745	0,10207	0,04789	0,11755
Округление z^*_i	—	0,085	0,048	0,117	0,127	0,102	0,048	0,118
1000 (z^*_i округл.—0,100)	—	—15	—52	17	27	2	—52	18

Сумма = —55; —55/7 = —7,9 = среднее = z^* (в условных единицах).
Сумма квадратов = 6979; 6979 — (—55)²/7 = 6547 = сумма квадратов отклонений.

$$\frac{6547}{6 \cdot 7} \approx 156 = s^2{}^*.$$

$\sqrt{156} \approx 12{,}5 = s^*$ (в условных единицах).
$|t_6|_{0,95} = 2{,}447$, (12,5) (2,447) = 30,6 = допуск (в условных единицах).

Иллюстрация 8.3.3

Результаты счета оценки доли (основные данные: итог 1920 г. = 22,919; log (итог 1920 г.) = 4,360; log итога = log (итог 1920 г.) + log доли)

Рассматриваемые величины		Найденные результаты	
название	обозначение	оценки	95 %-ные доверительные интервалы
Условные единицы	1000 ($z*$—0,100)	≈—7,9	от —38,5 до 22,7
log доли	$z*$	≈0,092	от 0,062 до 0,123
log итога	log (итог 1920) + $z*$	≈4,452	от 4,422 до 4,483
итог	antilog (log итога)	≈28300	от 26000 до 30400

Иллюстрация 8.4.1

Частоты пяти слов (на тысячу) в каждой из статей Гамильтона и Мэдисона (суммы квадратов и парных произведений для каждого автора и в целом)

А. Статьи Гамильтона

Номер	Группа j	Слова, i					Итог
		and 1	*in* 2	*of* 3	*the* 4	*to* 5	
60	1	16,1	35,3	63,9	98,3	38,4	252,0
69	2	32,2	24,5	78,2	110,0	31,4	276,3
36	3	24,3	23,5	64,7	90,8	42,3	245,6
73	4	18,0	27,2	59,6	86,8	35,9	227,5
26	5	20,6	26,9	61,4	83,6	39,5	232,0
7	6	21,8	17,4	73,1	90,4	35,6	238,3
112	7	27,9	23,1	61,9	85,4	41,3	239,6
11	8	28,5	26,1	71,3	74,5	33,3	233,7
35	9	28,9	20,9	56,9	82,7	44,9	234,3
34	10	21,3	25,0	60,4	82,2	47,7	236,6
66	11	18,5	30,7	72,7	109,3	36,6	267,8
Итоги		258,1	280,6	724,1	994,0	426,9	

Б. Статьи Мэдисона

Номер	Группа j	Слова, i					Итог
		and 1	*in* 2	*of* 3	*the* 4	*to* 5	
40	1	31,6	19,9	54,8	93,8	38,6	238,7
37	2	37,3	23,3	56,8	84,2	31,0	232,6
133	3	21,2	17,5	58,2	97,6	39,9	234,4
14	4	27,9	19,1	55,8	93,1	33,5	229,4
122	5	40,7	9,3	59,0	71,5	33,6	214,1
39	6	24,4	27,9	60,0	115,3	34,8	262,4
46	7	27,7	17,7	61,1	115,3	32,7	254,5
44	8	28,1	22,3	57,0	110,9	29,7	248,0
47	9	30,6	23,6	68,3	118,6	23,2	264,3

Продолжение

Номер	Группа *j*	Слова, *i*					Итог
		and 1	*in* 2	*of* 3	*the* 4	*to* 5	
42	10	33,9	21,8	64,9	93,7	33,6	247,9
132	11	23,3	31,4	34,8	94,3	49,6	233,4
Итоги		326,7	233,8	630,7	1088,3	380,2	

В. Гамильтон. Суммы квадратов и парных произведений отклонений

Слова	1	2	3	4	5
1	275,985				
2	—139,756	226,069			
3	95,181	—12,473	471,102		
4	—67.655	181,644	455,111	1267,465	
5	—31,396	—34,801	—260,843	—227,696	244,069

Г. Мэдисон. Суммы квадратов и парных произведений отклонений

Слова	1	2	3	4	5
1	351,920				
2	—173,050	334,287			
3	169,590	—212,312	719,265		
4	—514,910	381,708	364,715	2146,385	
5	—173,710	109,502	—479,175	—315,635	442,865

Д. Объединенные суммы квадратов и парных произведений отклонений

Слова	1	2	3	4	5
1	627,905				
2	—312,806	560,356			
3	264,771	—224,785	1190,367		
4	—582,565	563,352	819,826	3413,850	
5	—205,106	74,701	—740,018	—543,331	686,934

Е. Сумма для Гамильтона минус сумма для Мэдисона

—68,6	46,8	93,4	—94,3	46,7

Иллюстрация 8.4.2

Исходная дискриминантная функция $D_{общ}$ и 11 дискриминантов, $D_{(j)}$, построенных при отбрасывании последовательно всех пар статей Гамильтона и Мэдисона. Из них построены 11 псевдодискриминантов d_{*j} и их среднее D_*, которые тоже представлены здесь.

А. Исходная дискриминантная функция $D_{общ}$ и 11 дискриминантов $D_{(j)}$

	Коэффициент при					
	x_1	x_2	x_3	x_4	x_5	Постоянный член
$D_{общ}$	—0,01902	0,02851	0,05264	—0,01642	0,04056	—2,83668
$D_{(1)}$	—0,01904	0,03032	0,05295	—0,01660	0,04120	—2,89257
$D_{(2)}$	—0,02747	0,02559	0,04532	—0,01964	0,03163	—1,47660
$D_{(3)}$	—0,02884	0,01467	0,04928	—0,01479	0,03962	—2,14043
$D_{(4)}$	—0,00716	0,02874	0,05631	—0,01248	0,05243	—4,21632
$D_{(5)}$	—0,01790	0,02789	0,05348	—0,01642	0,04172	—2,94733
$D_{(6)}$	—0,01695	0,03151	0,05182	—0,01455	0,04145	—3,10996
$D_{(7)}$	—0,02053	0,03063	0,05166	—0,01757	0,03681	—2,55988
$D_{(8)}$	—0,01648	0,03338	0,05660	—0,01979	0,04184	—2,99265
$D_{(9)}$	—0,02350	0,02910	0,05002	—0,01670	0,03074	—2,20700
$D_{(10)}$	—0,01093	0,03047	0,05406	—0,01559	0,04523	—3,40123
$D_{(11)}$	—0,01983	0,02953	0,05521	—0,01616	0,04188	—3,06431

Б. Псевдодискриминанты D_{*j} и их среднее D_*

	Коэффициент при					
	x_1	x_2	x_3	x_4	x_5	Постоянный член
D_{*1}	—0,01886	0,01041	0,04956	—0,01463	0,03422	—2,27783
D_{*2}	0,06548	0,05770	0,12584	0,01582	0,12988	—16,43747
D_{*3}	0,07914	0,16695	0,08621	—0,03268	0,04996	—9,79921
D_{*4}	—0,13765	0,02623	0,01591	—0,05584	—0,07806	10,95966
D_{*5}	—0,03020	0,03477	0,04425	—0,01639	0,02902	—1,73025
D_{*6}	—0,03967	—0,00149	0,06081	—0,03513	0,03166	—0,10390
D_{*7}	—0,00389	0,00736	0,06242	—0,00490	0,07810	—5,60468
D_{*8}	—0,04443	—0,02015	0,01308	0,01731	0,02776	—1,27696
D_{*9}	0,02582	0,02262	0,07885	—0,01364	0,13878	—9,13354
D_{*10}	—0,09988	0,00897	0,03843	—0,02470	—0,00606	2,80880
D_{*11}	—0,01090	0,01831	0,02697	—0,01905	0,02736	—0,56044
D_*	—0,01955	0,03015	0,05476	—0,01671	0,04205	—3,01416

Иллюстрация 8.4.3

Значения дискриминантов $D_{общ}$ и D_* для каждой из 22 статей

Пара статей	$D_{общ}$		D_*		Пара статей	$D_{общ}$		D_*	
	Н	М	Н	М		Н	М	Н	М
1	1,170	0,039	1,206	0,024	7	0,822	—0,209	0,836	—0,227
2	0,833	—0,017	0,859	—0,033	8	1,246	—0,351	1,275	—0,374
3	1,001	0,338	1,023	0,333	9	0,668	—0,157	0,673	—0,167
4	0,764	—0,055	0,777	—0,075	10	1,235	0,380	1,263	0,381
5	1,000	—0,051	1,020	—0,080	11	1,203	—0,089	1,243	—0,107
6	1,052	0,171	1,073	0,172					
					Среднее	0,999	0,000	1,023	—0,014

Иллюстрация 8.5.1

Результаты применения дискриминантных функций $D_{(i)}$ к илл. 8.4.3 с отбрасываемыми для обеспечения перепроверки парами

i	$D_{(i)}$ в применении к		Сдвиг* относительно $D_{общ}$	
	H_i	M_i	H_i	M_i
1	1,205	—0,044	+0,035	—0,005
2	0,642	—0,004	—0,191	—0,013
3	1,025	0,510**	+0,024	—0,172
4	0,592	—0,130	—0,172	+0,075
5	0,993	—0,034	—0,007	—0,017
6	1,018	0,230	—0,034	—0,059
7	0,792	—0,253	—0,030	+0,044
8	1,363	—0,438	+0,117	+0,087
9	0,567	—0,091	—0,101	—0,066
10	1,269	0,459	+0,034	—0,079
11	1,256	—0,124	+0,053	+0,035
Среднее	0,975	+0,015	—0,024	—0,015

* Знак «+» означает лучшую дискриминацию, а знак «—» — худшую.
** Ошибка классификации.

Иллюстрация 8.6.1

Результаты счета псевдодискриминантных функций $D*_j(i)$ для групп i без двух пар и средние результаты для каждой пары, получаемые из дискриминантных функций «складного ножа» $D*(i)$ для тех пар статей H_i и M_i, которые не входят в их структуру.

I	Группа 1		Группа 2		Группа 3		Группа 4	
	H	M	H	M	H	M	H	M
1	—	—	0,492	—0,124	1,705	—0,388	—0,439	—0,716
2	1,250	0,167	—	—	0,855	0,409	—0,309	—1,170
3	1,723	—0,431	2,062	1,234	—	—	0,788	—0,127
4	2,598	—0,354	—0,535	0,180	0,976	0,882	—	—
5	1,446	—0,008	0,577	—0,213	0,958	0,429	1,283	0,133
6	0,785	—0,126	0,146	—0,234	0,773	—0,036	0,925	—0,007
7	1,058	0,428	1,463	—0,044	1,358	0,635	0,435	—0,030
8	0,738	0,395	0,213	—0,465	0,852	1,077	0,937	0,500
9	1,014	0,588	1,612	0,253	1,794	0,936	0,156	—0,301
10	1,383	—0,657	—0,957	—1,049	0,486	0,499	1,143	—0,211
11	0,535	0,102	0,976	—0,174	0,789	0,447	0,559	0,076
Среднее	1,253	0,010	0,605	—0,063	1,055	0,494	0,548	—0,185
$s^2 = 10\ s^2*$	0,353	0,164	0,903	0,338	0,180	0,188	0,347	0,218

j	Группа 5		Группа 6		Группа 7		Группа 8	
	Н	М	Н	М	Н	М	Н	М
1	0,582	—0,558	1,263	—0,701	1,301	0,222	1,402	—0,656
2	0,811	—0,852	—0,703	0,062	0,886	—1,769	2,487	—1,165
3	2,329	—0,003	2,026	0,138	0,681	—0,448	2,593	0,111
4	0,823	—0,019	1,127	0,206	0,793	—0,215	1,167	—0,327
5	—	—	1,021	—0,004	0,810	—0,562	1,604	—0,742
6	1,107	0,778	—	—	0,661	—0,130	0,732	0,082
7	1,018	0,498	1,329	0,418	—	—	0,969	—0,429
8	0,544	—0,479	0,761	0,904	0,322	1,312	—	—
9	1,197	0,938	1,314	0,039	1,634	—0,324	1,598	—0,636
10	1,036	—1,343	1,119	—0,084	0,243	—0,519	0,798	—0,571
11	0,841	0,493	1,394	—0,348	0,752	—0,225	0,696	—0,108
Среднее	1,029	—0,054	1,065	0,063	0,808	—0,265	1,405	—0,444
$s_2 = 10\ s^2*$	0,253	0,558	0,492	0,141	0,170	0,578	0,469	0,158

j	Группа 9		Группа 10		Группа 11	
	Н	М	Н	М	Н	М
1	0,828	0,844	1,532	0,653	2,646	—2,106
2	0,402	—0,190	0,795	0,789	1,342	—0,467
3	0,151	0,227	1,429	0,085	1,462	0,160
4	0,680	—0,086	1,261	0,222	1,206	0,091
5	0,711	—0,791	1,385	0,303	0,904	—2,121
6	1,164	—0,157	1,548	0,671	1,464	0,372
7	0,167	0,832	0,790	0,156	1,064	1,676
8	0,567	—0,151	2,284	0,785	1,270	1,836
9	—	—	1,020	0,318	1,117	—0,993
10	0,034	—0,468	—	—	0,862	—0,991
11	1,092	—0,651	0,943	0,401	—	—
Среднее	0,580	—0,059	1,299	0,438	1,334	—0,254
$s^2 = 10\ s^2*$	0,152	0,309	0,203	0,070	0,255	1,870

Иллюстрация 8.6.2
Отклонения эмпирических средних от 0,5 в масштабе $s*$.

Группа	Н			М		
	$\bar{y}_{Hi} - 0,5$	$s*$	$(\bar{y}_{Hi} - 0,5)/s*$	$0,5 - \bar{y}_{Mi}$	$s*$	$(0,5 - \bar{y}_{Mi})/s*$
1	0,753	0,188	4,0	0,490	0,128	3,8
2	0,105	0,300	0,3	0,563	0,184	3,0
3	0,555	0,134	4,1	0,006	0,137	0,0
4	0,048	0,186	0,3	0,685	0,148	4,6
5	0,529	0,159	3,3	0,554	0,236	2,3
6	0,565	0,222	2,5	0,437	0,119	3,7
7	0,308	0,130	2,4	0,765	0,240	3,2
8	0,905	0,217	4,2	0,944	0,126	7,5
9	0,080	0,123	0,6	0,559	0,176	3,2
10	0,799	0,142	5,6	0,062	0,084	0,7
11	0,834	0,160	5,2	0,754	0,432	1,7

Глава 9 ● ТАБЛИЦЫ С ДВУМЯ И БОЛЕЕ ВХОДАМИ

Большой и важный класс методов анализа начинается с данных, имеющих два или более воздействия. Вначале мы рассмотрим случай с двумя входами и представим данные как отклики, зависящие от двух факторов. В простейшем случае эти два воздействия могут меняться независимо. Здесь они выражаются местом каждого отклика в строке и столбце, номера которых и будут «значениями» наших двух факторов. Далее мы кратко остановимся на случае трех или большего числа независимых воздействий, задаваемых уже строками, столбцами и слоями — «значениями» трех или более факторов. (Мы можем, если действовать четко, так записать данные на обычном листе бумаги, что их структура будет ясна.) Обобщение на большее число независимых воздействий совсем не сложно.

Пусть модель в двухфакторном анализе имеет вид

предсказание = константа ПЛЮС строка ПЛЮС столбец,

где «константа» (общий член) — значение, относящееся к любой ячейке; «строка» символизирует значение, зависящее только от номера строки для данной ячейки; «столбец» — значение, зависящее только от номера столбца. И, как всегда,

остаток = данные МИНУС предсказание,

где «данные» — значения отклика, стоящие в той или иной конкретной ячейке. Так что можно записать окончательно

данные = константа ПЛЮС строка ПЛЮС столбец ПЛЮС остаток.

Позже мы сделаем еще один шажок, но все же за информацией о других моделях отсылаем читателя к *EDA*, гл. 11 и 12, и к работе [McNeil D. R., Tukey J. W. (1975)].

9.1. АДДИТИВНЫЙ АНАЛИЗ

Начнем с одновходового примера. Среднемесячная температура по Фаренгейту в Вашингтоне (округ Колумбия) от января до июля (Climatography of the States. Maryland, p. 9.) такова:

январь	февраль	март	апрель	май	июнь	июль
36,2	37,1	45,3	54,4	64,7	73,4	77,3

Медиана равна 54,4, среднее — 388,4/7 = 55,5. Теперь, собственно, можно записать данные вместе с «итогом» так:

36,2; 37,1; 45,3; 54,4; 64,7; 73,4; 77,3 | 54,4

либо так:

36,2; 37,1; 45,3; 54,4; 64,7; 73,4; 77,3 | 55,5

по схеме

значения данных | «итог»

Вертикальная черта служит для отделения итога от значений, к нему приводящих.

Следующий шаг состоит в использовании итога в качестве предсказания и замены данных остатками, вычитая из данных предсказания. Преобразовав данные, мы получаем и новые соотношения между двумя входами, поэтому будем теперь отделять итог двойной вертикальной чертой. Для наших медианы и среднего получим

—18,2; —17,3; —9,1; 0; 10,3; 19,0; 22,9 || 54,4

или

—19,3; —18,4; —10,2; —1,1; 9,2; 17,9; 21,8 || 55,5

в зависимости от того, что мы выберем: медиану или среднее.

Мы можем, если захотим, выбрать любое другое число, которое в том или ином смысле обобщает данные. Так, мы можем выбрать 55 или 50 (как круглое число) и получить

—18,8; —17,9; —9,7; —0,6; 9,7; 18,4; 22,3 || 55

или

—13,8; —12,9; —4,7; 4,4; 14,7; 23,4; 27,3 || 50

соответственно.

В каждом представлении мы раскладываем значения данных в сумму двух вкладов

данные = предсказание ПЛЮС остаток

или, как мы будем записывать в данном самом простом случае,

данные = константа ПЛЮС остаток.

Теперь мы хотим сделать то же самое, только представляя данные суммой не двух, а большего числа вкладов. Так, когда есть два воздействия, два фактора, чьи значения или наименования соответствуют числам, надо анализировать все отклики. Мы вернемся к этому в примере.

Двуфакторные описания. На илл. 9.1.1 (А и Б) приведены данные для трех городов на Восточном побережье о среднемесячных температурах с января по июль. Если теперь проделать все то, что уже было проделано выше, для одного из городов, то как раз и получится то, что показано в илл. 9.1.1А и Б. Сделав это, мы отражаем поведение данных гораздо более подробно. Так, из илл. 9.1.1А ясно, что в июле теплее, чем в январе, и в Ларедо (штат Техас) теплее, чем в Карибу (штат Мэн). Из илл. 9.1.1Б видим, что колебание температуры от января к июлю в Карибу сильнее, чем в Ларедо, а Вашингтон лежит между ними.

Мы получили эффект места для каждой строки, и теперь можно выделить «месячный» эффект в каждом столбце. На илл. 9.1.1В показаны результаты, записанные внизу, медианы значений в столбцах. На илл. 9.1.1Г эти значения вычтены из данных илл. 9.1.1В.

Теперь мы пришли к

двувходовому анализу,

где данные разбиты на члены

константа ПЛЮС строкаПЛЮС столбец ПЛЮС остаток,

представляющее каждое значение отклика как сумму четырех компонент. Для января в Карибу из анализа илл. 9.1.1Г получим

$$(54{,}4)+(-19{,}7)+(-18{,}3)+(-7{,}7)=8{,}7 \text{ (проверьте!)}.$$

Сумма дает исходное значение 8, 7, как и должно было быть. Для мая в Вашингтоне получим

$$(54{,}4)+(0{,}0)+(10{,}3)+(0{,}0)=64{,}7 \text{ (проверьте!)}$$

что снова совпадает с исходным значением 64,7.

В анализе с двумя входами возможных выборов для выделения любого эффекта строки (в данном примере — эффекта места) *и* любого эффекта столбца (здесь — эффекта месяца) гораздо больше, чем при анализе с одним входом, даже если и здесь много выборов при выделении любого значения, которое нас интересует.

Илл. 9.1.1Д демонстрирует альтернативный к предыдущему анализ тех же данных, выделяющий эффекты, округленные до 5 градусов (5° F). «Остатки» в этом случае не так малы, однако компоненты исключительно просты для анализа и суммирования. (Заметим, что для Карибу в январе получится: (55) + (— 20) + (— 20) + (— 6,3), что также приводит к 8.7, как и выше.)

Мы будем использовать структуру

остатки	эффекты строк
эффекты столбцов	общий член (константа)

в качестве стандартного и эффективного пути разложения таблиц с двумя входами. Особое внимание мы обращаем на следующие моменты:

1. Расчленение, выполненное аккуратно и количественно, дает то, что мы часто знаем качественно, из «разглядывания» таблицы.

2. Как правило, расчленение не только позволяет лучше обозреть ситуацию, но еще выпячивает остатки так, что можно «подсмотреть» в них, что бы еще стоило сделать, и, более того, можно (и обычно мы пользуемся этим) провести на такой основе дальнейший количественный анализ.

3. В нашей таблице теперь больше чисел,чем было вначале, что зачастую и происходит, особенно если первичные данные одновременно и анализируются, и сохраняются. На некотором этапе мы можем, конечно, преобразовать все остатки или некоторую их часть с помощью свертки и уменьшить объем данных, но так поступать имеет смысл лишь после тщательного изучения остатков.

4. Такой подход создает большие возможности варьирования при анализе. Стандартный вариант дает все основное, но не позволяет выбрать самый удобный вариант анализа.

Когда мы имеем дело с любым подобным разбиением данных, мы всегда можем называть части «вкладами» или более нейтрально «компонентами». Иногда, особенно при тщательном подборе модели, мы называем их «эффектами». Заметим, что, пользуясь «эффектами», мы пытаемся подняться над описанием данных и оценить настолько хорошо, насколько возможно, нечто лежащее в основе представленных нам конкретных чисел, может быть, даже глубокие причины или скорее гипотезы насчет мыслимой структурной модели, коэффициенты которой и подбираются по данным.

9.2. ЗНАКОМСТВО С ДВУВХОДОВЫМ АДДИТИВНЫМ АНАЛИЗОМ

Чтобы получить настоящее представление о двувходовом аддитивном анализе, надо ответить на следующие вопросы:

- что такое подходящая модель?
- чего желать от остатков?

Вскоре мы увидим, как ответ на первый вопрос помогает разрешить второй.

Рассмотрим альтернативный анализ, представленный на илл. 9.1.1Д, а именно

(остатки опущены)							—20 0 20
—20	—15	—10	0	10	20	25	55

Эту структуру можно записать многими способами, включая и такой:

(остатки опущены)							35 55 75
—20	—15	—10	0	10	20	25	Не используется или 0; любая интерпретация лучше

Такая запись дает предсказание для любой комбинации двух факторов (месяца и месторасположения) в виде суммы только двух компонент — одна задается местом (здесь: 35, 55 или 75), а другая — месяцем (здесь: от — 20 до + 25).

Начнем с построения графика с «месяцами» и «местами» по осям. Он представлен на илл. 9.2.1. Поскольку для обеих осей используется одна и та же шкала (в °F), мы можем смотреть на эффекты перемены места и месяца (расхождение во вкладах от места или месяца) не как на числа, а как на элементы графика.

Поэтому можно ввести значения остатков на график илл. 9.2.1 так, как это показано на илл. 9.2.2. Вы можете легко разглядеть общую структуру данных по таблице остатков. Может быть, будет лег-

че выявить ее, если для наглядности закодировать остатки. Вот вариант кодирования, который представляется наилучшим в нашей задаче, причем 5 ÷ 7 классов, как правило, вполне достаточно. Воспользовавшись методом представления остатков в виде «опоры и консоли», найдем медиану $M = -0{,}3$ и разность между верхним и нижним квартилями $I = 5{,}2$. Выберем в качестве центра интервал $M \pm \frac{1}{4} I$, затем возьмем следующий интервал $M \pm \frac{1}{2} I$ (кроме центрального интервала, включая его границы), далее — интервал $M \pm I$ и, наконец, то, что выпадает за все пределы. Введем такие условные знаки:

меньше $M-I$	от $M-I$ до $M-\frac{1}{2}I$	от $M-\frac{1}{2}I$ до $M-\frac{1}{4}I$	от $M-\frac{1}{4}I$ до $M+\frac{1}{4}I$	от $M+\frac{1}{4}I$ до $M+\frac{1}{2}I$	от $M+\frac{1}{2}I$ до $M+I$	больше $M+I$
◎	○	•	●	+	+	#

Эта символика, по-видимому, окажется удовлетворительной и для других ситуаций, хотя выбор самих интервалов скорее всего будет иным.

Использование этих знаков показано на илл. 9.2.3. Теперь гораздо лучше, чем на илл. 9.2.2, видно, что противоположные углы прямоугольника имеют остатки одного и того же знака и что середина таблицы аппроксимирована гораздо лучше. Это довольно широко распространенное свойство остатков от модели типа строка ПЛЮС столбец ПЛЮС константа.

9.3. ВЫЯВЛЕНИЕ ПРЕИМУЩЕСТВЕННЫХ УРОВНЕЙ

Наша модель имеет вид

константа ПЛЮС строка ПЛЮС столбец

(где общий член был у нас нулем). Рассмотрим все сочетания места и месяца, дающие какое-нибудь заданное значение отклика, скажем 55° F. Как видно из илл. 9.3.1, все эти сочетания попадают в точности на диагональ.

Теперь открывается следующая перспектива:

1. Построить некоторые линии равных значений отклика для других откликов, кроме 55° F.
2. Отметить «плюсами» точки наблюдений места и месяца.
3. Рассмотреть результат, обнаружить, насколько проще это выглядит после поворота на 45°, и произвести такой поворот.
4. На новой картинке не проводить контурных линий.
5. Вместо них ограничиться малым числом рисок на осях — по одной для каждого столбца и для каждой строки.

Илл. 9.3.2 показывает результат после первых двух шагов, а илл. 9.3.3 — после всех пяти.

Последняя картинка отличается от многих графических представлений тем, что в ней мы не пытаемся интерпретировать положение

по горизонтали; лишь положение на *вертикали*, которое оказывается значением отклика, представляется нам важным.

Остатки. Теперь, когда модель приведена к удобному виду, появляется естественная возможность представления остатков в каждой комбинации строк и столбцов — сочетаний места и месяца в нашем примере. Илл. 9.3.4 дает остатки, рассчитанные при альтернативном анализе и соответствующие каждой точке пересечения линий (места и месяца) на илл. 9.3.3. Теперь остатки представлены вполне наглядно и прекрасно чувствуются.

Иногда удобнее рассматривать остатки, кодированные иначе. Мы можем закодировать их значения и представить на рисунке, аналогичном илл. 9.3.3. Воспользуемся теми же обозначениями и интервалами, что и в илл. 9.2.3, и, кроме того, повернем сами символы

◎ ○ ∘ • × X ※ .

при повороте рисунка на 45°. Результат такой операции показан на илл. 9.3.5.

Поскольку при таком кодировании наши остатки оказываются хорошо структурированными и это видно из того, что

противолежащие углы имеют противоположные знаки,

нам приходится продолжить совершенствование модели.

Обдумывая это представление остатков, надо помнить, что: (1) такие данные в конце концов можно было бы прекрасно аппроксимировать и гораздо проще; (2) мы начинали работать с данными, имея «уйму» разной информации об эффектах места и месяца (в Ларедо — жарко, в Карибу — холодно, от января к июлю — теплеет, наиболее резкая смена температур — примерно в апреле). Мы могли бы упорядочить города и месяцы, не глядя на данные, и таким образом пройти изрядную часть пути до илл. 9.3.5. Для бóльших количеств данных мы не сможем знать столь много и будем вынуждены обращаться к двувходовой модели (например, вида илл. 9.3.3), чтобы «построить» картину остатков. Вертикальное расположение дает возможность легко представлять себе, что произойдет, если изменится месяц или место или и то и другое. Мы хорошо понимаем поведение всей модели.

Все проделанное относится к одной модели и одному набору данных. Однако все то же самое можно проделать и для любой другой модели вида

(константа ПЛЮС строка) ПЛЮС столбец

при любом множестве данных, так как мы можем «загнать» данные в модель и повторить в точности все то, что мы делали в примере.

9.4. ДОВОДКА АДДИТИВНЫХ МОДЕЛЕЙ

Когда для построения модели в таблицах с двумя входами используются средние, эффекты строк и столбцов можно найти одной итерацией. Последующие итерации не меняют результатов. Когда же

пользуются другими способами усреднения, такими, например, как медианы, то первого приближения может и не хватить. В настоящем параграфе мы обсудим эту ситуацию. Одновременно мы приобщим читателя к вычислению остатков и их исследованию как стандартным приемам анализа данных. Иллюстрируя итеративный подход с акцентом на анализ остатков, выясним, как средние и медианы влияют на остатки в таблицах с «выбросами» и в таблицах с пустыми ячейками.

И хотя иногда это может приводить к дополнительным вычислительным трудностям, накопленный опыт анализа данных требует вычислений и анализа остатков:

- для оценки модели как таковой;
- для изучения специфических свойств данных и интерпретаций;
- для сопоставления с другими способами анализа;
- для выявления грубых ошибок, просчетов.

Нешаблонный анализ данных мы рекомендуем начинать с медиан, но другие усреднения, например обычные средние, конечно же, могут оказаться более подходящими в некоторых задачах. Улучшить анализ способны и такие преобразования, как логарифмирование.

Основной подход. До этого момента мы анализировали таблицы с двумя входами, извлекая сначала эффект строки, затем эффект столбца, и на этом заканчивали анализ. Обычно этот первый шаг итерации в общем процессе построения модели имеет большое значение, но на нем работа не может окончиться либо потому, что используемый метод, по существу, требует итерации, либо из-за проблем округления, когда счет идет с заданным числом знаков, либо из-за «дыр» в таблице.

На каждом шаге итерации мы отделяем простой чертой данные (исходные или остатки) от их сумм. Двойную черту мы используем для отделения стадий расчета, т. е. сумм, от последующих остатков. Итак, последовательные шаги:

Первый	Второй	Третий	Четвертый	Пятый
Неразделенные исходные данные	Сумма по строкам или столбцам	Остатки для исходных данных	Сумма по строкам или столбцам	Новые остатки

и так далее. Когда мы сначала берем строки, а затем столбцы, удобным естественным расположением для таблиц с двумя входами будет «змееобразное»: слева —направо — вниз — влево — вниз — направо и так далее, по образцу:

Первый	Второй	Третий
Исходные данные	Сумма по строкам	Остатки от строк
		Четвертый: Сумма по столбцам
Седьмой	Шестой	Пятый
Остатки от строк	Сумма по строкам	Новые остатки от столбцов
Восьмой...		

Чтобы проиллюстрировать этот метод, на илл. 9.4.1 показан пример, в котором модель для медиан требует более одной итерации, чтобы закончить расчет. Напомним, что пример из илл. 9.1.1 завершился после первого цикла. Проще всего считать, что здесь приведены другие месячные температуры для тех же самых городов. Несмотря на предыдущее толкование, илл. 9.4.1 расшифровывается все же с трудом, и, по-видимому, имеет смысл напомнить этапы и детали.

Во-первых, начинаем с данных о температуре |
Во-вторых, получаем построчные медианы ||
В-третьих, вычитаем медианы (каждой строки) из ее ячеек
В-четвертых, получаем медианы столбцов
В-пятых, вычитаем медианы столбцов из остатков в ячейках этих столбцов |
В-шестых, берем медианы строк и отмечаем знаком √ (маркируем) медианы, равные нулю ||
В-седьмых, получаем остатки от строк
В-восьмых, получаем медианы столбцов и маркируем их
В-девятых, получаем остатки в столбцах |
В-десятых, получаем медианы строк и *стоп*,

поскольку все медианы как строк, так и столбцов оказываются маркированными, т. е. обращаются в нули. Этим завершается илл. 9.4.1А.

В илл. 9.4.1А введена идея полушага. Мы называем полушагом определение (вычисление и запись) медианы в строке (или столбце) с последующим подсчетом остатков. Таким образом, пары пунктов в приведенном выше списке как раз и образуют полушаги; таковы пункты второй и третий, четвертый и пятый и т. д. На илл. 9.4.1Б мы собрали для сопоставления данные по местам (городам), объединяя соответствующие члены из каждой итерации, в данном случае — из первого и третьего полушагов. Третья строка изменялась лишь однажды на третьем полушаге. Здесь мы могли бы остановиться либо же разложить сами медианы строк на две части: их общую медиану — медиану медиан и эффекты строк. Чтобы их найти, вычтем общую медиану 54,4, играющую при этом роль общего среднего, общего члена (константы) из всех медиан строк.

Данные, представленные на илл. 9.4.1В, для месячных прогнозов несколько труднее истолковать, поскольку они не упорядочиваются столь просто, как эффекты строк; лишь в мае и июне есть вклады на четвертом полушаге.

В нашем случае потребовалось четыре полушага. Опыт говорит, что четырех—восьми полушагов, как правило, хватает. Так как арифметические ошибки в таком итеративном процессе сами собой не корректируются и легко совершаются, было бы хорошо — если только это возможно, — вести счет на ЭВМ и распечатывать результаты. Хотя возможно медленное накопление ошибок, могущих проявиться при большом числе полушагов, вряд ли стоит обращать на них внимание. Ведь полная таблица потребует лишь нескольких полушагов скажем 4, 5 или 6, и на этом все кончится. (Таблицы с пропусками требуют несколько большего числа итераций.)

Модели для средних. Можно использовать тот же подход и для средних арифметических или любых иных видов усреднения. На илл. 9.4.2 показан для других температурных данных процесс доводки сначала для средних, а затем для медиан. Со средними мы сделали два полушага и остановились (при точности 0,1°), а с медианами — четыре.

Этот результат согласуется с теоретическим представлением о том, что анализ средних требует только двух полушагов. Читатель, может быть, удивится, когда узнает, что изредка анализ для средних занимает больше двух полушагов. Причина этого «противоречия» с теорией в том, что ее вывод относится к использованию точных значений: бесконечных десятичных или рациональных дробей. Но это невозможно, когда счет идет с фиксированным числом знаков и соответствующим округлением средних. Хотя в илл. 9.4.2 мы и округляли расчетные значения, анализ потребовал все же только двух полушагов, что и бывает, как правило, при анализе для средних.

Цель илл. 9.4.2 — сравнение двух наборов остатков. Если взглянуть на итоговые остатки илл. 9.4.2 (анализ для средних), то можно увидеть полный беспорядок, в котором 13 из 21 остатка по модулю больше единицы или равны ей. Лишь с крайним напряжением мы можем разглядеть в них какие-то информативные структуры. Однако среднее разбросано не сильно, следовательно, сумма квадратов отклонений мала, что мы еще покажем позже в этом параграфе.

Конечные остатки при анализе для медиан, как уже отмечалось, интерпретируются гораздо лучше. Так, лишь 6 остатков из 21 превышают по модулю единицу и все они (вместе с одним нулем) относятся к городу Ларедо. Так что Карибу и Вашингтон находятся в отличном согласии с моделью и надо подумать о Ларедо.

Если отделить остатки, попадающие в интервал от — 0,5 до + 0,5, то за его пределами еще останется для медиан 7 из 21, а для средних — 17 из 21.

На илл. 9.4.3 построены «опоры и консоли» для двух множеств остатков из илл. 9.4.2, сравниваемых между собой и с конечными остатками, полученными для полуразмаха в качестве среднего (полусумма крайних значений выборки). Вдобавок к «опоре и консоли» мы даем также три меры разброса: размах, сумму квадратов и сумму абсолютных значений. Каждый из трех способов анализа превосходит два остальных точно по одному из трех критериев, соотнесенных с разными мерами разброса.

Необычные значения («сюрпризы»). Чтобы продолжить обсуждение средних и медиан, мы исследуем их сравнительное действие в сильно упрощенной ситуации, когда в таблице с двумя входами все значения, кроме одного, равны между собой. Таблицу мы возьмем 3 × 7 и для наглядности сделаем все значения равными нулю, кроме центрального, которому припишем число 21 (чтобы удобно было делить и на 3, и на 7). Расположение в центре числа 21 создает симметрию и способствует визуальному контрасту, но не меняет ничего другого:

0	0	0	0	0	0	0
0	0	0	21	0	0	0
0	0	0	0	0	0	0

1. *Анализ для медиан.* Медианы всех строк и всех столбцов равны нулю, так что первоначальная таблица есть одновременно и таблица остатков. (Это вовсе не быстрота такого анализа, чтобы его рекомендовать, просто наш критерий не изменил бы любую таблицу.)

2. *Анализ для средних.* Здесь мы находим

0	0	0	0	0	0	0	V	0	0	0	0	0	0	0
0	0	0	21	0	0	0	3	−3	−3	−3	18	−3	−3	−3
0	0	0	0	0	0	0	V	0	0	0	0	0	0	0
								−1	−1	−1	6	−1	−1	−1
							V	1	1	1	−6	1	1	1
							V	−2	−2	−2	12	−2	−2	−2
							V	1	1	1	−6	1	1	1

Таким образом, начав с простой таблицы из 21 числа (20 одинаковых и одного, отличного от них), мы пришли к таблице с четырьмя классами распределенных значений.

Конечно, можно сказать, что, попадись такая таблица, ее бы сразу опознали и не стали бы применять средние арифметические. Это, однако, гораздо легче сказать, чем сделать. Когда данные громоздки и сложны, ситуации, аналогичные бросающейся в глаза здесь, могут быть хорошо замаскированы, и тогда понадобится множество тонких ухищрений для выявления выбросов.

Может возникнуть и другое положение, когда частичный успех анализа ведет лишь к перераспределению отклонений, так что сама структура остатков, полученных в анализе средних, оказывается приемлемой. А тот анализ, который вообще не перераспределяет отклонений, не может быть привлекательным.

Наша цель здесь не в том, чтобы сосредоточиваться на такого рода аргументах, а в том, чтобы прояснить, как два метода реагируют на выбросы. Тенденция медиан — оставлять их в «сохранности», а тенденция средних — сокращать разброс их отклонений. Сумма абсолютных значений остатков при анализе для средних в нашем примере равна 48. Значит, этот анализ ведет к увеличению более чем вдвое суммы абсолютных отклонений (48 против 21). Вместе с тем сумма квадратов остатков уменьшилась с $441 = (21)^2$ до $252 = (15{,}9)^2$. Если бы мы были уверены, что действительно хотим минимизировать сумму квадратов, то мы должны были бы воспользоваться при анализе средними, ну а если надо минимизировать сумму абсолютных значений, то гораздо больше пользы будет от медиан ($48/21 = 2{,}3$, а $21/15{,}9 = 1{,}3$).

Этот пример показывает, что анализ для средних лишь при малом числе больших выбросов позволяет скрыть, затушевать выбросы, которые могут быть выявлены другими методами гораздо проще (в диапазоне между явными выбросами и обычной нерегулярностью).

Пропуски. Иногда таблица имеет пустые ячейки, ячейки без наблюдений или же мы хотим «отложить» часть данных и посмотреть, что будет происходить в их отсутствие. Когда мы ведем итеративный счет двувходовой таблицы с несколькими пустыми ячейками, мы

рабатываем каждую строку (или столбец) в соответствии с числом имеющихся там ячеек. Находим итоги заполненных ячеек. Обычно это ведет к росту числа итераций.

В примере илл. 9.4.4 мы анализируем оба случая одновременно: с включением значения 1035 и без него, используя и медианы, и средние, — всего четыре варианта анализа. На илл. 9.4.4Б показано, как средние «расправляются» с выбросом, размазывая ошибку по всей таблице, и как медианы оставляют явно торчащее значение на месте. Когда же выброс из ячейки стерт, остатки, как показывает илл. 9.4.4В, для наших двух методов гораздо ближе друг к другу, хотя медианы дают 10 остатков, меньших десяти по модулю, а средние—всего лишь 6. Конечно, это особенность медиан — давать большее число как малых, так и больших остатков в сравнении с тем, что дают средние.

Собственно, наиболее важно изменение эффекта в столбце (обработке), обусловленное пропуском аномального значения. Оно таково: 0, — 12, 0, 14, для модели с медианами и 17, — 51, 17, 17 для модели со средними. Это показывает, сколь велики потери: изменения суммы квадратов от 313 доходят до 3468; одно аномальное значение существенно влияет на эффекты столбца (обработки), особенно при использовании средних.

Модели с одновременным использованием медиан и средних. Мы сравнивали достоинства методов анализа, основанных на среднем и медиане (и полуразмахе). Иногда, как показывает следующий пример, методы можно и комбинировать.

Пример. Сезонные колебания. На илл. 9.4.5 даны логарифмы индексов сезонной распродажи уцененных вещей универмагами по месяцам, с 1941 по 1945 г. (среднее за 1935—1939 гг., за базу принято 100%). В левой части илл. 9.4.5 в столбцах годов фактически стоят значения $1000\log_{10}$ (индекс /100). Логарифмы же взяты потому, что «финансовая» динамика гораздо чаще имеет мультипликативный, чем аддитивный, характер эффектов. (Жизненный опыт рекомендует логарифмы, хотя в нашем примере можно было бы приблизить значения строк и получше.) Обращаясь к илл. 9.4.5, мы видим, что числа возрастают от года к году. Чтобы заметить степень периодичности в годах, мы ранжируем данные для месяцев внутри каждого года (илл. 9.4.6), вычисляем медианы рангов для каждого месяца и размах рангов.

Для 6 месяцев размах рангов не превышает двух; октябрь, ноябрь и декабрь особенно выделяются по устойчивости высоких рангов. Январь и июль «борются» за наименьший уровень продаж. Февраль, март и август имеют максимальные размахи рангов. Отметим исключительно резкий разрыв между декабрем и следующим за ним январем. Это значит, что мы не можем надеяться на аппроксимацию поведения данных по годам гладкой кривой, если не учесть сезонных колебаний.

Для моделирования логарифмических данных с илл. 9.4.5 вычтем медианы, полученные для каждого месяца по данным пяти лет, и определим остатки во всех строках, как это сделано в правой части илл. 9.4.5. Мы собираемся построить регрессионную модель

$$y = \text{константа} + \text{месяц}\ (i) + \text{год}\ (j)$$

(без учета компонент ошибки), где «месяц (i)» и «год (j)» представляют соответственно эффекты месяца и года, i — январь, февраль, ..., декабрь; j — 1941, ..., 1945. Когда мы вычитаем медианы (помесячно, как на илл. 9.4.5), то находим оценку «эффекта месяца». (Это же мы могли бы сделать и со средними.) То, что останется после вычитания, может использоваться для оценивания среднего эффекта для каждого года отдельно.

Читатель увидит, что средние остатки для каждого года (приведенные внизу в правой части илл. 9.4.5) меняются по годам почти линейно. Появляется соблазн провести прямую по пяти точкам. Против этого выступает наше знание о том, что времена спада и процветания «скачут» вокруг явно нелинейных траекторий. А в таком случае, хотя прямая и может быть полезной для некоторых целей, в общем, мы все-таки хотим выделять эффекты для каждого года, или, может быть, подобрать какую-нибудь более сложную кривую, или, наконец, применить сглаживание. Следовательно, мы предпочитаем вычитать средние по столбцам из первых (уже сосчитанных) остатков и исследовать в дальнейшем их. (Средние вместо медиан берем только для показа многообразия возможных приемов анализа.)

На илл. 9.4.7 месяцы для всех пяти лет обозначены подряд числами 1, 2, ..., 60. Справа от номера месяца даются первые остатки (из илл. 9.4.5); конечные (или вторые) остатки получены вычитанием из остатков эффектов года; на илл. 9.4.7 они стоят справа от первых остатков. Для сглаживания используются скользящие медианы по трем соседним точкам. Теперь конечные сглаженные остатки наносим на график в зависимости от t, номера месяца, и представляем на илл. 9.4.8.

Для илл. 9.4.8 характерно, что малая изменчивость приходится на середину и большая для t, соответствующих двум первым и двум последним годам. Мы знаем из илл. 9.4.5 и 9.4.6, что средний год аппроксимируется очень хорошо, поскольку для 11 месяцев из двенадцати медианные значения приходились именно на средний год. Поэтому чем ближе к краям диапазона, тем хуже будет предсказание. Неясно, почему первые два года дали бóльшую изменчивость остатков, чем два последних, но сам факт из графика несомненен. Можно предполагать, что в первые годы второй мировой войны люди еще могли покупать более или менее то, что хотели, а в последние годы войны они покупали то, что могли купить.

9.5. ПОДБОР ЕЩЕ ОДНОГО ПОСТОЯННОГО КОЭФФИЦИЕНТА

Хотя аддитивная модель помогла понять таблицу среднемесячных температур (3 «места» × 7 месяцев), мы остановились на полпути. Самое малое, что еще следовало бы сделать, — это подобрать еще одну константу. Но что это может быть за константа?

Мы намерены добавить член, пропорциональный произведению эффектов строки и столбца, и тем самым ввести мультипликативный эффект. Это выражение мы запишем в форме

$$\text{константа} \times \frac{(\text{строка})\ (\text{столбец})}{\text{общий член}}$$

где «константа» — как раз и есть то, что мы хотим дополнительно оценить. Это приводит нашу модель к виду

$$\text{общий член ПЛЮС строка ПЛЮС столбец ПЛЮС константа} \times \frac{(\text{строка})\ (\text{столбец})}{\text{общий член}}\ .$$

Чтобы довести дело до конца, надо вспомнить, что в ячейках таблицы находятся

первые остатки = данные МИНУС (общий член ПЛЮС строка ПЛЮС столбец).

Мы собираемся ввести новый член — константа × (строка) (столбец/общий член)—в модель для этих первых остатков. Член, перед которым стоит искомая константа, мы можем назвать сопоставлением, где

$$\text{сопоставление} = \frac{(\text{строка})\ (\text{столбец})}{\text{общий член}}\ ,$$

причем правую часть равенства мы всегда можем вычислить. Это отношение можно представить графически как зависимость первых остатков от сопоставления или же, несколько более формально, прямой линией метода наименьших квадратов. На илл. 9.5.1А вычислены сопоставления для альтернативного анализа (см. илл. 9.1.1Д), а на илл. 9.5.1Б дана таблица упорядоченных попарно по убыванию сопоставлений (слева) и соответствующих им первых остатков. То, что знаки в парах хорошо «следят» друг за другом, говорит о зависимости между сопоставлениями и первыми остатками. Каждой группе равных между собой сопоставлений на илл. 9.5.1Б соответствует медиана группы первых остатков. Ясно видна сильная зависимость.

На илл. 9.5.2 приведен диагностический график с сопоставлениями на абсциссе и первыми остатками на ординате. Видно, что прямая с угловым коэффициентом, равным + 1,00, будет разумной моделью, если мы решим ограничиться прямой.

Илл. 9.5.3А и 9.5.3Б демонстрируют анализ, в котором также брался член

$$1{,}00\,\frac{(\text{строка})\ (\text{столбец})}{\text{общий член}}\ ,$$

который вычитался из первых остатков илл. 9.1.1Д. Мы видим, что (1) по величине новые остатки заметно меньше и (2) среди их знаков преобладает минус (16 из 21). «Опора и консоль» для новых остатков на илл. 9.5.3В подсказывает, что модель можно еще улучшить, если прибавить 1 (или даже 2) к новым остаткам и отнять 1 (или 2) от общего члена, что даст

$$54\ \text{ПЛЮС строка ПЛЮС столбец ПЛЮС}\ \frac{(\text{строка})\ (\text{столбец})}{55}$$

или же

$$53\ \text{ПЛЮС строка ПЛЮС столбец ПЛЮС}\ \frac{(\text{строка})\ (\text{столбец})}{55}\ .$$

В последнем случае максимум модуля остатков примерно равен 3,2 вместо 5,1 (в илл. 9.5.3) или 12,3 (в илл. 9.1.1). Так что включение в модель еще одного члена значительно улучшило адекватность данного конкретного описания.

Аддитивные и мультипликативные модели. Когда мы проводим такую дополнительную коррекцию модели, константа, стоящая при сопоставлении, часто хорошо интерпретируется (см. *EDA*, гл. 12). Эта специфически расширенная аддитивная модель с равным успехом может считаться и мультипликативной моделью, поскольку

$$\text{общий член ПЛЮС строка ПЛЮС столбец ПЛЮС } \frac{\text{(строка) (столбец)}}{\text{общий член}} \qquad (*)$$

эквивалентно выражению

$$\text{общий член} \left(1 + \frac{\text{строка}}{\text{общий член}}\right)\left(1 + \frac{\text{столбец}}{\text{общий член}}\right), \qquad (**)$$

в чем мы убедимся, если раскроем скобки в (**) и приведем подобные члены. Выражения в круглых скобках можно обозначить «строка*» и «столбец*», так как выражение в первой скобке зависит только от строки, а во второй — только от столбца, откуда имеем

$$\text{общий член УМНОЖИТЬ строка* УМНОЖИТЬ столбец*}.$$

9.6. ИСПОЛЬЗОВАНИЕ ПРЕОБРАЗОВАНИЙ

В этом параграфе мы покажем такое использование преобразований модели, которое требует одной дополнительной константы. Оно подгоняется к конкретному набору данных, откуда вовсе не следует, что данная константа в данном преобразовании будет хороша и для другого набора подобных данных. (Это существенно отличается от выбора между исходными данными и их логарифмами, где хороший выбор для одного набора данных может быть столь же хорошим и для других подобных наборов.) Как и в других случаях подбора отдельных констант, мы чувствуем здесь свободу выбора «любых испытанных значений». Если ничего лучшего не видно, то мы просто перебираем несколько значений константы и останавливаемся на том, что кажется лучше. (Когда мы пытаемся что-то сделать с выражением для откликов на скорую руку, мы берем согласованные преобразования, введенные в параграфе 5.5.)

Здесь же мы комбинируем преобразование откликов и двувходовой анализ. Большей частью в качестве преобразования работает логарифмирование счетных сумм, имеющих вид

$$\log((\text{общий член})\,(\text{строка}^*)\,(\text{столбец}^*)) \equiv \log \text{общий член} + \log \text{строка}^* + \log \text{столбец}^*.$$

Тогда мультипликативная модель превращается в модель аддитивную.

Эффективное преобразование откликов логарифмированием соответствует угловому коэффициенту $+1$ на диагностическом графике (см. илл. 9.5.2). Как следствие можно брать другие положительные уг-

лы в качестве допустимых преобразований в рамках использования логарифмов. В том же направлении идут прямые для квадратных и кубических корней, однако они менее круты, чем при логарифмах. А для обратных величин и соответствующих корней прямые идут так же, но уже более круто. Что же произойдет, если угол наклона диагностического графика станет отрицательным, — несколько менее частый, но возможный случай? Что будет, если отклики, взятые для преобразования, столь же часто отрицательны, как и положительны подобно температурам на илл. 9.1.1? А что было бы, если бы мы «сдвинули» наши города в илл. 9.1.1 на несколько сотен миль севернее?

Во многих случаях, когда угол наклона отрицателен, а отклики неотрицательны, мы можем с успехом преобразовать отклики, возводя их в квадрат или же в другую, большую двух, степень. Для откликов, которые, вообще говоря, могут принимать и отрицательные значения, подобные вещи не годятся, так как два значения y (одно положительное, а другое отрицательное) приводят к одному и тому же значению y^2 (кроме, конечно, нуля). Путаница в данных почти неизбежно ведет к путанице в результатах.

Простейшим преобразованием, одинаково применимым к отрицательным и к положительным значениям и имеющим график, как и для y^2, вогнутый к оси абсцисс, будет, по-видимому, экспоненциальное, введенное в параграфе 5.1

$$y \to e^{cy} = e^{y/d}.$$

Для него согласованной (по y_0) формой будет замена y на

$$de^{(y-y_0)/d} + (y_0 - d),$$

или, что то же самое,

$$(de^{-y_0/d})\, e^{y/d} + (y_0 - d).$$

Так что если в нашем «температурном» примере из илл. 9.1.1А мы выберем $y_0 = 55$ (поближе к центру данных, равному 54,4) и $d = 80$, то преобразование будет таким:

$$(80e^{-55/80})\, e^{y/80} + (55 - 80) = 40{,}23e^{y/80} - 25.$$

Зная несколько вычислительных «хитростей», считать можно на настольном калькуляторе.

На илл. 9.6.1 приведены результаты различных экспоненциальных преобразований наших данных о температурах на Восточном побережье (илл. 9.1.1А). Заметим, что по мере изменения d последовательно меняются и предсказания, и остатки.

Выбранные значения d либо кратны 10, либо дают простые значения для $120/d$, либо и то и другое. Число 120 фактически в преобразованиях и вычислениях илл. 9.6.1 не фигурирует. Выгода от выбора 120 и простоты $120/d$ заключается в удобстве построения графиков в зависимости от величин, обратных к d. Здесь, как мы ожидаем, — и в данном случае обоснованно, — более простым будет поведение в зависимости от обратных величин d (или кратных им), а не от самих d (или кратных им), особенно при больших значениях d, когда ситу-

ация приближается к исходной (что можно проверить, переходя к пределу). Если мы взглянем на некоторые процентили для остатков моделей при избранных d на илл. 9.6.1К, то становится очевидным, поскольку наши преобразования согласованы, что $d = 60$ лучше, чем $d = 120$ или $d = 40$. Какое же значение d самое лучшее? Чтобы приблизиться к ответу, рассмотрим графики на илл. 9.6.2 Н-разброса, Е-разброса, 2-разброса (размах между вторыми по величине от краев значениями) и 1-разброса (размах) для всех имеющихся значений d. Во всех случаях минимум «разброса» приходится на значения d около 60—70, причем $d = 60$ ближе для всех кривых, кроме Н-разброса, который ведет себя наиболее нерегулярно и минимум его ближе к $d = 70$.

Почему же минимум именно здесь? Обратимся к остаткам при $d = 60$:

Карибу	0,4	0	0	0	0	0	0
Вашингтон	0	—0,3	0	0	0,1	1,4	0,7
Ларедо	—3,7	0	1,5	3,1	0	—2,3	—5,3

и при $d = 70$:

Карибу	0	—0,1	0	0	0,4	0	0,7
Вашингтон	0	0	0	—0,4	—0,4	0,1	0
Ларедо	—1,9	2,1	2,9	3,6	0	—3,2	—5,6

Как мы видим, остатки для Карибу и Вашингтона очень малы, порядка десятых, что не так для Ларедо. Однако, просматривая илл. 9.6.1 сверху донизу, мы не видим лучшего согласия между Ларедо и либо Вашингтоном, либо Карибу, чем для $d = 60$ и для $d = 70$. Отсюда мы заключаем, что наилучшее экспоненциальное преобразование

- имеет d, заключенное в пределах от 60 до 70;
- дает хорошее согласие между Вашингтоном и Карибу;
- дает приемлемое согласие Ларедо с двумя другими городами, но не столь хорошее.

Это легко объяснимо, поскольку Ларедо, хотя и недалеко от Мексиканского залива, но это все же не совсем Восточное побережье, и было бы естественно для нас проверить более широко, как будет «работать» $d = 70$ для городов собственно Восточного побережья.

Заметим, что в таком экспоненциальном преобразовании мы вновь подбирали

одну дополнительную константу,

хотя в этот раз константа содержалась в преобразовании откликов, а не в самой модели (как в параграфе 9.5, где константа стояла перед сопоставлением, входящим в саму модель).

9.7. АНАЛИЗ ТАБЛИЦ С ТРЕМЯ И БОЛЕЕ ВХОДАМИ

Если мы хотим анализировать таблицы с числом входов, превышающим два, «арифметика» неизбежно становится тяжелее. Общий подход имеет много аналогичного с подходами к двувходовому анализу. Для таблиц с тремя входами (назовем их A, B и C) образуем парные

переменные — *AB*, *BC* и *CA*. Начнем с *AB*. Для каждой пары *AB* имеется строка входов, различающихся только по *C*. Сделаем полушаг, находя и вычленяя сумму для каждой такой строки. Затем обращаемся к паре *BC* и сделаем полушаг для остатков в направлении, где меняется только *A*. Затем обратимся к *CA* и далее последовательно опять к *AB*.

Этот подход описан более подробно и с примерами в *EDA*, гл. 13.

РЕЗЮМЕ. ТАБЛИЦЫ ОТКЛИКОВ С ДВУМЯ ВХОДАМИ

Анализ таблиц откликов в форме «общий член ПЛЮС строка ПЛЮС столбец ПЛЮС остаток» открывает общий подход, который может быть важным элементом моделирования.

Такое изучение расширяет наши возможности анализа по сравнению с визуальным подходом к остаткам.

Выработка в процессе анализа большего числа чисел, чем в начальных данных, — обычное явление, часто необходимое для успешного выяснения того, что надо бы делать дальше, чтобы продвинуться в анализе; такой подход не исключает дальнейшего обобщения.

Начиная с нуля или с результатов какого-нибудь другого анализа и последовательно исключая медианы, поочередно по строкам и по столбцам (доводка для медиан), придем к моделям «общий член ПЛЮС строка ПЛЮС столбец».

Такой анализ по медианам выпячивает редкие протуберанцы в данных с помощью остатков, тогда как анализ по средним, напротив, «размазывает» отдельные возмущения по всей таблице.

Такой анализ по медианам работает как для полных таблиц, так и в случае, когда некоторые из их ячеек стерты (потеряны, отсутствуют или выделены из набора данных для специального исследования).

Любую модель вида «строка ПЛЮС столбец ПЛЮС общий член» можно представить графически с координатами в виде двух семейств параллельных линий, где на вертикальной координате стоят значения самой модели.

Кодовые обозначения, показывающие одновременно знаки остатков и их примерные величины, делают картину остатков более выразительной.

Появление одинаковых знаков у остатков в противоположных углах либо хорошо организованной числовой таблицы, либо закодированной картинки остатков дает основание для попытки достроить модель за счет введения дополнительного параметра.

Модель рода

$$\text{общий член ПЛЮС строка ПЛЮС столбец ПЛЮС } 1{,}0\ \frac{(\text{строка})\ (\text{столбец})}{(\text{общий член})}$$

можно представить и так:

общий член УМНОЖИТЬ строка* УМНОЖИТЬ столбец*.

Графический анализ зависимости остатков от сопоставлений позволяет нам подобрать константу в выражении

$$\text{ПЛЮС (константа)}\ \frac{\text{(строка) (столбец)}}{\text{(общий член)}},$$

добавляемом к предыдущей модели вида «общий член ПЛЮС строка ПЛЮС столбец».

Часто преобразование y может служить альтернативой к модели с одной «лишней» константой. Чтобы найти наиболее подходящее преобразование для модели «общий член ПЛЮС строка ПЛЮС столбец», можем испытать класс преобразований откликов, меняя параметр класса.

Можно сравнивать результаты для таких различных моделей, используя согласованные преобразования вблизи центральных значений наших откликов и исследуя поведение разных простых мер разброса для остатков.

БИБЛИОГРАФИЯ

McNeil D. R. and Tukey J. W. (1975). Higher-order diagnosis of two-way tables, illustrated on two sets of demographic empirical distributions. — Biometrics, 31, 487—510.

EDA—Tukey J. W. (1977). Exploratory Data Analysis. Reading, Mass., Addison-Wesley.

ИЛЛЮСТРАЦИИ

Иллюстрация 9.1.1

Пример с двумя входами: среднемесячные температуры. Медианы используются для предсказания во всех частях примера, кроме Д.

А. ДАННЫЕ с медианами: данные | медианы

	янв.	фев.	март	апр.	май	июнь	июль	
Карибу	8,7	9,8	21,7	34,7	48,5	58,4	64,0	34 7
Вашингтон	36,2	37,1	45,3	54,4	64,7	73,4	77,3	54,4
Ларедо	57,6	61,9	68,4	75,9	81,2	85,8	87,7	75,9

Б. ОДНОВХОДОВЫЙ АНАЛИЗ: остатки ‖ медианы

Карибу	—26,0	—24,9	—13,0	0	13,8	23,7	29,3	34,7
Вашингтон	—18,2	—17,3	—9,1	0	10,3	19,0	22,9	54,4
Ларедо	—18,3	—14,0	—7,5	0	5,3	9,9	11,8	75,9

В. Медианы ДРУГОГО ВХОДА: остатки ‖ медианы строк

				медианы столбцов остатков		общая медиана		
Карибу	—26,0	—24,9	—13,0	0	13,8	23,7	29,3	34,7
Вашингтон	—18,2	—17,3	—9,1	0	10,3	19,0	22,9	54,4
Ларедо	—18,3	—14,0	—7,5	0	5,3	9,9	11,8	75,9
	—18,3	—17,3	—9,1	0	10,3	19,0	22,9	54,4

Г. Результаты ВЫДЕЛЕНИЯ: остатки | компоненты строки (остатки строковых медиан)

компоненты столбца (медианы остатков столбца) | константа (медиана строковых медиан)

	янв.	фев.	март	апрель	май	июнь	июль	
Карибу	—7,7	—7,6	—3,9	0	3,5	4,7	6,4	—19,7
Вашингтон	0,1	0	0	0	0	0	0	0,0
Ларедо	0	3,3	1,6	0	—5,0	—9,1	—11,1	21,5
	—18,3	—17,3	—9,1	0	10,3	19,0	22,9	54,4

Д. Альтернативный анализ

Карибу	—6,3	—10,2	—3,3	—0,3	3,5	3,4	4,0	—20
Вашингтон	1,2	—2,9	0,3	—0,6	—0,3	—1,6	—2,7	0
Ларедо	2,6	1,9	3,4	0,9	—3,8	—9,2	—12,3	20
	—20	—15	—10	0	10	20	25	55

Иллюстрация 9.2.1

Начальный график для модели из табл. Д илл. 9.1.1

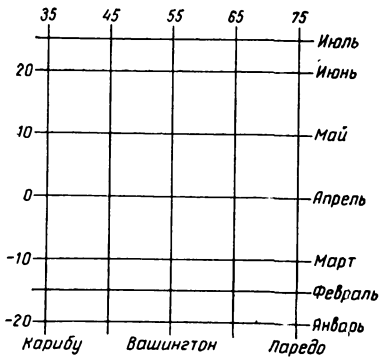


Иллюстрация 9.2.2

График для модели с остатками из табл. Д илл. 9.1.1

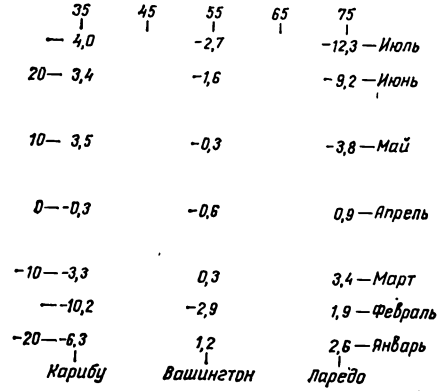


Иллюстрация 9.2.3

Закодированные остатки из илл. 9.2.2

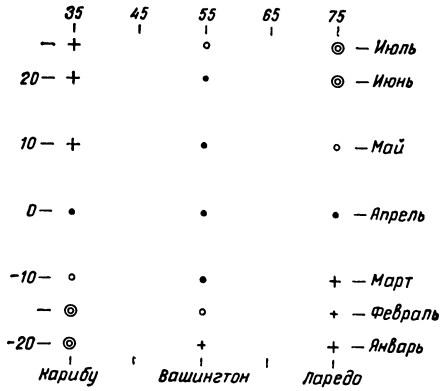


Иллюстрация 9.3.1.

Илл. 9.2.1 с прямой, соответствующей постоянной температуре 55° F

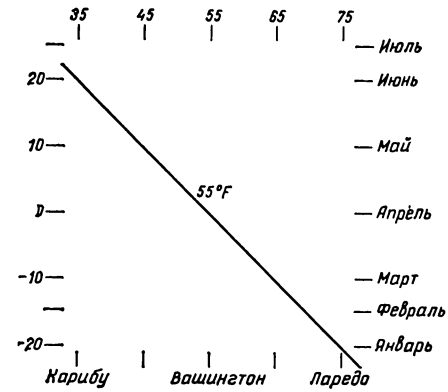


Иллюстрация 9.3.2

Илл. 9.3.1 с бо́льшим числом изотерм и с данными, отмеченными «плюсами»

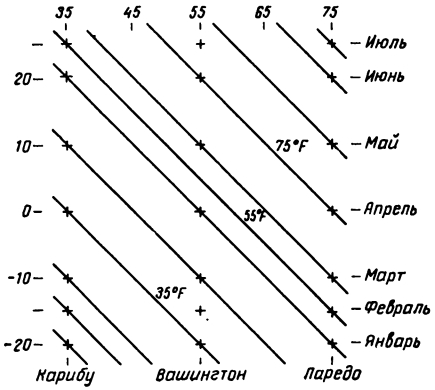


Иллюстрация 9.3.3

Преобразованная илл. 9.3.2: (3) поворот на 45°; (4) линии равного уровня обозначены по краям; (5) $3 \times 7 = 21$ точка пересечения, $3 + 7 = 10$ линий соответствует всем рассмотренным комбинациям городов и месяцев

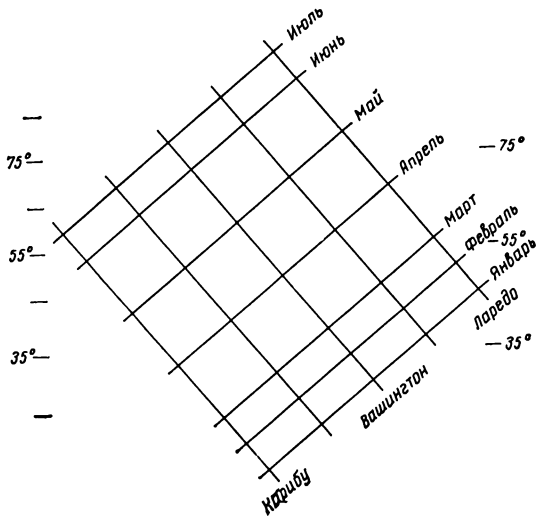


Иллюстрация 9.3.4

Илл. 9.3.3 с отстатками, пропорциональными вертикальным отрезкам

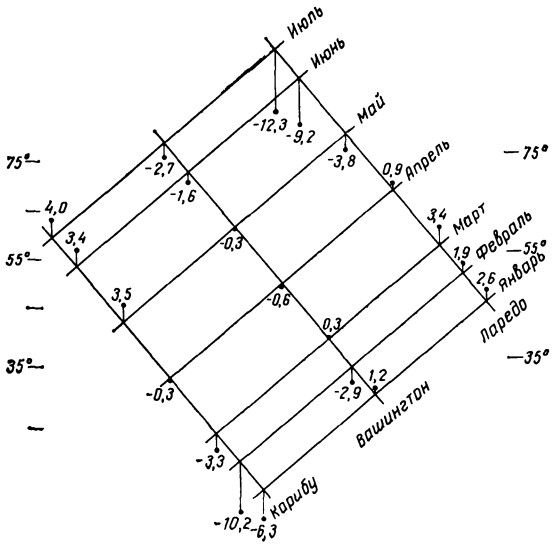


Иллюстрация 9.3.5

Илл. 9.3.3 с закодированными остатками

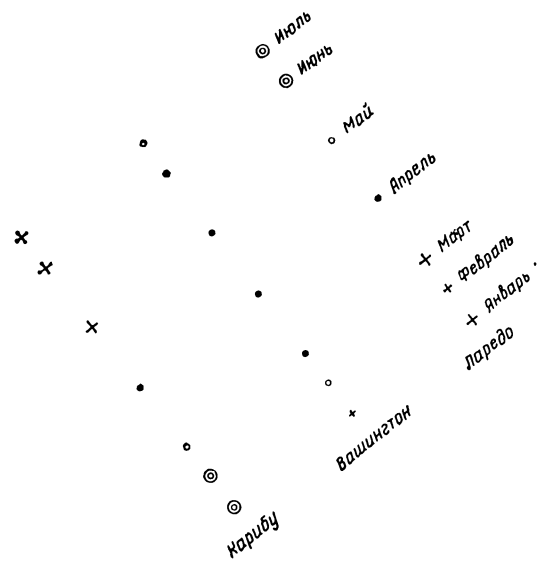


Иллюстрация 9.4.1

Медианно сглаженная аппроксимация. Первая таблица в части А представляет преобразованные температуры в зависимости от месяцев (столбцы) и расположения (строки). (Это значения экспоненциального преобразования с $d = 36$ илл. 9.6.1Ж.)

А. ДОВОДКА

28,9	29,3	33,3	39,5	49,1	58,6	65,2	39,5
40,4	40,9	46,5	54,4	66,1	79,0	85,9	54,4
57,7	62,6	71,2	83,3	93,5	103,7	108,3	83,3

—10,6	—10,2	—6,2	0	9,6	19,1	25,7
—14,0	—13,5	—7,9	0	11,7	24,6	31,5
—25,6	—20,7	—12,1	0	10,2	20,4	25,0
—14,0	—13,5	—7,9	0	10,2	20,4	25,7

3,4	3,3	1,7	0	—0,6	—1,3	0	$\surd$
0	0	0	0	1,5	4,2	5,8	$\surd$
—10,9	—6,5	—3,5	0,7	0,7	0,7	0	—0,7

3,4	3,3	1,7	0	—0,6	—1,3	0	$\surd$
0	0	0	0	1,5	4,2	5,8	$\surd$
—11,6	—7,2	—4,2	0	0	0	—0,7	$\surd$

$\surd$	$\surd$	$\surd$	$\surd$	0,7	0,7	$\surd$	

3,4	3,3	1,7	0	—1,3	—2,0	0	$\surd$
0	0	0	0	0,8	3,5	5,8	$\surd$
—10,9	—6,5	—3,5	0,7	0	0	0	$\surd$

Б. СОПОСТАВЛЕНИЕ ПРЕДСКАЗАНИЙ ДЛЯ МЕСТ

		Дает	Варианты	
39,5	0	39,5	—14,9	54,4
54,4	0	54,4	0	54,4
83,3	—0,7	82,6	28,2	54,4

В. СОПОСТАВЛЕНИЕ ПРЕДСКАЗАНИЙ ПО МЕСЯЦАМ

	—14,0	—13,5	—7,9	0	10,2	20,4	25,7
	0	0	0	0	0,7	0,7	0
дает	—14,0	—13,5	—7,9	0	10,9	21,1	25,7

Иллюстрация 9.4.2

Сглаживание средними и медианами. (Это значения экспоненциального преобразования с $d = 70$ илл. 9.6.1В.)

А. ДОВОДКА СРЕДНИМИ

21,1	21,7	28,5	37,4	48,8	58,5	64,6	40,1	—19,0	—18,4	—11,6	—2,7	8,7	18,4	24,5
38,5	39,2	45,9	54,4	65,4	76,0	81,3	57,2	—18,7	—18,0	—11,3	—2,8	8,2	18,8	24,1
57,6	62,2	69,8	79,4	86,8	93,7	96,7	78,0	—20,4	—15,8	—8,2	1,4	8,8	15,7	18,7
								—19,4	—17,4	—10,4	—1,4	8,6	17,6	22,4
							v	0,4	—1,0	—1,2	—1,3	0,1	0,8	2,1
							v	0,7	—0,6	0,9	—1,4	—0,4	1,2	1,7
							v	—1,0	1,6	2,2	2,8	0,2	—1,9	—3,7

Б. ДОВОДКА МЕДИАНАМИ

21,1	21,7	28,5	37,4	48,8	58,5	64,6	37,4	—16,3	—15,7	—8,9	0	11,4	21,1	27,2
38,5	39,2	45,9	54,4	65,4	76,0	81,3	54,4	—15,9	—15,2	—8,5	0	11,0	21,6	26,9
57,6	62,3	69,8	79,4	86,8	93,7	96,7	79,4	—21,8	—17,1	—9,6	0	7,4	14,3	17,3
								—16,3	—15,7	—8,9	0	11,0	21,1	26,9
0	0	0	0	0,4	0	0,3	0	0	0	0	0	0,4	0	0,3
0	0,1	0	—0,4	—0,4	0,1	—0,4	0,4	0,4	0,5	0,4	0	0	0,5	0
—1,9	2,2	2,9	3,6	0	—3,2	—6,0	—3,6	—5,5	—1,4	—0,7	0	—3,6	—6,8	—9,6
v	0,1	*v*	*v*	*v*	*v*	—0,4								
0	—0,1	0	0	0,4	0	0,7	*v*							
0	0	0	—0,4	—0,4	0,1	0	*v*							
—1,9	2,1	2,9	3,6	0	—3,2	—5,6	*v*							

Иллюстрация 9.4.3

«Опоры и консоли» для трех анализов одних и тех же данных (два из которых взяты в илл. 9.4.2)

	Анализ для средних		Анализ для медиан		Анализ для полуразмахов	
	6		**6**		**6**	
	5		**5**		**5**	
	4		**4**		**4**	
	3		**3**	6	**3**	
	2	128	**2**	09	**2**	5414
	1	276	**1**		**1**	37
	0	41872	**0**	004070001	**0**	0666
	—0	694	**—0**	104400	**—0**	356
	—1	023409	**—1**	9	**—1**	7053
	—2		**—2**		**—2**	1535
	—3	7	**—3**	2	**—3**	
	—4		**—4**		**—4**	
	—5		**—5**	6	**—5**	
	—6		**—6**		**—6**	
Максимум	2,8		3,6		2,5	
Минимум	—3,7		—5,6		—2,5	
Размах	6,5		9,2		5,0	
Сумма квадратов	51,04		71,57		58,57	
Сумма абсолютных значений	27,2		21,3		30,5	

Иллюстрация 9.4.4

Анализ данных[1] о влажности пшеницы (доля сухого вещества зерен × 1000) по средним и медианам с учетом и без учета значения 1035

А. Первичные данные

Номер блока	Не было	Ранняя	Средняя	Поздняя
1	718	732	734	792
2	725	781	725	716
3	704	1035	763	758
4	726	765	738	781

Б. С учетом значения 1035

Медианы					*Средние*				
—1	—38	1	33	—3	18	—78	12	48	—18
7	12	—7	—42	—4	32	—22	10	—21	—25
—30	250	15	—16	12	—67	154	—30	—57	53
1	—11	—1	16	3	18	—53	8	29	—10
—27	24	—13	13	749	—44	66	—22	0	762

В. Без учета значения 1035

Медианы					*Средние*				
—1	—26	1	19	0	1	—27	—5	31	—1
7	24	—7	—56	—1	15	29	—7	—38	—8
0	□	45	0	—15	—16	□	21	—6	2
0	0	—2	1	7	0	—3	—10	11	8
—27	12	—13	27	746	—27	15	—5	17	745

[1]Данные взяты с разрешения автора и Biometric Society из: Cochran W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. — Biometrics, 3, p. 27.

Иллюстрация 9.4.5

Вычитание значений месячных медиан из логарифмов индексов распродажи

Месяцы	1941	1942	1943	1944	1945	Медианы[1]	(Остаток)$_1$ = log y—медиана				
	log y										
Январь	—41	93	111	140	193	111	—152	—18	0	29	82
Февраль	—13	68	193	152	233	152	—165	—84	41	0	81
Март	45	146	158	230	328	158	—113	—12	0	72	170
Апрель	114	149	215	238	243	215	—101	—66	0	23	28
Май	121	124	193	250	262	193	—72	—69	0	57	69
Июнь	93	93	190	212	270	190	—97	—97	0	22	80
Июль	4	29	104	152	215	104	—100	—75	0	48	111
Август	117	104	146	196	225	146	—29	—42	0	50	79
Сентябрь	179	207	241	292	320	241	—62	—34	0	51	79
Октябрь	140	230	272	320	364	272	—132	—42	0	48	92
Ноябрь	199	274	330	394	438	330	—131	—56	0	64	108
Декабрь	364	418	438	507	548	438	—74	—20	0	69	110
Среднее остатков по столбцу							—102	—51	3	44	91

[1] Месячная медиана по данным за пять лет.

Иллюстрация 9.4.6

Ранги по месяцам внутри каждого года по данным илл. 9.4.5 вместе с медианными рангами и размахами

Месяцы	1941	1942	1943	1944	1945	Медиана	Размах
	ранг месяца внутри года						
Январь	1	3½	2	1	1	1	2½
Февраль	2	2	6½	2½	4	2½	4½
Март	4	7	4	6	9	6	5
Апрель	6	8	8	7	5	7	3
Май	8	6	6½	8	6	6½	2
Июнь	5	3½	5	5	7	5	3½
Июль	3	1	1	2½	2	2	2
Август	7	5	3	4	3	4	4
Сентябрь	10	9	9	9	8	9	2
Октябрь	9	10	10	10	10	10	1
Ноябрь	11	11	11	11	11	11	0
Декабрь	12	12	12	12	12	12	0

Иллюстрация 9.4.7

Таблица остатков$_1$, остатков$_2$ и сглаженных остатков$_2$, где остаток$_2$ = остаток$_1$ — среднее по остаткам$_1$ (остаток$_1$ = log y — медиана из илл. 9.4.5)

t	Остаток$_1$	Остаток$_2$[1]	Сглаженные остатки$_2$	t	Остаток$_1$	Остаток$_2$	Сглаженные остатки$_2$	t	Остаток$_1$	Остаток$_2$	Сглаженные остатки$_2$
1	—152	—50		25	0	—3	31 —3	49	82	—9	—9
2	—165	—63	—50	26	41	38	—3	50	81	—10	—9
3	—113	—11		27	0	—3		51	170	79	—10
4	—101	1		28	0	—3		52	28	—63	—22
5	—72	30	5	29	0	—3		53	69	—22	
6	—97	5		30	0	—3		54	80	—11	
7	—100	2	5	31	0	—3		55	111	20	—11
8	—29	73	40	32	0	—3		56	79	—12	
9	—62	40		33	0	—3		57	79	—12	
10	—132	—30	—29	34	0	—3		58	92	1	
11	—131	—29		35	0	—3		59	108	17	
12	—74	28		36	0	—3		60	110	19	
13	—18	33	28	37	29	—15					
14	—84	—33	33 28	38	0	—44	—15				
15	—12	39	—15	39	72	28	—21—15				
16	—66	—15		40	23	—21	13—21—15				
17	—69	—18		41	57	13	—21 4				
18	—97	—46	—24	42	22	—22	4				
19	—75	—24		43	48	4					
20	—42	9		44	50	6					
21	—34	17	9	45	51	7	6				
22	—42	9		46	48	4	7				
23	—56	—5	9	47	64	20					
24	—20	31	—3	48	69	25	20				

[1] Остаток$_2$=остаток$_1$ — среднее по остаткам$_1$.

Иллюстрация 9.4.8

Сглаженные остатки$_2$ из илл. 9.4.7, где остатки$_2$ = остатки$_1$ — средние по столбцам остатков$_1$

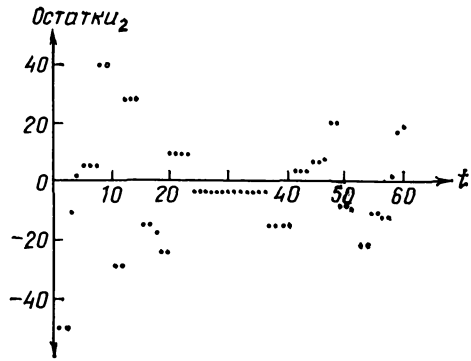


Иллюстрация 9.5.1

Вычисление сопоставлений для альтернативной модели из илл. 9.1.1Д и использование их для упорядочения первых остатков

А. Вычисление сопоставлений

		Месяцы						
		январь	февраль	март	апрель	май	июнь	июль
Столбец		—20	—15	—10	0	10	20	25
Столбец/общий член		—0,364	—0,273	—0,182	0	0,182	0,364	0,454
Сопоставление = (строка)·(столбец) / общий член	Строка —20:	—7,28	—5,46	—3,64	0	3,64	7,28	9,08
	0:	0	0	0	0	0	0	0
	20:	7,28	5,46	3,64	0	—3,64	—7,28	—9,08

Б. Сопоставления и первые остатки, упорядоченные попарно по убыванию сопоставлений

Сопоставления	Первые остатки	Сопоставления	Первые остатки	Сопоставления	Первые остатки
9,08	4,0	0	—0,6	9,08	4,0
7,28	2,6	0	—1,6	7,28	3,0
7,28	3,4	0	—2,7	5,46	1,9
5,46	1,9	0	—2,9	3,64	3,4
3,64	3,4	—3,64	—3,8	0	—0,3
3,64	3,5	—3,64	—3,3	—3,64	—3,6
0	1,2	—5,46	—10,2	—5,46	—10,2
0	0,9	—7,28	—9,2	—7,28	—7,8
0	0,3	—7,28	—6,3	—9,08	—12,3
0	—0,3	—9,08	—12,3		
0	—0,3				

Иллюстрация 9.5.2

Графическое представление зависимости остатков от сопоставлений по данным илл. 9.5.1Б. Кружком отмечена точка, в которой оказались два наблюдения

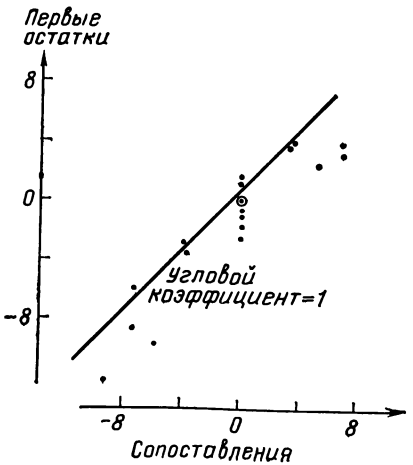


Иллюстрация 9.5.3

Подбор дополнительной константы к первым остаткам илл. 9.1.1Д.

А. СТРУКТУРНАЯ СХЕМА

Новые остатки	Вклад строк
$1{,}0 \times \frac{\text{(строка) (столбец)}}{\text{общий член}}$	
Вклад столбцов	Общий вклад

Б. РЕЗУЛЬТАТЫ

	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	
Карибу	1,0	—4,7	0,3	—0,3	—0,1	—3,9	—5,1	—20
Вашингтон	1,2	—2,9	0,3	—0,6	—0,3	—1,6	—2,7	0
Ларедо	—4,7	—3,6	—0,2	0,9	—0,2	—1,9	—3,2	20
Карибу	—7,3	—5,5	—3,6	0	3,6	7,3	9,1	
Вашингтон	0	0	0	0	0	0	0	
Ларедо	7,3	5,5	3,6	0	—3,6	—7,3	—9,1	
	—20	—15	—10	0	10	20	25	55

В. «ОПОРА и КОНСОЛЬ» для НОВЫХ ОСТАТКОВ

1	20		
0	933		
—0	1222336	Медиана	—0,6
—1	69	Верхний квартиль	—0,1
—2	79	Нижний квартиль	—3,2
—3	269	Межквартильный размах	3,1
—4	77		
—5	1		

Иллюстрация 9.6.1

Преобразованные значения из илл. 9.1.1Д (слева) и двувходовый анализ для медиан (справа). Восемь экспоненциальных преобразований (включая $d = \infty$, т. е. непреобразованные данные) и для сравнения — преобразование путем возведения в квадрат. (Все преобразования даются приведенными к значению 55.)

А. Непреобразованные данные или преобразованные при $d=\infty$, $120/d=0$

8,7	9,8	21,7	34,7	48,5	58,4	64,0		—7,7	—7,6	—3,9	0	3,5	4,7	6,4	—19,7
36,2	37,1	45,3	54,4	64,7	73,4	77,3		0,1	0	0	0	0	0	0	0
57,6	61,9	68,4	75,9	81,2	85,8	87,7		0	3,3	1,6	0	—5,0	—9,1	—11,1	21,5
								—18,3	—17,3	—9,1	0	10,3	19,0	22,9	54,4

Б. Преобразованные данные при $d=80$, $120/d=1{,}5$

19,8	20,5	27,8	37,1	48,8	58,5	64,5	0	—1,2	—0,8	0	0,8	0,1	1,1	—17,3
38,2	39,0	45,9	54,4	65,3	75,7	80,7	1,1	0	0	0	0	0	0	0
57,6	62,2	69,6	78,9	86,0	92,6	95,4	—0,2	2,5	3,0	3,8	0	—3,8	—6,0	20,7
							—17,3	—15,4	—8,5	0	10,9	21,3	26,3	54,4

В. Преобразованные данные при $d=70$, $120/d=1{,}7143$

21,1	21,7	28,5	37,4	48,8	58,5	64,6	0	—0,1	0	0	0,4	0	0,7	—17,4
38,5	39,2	45,9	54,4	65,4	76,0	81,3	0	0	0	—0,4	—0,4	0,1	0	0
57,6	62,3	69,8	79,4	86,8	93,7	96,7	—1,9	2,1	2,9	3,6	0	—3,2	—5,6	21,0
							—16,3	—15,6	—8,9	0	11,0	21,1	26,5	54,8

Г. Преобразованные данные при $d=60$, $120/d=2$

22,7	23,2	29,4	37,8	48,8	58,5	64,7	0,4	0	0	0	0	0	0	—16,6
38,9	39,5	46,0	54,4	65,5	76,5	82,0	0	—0,3	0	0	0,1	1,4	0,7	0
57,7	62,3	70,0	80,0	87,9	95,3	98,5	—3,7	0	1,5	3,1	0	—2,3	—5,3	22,5
							—15,5	—14,6	—8,4	0	11,0	20,7	26,9	54,4

Д. Преобразованные данные при $d=48$, $120/d=2{,}5$

25,3	25,7	31,0	38,4	48,9	58,5	64,9	1,9	1,6	0,2	0	—0,8	0	0	—16,0
39,4	40,1	46,2	54,4	65,7	77,4	83,4	0	0	—0,6	0	0	2,9	2,5	0
57,7	62,4	70,5	81,2	89,9	98,2	101,9	—5,4	—1,4	0	3,1	0,5	0	—2,7	23,7
							—15,0	—14,3	—7,6	0	11,3	20,1	26,5	54,4

Е. Преобразованные данные при $d=40$, $120/d=3$

27,6	27,9	32,4	39,1	49,0	58,5	65,1	2,9	2,6	1,3	0	—1,7	—2,4	0	—15,3
40,0	40,6	46,4	54,4	66,0	78,4	84,9	0	0	0	0	0	2,2	4,5	0
57,7	62,5	70,9	82,4	92,0	101,4	105,6	—7,5	—3,3	—0,7	2,8	0,8	0	0	25,2
							—14,4	—13,8	—8,0	0	11,6	21,8	26,0	54,4

Ж. Преобразованные данные при $d=36$, $120/d=3{,}33$

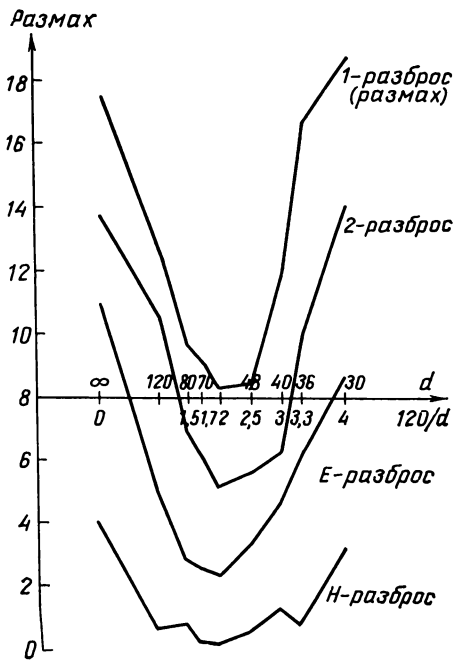
28,9	29,3	33,3	39,5	49,1	58,6	65,2	3,4	3,3	1,7	0	—1,3	—2,0	0	—14,9
40,4	40,9	46,5	54,4	66,1	79,0	85,9	0	0	0	0	0,8	3,5	5,8	0
57,7	62,6	71,2	83,3	93,5	103,7	108,3	—10,9	—6,5	—3,5	0,7	0	0	0	28,2
							—14,9	—13,5	—7,9	0	10,9	21,1	25,7	54,4

З. Преобразованные данные при $d=30$, $120/d=4$

31,4	31,6	34,9	40,2	49,2	58,6	65,5	4,6	4,3	2,4	0	—2,6	—5,2	—3,7	—14,2
41,0	41,5	46,7	54,4	66,5	80,4	88,1	0	0	0	0	0,5	2,4	4,7	0
57,7	62,8	71,9	85,2	96,8	108,8	114,2	—14,1	—9,5	—5,6	0	0	0	0	30,8
							—13,4	—12,9	—7,7	0	11,6	23,6	29,0	54,4

Иллюстрация 9.6.2

Зависимость от d «разбросов», вычисленных для остатков значений $de^{(y-y_0)/d} + (y_0 - d)$, преобразованных в илл. 9.6.1 среднемесячных температур



Глава 10 ● РОБАСТНЫЕ И УСТОЙЧИВЫЕ МЕРЫ ПОЛОЖЕНИЯ И МАСШТАБА

10.1. УСТОЙЧИВОСТЬ

В этой главе мы развиваем идеи робастных и устойчивых методов, беглое знакомство с которыми состоялось в гл. 1. Мы ссылались на них в гл. 9 и еще вернемся к ним более детально в параграфах 14.7, 14.8 и 14.9.

Устойчивость — это свойство, которое нам хотелось бы иметь для итоговых статистик. Если при изменении малой части от всего количества данных, возможно резком, значение такого итога может существенно измениться, то такая статистика неустойчива. Наоборот, если при изменении малой доли данных (неважно какой и сколь сильно) существенных изменений в суммирующей статистике не происходит, то она *устойчива*.

Арифметическое среднее — прототип неустойчивого итога. Если в равенстве

$$\frac{1+2+2+3+\ldots+23}{101}=9{,}58$$

мы заменим второе измерение, равное 2, числом 101002, то среднее арифметическое станет равным

$$\frac{1+2+101002+3+\ldots+23}{101}=1009{,}58.$$

Изменение 1/101 доли данных изменило это среднее катастрофически.

Медиана служит прототипом простого устойчивого итога. В нашем примере со 101 значением мы предполагали их следующими:

50 значений {1; 2; 2; 3; ...; 8; 9};
1 значение {9};
50 значений {9, 5; 10; ...; 23},

так что независимо от «многоточия» медиана равна 9. После замены 2 на 101002 получим

50 значений {1; 2; 3; ...; 8; 9; 9};
1 значение {9,5};
50 значений {10; ...; 23; 101002}

и медиану, равную 9,5, что не так уж далеко от 9.

Никаким изменением одного значения нельзя изменить медиану больше, чем она изменилась. Очевидно, что медиана устойчива.

Есть и другие устойчивые итоги. Среди них ряд важных определяются не явными формулами, а итеративно, как в приводимом ниже

случае, где требуются итерации, поскольку мы не можем вычислить y^*, пока мы не знаем весов w, и не можем вычислить веса, пока мы не знаем y^*:

$$y^* = \frac{\Sigma w_i y_i}{\Sigma w_i},$$

где

$$w_i = \begin{cases} \left(1 - \left(\frac{y_i - y^*}{cS}\right)^2\right)^2, & \text{если } \left(\frac{y_i - y^*}{cS}\right)^2 < 1; \\ 0 — \text{в противном случае}, \end{cases}$$

а

$$S = \text{медиана } \{|y_i - y^*|\}$$

или, возможно,

$$S = \frac{1}{2} \text{ (межквантильный размах)};$$

причем c здесь — константа, которая часто берется равной 9 или 6 (поскольку S есть оценка для примерно $\frac{2}{3}\sigma$, так что при $c = 6$ мы выкидываем остатки, превышающие 4σ, когда разброс измеряется наблюдениями скорее из середины распределения, чем с «хвостов»). Эта величина часто называется *бивес-оценкой* (см. гл. 14, где такие оценки исследуются подробнее).

10.2. РОБАСТНОСТЬ

Почему бы не удовлетвориться медианой? Зачем тратить значительные усилия для вычисления таких более сложных устойчивых оценок положения, как *бивес?*

Здесь мы соприкасаемся с эффективностью — с использованием итогов, которые извлекают из данных почти все, что мы можем узнать из них суммированием.

Как же измерить полноту извлечения? Нужно

- выбрать меру эффективности и
- выяснить, какая часть данных извлекается.

Во всех случаях, кроме малых выборок из распределений с очень растянутыми «хвостами», вполне удовлетворительной мерой эффективности служат дисперсии итогов (в предположении, что оценка в среднем близка к тому, что мы хотим оценить, — в некоторых особых случаях, например со специальной функцией потерь, могут действовать и другие правила). Для дисперсий естественная мера эффективности такова:

$$\text{эффективность} = \frac{\text{наименьшая возможная дисперсия}}{\text{фактическая дисперсия}}.$$

Обычно эффективность выражают в процентах. Что подразумевается тогда под 90%-ной эффективностью? Вот ответы:

- это очень хорошо;

● это то, что наши итоги извлекают из возможных 100% для самого «лучшего» итога;

● но это может быть слишком накладно для того, чтобы войти в практику.

Выигрыш одного процента в эффективности практически ни в чем особенном не проявляется.

Мы же хотим иметь

робастность эффективности,

т. е. высокую эффективность при широком варьировании ситуаций, а не при какой-то одной из них. Поэтому мы прежде всего выявляем наихудшую эффективность в некотором разумном множестве ситуаций. И если она достаточно высока, то можем с полным основанием считать, что получили хороший итог.

10.3. РОБАСТНЫЕ И УСТОЙЧИВЫЕ ОЦЕНКИ ПОЛОЖЕНИЯ

Как ведут себя по отношению к новому требованию робастности три обсуждаемые выше оценки положения? Об этом «рассказывает» илл.- 10.3.1.

Вывод, если пренебречь совсем малыми выборками, таков: *бивес-оценка* обладает всеми желаемыми свойствами. (Для совсем малых выборок, скажем, размера три, четыре или пять, лучше работать с медианой.)

Мы склонны рекомендовать для практики использовать следующее:

● медиану в предварительных исследованиях и тех ситуациях, где достаточно умеренной эффективности;

● бивес-оценку или что-нибудь в этом роде, когда требуется высокое качество;

● среднее арифметическое лишь после тщательного изучения, когда традиция области исследования или смысл задачи требуют этого, когда слишком трудно переделать вычислительную программу, когда наилучшие доверительные границы или методы оценки значимости разработаны лишь для среднего, когда нужна линейность или же, наконец, когда данные таковы, что «хвосты» коротки и нет выбросов.

Бывают распределения с «хвостами», «поджатыми» даже больше, чем у нормального или близко к нему, а вовсе не с большим разбросом. В таких случаях среднее эффективнее, чем бивес-оценка, а для очень коротких «хвостов» можно построить особые статистики положения, приписывающие большие веса наблюдениям, удаленным от центра. Так, исключительно короткие «хвосты» имеет равномерное распределение. Для выборки из него, как показывает теория, среднее двух крайних значений использует всю выборочную информацию о положении. Это поясняет, почему при укороченных «хвостах» распределений крайним значениям следует придавать больший вес.

10.4. РОБАСТНЫЕ ОЦЕНКИ МАСШТАБА

Здесь мы не знаем всего того, что нам бы хотелось. Известно, что и

$$\text{МАО} = \text{медиана абсолютных отклонений} = \text{медиана } |y_i - y^*|,$$

где y^* — устойчивая оценка положения, и межквартильный размах (разность между квартилями или 25%-ными точками),

$$I = \text{H-разброс} = \text{верхний квартиль} - \text{нижний квартиль}$$

будут устойчивыми оценками масштаба (разброса). Мы же выдаем эти две оценки за робастные к эффективности.

Размахи и оценки масштаба, основанные на размахах, по всей видимости, несколько менее робастны, чем $s = (\text{выборочная дисперсия})^{1/2}$. Но все они не конкурентноспособны с МАО и I.

Альтернатива. Давид Лакс [David Lax (1975)] проанализировал усложненную статистику, которая более чем в два раза лучше МАО в некоторых весьма реалистических ситуациях.

Пусть

$$y' = \text{медиана } y; \quad u_i = \frac{y_i - y'}{9\,(\text{МАО})}.$$

Он воспользовался для меры масштаба асимптотической дисперсией бивес-оценки, а именно выражением

$$\frac{n\Sigma'\,(y-y')^2\,(1-u^2)^4}{[\Sigma'\,(1-u^2)\,(1-5u^2)]^2},$$

где Σ' означает суммирование лишь по $u^2 \leqslant 1$. Грубо говоря, u_i задают веса. Когда u^2 малы, веса примерно равны и знаменатель сводится к n, так что все выражение стремится к $\Sigma\,(y - y')^2/n$, значит, такая оценка выглядит разумной.

Модификация приводит к

$$s^2 = \frac{1}{n-1}\,\Sigma\,(y-y')^2$$

и к несколько лучшему, как известно, виду

$$ns^2_{\text{bi}} = \frac{n\Sigma'\,(y-y')^2\,(1-u^2)^4}{[\Sigma'\,(1-u^2)\,(1-5u^2)][-1+\Sigma'\,(1-u^2)\,(1-5u^2)]}.$$

Заметим, что $(1 - u^2)^4 = w^2$, где w были введены в параграфе 10.1.

Неизбежно должны измениться и другие оценки.

10.5. РОБАСТНЫЕ И УСТОЙЧИВЫЕ ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

Слово «робастный» используется в статистике во многих смыслах и, вероятно, они будут еще множиться. Примерно до 1970 г. оно означало, как правило, устойчивость к предпосылкам, т. е. требование, чтобы 5% были действительно 5% при широкой вариации ситуаций. Это отличается от смысла, в котором данный термин употребляется здесь, но, отличаясь, не обязательно противоречит ему.

Так, Алан Гросс [Alan Gross (1976)] показал, что в простых процедурах получения доверительных интервалов для центров симметричных распределений можно иметь одновременно и робастность к эффективности, и робастность к предпосылкам. В последней работе Гросс сообщил, что это вполне возможно, если строить интервалы на бивес-оценке.

Усовершенствованная процедура использует $\hat{y}$ — бивес-оценку с весами

$$u_i = \frac{y_i - \hat{y}}{9\,(\text{МАО})},$$

а для оцениваемой дисперсии (асимптотически корректное) — выражение (см. параграф 10.4)

$$\widehat{\text{var}}\ \hat{y} = s^2_{\text{Bi}} = \frac{\Sigma' (y - \hat{y})^2 (1 - u^2)^4}{[\Sigma' (1 - u^2)(1 - 5u^2)]\,[\Sigma' (1 - u^2)(1 - 5u^2) - 1]}$$

(в сущности неважно, что используется: $\hat{y}$ или медиана y). Затем берется t-критерий Стьюдента с числом степеней свободы $\nu = 0{,}7 \times (n - 1)$, что приводит к интервалу

$$\hat{y} \pm t_\nu \sqrt{\widehat{\text{var}}\ \hat{y}}.$$

Эта аппроксимация хорошо работает при $n \geqslant 8$.

10.6. УСТОЙЧИВАЯ И РОБАСТНАЯ РЕГРЕССИЯ

Итеративное использование бивес-оценок в регрессии описано в параграфах 14.7, 14.8 и 14.9.

10.7. МНОГОМЕРНЫЕ ДАННЫЕ

При работе с многомерными (многокомпонентными) данными, — с векторами, как обычно говорят, — мы хотели бы иметь:

- устойчивые и робастные одновременно оценки положения, которые мы можем получить применением бивес-оценок покомпонентно;
- устойчивые и робастные одновременно «заменители» для дисперсий, ковариаций и корреляций.

Как же найти такие «заменители»?

Классически для вычисления оценок дисперсий и ковариаций чаще всего использовалась линейная регрессия. Мы сделаем наоборот; воспользуемся устойчиво-робастной линейной регрессией как инструментом для получения устойчиво-робастных оценок дисперсий и ковариаций.

Главная особенность дисперсий и ковариаций — это их *квадратичная природа*, лучше всего проявляющаяся в тождестве, которому они удовлетворяют: $\text{var}\,(fu + gv) \equiv f^2 \text{var}\,(u) + 2fg\,\text{cov}\,(u, v) + g^2 \text{var}\,(v)$, а более обще:

$$\operatorname{cov}(fu+gv,\ ju+kv) \equiv fj \operatorname{var}(u)+(fk+gj)\operatorname{cov}(u,\ v)+gk \operatorname{var}(v).$$

И ничего сверх этого не надо, чтобы понять, как они вычисляются. Наши обычные оценки var и cov дисперсий и ковариаций, конечно, удовлетворяют тем же тождествам.

Если же мы определяем оценки-аналоги для дисперсий и ковариаций, то мы хотим, чтобы и они им удовлетворяли. Но так как оценки-аналоги не будут квадратичными функциями от данных, то есть лишь один путь удовлетворить тождествам: сначала определить оценки для специальных компонент, а затем воспользоваться тождествами для расширения определения.

Итак, допустим, что мы начинаем со специально выбранных компонент x, $y - bx$, $z - cy - dx$, ... и что мы хотим:

1) иметь нуль в качестве оценки для аналога ковариации (который будем обозначать $\widehat{\text{cob}}$ вместо $\widehat{\text{cov}}$) между каждой парой этих специальных компонент;

2) взять квадрат устойчивой оценки масштаба в качестве аналога дисперсии (обозначим $\widehat{\text{bar}}$ вместо $\widehat{\text{var}}$) для каждой избранной компоненты.

Латинская буква «b» в обозначениях $\widehat{\text{cob}}$ и $\widehat{\text{bar}}$, которая заменяет букву «v» в обозначениях $\widehat{\text{cov}}$ и $\widehat{\text{var}}$, взята из слова robust (робастный), поскольку «r» было бы неблагозвучно («rar»). Заметим, кстати, что звуки «b» и «v» очень похожи.

Тождества дают нам, в частности, следующие равенства:

$$\widehat{\text{bar}}(y) \equiv \widehat{\text{bar}}((y-bx)+bx) = \widehat{\text{bar}}(y-bx)+b^2\,\widehat{\text{bar}}(x);$$

$$\widehat{\text{cob}}(x,\ y) \equiv \widehat{\text{cob}}(x,\ (y-bx)+bx) = b.\widehat{\text{bar}}(x);$$

$$\widehat{\text{bar}}(z) \equiv \widehat{\text{bar}}(z-cy-dx) + c^2\,\widehat{\text{bar}}(y-bx)+(d+bc)^2\,\widehat{\text{bar}}(x);$$

$$\widehat{\text{bar}}(2y+3z) \equiv 4\,\widehat{\text{bar}}\,y+6\,\widehat{\text{cob}}(y,\ z)+9\,\widehat{\text{bar}}\,z = 9\,\widehat{\text{bar}}(z-cy+dx)+$$

$$+(9c^2+6c+4)\,\widehat{\text{bar}}(y-bx)+(9(d+bc)^2+6(d+bc)\,b+4b^2)\,\widehat{\text{bar}}(x),$$

так что $\widehat{\text{bar}}$ и $\widehat{\text{cob}}$ определяются для всех линейных комбинаций x, y, z, ... Следовательно, нам остается:

1) выбрать константы b, c, d, ... и

2) решить, какие устойчиво-робастные оценки масштаба взять для $\widehat{\text{bar}}(x)$, $\widehat{\text{bar}}(y-bx)$, $\widehat{\text{bar}}(z-cy-dx)$, ...

Простейший выбор, вполне удовлетворяющий этому, таков:

1) величины $y - bx$, $z - cy - dx$, ... рассматриваем как остатки от нашей условной устойчиво-робастной регрессии y по x, z по x и y и

2) принимаем, что

$$\widehat{\text{bar}}(x) = ns^2_{\text{b}i}\ (\text{для значений } x); \qquad \widehat{\text{bar}}(y-bx) = ns^2_{\text{b}i}\ (\text{для значений } y-bx)$$

и т. д., где $ns^2_{\text{b}i}$ — это бивес-дисперсии из параграфа 10.4.

Другие робастные переменные можно выбрать как квадрат отклонений медианы от медианы или даже как квадрат разности между двумя порядковыми статистиками.

Замечание. Мы должны выбирать, сохраняя свойства устойчивости для $\widehat{\text{bar}}$ и $\widehat{\text{cob}}$, а именно:

● изменение малой части данных, каким бы большим оно ни было, мало меняет результаты.

Мы вынуждены признать (отчасти из-за того, что мы не знаем, как сохранить и робастность, и устойчивость) существование некоего свойства, которое может быть столь же важным, а именно:

● изменение порядка компонент, скажем $(z, x, y, \ldots)$ вместо $(x, y, z, \ldots)$, или же замена одного множества компонент другим эквивалентным набором их линейных комбинаций, скажем $(x + y, x - y, 2y - z, \ldots)$ вместо (x, y, z), не меняет результатов.

Мы поступаем так, как если бы инвариантность была отличной, но можно обойтись и без нее, между тем обойтись без устойчивости и робастности к эффективности никак нельзя. И это именно то, что мы думаем и рекомендуем.

Альтернативы. Если у нас всего 3 компоненты, то существует всего шесть упорядочений. Желающие установить порядок своих компонент могут за исходную взять любую из 6 перестановок и проделать расчеты устойчиво-робастных оценок для каждой из них, а затем объединить 6 групп значений, либо взяв среднее арифметическое, либо другой какой-нибудь робастный итог. С 10 компонентами упорядочиваний гораздо больше, и подобную процедуру не провести. Однако можно взять выборку в 10 или 20 случайных перестановок (сравни [Moses L. E. and Oakford R. V. (1963)]). И хотя средние (или другие итоги) 10—20 результатов не вполне свободны от эффектов, связанных с особенностями выборки, но соответствующие разбросы покажут своими величинами, каково влияние выбора начальных порядков на результаты. Мы обращаем внимание на эти возможности, но не рекомендуем пользоваться ими.

Необязательное замечание. Инвариантность к перестановкам или оценивание их роли может кое-кого не удовлетворить. Они захотят инвариантности ко всем вращениям или даже к любым аффинным преобразованиям. Если они удовлетворятся оценкой того, сколь сильны расхождения при вращениях либо при аффинных преобразованиях, они могут получить ее, «вытягивая» несколько случайных вращений или аффинных преобразований, применяя каждое к данной статистике и доводя до числа для получения специальных компонент. Тогда можно будет стабилизировать bar и cob для исходных данных начиная с каждого преобразованного множества начальных компонент. И опять разброс соответствующих результатов покажет, как велики различия, возникающие при выборе порядка компонент.

Следствия. Теперь, раз уж мы не собираемся в этом разделе употребить оценки-аналоги дисперсии и ковариации для расчета линейной регрессии, где они должны использоваться чаще всего (чего, увы, нет), то вычислим оценку аналога, например коэффициента корреляции:

$$\widehat{\text{corr}}\,(x, y) = \frac{\widehat{\text{cob}}\,(x, y)}{\sqrt{\widehat{\text{bar}}\,(x)\,\widehat{\text{bar}}\,(y)}} = \frac{b\,\widehat{\text{bar}}\,(x)}{\sqrt{\widehat{\text{bar}}\,(x)\left[\widehat{\text{bar}}\,(y - bx) + b^2\,\widehat{\text{bar}}\,(x)\right]}} =$$

$$= \pm \frac{1}{\sqrt{1 + \left(\frac{\widehat{\text{bar}}\,(y - bx)}{b^2\,\widehat{\text{bar}}\,(x)} \right)}},$$

где знак перед последним выражением совпадает со знаком b.

Другой распространенный подход — подбор линейных комбинаций исходных компонент, например, методом главных компонент или оценка свойства этих вновь избранных компонент методом канонических корреляций. Тогда мы можем обнаружить, что важно защищаться от пренебрежения инвариантностью. Это можно сделать весьма просто:

1) вычислить $\widehat{\text{bar}}$ и $\widehat{\text{cob}}$ по приведенным формулам;

2) распорядиться ими так, как если бы они были $\widehat{\text{var}}$ и $\widehat{\text{cov}}$ для вычисления новых компонент;

3) вернуться к первому шагу (1), используя новые компоненты вместо исходных;

4) повторить шаг (2), найдя поправку (обычно меньшую) к вычислениям шага (2);

5) использовать результат.

Этот план не позволяет избавиться от всех потерь инвариантности, но защищает от большинства из них.

Пример 1. Случай Эндрюса. На илл. 10.7.1 даны 11 пар значений x и y и соответствующие им значения $\widehat{\text{var}}$, $\widehat{\text{cov}}$, $\widehat{\text{bar}}$ и $\widehat{\text{cob}}$. Обычная парная корреляция равна:

$$\hat{r}\ (\text{классический}) = -0{,}7333.$$

Устойчиво-робастный аналог дает

$$\widehat{\text{corr}}\,(x, y) = 0{,}99992.$$

Мы видим, сколь огромна разница.

Пример 2. «Фокус с десяткой»*. На илл. 10.7.2 приведены 17 троек (x, y, z) и по порядку найдены:

1) столбец 6: остатки y для устойчивой подгонки прямой от x;

2) столбец 8: остатки от z для устойчивой подгонки прямой от x;

3) столбец 9: остатки столбца 8 для устойчивой подгонки прямой от столбца 6;

4) столбец 10: остатки от x для устойчивой подгонки зависимости от константы.

Значения s^2_{b1} и ns^2_{b1} для столбцов 6, 8 и 10 можно вычислить, ибо известно, что бивес-модель даст нуль в каждом из этих столбцов, что более чем хорошая аппроксимация.

Теперь находим b, c и d в выражениях x, $y - bx$ и $z - cy - dx$

*Здесь авторы прибегают к игре слов. В названии игрушки «Jack-in-a-box» (ящичка, из которого при открывании крышки выскакивает фигурка — Джек) первое слово заменено числительным десять (ten), что, видимо, указывает на воспроизведение в десятом столбце таблицы илл. 10.7.2 ее же первого столбца (в пределах ошибки). — *Примеч. ред.*

и $ns^2_{\widehat{\text{bi}}}$ для них (совпадающие с $\widehat{\text{bar}}$ для этих же выражений). Переходим далее, для сравнения с полной таблицей $\widehat{\text{var}}$ и $\widehat{\text{cov}}$, к расчету полной таблицы $\widehat{\text{bar}}$ и $\widehat{\text{cob}}$. Эти таблицы не имеют практически ничего общего.

Почему же «$\widehat{\text{bar}}$-$\widehat{\text{cob}}$»-таблица так отличается от «$\widehat{\text{var}}$-$\widehat{\text{cov}}$»-таблицы? Просто потому, что наш пример аккуратно нагружен восемью внешними точками, которые взяты очень близкими к углам ($\pm$ 1000; $\pm$ 1000; $\pm$ 1000) или вершинам большого куба и будут доминировать в ($\widehat{\text{var}}$, $\widehat{\text{cov}}$)- и слабо проявляться в ($\widehat{\text{bar}}$, $\widehat{\text{cob}}$)-данных. Если они доминируют, то дисперсии велики, приближаются к полумиллиону, и модель имеет форму, очень близкую к сферической (или кубической). Если углы в большом кубе убрать, то $\widehat{\text{bar}}$ станет невелико, около 435, и соответствующая форма будет напоминать плоскую сигару, эллипсоид с полуосями, приближенно равными 20,9; 4,1 и 1,8. Последняя фигура практически вполне определяется положением 9 внутренних точек. (Размеры еще возросли бы, если было бы только 9 точек, поскольку каждое из выражений ns_{bi} возрастает, когда точка становится очень маленькой и, наоборот, когда она растет, становится столь большим, что его уже можно назвать «скачком». Это происходит, когда величина $y—\hat{y}$ приближается к cS или превосходит ее.)

ЗАМЕЧАНИЕ

Такие фокусы с разными столбцами подстерегают вас на каждом шагу. Постройте или тщательно обдумайте упорядоченные точки из илл. 10.7.3 для примера «фокус с десяткой» из илл. 10.7.2. Вы сможете убедиться, что сама конфигурация, построенная для этого примера, явится для вас скорее всего неожиданной.

Пример 3. Альтернативное решение для случая Эндрюса. Вернемся к первому примеру и проведем анализ устойчивости для случая Эндрюса иначе, взяв в качестве остатков $y + x$, а не $y — x$. Илл. 10.7.4 содержит первые несколько шагов арифметических вычислений, приводящих к угловому коэффициенту — 1,0023 (что ближе к данным, чем — 1,0056 в классическом анализе). Итак, мы видим, что для этого набора данных можно использовать одну и ту же устойчивую процедуру подбора прямой линии для получения двух совершенно разных прямых: одной с коэффициентом 0,9942 (илл. 10.7.1), а другой с — 1,0023 (илл. 10.7.4).

Часто можно услышать, что прискорбно иметь два столь различных ответа в самом простом случае и следовало бы добиваться одного единственно верного. Однако если нанести данные с илл. 10.7.1 на график даже грубо, то станет ясно, что оба ответа разумны и что каждый может оказаться нужным нам в конкретных условиях. Какой из них взять, зависит от нашего отношения к последней точке (25, — 25). Все, что нам действительно нужно в этой ситуации, представлено обоими решениями и словами «если можешь выбирать, то прежде семь раз отмерь».

К несчастью, никто еще не написал программу для нахождения всех правдоподобных и различающихся решений и никто не подскажет нам, какое из решений выбрать*.

10.8. ЗАКЛЮЧИТЕЛЬНОЕ ЗАМЕЧАНИЕ

Мы сейчас видели, что устойчивые и робастные методы применимы для решения самых разнообразных вопросов. Чем шире использование, тем более изысканными становятся методы анализа и тем неизбежнее новое расширение. Однако уже известно не мало для того, чтобы сделать устойчиво-робастные методы обычными и естественными.

РЕЗЮМЕ. УСТОЙЧИВЫЕ И РОБАСТНЫЕ МЕТОДЫ

И устойчивость, и робастность к эффективности суть идеи, важные как для практики анализа данных, так и для его теории.

Мы считаем разумным использовать медиану для прикидок, а бивес-оценки — для точной работы, оставляя среднее арифметическое для весьма особых случаев.

Когда у нас есть основания считать, что «хвосты» разбросаны *меньше*, чем в гауссовском распределении, почти наверное, то выбор оценки ставит перед нами новые вопросы, которые здесь не обсуждаются.

Классические методы оценки масштаба (разброса) не очень робастны по эффективности. МАО (медиана абсолютных отклонений от медианы) или H-разброс (или межквартильный размах) — вот лучшие среди многих.

Несколько более сложная мера масштаба (разброса), обозначаемая нами «ns^2_{Bi}», и робастна к эффективности, и весьма устойчива.

Используя метод Гросса и комбинируя бивес с s^2_{Bi}, можно получить устойчивые робастные к эффективности и к предпосылкам доверительные интервалы.

Если нужны аналоги дисперсий и ковариаций для многомерных (многокомпонентных) данных, то можно следовать по этапам:

1) найти компоненты, которые мы хотели бы обрабатывать как ортогональные (получаемые как остатки от последовательных робастных и устойчивых регрессий);

2) оценить разброс для каждого (мы предлагаем использовать ns^2_{Bi});

3) распространить обычные тождества для дисперсий и ковариаций на определение $\widehat{\text{bar}}$ и $\widehat{\text{cob}}$ при всех линейных комбинациях (и исходных, и новых компонент).

Мы смогли оценить зависимость результирующих $\widehat{\text{bar}}$ и $\widehat{\text{cob}}$ от выбора компонентов, с которых начинали многокомпонентный анализ.

---

*Заключение этого параграфа удивительно созвучно с утверждениями, которые неоднократно формулировал В. В. Налимов (см., например: Налимов В. В. Планирование эксперимента. Найдут ли новые проблемы новые решения? — Журнал ВХО им. Д. И. Менделеева, 1980, т. 25, № 1, с. 3—4). — *Примеч. ред.*

Если поиск подстановок для дисперсий и ковариаций приведет к выбору каких-то новых компонентов, то мы можем повторить весь процесс получения bar-cob начиная с первого множества новых компонентов так, как будто они и были исходными. Это значительно уменьшит любую существенную зависимость окончательных ответов от того, какие начальные компоненты мы взяли.

БИБЛИОГРАФИЯ

Gross A. M. (1976). Confidence-interval robustness with long-tailed symmetric distribution. — J. Amer. Statist. Assoc., 71, 409—416.

Lax D. A. (1975). An interim report of a Monte Carlo study of robust estimators of width. — Technical Report No. 93, (Series 2). Department of Statistics, Princeton University.

Moses L. E. and Oakford R. V. (1963). Tables of Random Permutations. Stanford, California, Stanford University Press.

ИЛЛЮСТРАЦИИ

Иллюстрация 10.3.1

Устойчивость и робастность эффективности некоторых оценок положения

Оценка	Объем выборки	Устойчивость?	Гауссова эффективность	Эффективность при разбросанных «хвостах»	Робастность к эффективности
Арифметическое среднее	Малый	Нет	100%	Плохая	Плохая
	Большой	Нет	100%	Очень плохая	Очень плохая
Медиана	Малый	Да	Высокая	Достаточно высокая	Высокая
	Большой	Да	62%	Достаточно высокая	Умеренная
Бивес-оценка при $c=6$ или 9	Малый	Разумно	Неплохая	Достаточно высокая	Неплохая
	Большой	Да	>90%	>90%	Высокая

Иллюстрация 10.7.1

Случай Эндрюса: пример устойчивой корреляции

А. Данные

x	y	$y-x$	(4)	$x-5$	x—4,64
0,02	0,04	0,02	0,0093	—4,98	—4,62
0,99	1,03	0,04	0,0349	—4,01	—3,65
2,01	1,97	—0,04	—0,0391	—2,99	—2,63
2,98	2,96	—0,02	—0,0135	—2,02	—1,66
4,03	3,97	—0,06	—0,0474	—0,97	—0,61
5,01	4,98	—0,03	—0,0117	0,01	0,37
6,05	6,07	0,02	0,0443	1,05	1,41
6,98	7,03	0,05	0,0797	1,98	2,34

x	y	$y-x$	(4)	$x-5$	$x-4{,}64$
8,07	8,00	—0,07	—0,0340	3,01	3,37
9,03	8,96	—0,07	—0,0284	4,03	4,12
25,00	—25,0	—50,00	—49,8658	20,00	20,36
H		0,02	0,0221	2,50	2,86
H		—0,05	—0,0366	—2,50	—2,14
S		0,035	0,0294	2,50	2,50
cS		0,315	0,2641	22,50	22,50

Б. Результаты промежуточных вычислений

Бивес-зависимость $(y-x)$ от x: 0,0108 — 0,0058x. Столбец (4) : $(y-x)$ — — (0,0108 — 0,0058x) = y — 0,0108 — 0,9942x.

s^2_{bi} (для столбца (4), используя $cS = 0{,}2641$) = 0,00019136, ns^2_{bi} (для того же) = = 0,0021050 = $\widehat{\text{bar}}\,(y-bx)$.

Бивес-зависимость $(x-5)$ от константы: 0,36.

Столбец (5) : $(x-5) - (-0{,}36) = x - 4{,}64$.

s^2_{bi} (для $x - 4{,}64$, используя $cS = 22{,}5$) = 1,13587,

ns^2_{bi} (для того же) = 12,4945 = $\widehat{\text{bar}}\,x$.

$$\widehat{\text{corr}} = \left\{1 + \frac{0{,}0021050}{(0{,}9942)^2\,(12{,}4945)}\right\}^{-1/2} = 0{,}99992,$$

$$\hat{r}\ (\text{классическая}) = \frac{(-46{,}7004)}{\sqrt{(46{,}4410)\ (87{,}3414)}} = -0{,}7333.$$

Иллюстрация 10.7.2
Пример «фокус с десяткой», неупорядоченные данные

А (см. на с. 216).

Б. Подбор прямых для столбцов таблицы из А (везде используется бивес-оценка)

Подбор прямой для зависимости $(y-x)$ от x: 0,079 — 0,001261x.

Столбец (5): $y - x - (0{,}079 - 0{,}001261x) = y - 0{,}079 - 0{,}998739x$.

Подбор прямой для (5) от x: 0,001 — 0,000067x.

Столбец (6): столбец (5) — (0,001 — 0,000067x) = y — 0,080 — 0,998672x.

Подбор прямой для зависимости $(z-x)$ от x: —0,200 — 0,001235x.

Столбец (7): $z - x - (-0{,}200 - 0{,}001235\,x) = z + 0{,}200 - 0{,}998765x$.

Подбор прямой для зависимости (7) от x: —0,039 — 0,000003x.

Столбец (8): столбец (7) — (—0,039 — 0,000003x) = z + 0,239 — 0,998762x.

Подбор прямой для зависимости (8) от (6): —0,007 + 0,000132 (столбец 6).

Столбец (9): столбец (8) — (—0,007 + 0,000132(y — 0,080 — 0,998672x)) = = z + 0,246 — 0,000132y — 0,998630x.

Подбор константы для x: 0,117.

Столбец (10): x — 0,117.

В. Вычисление $\widehat{\text{bar}}$ и $\widehat{\text{cob}}$

Линейные комбинации:

	$\widehat{\text{bar}}$
x	434,914
$y-bx$	3,0577
$z-cy-dx$	17,0556,

(окончание на с. 217)

где $b = 0{,}998672$, $c = 0{,}000132$, $d = 0{,}998630$, так что

$$\widehat{\text{bar}}\,y = 3{,}0577 + (0{,}998672)^2\,(434{,}914) = 436{,}758;$$

$$\widehat{\text{cob}}\,(x, y) = (0{,}998672)\,(434{,}914) = 434{,}336;$$

$$\widehat{\text{bar}}\,z = 17{,}0556 + (0{,}000132)^2\,(3{,}0577) + (0{,}998672^*)^2\,(434{,}914) = 450{,}815;$$

$$\widehat{\text{cob}}\,(x, z) = (0{,}998672^*)\,(434{,}914) = 434{,}336;$$

А. Данные, остатки и разбросы

	x	y	z	$y-x$	(5)	(6)	$z-x$	(7)	(8)	(9)	(10)
	1001	999	—1001	—2	—0,817	—0,752	—2002	—2000,564	—2000,522	—2000,514	1000,883
	999	—1001	—1000	—2002	—2000,820	—2000,820	—1999	—1997,566	—1997,524	—1997,252	998,883
	15	14	33	—1	—1,060	—1,060	18	18,219	18,258	18,266	14,883
	—7	—6	—17	1	—0,929	0,927	—10	—9,808	—9,769	—9,762	—7,117
	1000	1001	999	1	2,182	2,247	—1	0,435	0,477	0,484	999,883
	—1001	1001	—1000	2002	2000,658	2000,590	1	—0,035	0,001	—0,255	—1001,017
	22	21	23	—1	—1,052	—1,052	1	1,227	1,266	1,274	21,883
	12	10	11	—2	—2,064	—2,065	—1	—0,785	—0,746	—0,738	11,883
	1	—1	0	—2	—2,078	—2,078	—1	—0,789	—0,760	—0,751	0,883
	—10	—11	—12	—1	—1,094	—1,094	—2	—1,812	—1,773	—1,765	—10,117
	—23	—22	—21	1	0,892	0,889	2	2,172	2,111	2,218	—23,117
	1001	—1000	999	—2001	—1999,817	—1999,752	—2	—0,564	—0,523	—0,250	1000,883
	—1001	—999	—1000	2	0,658	0,590	1	—0,035	0,001	0,009	—1001,117
	6	8	16	2	1,928	1,927	10	10,208	10,247	10,254	5,883
	—155	—14	—33	1	0,902	0,900	—18	—17,818	—17,779	—17,772	—15,117
	—1000	1001	999	2001	1999,660	1999,592	1999	1997,964	1998,022	1997,746	—1000,117
	—1001	—999	1000	2	0,658	0,590	2001	1999,963	2000,001	2000,009	—1001,117
H	22			2	0,902	0,902	2	2,170	2,111	2,218	21,883
H	—23			—2	—1,092	—1,094	—2	—1,814	—1,773	—1,765	—23,117
S	22,5			2	0,997	0,997	2	1,992	1,992	1,992	22,5
cS	202,5			18	8,973	8,973	18	17,928	17,928	17,928	202,5
s^2_{bI}						0,1799				1,0033	25,583
ns^2_{bI}						3,0577				17,0556	434,914

$\widehat{\text{cob}}\,(y, z) = (0{,}00132)\,(3{,}0577) + (0{,}998672)\,(0{,}998672^*)\,(434{,}914) = 433{,}760.$

Г. Сравнение $\widehat{\text{var}}/\widehat{\text{cov}}$ и $\widehat{\text{bar}}/\widehat{\text{cob}}$

	$\widehat{\text{var}}$ и $\widehat{\text{cov}}$				$\widehat{\text{bar}}$ и $\widehat{\text{cob}}$	
500362,43	—81,49	—97,70	x	434,915	434,336	434,336
	500221,61	332,59	y		436,758	433,760
		500217,57	z			450,815

* Здесь $d + bc = 0{,}998672$, а не $b = 0{,}998672$.

Иллюстрация 10.7.3
Пример «фокус с десяткой», упорядоченные данные

x	y	z	x	y	z
1000	1000	—1000	0	0	0
1000	—1000	—1000	—11	—11	—11
14	14	32	—22	—22	—22
—7	—7	—16	1000	—1000	1000
1000	1000	1000	—1000	—1000	—1000
—1000	1000	—1000	7	7	16
22	22	22	—14	—14	—32
11	11	11	—1000	1000	1000
			—1000	—1000	1000

Иллюстрация 10.7.4
Другой взгляд на пример Эндрюса

x	y	$y+x$	(4)	(5)	(6)
0,02	0,04	0,06	—7,50	—8,06	—8,118
0,99	1,03	2,02	—5,54	—6,10	—6,155
2,01	1,97	3,98	—3,58	—4,14	—4,193
2,98	2,96	5,94	—1,62	—2,18	—2,230
4,03	3,97	8,00	0,44	—0,12	—0,169
5,01	4,98	9,99	2,43	1,87	1,824
6,05	6,07	12,12	4,56	4,00	3,956
6,98	7,03	14,01	6,45	5,89	5,848
8,01	8,00	16,01	8,45	7,89	7,850
9,03	8,96	17,99	10,43	9,87	9,833
25,00	—25,00	0	—7,56	—8,12	—8,121
H		13,06	5,50	4,94	4,902
H		3,00	—4,56	—5,12	—5,174
S		5,03	5,03	5,03	5,038
cS		45,27	45,27	45,27	45,34

Бивес $y + x$: 7,56.
Столбец (4): $y + x - 7{,}56$.
Бивес-оценка (4) (использующая $cS = 45{,}27$): 0,56.
Столбец (5): $(y+x - 7{,}56) - 0{,}56 = y+x - 8{,}12$.
Бивес-оценка зависимости (5) от x (использующая $cS = 45{,}2$): $0{,}058 + 0{,}0023x$.
Столбец (6): $(y+x - 8{,}12) - (0{,}058 - 0{,}0023x) = y - 1{,}0023x - 8{,}062$.

Глава 11 ● НОРМИРОВАНИЕ ДАННЫХ ДЛЯ СРАВНЕНИЙ

Для откликов, представимых в долях или в процентах, мы прежде всего хотели бы сопоставить эффекты различных воздействий (обработок). Когда же все доли сосчитаны и сравнены, мы обычно ищем новой информации. Например, найдя, что одна форма обучения приводит к более высокой грамотности учащихся, чем другая, скептически настроенный исследователь немедленно спросит о правомерности сравнения тестируемых групп. Аналогично, если новый антисептик, применяемый при полостных операциях, демонстрирует меньшую токсичность, чем старый, то мы хотели бы быть уверенными в том, что пациенты, на которых опробовали новый препарат, не были более здоровыми с самого начала.

Использование рандомизированных планов эксперимента в значительной степени снимает остроту таких вопросов. Даже то, что случай может сыграть с нами злую шутку, мы пытаемся обратить в благо. Наиболее важно обеспечение «одинаковости» групп при пассивном эксперименте или плохой управляемости опытов. Сравнения часто можно улучшить за счет методов нормирования, описываемых в этой главе.

И хотя здесь рассматриваются лишь доли или проценты, возникают такие же вопросы и иногда требуются похожие решения, полезные в случаях, когда мы имеем дело со средними или медианами.

11.1. ПРОСТЕЙШИЙ СЛУЧАЙ

Пример 1. Два варианта, дихотомические* совокупности. Начнем с простейшей задачи. Используются два варианта (например, лечения): один для первой группы, другой — для второй; кроме того, каждая группа распадается на два типа (людей, например) — тех, что *легко* достигают успеха («выживают», например), и тех, что достигают успеха *с трудом*. Причем классификация на «легких» и «трудных» определяется по критериям, *не связанным с самими* вариантами. Отчет о результатах исследования показывает, что вариант I дает 60% успехов, а вариант II — 44% (илл. 11.1.1А). В демографической статистике эти числа назывались бы «примерными долями успехов». Они действительно грубы, приблизительны, поскольку не учтена срав-

*Т. е. смесь из двух однородных совокупностей. — *Примеч. пер.*

нимость групп, подвергнутых разным вариантам воздействия, и поскольку они, в сущности, игнорируют дополнительную информацию о классификации внутри типа.

По данным илл. 11.1.1А примерная доля успехов, например, варианта I равна:

$$\frac{0{,}7\cdot 800 + 0{,}2\cdot 200}{1000} = \frac{600}{1000} = 0{,}60, \text{ или } 60\,\%.$$

Мы замечаем одно тревожное обстоятельство: хотя примерная доля успехов варианта I выше, вариант II оказывается лучше как для «легких», так и для «трудных» (см. числа в илл. 11.1.1В). В этом примере сравнение приближенных долей успехов ведет к заблуждению насчет относительной ценности вариантов. В варианте II получилось почти на четверть меньше успехов, поскольку ему «досталось» гораздо больше трудных случаев.

Отметив это, можем теперь подумать о том, что же с этим делать. Конечно, совокупность четырех результатов (в %) хорошо подытоживает ситуацию, и, зная эти числа, можно отвечать на разные вопросы. Вот четыре из них:

1) каким было бы сравнение долей, если бы оба варианта испытывались в группах того же состава, что был в группе варианта I?

2) тот же вопрос, но при составе группы варианта II;

3) что можно сказать насчет усредненной совокупности (половина из варианта I и половина из варианта II)?

	Распределение объектов по вариантам	
	I	II
«Легкие»	450	450
«Трудные»	550	550

4) что можно сказать насчет произвольной совокупности? Сведем ответы в табличку.

	Процент успехов		Разность II—I
	вариант I	вариант II	
Ответы:			
на вопрос 1	60	72	12
на вопрос 2	25	44	19
на вопрос 3	42,5	58	15,5

Вопрос 4. Переходя к 4-му вопросу, мы из илл. 11.1.1В видим, что различия где-то в пределах между 10% (80—70) и 20% (40—20)

в пользу варианта II в зависимости от сочетания «легких» и «трудных». Для групп из одних «легких» вариант II дает преимущество в 10%, а для групп из одних «трудных» — в 20%. В группах с равным количеством тех и других преимущество оказывается равным 15% (60—45), как раз посередине между 10% и 20%.

Какую же из всех этих совокупностей стоит выбрать? Конечно, ответ зависит от нашей цели, которая может быть и просто обзором возможностей (что мы сейчас и делали) для определения чувствительности результатов к изменению пропорций в смеси. В этом смысле мы сделали уже вполне достаточно. Однако хотелось бы описать и какую-нибудь конкретную ситуацию.

Для этого надо получить лучшую оценку вида совокупности, с которой мы собираемся работать, и считать на ее основе, быть может, с учетом возможных отклонений от нее. Назовем эту выбранную совокупность *эталонной*, а это значит, что сравнения для нее нормированы. Любую смесь можно принять за эталон. Теперь возьмем второй пример.

Пример 2. Перекрестные проценты. Пусть доли успехов для «легких» и «трудных» типов таковы:

	Успехи по вариантам, %	
	I	II
«Легкие»	70	80
«Трудные»	40	10

Тогда, если все испытуемые будут «легкими», вариант II даст преимущество в 10%, но для группы из одних «трудных» он уже проигрывает 30%, и таким образом обнаруживается, что не все равно, какой взять эталон. В примере 1 мы просто не встретились с этим случаем, но теперь, если возможен лишь один вариант, то наш выбор будет зависеть от той пропорции, которая существует в планируемой совокупности. При более 75% «легких» мы выберем вариант II, а если их будет меньше — вариант I.

Конечно, если бы можно было задать класс «объектов» заранее и взять самый успешный вариант, это было бы самым лучшим. Попади мы в столь благоприятные обстоятельства, интерес к эталонам мгновенно бы пропал.

Нет нужды говорить, что такой подход распространяется на несколько вариантов и несколько классов. Обычно он называется *прямым нормированием*.

Результаты прямого нормирования часто рассчитывают, как мы только что поступили, не обращая внимания на показанную изменчивость, но все же какая-нибудь индикация изменчивости желательна. Выше, в вопросе 1 примера 1, мы имели расчетную долю успеха:

$$p_{\text{стд}} = 0{,}8p_{\text{лег}} + 0{,}2p_{\text{труд}}.$$

Считая 0,8 и 0,2 постоянными, используя равенство

$$\widehat{\operatorname{var}}\, p = \frac{pq}{N}$$

(где $q = 1 - p$, N — объем группы) и обозначая через u и v наблюденные доли, получим

$$\widehat{\operatorname{var}}\,(0{,}8u + 0{,}2v) = 0{,}64\,\widehat{\operatorname{var}}\, u + 0{,}32\,\widehat{\operatorname{cov}}\,(u, v) + 0{,}04\,\widehat{\operatorname{var}}\, v,$$

что позволяет нам вычислять искомые дисперсии.

Пусть $p_{\text{стд,I}}$ — наблюдаемая доля успехов, когда к эталону (0,8 «легких» и 0,2 «трудных») применяются доли успехов варианта I. Аналогично $p_{\text{стд,II}}$ — результат применения варианта II к тому же эталону. Тогда $(p_{\text{стд,I}} - p_{\text{стд,II}})$ — разность наблюдений и хотелось бы найти ее дисперсию. Предполагая нулевую корреляцию между данными, получим

$$\widehat{\operatorname{var}}\, p_{\text{стд. I}} = 0{,}64\,\frac{(0{,}70)\,(0{,}30)}{800} + 0{,}04\,\frac{(0{,}20)\,(0{,}80)}{200} =$$

$$= 0{,}000200 = (0{,}014)^2 = (1{,}4\%)^2;$$

$$\widehat{\operatorname{var}}\, p_{\text{стд. II}} = 0{,}64\,\frac{(0{,}80)\,(0{,}20)}{100} + 0{,}04\,\frac{(0{,}40)\,(0{,}60)}{900} =$$

$$= 0{,}001035 = (0{,}032)^2 = (3{,}2\%)^2;$$

$$\widehat{\operatorname{var}}\,(p_{\text{стд. I}} - p_{\text{стд. II}}) = 0{,}64\left(\frac{(0{,}70)\,(0{,}30)}{800} + \frac{(0{,}80)\,(0{,}20)}{100}\right) +$$

$$+ 0{,}04\left(\frac{(0{,}20)\,(0{,}80)}{200} + \frac{(0{,}40)\,(0{,}60)}{900}\right) = 0{,}001235 = (0{,}035)^2 = (3{,}5\%)^2$$

или проще, поскольку между вариантами нет взаимной корреляции,

$$\widehat{\operatorname{var}}\,(p_{\text{стд. I}} - p_{\text{стд. II}}) = (1{,}4\%)^2 + (3{,}2\%)^2 = (3{,}5\%)^2.$$

Так что разность в 12% (в вопросе 1) имеет стандартную ошибку 3,5% и может рассматриваться лишь как указатель знака различия, ибо, как известно, эту ошибку надо еще умножить на два.

11.2. ПРЯМОЕ НОРМИРОВАНИЕ

Допустим, что заданы J вариантов и K типов. В эталонной совокупности мы имеем доли W_k, $k=1, 2, \ldots, K$, которые удобно считать весами, так что $\Sigma W_k = 1$. Доли успехов равны P_{jk} соответственно для j-го варианта и k-го типа. В примере 1 типами были «легкие» и «трудные», а P_{jk} даны внизу таблицы илл. 11.1.1.

Тогда D_j — доля успехов j-го варианта в эталонной совокупности есть

$$D_j = \frac{\sum_{k=1}^{K} W_k P_{jk}}{\sum_{k=1}^{K} W_k},$$

что упрощается при $\Sigma W_k = 1$, переходя в

$$D_j = \sum_{k=1}^{K} W_k P_{jk}.$$

Пример 3. Смертность в штатах Мэн и Южная Каролина. Теодор Вулси (T. D. Woolsey) в описанных им откорректированных сводках данных о смертности приводит характерный пример того, как грубое определение коэффициентов смертности порождает ошибочные результаты при сравнении смертности в штатах Мэн и Южная Каролина в 1930 г. На илл. 11.2.1 дана таблица основных данных. Соль примера в том, что он похож на наш 1-й пример: Южная Каролина имеет более высокую смертность, чем Мэн, во всех возрастных группах, кроме одной, но даже в ней коэффициенты смертности примерно равны. Несмотря на это, приближенный коэффициент смертности в целом по Южной Каролине оказался меньше. Причина же в том, что население Мэн было вообще более старым, чем Южной Каролины. Мэн имеет показатель смертности примерно 1390,8 на 100 тыс. жителей, тогда как Южная Каролина — лишь 1288,8.

Для лучшего сравнения общих коэффициентов смертности Вулси не стал выбирать возрастное распределение населения какого-либо из этих штатов. Вместо этого он взял как эталон возрастное распределение в целом по Соединенным Штатам и соотнес с ним показатели смертности для каждого возрастного интервала в Мэн ($P_{\text{Мэн}k}$) и в Южной Каролине ($P_{\text{Ю.К}k}$). Эталон он нормировал на 1 млн. человек, распределенных по возрастным группам. Штат «соответствует» варианту, возрастная группа — типу. Вычисления приведены на илл. 11.2.2*.

11.3. ТОЧНОСТЬ РЕЗУЛЬТАТОВ ПРЯМОГО НОРМИРОВАНИЯ

Нам нужна какая-нибудь идея об оценке возможной точности показателей смертности, установленных методом прямого нормирования. Однако нам не нужна точность, максимально возможная в принципе. Итак, будем развивать приближенный (определяется ниже) и прикидочный расчеты, так же как и относительно точный.

Пусть N — объем совокупности и p — доля умерших. Тогда, предполагая биномиальное распределение и пользуясь тем, что p мало, $q = 1 - p$ близко к 1, мы получим

$$\text{var (фактическая смертность)} = Npq \approx Np \approx \text{фактическая смертность}.$$

Метод 1. Прикидочный биномиальный. Для Южной Каролины фактическая смертность равна 22401. Эта оценка как для среднего Np, так и для дисперсии Npq, поскольку последняя близка к Np.

Для Мэн соответствующие оценки равны 11082.

*Здесь авторы демонстрируют возможности своих подходов в еще одной широкой области приложений — демографии. Желающие углубиться в эту область могут начать, например, с работ: Статистический анализ в демографии. М., Статистика, 1980; Имитационные модели в демографии. М., Статистика, 1980. — *Примеч. ред.*

Метод 2. Нормированный биномиальный (приближенная стандартная ошибка). Если D — число смертей, то пусть KD — нормированная смертность, где K рассматривается как константа. Тогда среднее равно KNp и дисперсия $K^2Npq \approx K^2Np$. Величину K можно определить равенством

$$K = \frac{\text{нормированная смертность}}{\text{фактическая смертность}} \approx \frac{\text{нормированная смертность}}{Np}.$$

С другой стороны, улучшенная аппроксимация дисперсии нормированной смертности есть

$$\frac{(\text{нормированная смертность})^2}{\text{фактическая смертность}}, \text{ так как } q \approx 1.$$

Если работать более аккуратно, то надо бы умножить на q.

Для Южной Каролины, принимая $q = 1$, имеем

$$\frac{(17163)^2}{22401} = 13150;$$

для Мэн соответствующая оценка равна:

$$\frac{(12033)^2}{11082} = 13066.$$

Используя эти оценки дисперсий для нормированной разности, найдем

$$\text{Южная Каролина МИНУС Мэн} = 17163 - 12033 = 5130,$$

оценку числа избыточных смертей в Южной Каролине, и получим оценку дисперсии (этой оценки избытка)

$$13150 + 13066 = 26216 = (162)^2.$$

Так что разность составляет примерно 30 стандартных отклонений, и маловероятно, чтобы в таком случае потребовалось более точное стандартное отклонение, однако ради иллюстрации метода мы двинемся дальше.

Метод 3. Биномиальный стратифицированный. Для улучшения расчета нужно, собственно, применить формулу

$$\frac{(\text{нормированная смертность})^2}{\text{фактическая смертность}}$$

к каждому типу (здесь — возрастной группе) отдельно и просуммировать. Для Южной Каролины:

$$\frac{(1916)^2}{4905} + \frac{(150)^2}{446} + \frac{(164)^2}{410} + \ldots,$$

или

$$748{,}4 + 50{,}4 + 65{,}6 + 176{,}7 + 300{,}7 + 1043{,}8 + 1257{,}7 + \\ + 1928{,}0 + 2653{,}0 + 3560{,}2 + 3414{,}2 = 15198{,}7,$$

а для штата Мэн:

$$1758{,}0 + 154{,}1 + 150{,}2 + 285{,}5 + 474{,}4 + 973{,}3 + 1043{,}6 + \\ + 1666{,}6 + 1811{,}1 + 2622{,}6 + 2462{,}1 = 13401{,}5.$$

Таким образом, наша приближенная оценка дисперсии разности равна теперь

$$15198,7+13401,5=28600,2=(169)^2$$

и менее чем на 5% превышает оценку в методе 2.

Учет q. Для того чтобы еще раз уточнить оценку, мы должны вспомнить о q в Npq и использовать формулу

$$\frac{(\text{нормированная смертность})^2}{\text{фактическая смертность}}\ (1-\text{вероятность смерти})$$

для каждого типа. Так, для первых двух типов Южной Каролины получим

$$\frac{(1916)^2}{4905}\ (1-0,2392)=730,5 \text{ и } \frac{(150)^2}{446}\ (1-0,00185)=50,3;$$

в целом по Южной Каролине имеем

$$730,5+50,3+65,5+175,9+298,8+1034,7+1242,1+$$
$$+1889,6+2565,1+3341,4+2931,6=14325,5.$$

Аналогично для штата Мэн

$$1721,9+153,8+150,0+284,9+472,6+969,5+1037,9+$$
$$+1648,5+1774,2+2485,7+2126,1=12825,1.$$

Это для оценки дисперсии разности дает

$$14325,5+12825,1=27150,6=(165)^2.$$

СРАВНЕНИЕ ОЦЕНОК

В этом примере несущественно, какую ошибку при разности в смертностях, равной 5130, для эталонной совокупности мы берем: $\pm$ 162, $\pm$ 169, $\pm$ 165, тем более, что надо еще считаться с ошибкой биномиальной аппроксимации. Например, вариация от года к году в разности смертностей между Южной Каролиной и Мэн может вносить вклад в нашу вариацию. Но в других более крайних случаях различие между оценками и точной величиной Npq может оказаться важным.

Когда мы сталкиваемся с малым числом умерших или живущих, иной раз приходится менять формулы; особенно в крайнем случае, когда наблюденное число умерших (или живущих) равно нулю. В таких случаях величина pq в старых формулах исчезает, и если мы не учтем этого, то получим расчетную дисперсию, равную нулю. Но эту оценку очень хотелось бы повысить, ибо ясно, что дисперсия здесь никак не нуль. Мы не знаем в точности, что же делать, хотя есть подход, предложенный Сазерлендом, Холлендом и Файнбергом [Sutherland M., Holland P. and Fienberg S. E. (1974)] и призванный помочь в особо сложных случаях. Мы предлагаем здесь метод, который не использует значений наших факторов. Однако он не состоятелен, поскольку нет гарантии, что коррекция будет ближе к истинной величине, чем предварительная прикидка.

Предлагаемая коррекция состоит в добавке 1/6 к расчету каждой категории

$$p^* = \frac{\text{фактически умершие} + 1/6}{\text{объем ячейки} + 1/3} \quad \text{и} \quad q^* = \frac{\text{фактически живущие} + 1/6}{\text{объем ячейки} + 1/3},$$

где фактические числа умерших и живущих берутся для той ячейки, для которой вычисляется дисперсия, что приводит к формуле

$$\text{оцениваемая дисперсия} = \left(\frac{\text{нормированная совокупность}}{\#\ \text{испытуемые}}\right)^2 \times$$

$$\times\ (\text{фактически умершие} + 1/6) \times \left(1 - \frac{\text{фактически умершие} + 1/6}{\text{объем ячейки} + 1/3}\right),$$

которая относится ко всем ячейкам, включая и те, где не было умерших.

11.4. ТРУДНОСТИ ПРЯМОГО НОРМИРОВАНИЯ

Начав с того, что неопределенностей в величинах P нет, мы пришли к некоторому расчету возможных неопределенностей. В примере 3 показатели смертности для каждого возраста, называемые возрастными показателями смертности, определены довольно хорошо. Однако часто наши показатели основаны на исследованиях, и опыт показывает, что отдельные ячейки (пары вариант—тип) могут содержать малое число случаев, так что величины P будут оцениваться со значительной неопределенностью. Когда же такая неопределенность сочетается со значительным весом ячейки, результаты могут быть чрезмерно чувствительны к особенностям выбросов.

Пример 4. Плохо определенные доли. Давайте возьмем пример, в котором этот эффект легко увидеть. Если мы работаем с показателями смертности, связанными с определенной хирургической операцией, проводимой двумя способами (вариантами), то мы можем рассмотреть три класса пациентов: с отличным, удовлетворительным и плохим состоянием здоровья (О-, У- и П-типы). Распределение пациентов показано на илл. 11.4.1.

Заметим, что умер лишь один пациент в ячейке [П-тип, способ I]. Если бы он выжил, то доля в ней была бы О/1 = О, показатель смертности на 1000 сразу же упал бы с 26,48 до 1,47, и способ I выглядел бы явно предпочтительнее для нормированной совокупности. Вместо того, чтобы быть в 6 раз хуже, он был бы в 3 раза лучше способа II.

Как это зафиксировать? Но именно такие вещи как раз неуловимы. В этом примере мы можем, конечно, попробовать разузнать, не страдал ли этот единственный умерший пациент чем-нибудь настолько ужасным, что был *обречен* с самого начала и, возможно, был на «безнадежном» краю в типе «плохих». Но в этой же категории могли находиться умершие и при способе II. Подобные трудности требуют тщательного обсуждения и лежат за гранью статистического подхода.

Этот эффект появляется тогда, когда малые вероятности по воле случая получают большие оценки, завышенные иногда в 10 или 100 раз, и притом имеют значительный вес в нормированной со-

вокупности. В таких ситуациях метод прямого нормирования может давать «дикие» результаты. Следовательно, важно ревизовать все методы прямого нормирования применительно к столь трудным обстоятельствам. Одна из серьезнейших опасностей состоит в том, что число ячеек может быть очень велико, а вычисления происходят без выдачи промежуточных результатов, что может ввести в заблуждение. Всегда, как только метод прямого нормирования привел к «дикому» результату, надо проверить, не вызвано ли это малой вероятностью для ячейки с большим весом. Если же ничего такого «дикого» не попалось, то мы можем удивиться тому, что его нет. И проверка, и удивление равно помогают нам в поисках подходящей оценки стандартного отклонения.

Если мы рассчитаем стандартную ошибку по измененной формуле, используя p^* и q^*, то найдем для ожидаемой смертности в ячейке [«плохое» — способ I) следующие значения:

$$p^* = \frac{7/6}{4/3} = \frac{7}{8} \text{ и } q^* = \frac{1/6}{4/3} = \frac{1}{8},$$

$$\text{ожидаемая дисперсия} = \frac{(500)^2}{1} \cdot \frac{7}{6}\left(1 - \frac{7}{8}\right) = (191)^2.$$

Это вносит вклад $\pm$ 9,55 в оценку стандартного отклонения, равную 26,48, что достаточно, чтобы предостеречь нас. (Разность превышает свое стандартное отклонение всего в 2,3 раза.)

Если нам нужна машинная программа, которая распечатывает (а) оценку дисперсии суммарной ожидаемой смертности, (б) наибольший вклад отдельной ячейки, то их надо как-то обозначить, так как они равны:

$$36505 = (191)^2 \text{ и } 36458 = (191)^2.$$

Здесь

$$36505 = \frac{(500)^2}{1}\left(\frac{7}{6}\right)\left(1 - \frac{7}{8}\right) + \left(\frac{5000}{3000}\right)^2 9\frac{1}{6}\left(1 - \frac{9\frac{1}{6}}{3000\frac{1}{3}}\right) +$$

$$+ \left(\frac{14500}{10000}\right)^2 \left(10\frac{1}{6}\right)\left(1 - \frac{10\frac{1}{6}}{10000\frac{1}{3}}\right) = 36458{,}33 + 25{,}39 + 21{,}35.$$

(Если мы добавим 1/2, как иногда делается, вместо 1/6, то найдем

$$p^* = \frac{3/2}{2} = \frac{3}{4} \text{ и } q^* = 1 - p^* = \frac{1}{4},$$

$$\text{ожидаемая дисперсия} = \frac{(500)^2}{1}\left(\frac{3}{2}\right)\left(1 - \frac{3}{4}\right) = (306)^2,$$

что увеличивает оцениваемый вклад в стандартную ошибку почти в $\sqrt{3}$ раз. Разница между использованием 1/6 и 1/2, хотя иногда и велика, редко имеет значение. Вот если бы мы прибавили 0,01 вместо лю-

бого из этих чисел, то оцениваемая дисперсия стала бы $(49)^2$, и это уже могло иметь значение.)

Ни одно из этих замечаний или расчетов не снимает принципиальной неопределенности, скрытой в подобных методах, из-за существования одной ячейки с одним наблюдением, которому в силу этого придается слишком большой вес. Конечно, может быть и еще хуже — наблюдения могут отсутствовать вовсе.

Эта особенность метода прямого нормирования привела к развитию другого подхода, к рассмотрению которого мы сейчас и приступаем.

11.5. КОСВЕННАЯ НОРМАЛИЗАЦИЯ

Для прямого нормирования нам приходится либо искать эталон «на стороне», либо конструировать его как некое среднее из тех двух совокупностей, что есть в наличии. Мы используем свои доли для каждого варианта эталона. При косвенном нормировании мы идем в обход. Мы берем множество нормированных конкретных долей и испытываем его на совокупностях, относящихся к каждому варианту. При этом смотрим, как получившаяся частота успехов соотносится с примерной долей в совокупности. Затем мы сравниваем два отношения. Флисс [Fleiss J. L. (1973)] вызвал интересную дискуссию об опасностях и аномалиях при косвенном нормировании на примерах из медицины и на показателях рождаемости и смертности.

Где же взять нормированные доли для *косвенной* нормализации? Для задачи, аналогичной той, что мы обсуждали в примере 3, когда сравнивали штат Мэн с Южной Каролиной, Вулси выбрал эталонные показатели смертности по данным всех штатов. Тем самым возрастное распределение населения США рассматривается как начало отсчета, когда Вулси берет его в качестве эталона прямым методом.

Иногда при прямой нормализации за эталон выбирают объединение тех совокупностей, которые наличествуют. Соответствующий выбор для косвенной нормализации мог бы строиться на нормированных долях некоторой взвешенной комбинации долей, имеющихся в каждом из испытываемых вариантов. Безусловно, один из заманчивых подходов к этой цели—«взвешивание пропорционально объемам типов». Использование таких весов эквивалентно использованию долей, объединенных по различным вариантам. Вернемся к данным примера 1 и взглянем на взвешивание в числах.

Пример 5. Косвенная нормализация для примера 1 (илл. 11.5.1). Сначала мы рассчитаем нормированные доли для каждого типа отдельно, а затем учтем число объектов в двух типах данного варианта для каждого варианта отдельно. Это даст нам упомянутый процент успеха для каждого варианта. (Эти проценты нельзя сравнивать, а надо соотносить с их же собственными примерными долями успеха.)

Результаты допускают, что вариант II обеспечивает относительно больший процент успехов, нежели вариант I. Относительно чего? Вариант II на 10% лучше своего эталона, а вариант I на 7% хуже своего.

Принципы приближенной аппроксимации для дисперсии (биномиальной), использованные выше для прямого нормирования, применимы и здесь. Самый грубый подход: *считать* $n_{\text{стд}}$ *известным* и вычислить

$$\frac{n_{\text{набл}} - n_{\text{стд}}}{\sqrt{n_{\text{стд}}}} \text{ или } \sqrt{4n_{\text{набл}}+2} - \sqrt{4n_{\text{стд}}+1},$$

принимая их приближенно за нормальные стандартные отклонения. Действительно, если только $n_{\text{набл}} = 0$, то $\sqrt{4n_{\text{набл}} + 2}$ достаточно близко к $\sqrt{n_{\text{набл}}} + \sqrt{n_{\text{набл}} + 1}$, что приближенно нормально [Freeman M. F. and Tykey J. W. (1950)]. Для $n_{\text{набл}} = 0$ заменим $\sqrt{4 \cdot 0 + 2}$ на 1. В нашем примере было 600 успехов $= n_{\text{набл}}$ и 642 нормированных успехов $= n_{\text{стд}}$, так что

$$Z_{\text{I}} = \frac{600 - 642}{\sqrt{642}} = \frac{-42}{25{,}3} = -1{,}66, \text{ или } \sqrt{2402} - \sqrt{2569} = 49{,}01 - 50{,}68 = -1{,}67.$$

Знак минус свидетельствует о том, что наблюдаенное число успехов меньше нормированного. Для варианта II имеем

$$Z_{\text{II}} = \frac{n_{\text{набл}} - n_{\text{стд}}}{\sqrt{n_{\text{стд}}}} = \frac{440 - 399}{\sqrt{399}} = \frac{41}{20{,}0} = 2{,}05 \text{ или } \sqrt{1762} - \sqrt{1597} = 2{,}02.$$

Итак, величина, которая нас интересует, равна:

$$\frac{Z_{\text{I}} - Z_{\text{II}}}{\sqrt{2}} = \frac{-1{,}66 - 2{,}05}{1{,}414} = -2{,}62, \text{ или } \frac{-1{,}67 - 2{,}02}{\sqrt{2}} = -2{,}61,$$

что указывает на существенные различия двух вариантов, поскольку при одинаково эффективных вариантах эта статистика будет приближенно нормально распределенной величиной с единичной дисперсией и нулевым средним.

Пример 6. Другое множество нормированных долей. Нормированные доли не обязаны быть взвешенными средними; вместо этого мы можем взвешивать по всем вариантам, которые использовались, что приводит к величинам

«легкие» — 75% $= [(1000 \times 0{,}70) + (1000 \times 0{,}80)]/2000;$

«трудные» — 30%

учитывая эти доли, найдем:

		Вариант I — 66%	Вариант II — 34,5%
Отношения:	$\frac{\text{примерные}}{\text{нормированные}}$	$\frac{600}{660} = 0{,}91$	$\frac{440}{345} = 1{,}28$

и снова вариант II дал наивысшую долю.

ОБСУЖДЕНИЕ

В итоге разность в примере 6 составляет

$$(-9\%) - (28\%) = -37\%$$

вместо

$$(-7\%)-(10\%)=-17\%$$

для примера 5. (И даже $(Z_I - Z_{II})/\sqrt{2}$ стало почти вдвое больше первоначальной величины.) Какова бы ни была оценка неопределенности, что же нам все-таки думать об этих двух ответах, которые хотя оба и отрицательны, но существенно различаются численно? Если бы эталон был известен точно и не зависел от данных, то мы могли бы им спокойно пользоваться. Но что же делать, если это не так? Ясно, что первым делом надо бы понять, если, конечно, это удастся, почему типы так несбалансированны по двум вариантам. Иногда это может подсказать нам, какие эталоны взять. Но как быть, если и это для нас недостижимо?

Самые крайние эталонные доли, которые еще «состоятельны» для имеющихся данных вместе с соответствующими долями, таковы:

«Легкие»	«Трудные»	Отношение: примерные успехи / эталонные успехи		
		I	II	II—I
70%	20%	1,00	1,76	0,76
70%	40%	0,94	1,02	0,08
80%	20%	0,88	1,69	0,81
80%	40%	0,83	1,00	0,17

Все они согласуются с тем, что вариант I приводит, видимо, к меньшей доле смертей, нежели вариант II, — в этом отношении любой из эталонов гораздо лучше прикидки, которая дает даже неверный знак. Вместе с тем эталоны довольно здорово отличаются друг от друга, напоминая нам «при всем при том», сколь ощутимо значение используемых нами эталонов.

Такая неопределенность больше и вне пределов неопределенности, рассмотренной выше, когда эталоны полагались известными.

Часто несбалансированность совсем не так велика, как здесь. Соответственно и точный выбор эталона не так уж существен. В некоторых исследованиях мы можем найти основания для выбора эталона, в других — единственное, что остается, — брать средние (или медианы) по всем вариантам (включая контрольные). В нашем примере, конечно же, это опять приведет к 75% и 30%, которые мы получили, когда учли веса в целом по варианту.

Пример 7. Косвенное нормирование для плохо определенных долей. Давайте приложим косвенный метод к примеру 4, где есть одна существенно плохо определенная доля. Условимся считать эталоном объединенную совокупность, как показано на илл. 11.5.2.

Этот подход предполагает, что вариант II имеет более высокий показатель смертности, однако за это нельзя поручиться. Так удается подавить эффект единичного наблюдения в ячейке [вариант I — «пло-

хое»] настолько, насколько позволяет индикация, но он сохраняется в оценках неопределенности.

Взглянув на илл. 11.5.2, мы видим также, что увеличение объема данных в 10 или 100 раз при сохранении их структуры сделало бы это обстоятельство явным.

ОБСУЖДЕНИЕ

Нам видятся два явных огреха в косвенном подходе:

1) даже если группы распределены равномерно по вариантам, но зато сами имеют разные веса в совокупности, косвенный метод даст различие в суммарных долях успехов, как правило, слабое, но досадное;

2) Вудворт (G. G. Woodworth) обнаружил пример, показывающий, что при сравнении трех или более вариантов один из них может иметь более высокие доли успехов в каждой группе, чем второй вариант, и в то же время более низкую нормированную долю. Это, видимо, бóльшая беда, чем предыдущая трудность.

Илл. 11.5.3 содержит пример Вудворта. Заметим, что вариант I имеет более высокие доли успехов, чем вариант II, и в 1-м, и во 2-м типах. Однако после косвенной нормализации оказывается 1,02 и 1,08 соответственно для вариантов I и II. Таким образом, первый вариант, который вначале имел *относительно* бóльшую на 10—11% долю успехов, в итоге имеет ее на 6% ниже. Вудворт отмечает, что если все сравнения делать попарно, то этого эффекта не будет.

Этот пример заслуживает дальнейшего внимания. Мы стоим перед явной несогласованностью: два варианта имеют большие доли успехов в 1-м типе, тогда как третий — во 2-м. Если мы проанализируем таблицу долей способом, описанным в гл. 9, то получим

0	0	—22	10
0	0	22	—10
10	0	—10	100

для самих долей (умноженных на 1000) и

0	0	—10	4
0	—1	10	—4
4	0	5	—100

для их логарифмов (умноженных на 100). И та и другая таблицы объясняют, почему вариант III не похож на остальные.

Вероятно, даже более важно понять это различие между вариантом III и первыми двумя, чем уточнять сравнение вариантов I и II.

Используй мы в этом примере правило медианы, и наши нормированные доли дали бы 0,115 и 0,095, так что нормирование привело бы к верному знаку для сравнения вариантов I и II. Зачастую такой прием срабатывает, но, как это ни печально, не всегда. (Так, если взять новый пример, заменив вариант III тремя, скажем, IIIa, IIIb и IIIc, ведущими себя точно так же, как и III, и с тем же суммарным объемом, то анализ с помощью таблицы с двумя входами выявит еще боль-

шую противоречивость.) Ничто не заменяет анализ таблицы с двумя входами для отдельных долей!

Вот еще раз основные моменты:

1. Одно число не может выполнять «по-честному» работу множества чисел.

2. Исходные таблицы исследовать стоит столь тщательно, сколь возможно.

3. Иногда мы вынуждены применять, несмотря ни на что, статистическую свертку даже к единичному наблюдению, точнее, считаем желательным так поступать.

11.6. ПЕРЕСТРОЙКА ДЛЯ НЕПРЕРЫВНЫХ КАТЕГОРИЙ

До сих пор мы рассматривали типы так, как будто они монолитны. Но надо считаться с тем, что типы будут гораздо разнообразнее по составу, чем нам бы хотелось.

Давайте вернемся к примеру 1 (на илл. 11.5.1, в частности), где участвуют всего два типа: «легкие» и «трудные». Пока мы не способны (либо из-за отсутствия записей, либо из-за отсутствия проницательности) классифицировать случаи более тонко, мы будем в полном недоумении насчет того, какие из «легких» легче или какие из «трудных» труднее.

Маловероятно, чтобы какой-нибудь механизм в условиях, когда в варианте I участвует 80% «легких» , а в варианте II — 90% «трудных», выявил бы оттенки «легкости» или «трудности». Если бы мы могли разделить данные на 4 типа вместо двух, то можно было бы ожидать нечто вроде илл. 11.6.1. Различие между вариантами стало несколько сильнее, чем на илл. 11.5.1. Если мы воспользуемся косвенным методом нормирования, то снова они в этом случае будут различаться сильнее. Наличие четырех типов вместо двух может привести к довольно большой коррекции. Какова же мораль сей басни?

Прежде всего нельзя считать, что нормирование для тех типов, которые нам попались под руку, исключает влияние любых факторов, лежащих в основе нашего разбиения на типы. Мы избавляемся лишь от некоторых, но не от всех эффектов, обусловленных этими факторами. Надо всегда быть готовым к новым перестройкам.

Более того, надо тщательно искать подходящий метод для новой перестройки. Он может оказаться полезным и тогда, когда его основания звучат невероятно. Применять его имеет смысл, даже если он указывает только порядок величин типов, требуемых при дальнейшей перестройке. (А обычно мы можем сделать значительно больше).

Учет эффектов непрерывных категорий. Когда мы приспосабливаемся к непрерывным категориям, таким, как «легкие» (легкоизлечимые) и «трудные» (трудноизлечимые) из примера в параграфе 11.1, мы представляем себе эти две категории как непрерывные типы с непрерывно меняющейся трудностью достижения успеха. И мы подразумеваем конкретных индивидуумов, распределенных как-то вдоль этого континуума. В нашем примере при варианте I 80% легко вылечиваются, а 20% — трудно. И мы рисуем себе распределение примерно так, как показано на илл. 11.6.2. Граница делит кривую распределе-

ния на две части: «легкие» и «трудные» случаи. Площади под кривой слева и справа задают процентные соотношения двух типов 80% и 20% соответственно. Если у нас есть такая шкала «трудности» и такое распределение, то мы в принципе можем найти центры тяжести для «легких» и «трудных». На илл. 11.6.2 они показаны знаком опоры ▲ под горизонтальной осью. Эти центры тяжести можно рассматривать как меры средних трудностей достижения успехов для варианта я двух типах. Не будем пока беспокоиться о том, откуда взялись именно такая шкала и такое распределение, — об этом скажем позже.

Далее, для варианта II соответственно в данном примере есть 10% «легких» и 90% «трудных», как это и показано на илл. 11.6.3. Снова отмечены центры тяжести двух типов.

Мы хотим знать, как доли успехов (илл. 11.1.1В) связаны с центрами тяжести для обоих вариантов и двух типов индивидуумов, к которым эти варианты применяются. Для этого мы строим илл. 11.6.4, где знаком «×» показаны доли успехов, соответствующие центрам тяжести четырех типов. Одна из важных находок заключается в том, что два типа «легких» не согласуются на шкале трудности. Аналогично не совпадают на шкале и типы «трудных». Наша цель состоит в том, чтобы приспособиться к этим несоответствиям в трудностях.

Для этого надо усреднить трудности и привести доли успехов в соответствие со средней трудностью. Можно взять среднее арифметическое или медиану (если есть несколько вариантов и несколько типов), может быть, взвешивая по объемам типов. Если прямые, соединяющие точки на рисунке илл. 11.6.4, примерно параллельны, то различия в выбранных средних будут незначительными. Пока давайте согласуем средние трудности отдельно для «легких» и для «трудных» типов. Геометрически на илл. 11.6.5 пустым кружком обозначены точки, абсциссы которых представляют средние нормированные трудности для «легких» и «трудных» типов.

Чтобы найти окончательный результат для разности между долями успехов двух вариантов, согласованных на шкале трудности, мы можем усреднить разности между «легкими» и «трудными» типами, быть может, взвешивая по объемам исходных типов или по ожидаемым объемам совокупности, с которой предстоит работать.

Мы набросали общий план работы для задачи с двумя вариантами и двумя типами с одним лишь пропуском, — мы опустили вопрос о распределении и его связи со шкалой трудностей. И хотя можно выбирать различные распределения, было бы желательно иметь такое, чтобы легко рассчитывались центры тяжести сегментов. Если мы возьмем такое распределение, вид которого не противоречит нашим общим взглядам на то, каким должно быть распределение «трудностей» в популяции, то едва ли имеет большое значение, какое именно конкретное распределение взять. Логистическое распределение напрашивается прежде всего. Оно имеет привлекательную форму, симметричную относительно единственной моды и не слишком остроконечную. Формула его удобна, и, кроме того, есть таблицы, облегчающие расчет центров тяжести (см. илл. 5.6.2). Идея применения для центров тяжести балльных оценок изложена в параграфе 5.8 с необходимыми формулами,

Логистическое распределение. Общий вид плотности нормированного логистического распределения дан на илл. 11.6.6. Оно симметрично относительно нуля. Определяющее свойство его в том, что абсцисса точки, отсекающей площадь под кривой, равную p, слева от себя, равна:

$$x = d = \log \frac{p}{1-p}.$$

Отметим, что $d = 0$ при $p = 1/2$. Обозначение d используется нами как мнемоническое для шкалы «трудности» (difficulty).

Мы избрали логистическое распределение и его преобразования для представления распределения степеней трудности в испытываемых совокупностях. Мы хотим установить граничную точку на нуль и тем самым отразить распределение на интервал от $\log [p/(1-p)]$ до 0, или, что то же самое, мы мерим расстояние центра тяжести от точки $\log [p/(1-p)]$.

Основная идея заключена в том, что трудность успешного «лечения» в совокупности можно приближенно представлять колоколообразной кривой, центр тяжести которой сдвигается вправо, когда в совокупности оказывается больше трудноизлечимых больных.

Чтобы дать представление о шкале нормированного логистического распределения, отметим, что симметричный относительно центра интервал, накрывающий 95% вероятности, задается границами от $-3{,}23$ до $+3{,}23$.

Определим, наконец, центр тяжести такого интервала. Из параграфа 5.8 мы знаем, что при верхнем B и нижнем A уровнях накопленной вероятности, определяющих интервал, его центр тяжести вычисляется по формуле

$$\mathrm{CG} = \frac{[B \log B + (1-B) \log (1-B)] - [A \log A + (1-A) \log (1-A)]}{B-A}.$$

Когда у нас только два типа, то один расположен от 0 до p, другой — от p до 1. Положив $A = 0$, $B = p$, мы получаем центр тяжести левого интервала нормированного логистического распределения от 0 до p

$$\mathrm{CG} = \frac{p \log p + (1-p) \log (1-p)}{p} = \frac{\varphi(p)}{p}.$$

Так же, полагая $A = p$, $B = 1$, для правого интервала найдем от p до 1:

$$\mathrm{CG} = -\frac{p \log p + (1-p) \log (1-p)}{1-p} = \frac{-\varphi(p)}{1-p} = \frac{\varphi(1-p)}{1-p},$$

причем таблица $\varphi(p)$ приведена на илл. 5.6.2 параграфа 5.6. Вспомним, что, преобразуя распределение, мы измеряем расстояния от этих центров тяжести до граничной точки, разделяющей типы «трудных» и «легких»

$$\log \frac{p}{1-p},$$

так что разумно принять эту точку за нуль. Теперь мы готовы сосчитать пример, данный в начале главы. Все, что нам понадобится, приведено на илл. 11.6.7.

Численный пример. Проведем сначала расчеты для варианта I. Доля типа «легкие» $p = 0{,}8$ и точка деления на нормированном логистическом распределении равна:

$$\log_e \frac{0{,}8}{1-0{,}8} = \log_e 4 = 1{,}386.$$

Для расчета двух центров тяжести по таблице илл. 5.6.2 находим (при $p = 0{,}8$), что $p \log_e p + (1-p) \log_e (1-p) = -0{,}5004$. Центры тяжести равны соответственно:

$$\text{от 0 до 0,8: } CG = \frac{-0{,}5004}{0{,}8} = -0{,}6255 \approx -0{,}626;$$

$$\text{от 0,8 до 1: } CG = \frac{-(-0{,}5004)}{0{,}2} = +2{,}502.$$

Желая найти положение центров тяжести, когда распределение сдвинуто так, что граничная точка приходится на нуль, мы должны вычесть из каждого центра по 1,386. Это дает следующие результаты.

Вариант I. Результаты для сдвинутого распределения, когда нуль в точке $p = 0{,}8$:

«легкие»	от 0 до 0,8	$-2{,}012 = (-0{,}626 - 1{,}386)$;
«трудные»	от 0,8 до 1	$1{,}116 = (2{,}502 - 1{,}386)$;
граница		$0 = (1{,}386 - 1{,}386)$.

Проделав соответствующие расчеты для варианта II, где было 10% «легких» и 90% «трудных», получим (используя числитель (— 0,3250) из илл. 5.6.2), что центры тяжести лежат в точках (— 3,250) и (0,361), а точка деления есть (— 2,187).

Вариант II. Результаты для сдвинутого распределения, когда нуль в точке $p = 0{,}1$:

«легкие»	от 0 до 0,1	$-1{,}063 = (-3{,}250 + 2{,}187)$;
«трудные»	от 0,1 до 1	$2{,}548 = (0{,}361 + 2{,}187)$;
граница		$0 = (-2{,}187 + 2{,}187) \log(0{,}1/0{,}9) = -2{,}187$.

Среднее трудностей по двум типам в каждой категории:

тип «легкие»	$1/2 \quad (-2{,}012 - 1{,}063) = -1{,}538$;
тип «трудные»	$1/2 \quad (1{,}116 + 2{,}548) = 1{,}832$.

Мы можем прочесть эти скорректированные значения на графике на илл. 11.6.8. или проинтерполировать после того, как получим угловые коэффициенты прямых, как показано ниже.

	Скорректированные доли успехов
Для типа «легких»:	
вариант I	70—15,98 (—1,538—(—2,012)) = 62,4%
вариант II	80—11,08 (—1,538—(—1,063)) = 85,3%
Для типа «трудных»:	
вариант I	20—15,98 (1,832—1,116) = 8,6%
вариант II	40—11,08 (1,832—2,548) = 47,9%

Превышения варианта II над вариантом I таковы:

«легкие»	85,3—62,4 = 22,9
«трудные»	47,9— 8,6 = 39,3
Среднее превышение	31,1

Наш расчет среднего превышения процента успехов при переходе от варианта I к варианту II дает 31,1% по сравнению с аналогичной величиной при простом усреднении превышений: $15\% = \frac{1}{2}(10\% + 20\%)$ и с «превышением» примерного подсчета успехов: $-16\% = (44\% - 60\%)$.

Этот пример показывает, что учет распределения трудности может приводить к существенно иным результатам.

Эталонная совокупность. Если мы хотим взять в качестве эталона совокупность, состоящую из 45% «легких» и 55% «трудных», т. е. среднюю из двух типов, используемых в вариантах I и II, то должны взвесить полученные разности (взять с весами 0,45 и 0,55), что дает

$$0{,}45\,(22{,}9) + 0{,}55\,(39{,}3) = 31{,}9\%,$$

а для сравнения проведем взвешивание исходных разностей:

$$0{,}45\,(10) + 0{,}55\,(20) = 15{,}5\%.$$

КОММЕНТАРИЙ

Корректировка, которую мы проделали, может быть точной (для очень больших выборок), если:

1) процент успехов линейно зависит от CG соответствующего распределения «трудности»;

2) распределения трудностей имеют одинаковые разбросы.

Первое из этих требований верно только тогда, когда процент успехов линейно входит в нормированную «трудность». Но поскольку для корректировки требуется лишь приблизительно хорошее поведение, то этот вопрос, по-видимому, не столь уж важен для практики.

Следствия того, что распределения «трудности» должны иметь один и тот же разброс, по-видимому, дают больше оснований для волнения. Этого, правда, можно избежать путем расчета процента успехов *по отношению к граничной точке*, который только и требует, чтобы процент успехов линейно зависел от центров тяжести нормированной «трудности». Для нашего примера такой расчет приводит к значениям:

$$\text{для варианта I} \quad \frac{1{,}116}{3{,}128}(70\%) + \frac{2{,}012}{3{,}128}(20\%) = 37{,}84\%;$$

$$\text{для варианта II} \quad \frac{2{,}548}{3{,}611}(80\%) + \frac{1{,}063}{3{,}611}(40\%) = 68{,}22\%,$$

где 3,128 = 1,116 + 2,012; 3,611 = 2,548 + 1,063. Разность в 30,38% полностью перекрывает разности, полученные другими путями, рассмотренными выше, и изображена на илл. 11.6.8 как разность между вариантами II и I в точке, где нормированная «трудность» равна нулю для обоих вариантов.

11.7. БОЛЬШЕ ДВУХ НЕПРЕРЫВНЫХ КАТЕГОРИЙ

Случай, рассматриваемый в предыдущем параграфе, фактически наиболее сложный, ибо имеет *только одну граничную точку*. Наличие большего числа точек деления, в сущности, упрощает шкалирование, и наша задача распадается на две части:

- все категории, кроме двух, внутренние, с двумя границами, их мы можем закрепить просто и надежно;
- две категории будут внешними и часто включают малую долю случаев, что, однако, по причинам, описанным ниже, не создает больших проблем.

Правда, есть много альтернатив, которые надо обсуждать, но работа с числом категорий больше двух действительно намного робастнее, поэтому дает более удовлетворительные результаты.

Давайте подытожим некоторые узловые моменты общей ситуации:

- случаи различаются «трудностью»;
- квалифицированное мнение делит их на три или более типа по «трудности»;
- предполагается, что «трудность» в действительности изменяется непрерывно, поэтому разумно обрабатывать имеющиеся типы так, как если бы существовала непрерывная шкала «трудностей», разделенная на несколько частей;
- один путь приписывания чисел «трудностям» (безотносительно к наблюдаемым результатам) — предположить, что вид некоторых или всех распределений «трудностей» известен; для этой цели удобно логистическое распределение;
- для внутренних категорий, как мы увидим, выбор вида распределения не очень важен;
- для внешних категорий, как будет показано, логистическое распределение часто ведет себя нейтрально.

За всем этим кроются два вопроса: (1) насколько разумно ожидать, что логистическая шкала удобна для работы? Если же она нас удовлетворяет, то (2) каковы наши соображения относительно связи с ней процента успехов? На этих вопросах мы сейчас и остановимся.

Если наблюдаемые количества в паре совокупностей лишь слегка разнятся, то два распределения «трудностей» (с неизвестной, но вполне определенной шкалой) будут совершенно подобны. В таком случае численное выражение «трудностей», согласованное с логистическим распределением для одного из них, будет согласовано и с другим. Если же, напротив, различие «трудностей» в двух совокупностях, соответствующих двум вариантам, столь велико, что они почти не перекрываются, то результат замены каждого из распределений логистическим может противоречить другому лишь в короткой области перекры-

тия, так что мы снова можем подбирать оба распределения достаточно близкими к логистическому с общей шкалой. Если здесь и есть опасность, то она может возникнуть в промежуточных случаях между этими двумя крайними, т. е. там, где различие в распределениях «трудностей» существенно, но есть и значительное перекрытие.

Вопрос об отношении отклика к шкале «трудностей» и обсуждается ниже.

ВНУТРЕННИЕ КАТЕГОРИИ И ЦЕНТРЫ ТЯЖЕСТИ

По крайней мере некоторые соображения о центрах тяжести (CG) внутренних категорий просты и не слишком зависят от выбора нормированного распределения для «трудностей». Пусть, для примера, имеется категория с шириной в 30%. Если она располагается между 35 и 65% на шкале качества «трудностей», то едва ли мы будем колебаться в приписывании ее центру тяжести средней точки, особенно для симметричного распределения величин «трудности».

Если категория простирается от 40 до 70% (или от 50 до 80%), то, лишь немного сомневаясь, мы поместим центр тяжести между средней точкой и границей 40% (50%) интервала. Это происходит для всякого симметричного одновершинного распределения, так что характер положения центра тяжести (слева от середины, в середине или справа от нее) очевиден.

Испытание нескольких альтернативных распределений — гауссовского, логистического и даже Коши — показало, что численные значения центров тяжести для внутренних категорий практически одни и те же, следовательно, любое из распределений подходит и мы можем взять наипростейшее.

Рассмотрим теперь второй вопрос. Определение степени изменения процента успехов с изменением положения центра тяжести много сложнее, чем локализация центра тяжести. Предполагаем лишь, что наша коррекция для данной категории всегда производится в правильном направлении и лежит между 0,5—1,5 истинной величины. Тогда после коррекции

$$|\text{остаточная ошибка}| < \frac{1}{2} |\text{исходная ошибка}|,$$

и мы имеем реальный прогресс.

Давайте спланируем коррекцию середин интервалов для категорий. Чтобы избрать правильное направление, нам надо только не ошибиться насчет области расположения центров тяжести и в направлении изменения процента успехов в зависимости от «трудности». Это, собственно, простейшие вещи, которые приходится делать. Мы можем делать их почти безошибочно. Причем нам не нужна высокая точность численной локализации центра тяжести или углового коэффициента зависимости между процентом успехов и количественной «трудностью». Посмотрим, как это выглядит в числах.

В примере, приведенном на илл. 11.6.1, есть четыре категории и два варианта. На илл. 11.7.1 приведены расчеты по определению

центров тяжести сегментов, использующие обычные формулы и таблицу из илл. 5.6.2 параграфа 5.6.

Для примера давайте остановимся на категории «полулегких» и варианте I. То, что 80% находится по одну сторону от границы, видно из илл. 11.6.1Б. Илл. 11.6.4 этой главы для 80% дает — 0,5004. Чтобы найти центр тяжести (в предположении нормированного логистического распределения), надо разделить на 0,4 разность — 0,5004 — — (— 0, 6730). Столбец разностей пригодится впоследствии для вычисления угловых коэффициентов.

В часть Б илл. 11.1.1 мы переносим доли успехов из илл. 11.6.1В. Теперь можно нанести на график эти доли в зависимости от величин «трудности», измеряемых центрами тяжести, что и проделано на илл. 11.7.2 для варианта I. Рисунок полезен, поскольку подсказывает, какую хорду надо интерполировать и в каком направлении.

Для варианта I и категории «полулегких» нормированные «трудности» таковы:

один край	—0,406.
CG	0,432 (середина 0,490, поправка 0,058).
другой край	1,386.

Здесь $0{,}490 = \frac{1}{2}(-0{,}406 + 1{,}386)$ и направление + 0,058 вполне очевидно. Так как центр тяжести «полутрудных» равен 2,013, то поправка равна

$$\frac{0{,}058}{2{,}013 - 0{,}432} = 0{,}0367$$

от (50% — 23,3%) = 26,7% разности успехов. Эта коррекция дает (вариант I) успех «полулегких» в средней точке 50% — 0,98% = = 49,02%, который представлен в части Б илл. 11.7.1 и показан кружком на рисунке илл.11.7.2. Для варианта II подобный расчет приводит к значению 77,39% = 75% + 2,39%.

Первую коррекцию мы можем записать в виде

(+0,058 от нормированной «трудности»)·(16,9% на единицу нормированной «трудности»),

где

$$16{,}9\% = \frac{50\% - 23{,}3\%}{2{,}013 - 0{,}432}$$

— угловой коэффициент пунктирной прямой на илл. 11.7.2. Здесь + 0,058 очено мало зависит от выбора распределения, которое накрывает только этот интервал, что и дает 0,058. Однако величина 16,9 будет зависеть от распределения несколько больше, хотя и не слишком сильно (здесь используемые распределения влияют на две категории: «полулегких» и «полутрудных»). Мы удивились бы любому из отклонений такого рода, как 0,05 и 12%, дающему 0,6% поправки, или 0,065 и 20%, дающему 1,3% поправки. Поправка в 0,98% невелика, но не может облегчить наш жребий. Это именно то, чего мы ожидаем

от процесса коррекции для категорий с непрерывным изменением свойств.

Нас привлекает такой сорт корректировки для каждого внутреннего класса, так как при его использовании не требуется приближенного равенства разбросов в распределениях нормированных «трудностей» для различных вариантов, нужно разумное поведение лишь в одном классе или в двух смежных.

Поскольку внешние категории находятся в несколько ином положении (оно обсуждается ниже), мы не можем полагаться всерьез на угол наклона между внутренними и внешними категориями для расчета скорректированных процентов для внутренних классов, в нашем случае «полутрудных». Вместо этого надежнее использовать угол наклона между внутренними категориями для корректировки внутренних значений. Этот угол наклона тоже изображен на илл. 11.7.2.

ВНЕШНИЕ КАТЕГОРИИ

Однако нам еще могут доставить неприятность внешние категории. Мы не очень сильно стеснены в действиях, когда сталкиваемся только с одной границей. Но лучшее, на что мы можем обычно надеяться, это работать с двумя границами, рассматривая две категории — внешнюю и соседнюю с ней, внутреннюю.

И все же дела не так плохи, как могло бы показаться. Расстояние между центром тяжести внешней категории и границей, которая отделяет эту категорию, ведет себя так, что это часто нам весьма полезно. На илл. 11.7.3 мы видим, что для внешних категорий, на которые приходится не более 10% всех случаев, центр тяжести находится на расстоянии около 1,0 от границы.

Таким образом, если разброс «хвостов» распределений «трудностей» одинаков для разных вариантов и обе крайние категории малы, то все CG будут находиться на расстоянии около одной единицы от своих границ и не смогут сильно разниться друг от друга, так что требуемая коррекция будет мала.

Это — характерное свойство распределений вроде логистического, с приблизительно экспоненциальными «хвостами». Такого рода «нейтральность» поведения, приводящая к небольшим поправкам там, где обе крайние категории малы, дает, по-видимому, возможность выбора и тогда, когда мы не знаем точного типа распределения, а это практически всегда так.

Иногда, как в нашем, уже образцовом, примере, не все крайние классы малы. В таком случае можно взять любой из методов, примененных выше для двух категорий, т. е. в случае когда обе категории крайние. Можно использовать менее робастный подход и найти скорректированные проценты успехов для крайних категорий, работая с ближайшей границей, или же можно, что более устойчиво, интерполировать к этой точке. (Иногда удобно давать скорректированные проценты успехов для каждой границы и для середины каждой внутренней категории, хотя мы не будем брать это на вооружение.) Заметим, что

в одной ячейке мы получили скорректированное значение, превышающее 100% (1-я строка, предпоследний столбец илл. 11.7.1Б).

Наконец, нам нужен эталонный набор весов. Мы выбираем объединенные значения в группах для обоих вариантов (они показаны в последнем столбце илл. 11.7.1Б) и, взвешивая ими, рассчитываем средние взвешенные разностей для скорректированных долей успехов и для исходных долей. Для исходных процентов эта разность равна 20,7%, а для скорректированных — 24,4%. Так что различие меньше, чем было получено для случая двух категорий, но оно тем не менее существенно. Илл. 11.7.4 подытоживает разности (вариант II МИНУС вариант I) долей успехов, уже рассчитанных для примера с двумя и четырьмя типами, а также пример с одним типом, игнорирующий деление на «легких» и «трудных» и дающий только примерную долю успехов. Группа, скорректированная взвешенным средним, в случае двух типов была обсуждена в параграфе 11.6, когда изучался выбор эталонной совокупности.

Прямое нормирование. Способ корректировки, проведенный выше, соответствует прямому нормированию, в котором мы пытались рассчитать доли успехов, соответствующие каждому варианту (для каждого класса), затем анализировали разности и, наконец, сочетали анализ разностей с взвешиванием. Если бы мы вначале сочетали скорректированные доли успехов с весами, а затем брали разности (что представляет эквивалентный расчет), то мы могли бы провести точную параллель с прямой корректировкой, если пренебречь непрерывностью категорий.

РЕЗЮМЕ. НОРМИРОВАНИЕ ДАННЫХ ДЛЯ СРАВНЕНИЯ

Часто оказывается, что отклики, выражаемые как доли, проценты или пропорции, удобнее сравнивать после нормирования совокупности относительно некоторого фундаментального фактора.

Можно нормировать проценты (или доли) относительно этих факторов прямо.

Можно рассчитывать, используя различные компромиссы между простотой и качеством аппроксимации, оцениваемым квадратичной ошибкой прямого нормирования.

Плохо определенные доли, даже в одной ячейке, могут создавать препятствия для прямого нормирования, и нам нужны (1) сигнал тревоги для тех случаев, когда это может случиться; (2) метод нормирования, устраняющий эти трудности.

Расчет стандартных ошибок (возможно, общих, но лучше в терминах наибольших вкладов от одной либо нескольких ячеек) предупреждает о присутствии плохо определенных величин и необходимости другого подхода.

В тех случаях, когда это более удобно, можно воспользоваться косвенной нормализацией.

Такая нормализация имеет свои преимущества и обладает разумной логикой и простыми вычислительными схемами.

Нормирование относительно фундаментального фактора и непрерывных категорий не всегда достаточно для исключения всех смещений, связанных с этим фактором, и часто требует дополнительных данных.

С целью дальнейшего углубления результатов нормирования для непрерывных категорий мы можем использовать один метод для внешних категорий (следовательно, этот метод нужен лишь тогда, когда фундаментальная переменная имеет больше чем две категории) и другой метод для внутренних категорий.

Логистическое распределение открывает удобный путь описания центров тяжести непрерывных категорий — путь, который отражает обычно действительные разности в центрах тяжести различных частей (для различных вариантов) сравниваемых совокупностей.

БИБЛИОГРАФИЯ

Fleiss J. L. (1973). Statistical Methods for Rates and Proportions. New York, Wiley and Sons. Chapter 13.

Freeman M. F. and Tukey J. W. (1950). Transformations related to the angular and the square root. — Ann. Math. Stat., 21, 607—611.

Sutherland M., Holland P. and Fienberg S. E. (1974). Combining Bayes and frequency approaches to estimate a multinomial parameter. B: Fienberg S. E. and Zellner A. (Eds.) Studies in Bayesian Economics and Statistics. Amsterdam, North-Holland, 585—617. Или: Bishop Y. M. M., Fienberg S. E. and Holland P. (1975). Discrete Multivariate Analysis. Cambridge, Mass, MIT Press, 429—433.

ИЛЛЮСТРАЦИИ

Иллюстрация 11.1.1

Грубое вычисление относительного «успеха»; два варианта и два слоя

А. Число успехов

Типы	Варианты	
	I	II
«Легкие»	560	80
«Трудные»	40	360
Всего	600	440

Б. Число испытуемых

«Легкие»	800	100
«Трудные»	200	900
Всего	1000	1000

B. Доля успехов

«Легкие»	70%	80%
«Трудные»	20%	40%
Примерная доля	60%	44%

Иллюстрация 11.2.1

Повозрастные смертность и численность населения в штатах Мэн и Южная Каролина на 1930 г.

Данные по Woolsey T. D. (1959), Chapter 4, p. 67. В: Linder F. E. and Grove R. D. Vital Rates in the United States, 1900—1940. National Office of Vital Statistics, U. S. Government Printing Office, Washington, D. C., 1959.]

Возраст (в годах)	Мэн		Южная Каролина		Процентное распределение численности	
	Удельный коэффициент смертности (число смертей на 100 тыс. чел.)	Численность населения	Удельный коэффициент смертности (число смертей на 100 тыс. чел.)	Численность населения	Мэн	Южная Каролина
0—4	2056	75037	2392	205076	9,4	11,8
5—9	186	79727	185	240750	10,0	13,9
10—14	140	74061	184	222808	9,3	12,8
15—19	223	68683	426	211345	8,6	12,2
20—24	370	60575	645	166354	7,6	9,6
25—34	391	105723	871	219327	13,3	12,6
35—44	545	101192	1242	191349	12,7	11,0
45—54	1085	90346	1994	143509	11,3	8,3
55—64	2036	72478	3313	80491	9,1	4,6
65—74	5219	46614	6147	40441	5,8	2,3
75+	13645	22396	14136	16723	2,8	1,0
Всего	...	796832	...	1738173	99,9	100,1
Примерный коэффициент смертности (на 100 тыс. чел.)	1390,8	...	1288,8	...	...	...

Замечание. Из данных исключены умершие и живые неизвестного возраста.

Иллюстрация 11.2.2

Расчет по данным илл. 11.2.1 методом прямого нормирования показателей смертности для штатов Мэн и Южная Каролина (результаты сравниваются).

Возраст (в годах)	Возрастное распределение (на 1 млн. чел.) в США по данным 1940 г.	Мэн				Южная Каролина			
		численность населения на 1930 г.	смертность в 1930 г.	удельная смертность (на 100 тыс. чел.)	ожидаемая смертность в эталонной группе для США	численность населения на 1930 г.	смертность в 1930 г.	удельная смертность (на 100 тыс. чел.)	ожидаемая смертность в эталонной группе для США
	(1)	(2)	(3)	(4) = (3) ÷ (2)	(5) = (4) × (1)	(6)	(7)	(8) = (7) ÷ (6)	(9) = (8) × (1)
0—4	80100	75037	1543	2056	1647	205076	4905	2392	1916
5—9	81100	79727	148	186	151	240750	446	185	150
10—14	89200	74061	104	140	125	222808	410	184	164
15—19	93700	68683	153	223	209	211345	901	426	399
20—24	88000	60575	224	370	326	166354	1073	645	568
25—34	162100	105723	413	391	634	219327	1910	871	1412
35—44	139200	101192	552	545	759	191349	2377	1242	1729
45—54	117800	90346	980	1085	1278	143509	2862	1994	2349
55—64	80300	72478	1476	2036	1635	80491	2667	3313	2660
65—74	48400	46614	2433	5219	2526	40441	2486	6147	2975
75+	20100	22396	3056	13645	2743	16723	2364	14136	2841
Всего	1000000	796832	11082	...	12033	1738173	22401	...	17163

$$D_{\text{Мэн}} = \frac{12033}{1000000} \times 1000 = 12{,}03;\quad D_{\text{Ю.К}} = \frac{17163}{1000000} \times 1000 = 17{,}16;\quad D_{\text{Ю.К}}/D_{\text{Мэн}} = \frac{17{,}16}{12{,}03} = 1{,}43.$$

Замечания.

1. Формула дает показатель смертности на 1000 человек. Если же предпочтительнее доли, то $D_{\text{Мэн}}$ и $D_{\text{Ю.К}}$ делятся на 1000.

2. Таблица включает вычисления показателей по всем группам, имеющимся в илл. 11.2.2

Иллюстрация 11.4.1

Таблица с ненадежной ячейкой, имеющей большой вес

А. Смертность

Состояние	Способ	
	I	II
Отличное	10	4
Удовлетворительное	9	20
Плохое	1	2

Б. Число испытуемых (объемы ячеек)

Состояние	Способ	
	I	II
Отличное	10000	5000
Удовлетворительное	3000	4000
Плохое	1	20

В. Смертность в расчете на 1000 пациентов

Состояние	Способ		Нормированная совокупность	Ожидаемое число смертей	
	I	II		I	II
Отличное	1	0,8	14 500	14,5	11,6
Удовлетворительное	3	5	5 000	15,0	25,0
Плохое	1000	100	500	500,0	50,0
Всего			20 000	529,5	86,6
Смертность на 1000 пациентов				26,48	4,33
Смертность на 1000 пациентов в случае, когда в ячейке «Плохое — способ I» стоит 0				1,48	4,33

Иллюстрация 11.5.1
Косвенное нормирование в примере 1

А. Число успехов

Состояние	Вариант		Всего
	I	II	
«Легкие»	560	80	640
«Трудные»	40	360	400
Всего	600	440	

Б. Число испытуемых

«Легкие»	800	100	900
«Трудные»	200	900	1100
Всего	1000	1000	

В. Доля успехов

			Нормированные доли
«Легкие»	70%	80%	$71{,}1\%\ \left(=\frac{640}{900}\right)$
«Трудные»	20%	40%	$36{,}4\%\ \left(=\frac{400}{1100}\right)$
Примерная доля	60%	44%	

Применяем нормированные доли — Эталонные доли успехов

к варианту I: $\dfrac{71{,}1\times 800+36{,}4\times 200}{1000} = 64{,}2\%;$

к варианту II: $\dfrac{71{,}1 \times 100 + 36{,}4 \times 900}{1000} = 39{,}9\%$.

Отношение нормированных успехов $= \dfrac{\text{примерная доля}}{\text{эталонная доля}}$;

для варианта I: $\dfrac{60\%}{64{,}2\%} = 0{,}93$;

для варианта II: $\dfrac{44\%}{39{,}9\%} = 1{,}10$.

Иллюстрация 11.5.2
Косвенное нормирование для плохо определенных долей

А. Число успехов

Состояние	Вариант		Всего
	I	II	
Отличное	10	4	14
Удовлетвори-тельное	9	20	29
Плохое	1	2	3
Всего	20	26	

Б. Число испытуемых

Отличное	10 000	5000	15 000
Удовлетвори-тельное	3000	4000	7000
Плохое	1	20	21

В. Доля успехов на 1000 пациентов

			Эталон
Отличное	1	0,8	$\dfrac{14}{15\,000}$
Удовлетвори-тельное	3	5	$\dfrac{29}{7000}$
Плохое	1000	100	$\dfrac{3}{21} = \dfrac{1}{7}$

Г. Нормированное число успехов

Отличное	9,33	44,67
Удовлетвори-тельное	12,43	16,57
Плохое	0,14	2,86
Всего	21,90	24,10

Отношения $=\dfrac{\text{примерная доля успехов}}{\text{эталонная доля успехов}}$: I вариант 0,913, II вариант 1,08.

Приближенная оценка мощности

$$\left.\begin{array}{l}\sqrt{4\,(20)+2}-\sqrt{4\,(21{,}90)+1}=9{,}06-9{,}41=-\,0{,}35\\ \sqrt{4\,(26)+2}-\sqrt{4\,(24{,}10)+1}=10{,}30-9{,}87=0{,}43\end{array}\right\}\frac{-0{,}35-0{,}43}{\sqrt{2}}=-0{,}55.$$

Если бы у нас было в 10 раз больше данных, то конечный результат умножился бы примерно на $\sqrt{10}\approx 3{,}16$, а если в 100 раз, то примерно на 10. (Так что надо много больше данных, чтобы достичь приемлемого уровня значимости изменений разности.)

Иллюстрация 11.5.3

Пример Вудворта о «перестановке» в величинах успеха при косвенном нормировании. [Данные взяты с разрешения автора из неопубликованной рукописи Woodworth G. G. (1971). Standardized Mortality Comparisons: Sketch of a Review. Unpublished manuscript. Used with permission of the author.]

А. Успехи

Состояние	Вариант			Всего
	I	II	III	
1	12	99	39	150
2	90	9	51	150
Всего	102	108	90	300

Б. Число испытуемых

1	100	900	500	1500
2	900	100	500	1500
Всего	1000	1000	1000	

В. Доля успехов на 1000 испытуемых

				Эталон
1	0,120	0,110	0,078	0,1
2	0,100	0,090	0,102	0,1

Г. Нормированное число успехов

1	10	90	50
2	90	10	50
Всего	100	100	100

Отношение $=\dfrac{\text{примерная доля успехов}}{\text{нормированная доля успехов}}=$

1,02 1,08 0,90

Разность между примерной и нормированной долями успеха в единицах стандартного отклонения (оценка через $\sqrt{4\,n_{\text{набл}}+2}-\sqrt{4\,n_{\text{норм}}+1}$):

0,22 0,81 —1,00

Иллюстрация 11.6.1

Данные илл. 11.5.1, разбитые на подклассы (гипотетически)

А. Успехи

Состояние	Вариант		Всего
	I	II	
«Легкие»	360	20	380
«Полулегкие»	200	60	260
«Полутрудные»	35	250	285
«Трудные»	5	110	115

Б. Число испытуемых

«Легкие»	400	20	420
«Полулегкие»	400	80	480
«Полутрудные»	150	500	650
«Трудные»	50	400	450

В. Доли

«Легкие»	90%	100%	90,5%
«Полулегкие»	50%	75%	54,2%
«Полутрудные»	23,3%	50%	43,8%
«Трудные»	10%	27,5%	25,6%

Иллюстрация 11.6.2

Иллюстративное распределение объектов на шкале «трудности» с 80% «легких» и 20% «трудных» как в варианте I из параграфа 11.1. Знаком ▲ отмечены центры тяжести соответствующих групп

Иллюстрация 11.6.3

Гипотетическое распределение «трудности» для испытуемых варианта II из параграфа 11.1

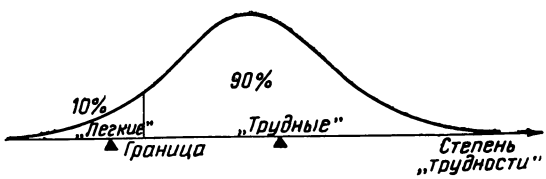


Иллюстрация 11.6.4

График зависимости долей успеха от центров тяжести (для двух «легких» и двух «трудных» групп)

Здесь знаком «×» показаны точки с координатами центров тяжести и долей успехов для 4 групп. Мы видим, что группы «легких» так же, как и «трудных», не согласуются на шкале «трудностей».

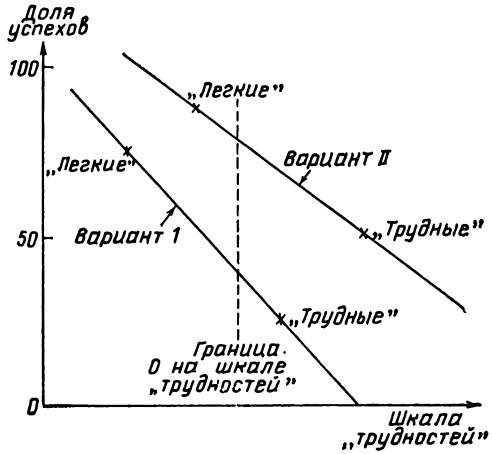


Иллюстрация 11.6.5

График со средними нормированными «трудностями» в зависимости от долей успехов для «легких» и «трудных» групп

Абсциссы кружков соответствуют средним нормированным значениям «трудности» для «легких» групп и для «трудных». Стрелками показаны для каждой из четырех групп требуемые сдвиги, чтобы согласовать доли успехов со средними «трудностями». Длины вертикальных отрезков оценивают различия в долях успехов для двух вариантов после перестройки шкалы «трудности».

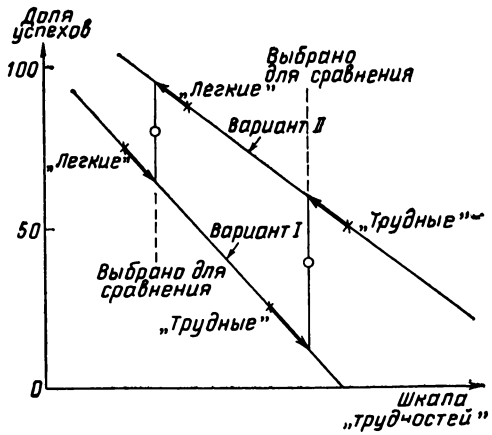


Иллюстрация 11.6.6

Плотность стандартного логистического распределения. Абсцисса log [p/ (1—p)], если через p обозначить накопленное значение функции распределения в точке x

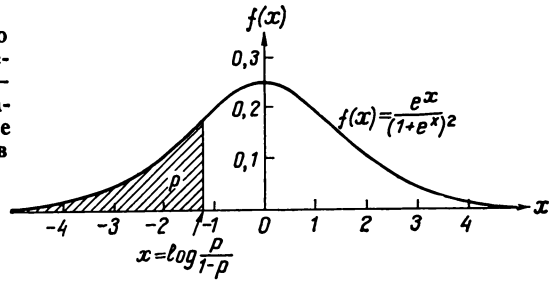


Иллюстрация 11.6.7

Примерный процент успехов; два варианта, два слоя воспроизводится из илл. 11.1.1Б и 11.1.1В

А. Число испытуемых

Состояние	Вариант I	Вариант II
«Легкие»	800	100
«Трудные»	200	900
Всего	1000	1000

Б. Процент успехов

Состояние	Вариант I	Вариант II
«Легкие», %	70	80
«Трудные», %	20	40
Примерный процент успехов	60	44

Иллюстрация 11.6.8

Уровни успехов после коррекции для двух непрерывных категорий (отмечены жирными стрелками)

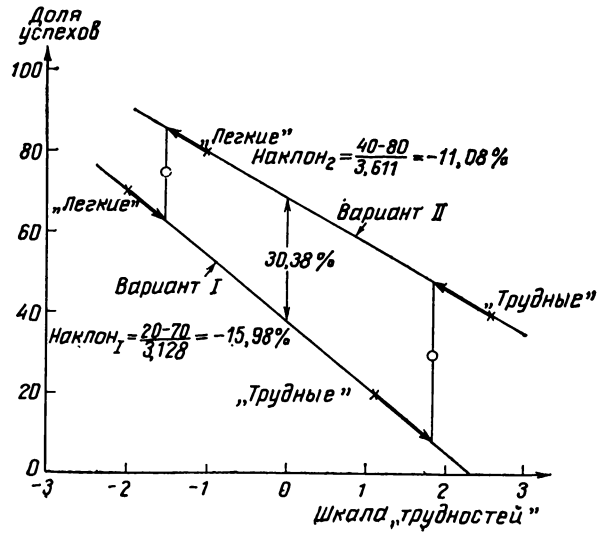


Иллюстрация 11.7.1

Вычисления центров тяжести для илл. 11.6.1

А. Вычисления положения центров тяжести на оси «трудности»

Состояние	Граница, %	$\varphi(p)$*	$P-p$	Нормированный центр тяжести	Разность	Значение** $\log(p/(1-p))$	Сдвиг (крайний)	Коррекция для ***
				Вариант I				
	0	0						
«Легкие»			0,4	—1,682			—1,276	—0,134
	40	—0,6730			2,114	—0,406		
«Полулегкие»			0,4	0,432		(0,490)		—0,058
	80	—0,5004			1,581	1,386		
«Полутрудные»			0,15	2,013		(2,165)		—0,152
	95	—0,1985			1,957	2,944		
«Трудные»			0,05	3,970			1,026	—0,125
	100	0						
				Вариант II				
	0	0						
«Легкие»			0,02	—4,900			—1,008	+0,134
	2	—0,0980			2,061	—3,893		
«Полулегкие»			0,08	—2,839		(—3,045)		+0,206
	10	—0,3251			2,143	—2,197		
«Полутрудные»			0,50	—0,696		(—0,896)		+0,200
	60	—0,6730			2,378	0,405		
«Трудные»			0,40	1,682			1,276	+0,125
	100	0						
				Медиана				
«Легкие» «Полулегкие»							—1,142±0,134	
«Полутрудные», «Трудные»							1,151±0,125	

* $\varphi(p) = p\log_e p + (1-p)\log_e(1-p)$.

** Значение в скобках есть середина между соседними значениями $\log(p/(1-p))$.

*** Для крайних категорий: значение нормированного центра тяжести МИНУС медиана той же категории (по вариантам). Для внутренних категорий: нормированное значение центра тяжести МИНУС середина (они приведены в скобках).

Б. Процент успехов по вариантам

Объединенные проценты	Категория	Вариант I	Вариант II	Разность	Скорректированные проценты I	Скорректированные проценты II	Разность
21	«Легкие», %	90	100	10	87,5	101,6	14,1*
	Угловой коэффициент	—18,92	—12,13				
24	«Полулегкие», %	50	75	25	49,0	77,4	28,4
	Угловой коэффициент	—16,89	—11,67				
32,5	«Полутрудные», %	23,3	50	26,7	20,7	52,3	31,6
	Угловой коэффициент	—6,80	—9,46				
22,5	«Трудные», %	10	27,5	17,5	9,2	28,7	19,5
	Взвешенное среднее			20,7	Взвешенное среднее		24,4

* Заметим, что только 2% всего варианта I приходятся на эту строчку.

Иллюстрация 11.7.2

График долей успехов для варианта I в зависимости от центров тяжести, измеряющих «трудности»

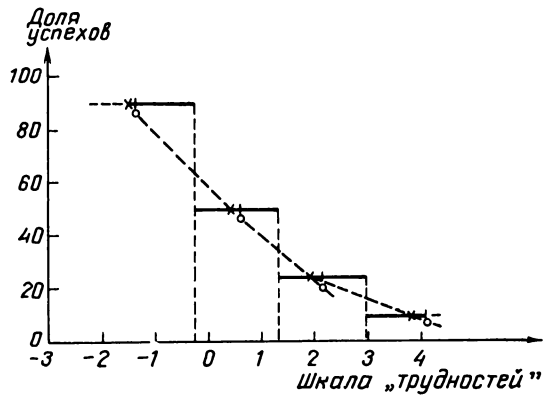


Иллюстрация 11.7.3

Фиксация центров тяжести по граничной точке для внешних категорий на шкале «трудностей»

Содержание крайней категории, %	Координата центра тяжести	Содержание крайней категории, %	Координата центра тяжести
99	4,652	40	1,277
98	3,991	20	1,116
95	3,153	10	1,054
90	2,012	5	1,026
80	2,557	2	1,008
60	1,627	1	1,005

Иллюстрация 11.7.4

Разности (вариант II МИНУС вариант I) в процентах успехов («Без учета поправок» означает здесь, что разности в «трудностях» среди групп «легкие» и среди групп «трудные» не принимались в расчет, а «С учетом поправок» — что принимались).

Пример одной общей группы: 44%—60% = —16%.

Пример двух групп (%):

	Без учета поправок	С учетом поправок
Среднее	15	31,1
Взвешенное среднее	15,5	31,9

Пример четырех групп (%):

	Без учета поправок	С учетом поправок
Среднее	19,8	23,4
Взвешенное среднее	20,7	24,4

Приложение (к гл. 6) ○ ПОДРОБНОСТИ НАСЧЕТ ТОГО, НУЖНЫ ЛИ НАМ ПРЕОБРАЗОВАНИЯ

А. ОБЩИЙ СЛУЧАЙ

Перед нами стоит вопрос: «Надо ли нам переходить от малопригодного выражения для x (называемого ниже «исходным носителем») к более подходящему выражению (называемому ниже «спрямляющим носителем»)?»

Мы полагаем заведомо ясным, что спрямляющий носитель должен быть лучше, и вопрос заключается лишь в том «*достаточно* ли этого улучшения для оправдания хлопот, которые могут быть связаны с вычислениями, но прежде всего с интерпретацией и описанием данных?». Часто мы стремимся получить ответ на этот вопрос, опираясь на опыт, который удалось приобрести при построении регрессии на исходном носителе, чтобы не связываться с построением (пусть несколько лучшей) регрессии на спрямляющем (более подходящем) носителе.

Тогда основой для решения служат такие данные:

- насколько мы согласны ухудшить модель, сохраняя исходный носитель?
- похожи ли исходный и спрямляющий носители?
- сколь хороша была модель для исходного носителя?
- сколько было экспериментальных точек?

Мы измеряем первые три из них таким образом.

Ухудшение модели — через δ, где средний квадрат ухудшения модели (разновидность среднего квадрата ошибки, исследуемая в параграфе Д) в $(1+\delta)$ раз больше, когда вместо спрямляющего носителя используется исходный.

Близость между исходным и спрямляющим носителями либо через $r_{\text{носит}}$, коэффициент парной корреляции между ними, либо через ε, где

$$r^2_{\text{носит}} = \text{корреляция}^2 (x_{\text{исх}},\ x_{\text{спрямл.}}) = \frac{1}{1+\varepsilon^2}$$

(величина ε часто служит простейшим основанием для табулирования или обсуждения. Ее интерпретация проводится в параграфе Д).

Качество подгонки для исходного носителя — через соответствующий коэффициент парной корреляции $r_{\text{исх. мод}}$, где

$$r^2_{\text{исх. мод}} = \text{корреляция}^2 (y,\ x_{\text{исх}}),$$

$$1 - r^2_{\text{исх. мод}} = \frac{\text{дисперсия остатков от модели для } x_{\text{исх}}}{\text{дисперсия откликов до регрессии}}.$$

Теперь три первых вопроса звучат так:

- как мы судим о том, что δ приемлема?
- как практически оценить $r^2_{\text{носит}}$ (или ε)?
- как найти действительное δ, используя, скажем n, $r^2_{\text{исх.мод}}$ и ε?

Как уже упоминалось, мы отвечаем на второй вопрос для наиболее часто встречающихся преобразований величины $r^2_{\text{носит}}$ так:

- (чаще всего) оценкой служит отношение «x наибольшее/x наименьшее» для исходного носителя;
- (несколько вариантов) сравнить значения x для нормального распределения обычного, с «распущенными» или с «поджатыми» «хвостами» (с тем, что было бы, например, для равномерного распределения).

Наиболее важная зависимость для $r^2_{\text{носит}}$ — это отношение наибольшего x к наименьшему, но нам следует учитывать и некоторые вариации формы распределения значений x. Мы еще продолжим обсуждение этого отношения в параграфах В и Г.

Можно представить δ, как говорилось, в терминах ε^2, n и $r^2_{\text{исх. мод}}$.

$$\delta = \frac{\varepsilon^2}{1+\varepsilon^2}\left(\frac{(1+(n-1)\,\varepsilon^2)\,r^2_{\text{исх. мод}}}{1-(1+\varepsilon^2)\,r^2_{\text{исх. мод}}} - 1\right).$$

Рассмотрим три значения δ, а именно:

$\delta = 1$, когда мы не слишком заботимся о качестве модели и не морочим себе голову средним квадратом ухудшения;

$\delta = 0{,}1$, когда мы живо интересуемся качеством модели, но легко допускаем прирост среднего квадрата ухудшения до 10%;

$\delta = 0{,}01$, когда мы крайне озабочены качеством модели и лишь миримся с приростом среднего квадрата ухудшения на 1%.

Мы считаем два крайних значения $\delta = 1$ и $\delta = 0{,}01$ предельными, а промежуточное значение $\delta = 0{,}1$ умеренным.

Иногда вместо того, чтобы считать δ, мы предпочитаем просмотр таблицы значений $r^2_{\text{исх.мод}}$ для трех различных значений δ. Она представлена на илл. А.1.

Некоторые примеры. Пусть $n = 100$ и $r^2_{\text{носит}} = 0{,}9412$, так что $\varepsilon = 0{,}25$. ($И$ ($1 - r^2_{\text{носит}}$), и ε^2 показывают, что если бы мы перешли от исходного носителя к спрямляющему, то выиграли бы около 6% дисперсии). Допустим еще, что в совокупности $r^2_{\text{исх.мод}}$ примерно равно 0,50. Что мы могли бы извлечь из этих данных?

Если взглянуть в таблицу илл. А.1 для $n = 100$ и $\varepsilon = 0{,}25$, то мы увидим, что: (А) надо взять $r^2_{\text{исх. мод}}$ равным 0,68, чтобы получить $\delta = 1$ и $1 + \delta = 2$; (Б) для $r^2_{\text{исх. мод}}$ 0,27 получается $\delta = 0{,}1$ и $1 + \delta = 1{,}1$. Если же нам захочется допустить рост среднего квадрата ухудшения до $\delta = 1$, то нет никакой нужды в преобразованиях; зато если нам не хочется выходить за границу $\delta = 0{,}1$, то преобразование неизбежно.

Если мы решим сосчитать δ, то найдем

$$\delta = \frac{0{,}0625}{1{,}0625}\left(\frac{(1+(100-1)\,(0{,}0625))\,(0{,}50)}{1-(1+0{,}0625)\,(0{,}50)} - 1\right) = 0{,}39.$$

Для обычных моделей среднего качества (то, что для одной области практики будет отличным, в другой может, конечно, не подойти вовсе) нельзя принять $\varepsilon = 0{,}25$, если мы хотим, чтобы модель была работоспособной.

А что, если бы ε было 0,1 и наблюденное значение $r^2_{\text{исх. мод}}$ около 0,50? При $n = 100$, из илл. А.1 мы найдем, что истинные значения $r^2_{\text{исх. мод}}$ равны 0,971; 0,84 и 0,50 соответственно для δ1; 0,1 и 0,01. Значит, наше предполагаемое наблюденное значение $r^2_{\text{исх. мод}}$ падает как раз на $\delta = 0{,}01$. Если мы не хотим, чтобы модель ухудшилась более чем на 1%, то использование исходного носителя вполне допустимо. (Если бы наша исходная модель была чуть лучше с $r^2_{\text{исх. мод}} = 0{,}6$ или 0,7, то мы могли бы угодить между этими двумя значениями, при ухудшении более чем 1%, но менее чем 10% для исходного носителя; ну, а если бы $r^2_{\text{исх. мод}} = 0{,}84$, то мы бы имели дело с ухудшением более 10%.)

Ограничившись исходным носителем или найдя лучшие, все равно мы должны сделать модель достаточно эффективной.

ОТНОШЕНИЕ К ЗНАЧИМОСТИ-НЕЗНАЧИМОСТИ

Понятно, что повысить качество модели, измеряемое величиной $r^2_{\text{исх. мод}}$, это то же самое, что увеличить значение δ. Значит, при данном $r^2_{\text{носит}}$ или ε, а также при известном n получения большого значения $r^2_{\text{исх. мод}}$ достаточно, чтобы сде-

лать применение исходного носителя неприемлемым. Совсем другие вещи получаются, если начать с $r^2_{\text{исх. мод}} = 0$ и постепенно его увеличивать и увеличивать. В частности, модель, искони незначимо отличная от нуля, сначала станет значимой на 5%-ном, а потом и на 1%-ном уровне значимости.

Если модели не принимать всерьез, пока они в пределах или почти в пределах значимости, то можно наметить такую последовательность их проявления:

- сначала модели не принимаются всерьез, поскольку они не выходят за границы значимости;
- затем модели становятся более содержательными, но не имеют предсказательной силы;
- наконец, модели обретают предсказательную силу.

Эти три варианта возникают в том случае, если значимость наблюдается при меньшем $r^2_{\text{исх. мод}}$, чем тот, который дает существенный прирост предсказательной силы.

С другой стороны, если для демонстраций предсказательной силы нужен сам по себе меньший $r^2_{\text{исх. мод}}$, чем тот, который обеспечивает значимость, то средний вариант пропадает, и нам остается выбор лишь между моделями, которые «не стоит принимать всерьез» и «способными прогнозировать с самого начала».

Таким образом, довольно важно, что получится при меньших $r^2_{\text{исх. мод}}$ — значимость или способность прогнозировать. И если способность прогнозировать придет раньше, чем значимость, то любую модель, с какой бы тщательностью мы ее ни построили, следовало бы переделать заново со спрямляющим носителем.

Давайте подытожим наши мысли. Зафиксируем временно (1) объем выборки и (2) ε^2 или $r^2_{\text{носит}}$ и посмотрим, что случится при изменениях $r^2_{\text{исх.мод}}$. Если $r^2_{\text{исх.мод}}$ достаточно мал, и модель с исходным носителем будет, видимо, незначима, то остается предположить, что либо мы вовсе пренебрежем такой моделью, либо попытаемся применить спрямляющий носитель и посмотреть, не будет ли он давать значимую модель. Если же, наоборот, $r^2_{\text{исх.мод}}$ велик (близок к 1), то и значение δ тоже станет большим, а величина $r^2_{\text{исх.мод}}$ может столь возрасти, что соответствующее значение δ окажется просто недопустимым. Следовательно, интервал значений $r^2_{\text{исх.мод}}$, для которых имеет смысл оставлять исходный носитель, ограничен с двух сторон—с одной стороны из-за незначимости, а с другой — по причине непригодности δ. По мере смещения требований к нетерпимости, снижая δ, мы можем допустить, чтобы этот рабочий интервал начал сжиматься и, наконец, совсем исчез.

Возможны случаи, где значения $r^2_{\text{исх.мод}}$ бесполезны. Они обозначены на илл. А.1 буквой N, что означает «незначим или неприемлем» для 5%-ного уровня значимости, ограничивающего интервал пригодных значений. Еще мы заключали в скобки значения, для которых интервал пригодных значений вырождается, когда граница значимости снижается до 1%.

Конечно, в жизни мы знаем только выборочную оценку $r^2_{\text{исх.мод}}$ и вынуждены брать именно ее при пользовании илл. А.1, в которой использовались генеральные (истинные) значения $r^2_{\text{исх.мод}}$. Но это кажется достаточно корректным отчасти потому, что наш выбор δ редко действительно точен. Вот почему мы рекомендуем пренебрегать этим обстоятельством при работе с илл. А.1.

Из данных илл. А.1 и появления в ней буквы N и скобок можно извлечь несколько простых ориентировочных правил.

1. Если приемлемо удвоенное значение среднего квадрата ухудшения для исходного носителя

- и если выявлена безусловная зависимость между откликом и исходным носителем, значимая лишь на 5%-ном уровне, то мы можем работать даже с таким большим значением ε, как 0,6 (что соответствует $r^2_{\text{носит}}$ порядка 0,74) (это дает объемы выборок где-то между 10 и 100);
- но если мы хотим лучше подогнать исходный носитель, то при больших $r^2_{\text{исх.мод}}$ нельзя работать со столь большими значениями ε и мы обязаны взять $r^2_{\text{носит}}$ более близким к 1;

● и если модель, как мы можем судить по многим опытам, с точки зрения качества представляется «не слишком хорошей» для физика-новичка, поскольку, скажем, ее остатки колеблются около 1% и имеют тот же порядок, как разности $(y - \bar{y})$, так что $1 - r^2_{\text{исх.мод}} \approx 0{,}0001$, то надо снижать ε до величины порядка 0,007 (для $r^2_{\text{носит}} \approx 0{,}99995$).

2. Если мы хотели бы иметь модель умеренного качества ($\delta = 0{,}1$).

● когда близость между откликом и исходным носителем безусловно значима на 5%-ном уровне (этого достаточно, чтобы определить модель), мы, может быть, будем готовы дойти до значений ε, несколько меньших, чем 0,3, что дает для $r^2_{\text{носит}}$ число 0,92 (это приводит к выборкам объемом между 30 и 100);

● где близость выше, то мы должны уменьшать ε, может быть, до 0,1 (и $r^2_{\text{носит}}$ между исходным и спрямляющим носителями будет 0,99);

● когда близость такова, как в странном эксперименте нашего физика-новичка, и соответствует, скажем, остаткам порядка 0,1%, как и у разностей $(y - \bar{y})$, то мы будем вынуждены понизить ε до 0,0003 или даже еще ниже (что отвечает $r^2_{\text{носит}} \approx 0{,}999999$ или еще выше для корреляции между исходным и спрямляющим носителями).

3. Если, наконец, нам нужна очень хорошая модель ($\delta = 0{,}01$),

● где близость безусловна, то мы должны быть готовы идти до значения $\varepsilon = 0{,}15$, $r^2_{\text{носит}} = 0{,}98$ (и снова $n \approx 30 - 100$);

● когда близость еще выше, то надо останавливаться на $\varepsilon = 0{,}05$ и, как и выше, $r^2_{\text{носит}} = 0{,}997$.

Лучшее, что можно попытаться сделать, т. е. уменьшить δ, насколько возможно, и улучшить подгонку (увеличить $r^2_{\text{исх. мод}}$), значит, начать с наилучшего, на что мы только способны. Тогда мы должны держать меньшим ε или $(1 - r^2_{\text{носит}})$ для связи между исходным и спрямляющим носителями, если мы сочтем разумным, чтобы работал исходный носитель.

Б. СЛУЧАЙ ОЧЕНЬ ХОРОШЕЙ МОДЕЛИ

Если мы сталкиваемся с очень хорошими моделями (с большими $r^2_{\text{исх.мод}}$) и, следовательно, с малыми ε, то при данном δ отношение

$$\frac{1 - r^2_{\text{исх. мод.}}}{1 - r^2_{\text{носит}}}$$

оказывается практически неизменным.

На илл. Б.1 приведены значения этого отношения для малых ε и различных n при $\delta = 1$, 0,1 и 0,01. Для получения очень хорошей аппроксимации

● мы должны быть уверены, что $(1 - r^2_{\text{носит}})$ составляет не более половины от $(1 - r^2_{\text{исх. мод}})$, и тогда будет не хуже, чем удвоение при использовании исходного носителя;

● мы должны держать $(1 - r^2_{\text{носит}})$ не выше, чем 1/10 (а практически около 1/11) от $(1 - r^2_{\text{исх. мод}})$, и тогда будет не хуже, чем 10% превышения;

● мы должны держать $(1 - r^2_{\text{носит}})$ не свыше 1/100 от $(1 - r^2_{\text{исх. мод}})$, и тогда будет не хуже, чем 1% превышения.

В области хороших моделей все легко описывается. Мы видим, что часто надо иметь

$$1 - r^2_{\text{носит}},$$

которые несколько или весьма существенно меньше, чем

$$1 - r^2_{\text{исх, мод}},$$

если мы не собираемся переходить от $x_{\text{исх}}$ к $x_{\text{спрямл}}$.

В. НУЖНЫ ЛИ НАМ ЛОГАРИФМЫ?

Пока мы получаем ответы в терминах $r^2_{\text{носит}}$, еще можно как-то существовать. Но часто нужны даже более простые ответы на вопросы такого сорта: если мы убеждены в пригодности $\log x$, то стоит ли брать сам x? Если мы верим в $\sqrt{x}$, то стоит ли брать x? Если мы ставим на $-1/x$, то нужен ли x?

Относительно простые ответы на подобные вопросы содержит отношение наибольший x/наименьший x по крайней мере для хорошо себя ведущих множеств x.

Будучи уверенными в этом, мы можем лишь подсчитать $r^2_{\text{носит}}$ между множеством из n штук x и соответствующим множеством их логарифмов. Результаты будут зависеть от размаха n значений и от вариации формы распределения в выборке. Мы остановились на трех основных вариантах распределений:

- равномерном (стиснутом концами);
- приближенно нормальном (натянутом между i-м и $(i+1)$-м порядковыми значениями пропорционально $1/i\ (n-i)$);
- с растянутыми «хвостами» (значениями тангенсов подходящей равномерной выборки).

Каждое из них может использоваться для описания как распределения x, так и для $\log x$, таким образом несколько эксплуатируя эффект эксцесса.

На илл. В.1 представлены выборки объемом 10 и 20. Разнообразие распределений бросается в глаза. Крайние случаи, видимо, гораздо более экстремальны, чем те примеры, которые будут появляться на практике.

На илл. В.2 показаны очевидные зависимости $r^2_{\text{носит}}$ от отношения наибольший x/наименьший x для трех групп ситуаций, причем для каждой представлены только наиболее удаленные от центра кривые, а именно от NW к SE:

- распределение с растянутыми «хвостами» симметрично для $\log x$ при $n = 10$, а для x оно симметрично и при $n = 10, 20, 40$ (пунктирные границы);
- распределение с обычными «хвостами» ($\sim$ нормальное) симметрично для $\log x$ при $n = 10$, а для x оно симметрично и при $n = 5, 10, 20, 40$ (сплошные границы);
- распределение, равномерное на отрезке, симметрично для $\log x$ при $n = 10$, а для x оно симметрично и при $n = 10, 20, 40$ (точечные границы).

Илл. В 2 охватывает диапазон отношений примерно от 15 до 1,05 и соответствующие им значения $r^2_{\text{носит}}$ от порядка 0,85 до 0,9999. Значения $r^2_{\text{носит}}$ для разных n очень близки, а их изменения от $n = 20$ до $n = 40$ так малы, что мы не чувствуем никаких расхождений между ними на графике илл. В.2.

Для конкретной демонстрации ухудшения (в зависимости от δ) и тесноты подгонки (измеряемой $r^2_{\text{исх. мод}}$), рассмотренной в параграфе А, мы установили, что когда $\log x$ — хороший носитель, использовать:

1. Если допустимо удвоение

- и надо рассматривать безусловные модели, то (поскольку мы видим, что $r^2_{\text{носит}}$ может снизиться до 0,74) мы можем пройти весь диапазон графика илл. В.2 до отношений, несколько больших, чем 20, не испытывая нужды в преобразованиях;
- но когда мы хотим сохранить исходный носитель при более тесной подгонке, то нам надо добиваться (см. параграф А) для $r^2_{\text{носит}}$ чего-нибудь близкого к 0,8 или 0,9, и мы попадем в верхнюю часть рисунка на илл. В.2, где можно допустить отношение (наибольший x)/(наименьший x) около 5;
- и когда мы стремимся к качественной модели, может быть, и «не слишком хорошей» для многих физиков-новичков, имеющей $r^2_{\text{носит}} = 0{,}99995$ (см. параграф А), тогда мы оказываемся в нижней части картинки В.2 и должны иметь неудобные значения отношения (наибольший x)/(наименьший x) вроде 1,05 или 1,1.

2. Если желательна умеренная эффективность ($\delta = 0{,}1$)

- для безусловно определяемых моделей, то мы работаем в верхней части картинки В.2 ($r^2_{\text{носит}} = 0{,}92$ или $0{,}93$) и удовлетворимся отношениями порядка 8 (для нормального распределения), 10 (для равномерного на отрезке) и 4 (для растянутых «хвостов»);

● для более тесных подгонок, то мы попадаем примерно в середину картинки илл. В.2 ($r^2_{\text{носит}} = 0{,}99$) и можем работать с отношениями около 2 (нормальное), 2,5 (равномерное) и 1,8 (растянутое);

● когда близость такова, как в «блестящем» эксперименте физика-новичка, т. е. (см. параграф А) нам нужны $r^2_{\text{носит}}$ порядка 0,999999, то мы оказываемся левее картинки илл. В.2 и можем настаивать на исходном носителе, даже если отношение дойдет до 1,02.

3. Если нужна очень хорошая модель ($\delta = 0{,}01$),

● где близость безусловна, то мы используем часть графика илл. В.2, которая чуть ниже вершины ($r^2_{\text{носит}} = 0{,}98$), и можем существовать с отношениями порядка 2,5 (нормальное), 3 (равномерное) и 2 («хвостатое»);

● с гораздо бо́льшей теснотой, чем безусловная, то мы оказываемся чуть ниже середины графика илл. В.2 ($r^2_{\text{носит}} = 0{,}997$) и нас устраивают отношения порядка 1,4 (нормальное), 1,5 (равномерное) или 1,3 («хвостатое»).

Понятно, что подходящее отношение (наибольший x)/(наименьший x) весьма зависит от того, сколь хорошо подогнанную модель мы хотим иметь, и, в несколько меньшей степени, от того, сколь точной она должна быть, т. е. насколько малым мы хотим сохранить δ.

Г. А КАК НАСЧЕТ $\sqrt{x}$ И $-1/x$?

Мы довольно подробно рассмотрели вопрос о том, когда «нужно» использовать $\log x$. Ну а как быть, если «нужно» брать $\sqrt{x}$ или $-1/x$?

На илл. Г.1 сопоставлены кривые для всех трех случаев. Для простоты показаны только равномерное и «хвостатое» распределения. В свете нашего опыта с логарифмами, который показал, что объем выборки не слишком важен, мы ограничились на рисунке $n = 10$. Чтобы было легче сравнивать с илл. А. 1. и обсуждать результаты в связи друг с другом, вертикальная ось проградуирована в величинах ε (более удобных, чем $r^2_{\text{носит}} = 1/1(+\varepsilon^2)$). Этот рисунок вместе с илл. А.1 при их совместном обсуждении показывает, сколь важным может быть преобразование x, если только знать, что оно должно быть испытано.

Так, для примера, если $\varepsilon = 0{,}1$— искомое качество, то мы, вероятно, можем пренебречь преобразованием для отношений (наибольший x)/ (наименьший x не больших, чем

● от 1,3 («хвостатое») до 1,4 (равномерное), если испытывается $-1/x$;

● от 1,7 («хвостатое») до 2,0 (равномерное), если испытывается $\log x$;

● от 3,5 («хвостатое») до 5 (равномерное), если испытывается $\sqrt{x}$.

Возвращаясь к илл. А.1, мы видим, что $\varepsilon = 0{,}1$ соответствует:

● удвоению среднего квадрата ухудшения для исходного носителя свыше $r^2_{\text{исх. мод}} = 0{,}98$ ($n \approx 5$ или 10), 0,97 ($n \approx 100$) и 0,95 ($n \approx 300$);

● увеличению на 10% среднего квадрата ухудшения свыше $r^2_{\text{исх. мод}} = 0{,}90$ ($n \approx 10$), 0,84 ($n \approx 100$) и 0,73 ($n \approx 300$);

● увеличению на 1% среднего квадрата ухудшения свыше $r^2_{\text{исх. мод}} = 0{,}60$ ($n \approx 30$), 0,50 ($n \approx 100$) и 0,33 ($n \approx 300$). Короче говоря, у нас есть приемлемое руководство к действию почти в любой ситуации.

В частности:

● хотя метод был описан для регрессии с одним носителем, мы не испытываем угрызений совести, используя это руководство и для множественной регрессии;

● если нам нужно $\varepsilon < 0{,}01$, то мы можем преобразовывать, не глядя на илл. Г.1;

● хотя и могут встретиться некоторые крайние случаи, при использовании ε, большего, чем 0,5, мы не станем ждать такого случая, чтобы им воспользоваться; просто если нет $\varepsilon < 0{,}5$, то мы будем хоть как-то преобразовывать;

● когда анализ илл. Г.1. оставляет нас в сомнениях, мы их отбрасываем и преобразуем носитель.

Д. ОБОСНОВАНИЕ

Теперь мы вернемся к исходной ситуации и выведем формулы. Давайте допустим, что:

1) случайные величины x и z имеют нулевые средние, равные дисперсии (скажем, единичные) и нулевую ковариацию;

2) в качестве носителя следовало бы взять x, но мы собираемся использовать вместо него $x + \varepsilon z$, где ε — меры числа введенных несоответствий. Тогда $x = x_{\text{спрямл}}$, а $x + \varepsilon z = x_{\text{исх}}$ в наших старых обозначениях;

3) тогда сколько мы теряем в качестве оценивания регрессионного коэффициента b при использовании $x + \varepsilon z$ вместо x?

Мы не теряем в общности, полагая, что среднее 0 и дисперсия 1 одновременно для x и для z и, далее, что откликом в регрессии служит сам x, отягченный независимыми ошибками σe_i со средним 0 и дисперсией σ^2, т. е. что желаемое значение коэффициента регрессии b есть 1,00. Таким образом, отклик

$$y_i = x_i + \sigma e_i, \quad i = 1, 2, \ldots, n.$$

Мы будем судить о качестве оценки b не по среднему квадрату ошибки, которая есть

$$(\text{смещение } b)^2 + (\text{действительная дисперсия } b),$$

а по величине, которую мы назвали *средним квадратом ухудшения*

$$(\text{смещение } b)^2 + \text{средн. арифм. (кажущаяся дисперсия } b),$$

поскольку мы прогадываем, когда думаем, что дисперсия b больше, чем она есть на самом деле. Выражение «средн. арифм. ( )» означает операцию усреднения (взятия математического ожидания) для стоящего в скобках.

Мы берем среднюю кажущуюся дисперсию, ибо любая систематическая зависимость y от x, которую нельзя свести к линейной зависимости, вносит вклад в остатки $(y - \hat{y})$ и, следовательно, в сумму их квадратов. Отсюда величина s^2, служащая нам условной оценкой для var $\{b\}$ — дисперсии b, должна основываться на

$$\text{средн. арифм } \{s^2\} = \sigma^2 + \frac{1}{n+1} \sum (\text{систематич. отклонен.})^2,$$

и мы будем обычно иметь

$$\text{средн. арифм. } \{s^2\} > \sigma^2,$$

где σ^2 — средняя дисперсия y (относительно суммы для прямой и систематических отклонений).

Поскольку на самом деле мы не знаем, как разделяется среднее s^2 между σ^2 и систематической составляющей, нам приходится удовлетвориться кажущейся дисперсией b, среднее значение которой больше, чем фактическая дисперсия b. Средний квадрат ухудшения дает некоторую меру этого несоответствия (а средний квадрат ошибки не дает).

Теперь мы получим выражение для среднего квадрата ухудшения при наших гипотезах. По определению имеем

$$b = \frac{\Sigma (x_i + \sigma e_i)(x_i + \varepsilon z_i)}{\Sigma (x_i + \varepsilon z_i)^2}.$$

Если множества x и z заданы, то из наших предположений для средних, дисперсий и ковариаций следует, что

$$\Sigma x_i = \Sigma z_i = 0; \quad \frac{\Sigma x_i^2}{n-1} = \frac{\Sigma z_i^2}{n-1} = 1; \quad \Sigma x_i z_i = 0.$$

Для получения ожидаемого значения b мы просто перемножаем и складываем числа в числителе и знаменателе, а затем берем математическое ожидание числителя. Тогда

$$\text{средн. арифм. } b = \text{средн. арифм.} \left[\frac{\sum x_i^2 + \sigma \sum x_i e_i + \varepsilon \sum x_i z_i + \sigma\varepsilon \sum e_i z_i}{\sum x_i^2 + 2\varepsilon \sum x_i z_i + \varepsilon^2 \sum z_i^2} \right] =$$

$$= \frac{\sum x_i^2}{\sum x_i^2 + \varepsilon^2 \sum z_i^2} = \frac{1}{1+\varepsilon^2}.$$

Поскольку мы приняли за истинное значение коэффициента единицу,

$$\text{смещение в } b = 1 - \text{средн. арифм. } b = \frac{\varepsilon^2}{1+\varepsilon^2}.$$

В гл. 14 (вып. 2) мы узнаем, что дисперсия регрессионного коэффициента равна остаточной дисперсии, деленной на сумму квадратов носителя, и значит

$$\text{var } b = \frac{\sigma^2}{\Sigma (x_i + \varepsilon z_i)^2} = \frac{\sigma^2}{(n-1)(1+\varepsilon^2)}.$$

Остатки. Остатки порождаются выражением

$$x_i + \sigma e_i - b(x_i + \varepsilon z_i) = (1-b) x_i + \sigma e_i - b\varepsilon z_i.$$

Его надо возвести в квадрат, взять математическое ожидание и просуммировать по i. После возведения в квадрат получаем

$$(1-b)^2 x_i^2 + \sigma^2 e_i^2 + b^2 \varepsilon^2 z_i^2 + 2\sigma (1-b) x_i e_i - 2(1-b) b\varepsilon x_i z_i - 2\sigma b \varepsilon_i z_i. \quad (1*)$$

Взятие математического ожидания и сложение первых трех членов дают

$$\sum x_i^2 \text{ средн. арифм. } (1-b)^2 + \sigma^2 \text{ средн. арифм. } \sum e_i^2 + \varepsilon^2 \sum z_i^2 \text{ средн. арифм. } b^2. \quad (2*)$$

Среди трех последних членов в (1*) средний есть сумма нулей, так как $\Sigma x_i z_i = 0$, в силу нашего постулата о ковариации. Остальные члены можно перегруппировать так:

$$2\sigma x_i e_i - 2\sigma (x_i + \varepsilon z_i) b e_i.$$

Член $2\sigma x_i e_i$ имеет нулевое ожидание, в связи с тем что $e_i = 0$, а $x_i =$ заданные константы. У нас остается теперь только

$$-2\sigma (x_i + \varepsilon z_i) b e_i.$$

Ожидание суммы можно записать так:

$$-2\sigma \sum (x_i + \varepsilon z_i) \text{ cov } (e_i, b), \quad (3*)$$

где член, содержащий произведение средних e_i и b исчезает, поскольку $\bar{e}_i = 0$.

Обобщая наши результаты, мы получим ожидаемую сумму квадратов в виде

$$\sum x_i^2 \text{ средн. арифм. } (1-b)^2 + \sigma^2 \text{ средн. арифм. } \sum e_i^2 - 2\sigma \sum (x_i + \varepsilon z_i) \text{ cov } (e_i, b) + \varepsilon^2 \left(\sum z_i^2\right) \text{ средн. арифм. } b^2. \quad (4*)$$

Давайте рассмотрим члены по одному. В первом из них $\Sigma x_i^2 = n - 1$, и мы всегда знаем математическое ожидание b и его дисперсию

$$\overline{(1-b)^2} = \{\overline{(1-b)^2}\} + \text{var } b = \left(\frac{\varepsilon^2}{1+\varepsilon^2}\right)^2 + \text{var } b.$$

И, значит, первый член из (4*) сводится к

$$(n-1)\left[\left(\frac{\varepsilon^2}{1+\varepsilon^2}\right)^2 + \frac{\sigma^2}{(n-1)(1+\varepsilon^2)}\right]. \quad (1)$$

Во втором члене из (4*) $\bar{e_i^2} = 1$, а, значит, сам член обращается в

$$n\sigma^2. \tag{2}$$

Чтобы найти третий член из (4*), нам нужно выражение для cov (e_i, b). Когда мы раскрываем сумму в числителе для b, получается сумма такого вида:

$$x_j^2 + \sigma e_j x_j + \varepsilon x_j z_j + \sigma\varepsilon e_j z_j,\ j = 1, 2, \ldots, n.$$

Это надо умножить на e_i и взять математическое ожидание. Только те значения, которые влияют на результат, а именно второе и четвертое, имеют $j = i$. Они дают, поскольку $\bar{e_i^2} = 1$,

$$\sigma(x_i + \varepsilon z_i),$$

откуда искомая ковариация есть

$$\operatorname{cov}(e_i, b) = \frac{(x_i + \varepsilon z_i)\,\sigma}{\Sigma (x_i + \varepsilon z_i)^2}.$$

С ее помощью мы сведем третий член из (4*) к выражению

$$-2\sigma^2. \tag{3}$$

Наконец, четвертый член из (4 *) дает

$$(n-1)\,\varepsilon^2\left(\frac{\sigma^2}{(n-1)(1+\varepsilon^2)} + \left(\frac{1}{1+\varepsilon^2}\right)^2\right). \tag{4}$$

Объединяя (1), (2), (3) и (4) и приводя подобные члены, получим ожидаемую сумму квадратов остатков

$$(n-1)\left(\frac{\varepsilon^2}{1+\varepsilon^2} + \sigma^2\right).$$

Если s^2 — выборочная дисперсия остатков, то ее ожидаемое значение равно:

$$\text{средн. арифм. } s^2 = \sigma^2 + \frac{\varepsilon^2}{1+\varepsilon^2},$$

а ожидаемое значение оцениваемой дисперсии b есть

$$\text{средн. арифм. } \frac{s^2}{\Sigma (x_i + \varepsilon z_i)^2} = \frac{\sigma^2 + \dfrac{\varepsilon^2}{1+\varepsilon^2}}{(n-1)(1+\varepsilon^2)}.$$

И наконец,

средний квадрат ухудшения (СКУ) = (смещение b)^2 + средн. арифм. (оцениваемая дисперсия b) $= \left(\dfrac{\varepsilon^2}{1+\varepsilon^2}\right)^2 + \dfrac{\sigma^2 + \dfrac{\varepsilon^2}{1+\varepsilon^2}}{(n-1)(1+\varepsilon^2)}$. (СКУ)

Размер неучтенного вклада. Один из возможных путей осмысления среднего квадрата ухудшения заключается в том, чтобы считать его пропорциональным изменению от того, что мы могли бы иметь «в идеале», т. е. при $\varepsilon = 0$. Тогда средний квадрат ухудшения был бы равен $\sigma^2/(n-1)$. Мы можем считать, что средний квадрат ухудшения есть эта величина, взятая $(1+\delta)$ раз, и тогда δ будет мерой, пропорциональной ухудшению.

Записывая

$$(1+\delta)\frac{\sigma^2}{n-1} = \left(\frac{\varepsilon^2}{1+\varepsilon^2}\right)^2 + \frac{\sigma^2 + \dfrac{\varepsilon^2}{1+\varepsilon^2}}{(n-1)(1+\varepsilon^2)},$$

мы можем получить отсюда σ^2, что после упрощения даст

$$\sigma^2 = \frac{n-1+\dfrac{1}{\varepsilon^2}}{\delta\left(\dfrac{1+\varepsilon^2}{\varepsilon^2}\right)^2 + \dfrac{1+\varepsilon^2}{\varepsilon^2}}. \qquad (5*)$$

Корреляции. Мы хотим найти долю дисперсии y, объясняющую для заданной регрессии величину ε^2. Чтобы вывести это соотношение, сначала вспомним, что без ошибки дисперсия x равна 1, а с учетом ошибки она равна $(1+\sigma)^2$. Отсюда мы выводим

средний квадрат остатка $= (1 - r^2_{\text{исх. мод}})$ (дисперсия y)

или

$$\sigma^2 + \frac{\varepsilon^2}{1+\varepsilon^2} = (1 - r^2_{\text{исх. мод}})(1+\sigma^2).$$

Решая это относительно $r^2_{\text{исх. мод}}$, имеем

$$r^2_{\text{исх, мод}} = \frac{1}{(1+\varepsilon^2)(1+\sigma^2)}, \qquad (6*)$$

где σ^2 — функция от n, δ и ε, как видно из (5*).

Выражение для δ. Решая (5*) относительно δ, получим

$$\delta = \left(\frac{(n-1)+\dfrac{1}{\varepsilon^2}}{\sigma^2} - \frac{1+\varepsilon^2}{\varepsilon^2}\right)\left(\frac{\varepsilon^2}{1+\varepsilon^2}\right)^2 = \frac{\varepsilon^2}{1+\varepsilon^2}\left(\frac{(n-1)\varepsilon^2+1}{(1+\varepsilon^2)\sigma^2} - 1\right).$$

Решая (6*) относительно σ^2, найдем

$$\sigma^2 = \frac{1 - r^2_{\text{исх,мод}} - \dfrac{\varepsilon^2}{1+\varepsilon^2}}{r^2_{\text{исх. мод}}}.$$

Отсюда

$$(1+\varepsilon^2)\sigma^2 = (1+\varepsilon^2)\frac{1 - r^2_{\text{исх.мод}} - \dfrac{\varepsilon^2}{1+\varepsilon^2}}{r^2_{\text{исх. мод}}} = \frac{1 - (1+\varepsilon^2) r^2_{\text{исх. мод}}}{r^2_{\text{исх. мод}}}.$$

Следовательно, мы можем записать

$$\delta = \frac{\varepsilon^2}{1+\varepsilon^2}\left(\frac{(1+(n-1)\varepsilon^2) r^2_{\text{исх.мод}}}{1-(1+\varepsilon^2) r^2_{\text{исх. мод}}} - 1\right), \qquad (7*)$$

или иначе

$$\delta = (1 - r^2_{\text{носит}})\left(\frac{1+(n-1)(1-r^2_{\text{носит.}})}{r^2_{\text{носит}} - r^2_{\text{исх. мод}}}\right) r^2_{\text{исх. мод}}.$$

РЕЗЮМЕ. КОГДА ЖЕ ВЫГОДНЫ ПРЕОБРАЗОВАНИЯ?

Мы можем оценить, оправданы ли хлопоты по преобразованию x в $\log x$ (или в $\sqrt{x}$, или в $-1/x$), когда мы знаем, что такое преобразование носителя, вообще говоря, хорошо. Вопросы, на которые приходится отвечать, таковы: на-

сколько это хорошо? Не будет ли это достижение, как большей частью и бывает, лишь достижением подгонки?

Мы подходим к первому вопросу с помощью двух коэффициентов корреляции: $r_{\text{носит}}$ — корреляции между x и $\log x$ (или другим предпочтительным носителем) и $r^2_{\text{исх. мод}}$ — корреляции между x (или несколькими x) и y.

Самый общий ответ таков : «Цена резко меняется, если только $1 - r^2_{\text{носит}}$ не будет составлять лишь малую долю от $1 - r^2_{\text{исх. мод}}$.

Подробности содержит илл. А.1, которая может быть нам полезна при принятии решений, когда это правило не совсем подходит.

Мы довольно хорошо оцениваем величину $r^2_{\text{носит}}$, пользуясь величиной отношения (наибольший x) /(наименьший x) и обращаясь к илл. В.2 (только для $\log x$) или к илл. Г.1 (в других рассмотренных случаях).

ИЛЛЮСТРАЦИИ

Иллюстрация А1.

(Истинные) значения $r^2_{\text{исх.мод}}$ между откликом и исходным носителем, соответствующие таким средним квадратам ухудшения, которые угловой коэффициент (А) удваивают, (Б) увеличивают на 10%, (В) увеличивают на 1%.

$\varepsilon=0{,}6$ $\varepsilon=0{,}5$ $\varepsilon=0{,}4$ $\varepsilon=0{,}3$ $\varepsilon=0{,}25$ $\varepsilon=0{,}2$ $\varepsilon=0{,}15$ $\varepsilon=0{,}1$ $\varepsilon=0{,}06$ $\varepsilon=0{,}03$ $\varepsilon=0{,}01$

А. Средний квадрат ухудшения — умножается на 2 $(=1+\delta)$

$n=5$	N	N	N	(0,84)	(0,88)	0,923	0,956	0,980	0,9928	0,9982	0,999800
10	N	(0,56)	(0,69)	0,81	0,87	0,917	0,954	0,980	0,9928	0,9982	0,999800
30	N	(0,38)	0,54	0,73	0,82	0,893	0,945	0,978	0,9925	0,9982	0,999800
100	N	(0,18)	0,31	0,54	0,68	0,82	0,916	0,971	0,9916	0,9981	0,999799
300	N	N	0,14	0,31	0,46	0,66	0,841	0,953	0,9891	0,9980	0,999797
1000	N	N	N	0,12	0,22	0,39	0,654	0,895	0,9804	0,9974	0,99790

Б. Средний квадрат ухудшения — умножается на 1,1 $(=1+\delta)$

$n=5$	N	N	N	N	N	N	(0,82)	(0,906)	0,963	0,9902	0,99890
10	N	N	N	N	(0,61)	(0,71)	0,81	0,902	0,962	0,9902	0,99890
30	N	N	N	(0,37)	0,48	0,61	0,76	0,888	0,960	0,9901	0,99890
100	N	N	N	(0,18)	0,27	0,41	0,62	0,84	0,952	0,9895	0,99889
300	N	N	N	N	(0,12)	0,22	0,41	0,73	0,930	0,9879	0,99887
000	N	N	N	N	N	0,08	0,19	0,50	0,860	0,9825	0,99880

В. Средний квадрат ухудшения — умножается на 1,01 $(=1+\delta)$

$n=5$	N	N	N	N	N	N	N	N	N	0,923	0,9901
10	N	N	N	N	N	N	N	(0,64)	0,78	0,922	0,9901
30	N	N	N	N	N	(0,36)	0,46	0,61	0,77	0,921	0,9901
100	N	N	N	N	N	(0,20)	0,31	0,50	0,73	0,917	0,9900
300	N	N	N	N	N	N	0,16	0,33	0,64	0,904	0,9898
11000	N	N	N	N	N	N	(0,06)	0,15	0,45	0,864	0,9891

Значения N почти всегда, а значения в скобках — как правило указывают на то, что исходная модель не годится (либо потому, что она слишком слаба (незначима), либо потому, что модель с линеаризующим носителем лучше). (См. раздел «Отношение к значимости — незначимости», с. 253, где даны дальнейшие разъяснения.)

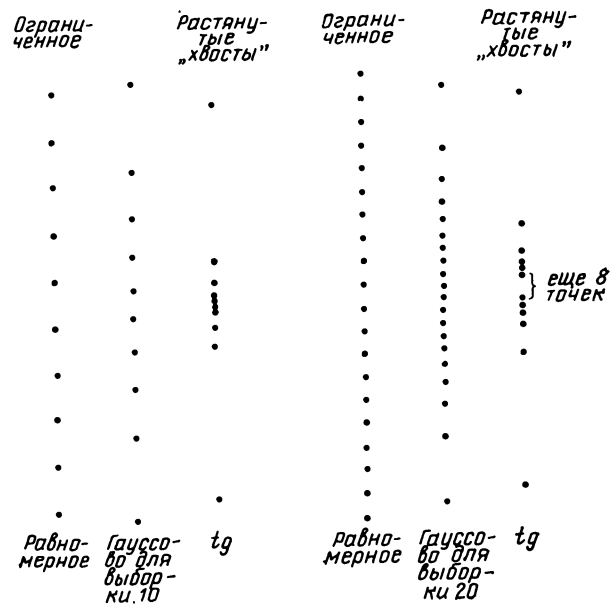
Иллюстрация Б.1

Отношение $(1-r^2_{\text{исх. мод}})$ к $(1-r^2_{\text{носит}})$ для малых ε

	$\varepsilon=0{,}01$	$\varepsilon=0{,}003$	$\varepsilon=0{,}001$
А. Для удвоения, $\delta=1$.			
$n=5$	2,00	2,00	
1000	2,10	2,01	2,00
3000	2,23	2,02	
Б. Для $\delta=0{,}1$			
$n=5$	10,98	11,00	
1000	11,97	11,09	11,00
3000	13,97	11,27	
В. Для $\delta=0{,}01$			
$n=5$	99,0	100,8	101,0
1000	108,7	101,8	101,2
3000	128,0	103,6	

Иллюстрация В.1

Три формы выборочных распределений для выборок объема 10 и 20. (Относительные масштабы не имеют значения.)



Первый столбец: x распределены равномерно.
Второй столбец: x распределены пропорционально:

Разности	1/9	1/16	1/21	1/24	1/25	1/24
Значения	−0,2829	−0,1718	−0,1093	−0,0617	−0,0200	0,0200	0,0617

	1/21	1/16	1/9
	0,1093	0,1718	0,2829

Третий столбец: x — пропорциональны:

Разности	tg 85°	tg 66,11°	tg 44,22°	tg 28,33°	tg 9,44°
Значения	11,4301	2,2578	1,0807	0,5392	0,1663

и их отрицательные значения.

Четвертый столбец: x — распределены равномерно.
Пятый столбец: x распределены пропорционально:

Разности	1/100	1/99	1/96	1/91	1/84	1/75	1/64
Значения	±0,00500	±0,01510	±0,02552	±0,03651	±0,04841	±0,06174	±0,07737

	1/51	1/36	1/19
	±0,09698	±0,12476	±0,17739

Шестой столбец: x распределены подобно:

Разности $\quad \operatorname{tg} 85°, \quad \operatorname{tg}\left(\frac{17}{19}\right) 85°, \ldots$

Значения $\quad \pm 11{,}4301, \quad \pm 4{,}0265, \quad \pm 2{,}3679, \quad \pm 1{,}6102, \quad \pm 1{,}1589, \quad \pm 0{,}8470$

В пропущенных на рисунке точках стоят числа (±0,6084; ±0,4115; ±0,2386; ±0,0782).

Заметим, что все масштабы не фиксированы и не согласованы.

Иллюстрация В.2

Зависимость $r^2_{\text{исх.мод}}$ от (наибольший x)/(наименьший x), когда спрямляющим носителем служит $\log x$. (Сплошные линии — обычные «хвосты»; пунктирные линии — растянутые «хвосты»; точечные линии — сжатые «хвосты»; подробности в тексте.)

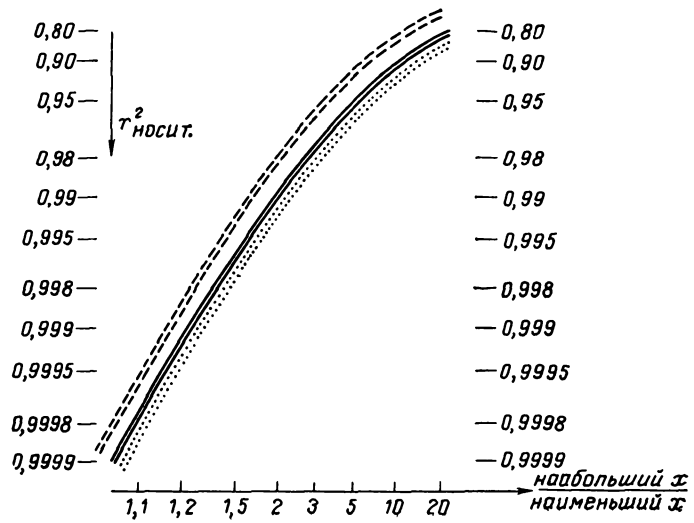
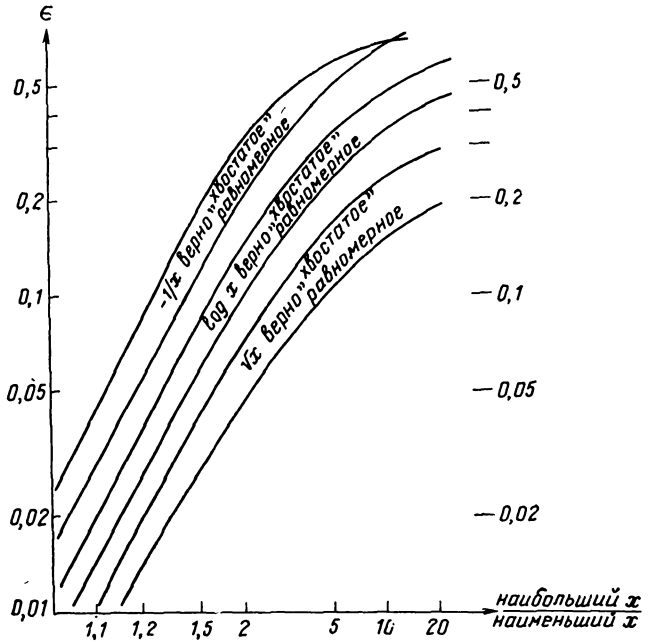


Иллюстрация Г. 1

Зависимость между ε и отношением (наибольший x)/(наименьший x), когда верны преобразования — $1/x$, $\log x$ или $\sqrt{x}$. (Графики для $n=10$, но они могут быть полезны и для других объемов выборки n.)



● ЗАДАНИЯ ДЛЯ УПРАЖНЕНИЙ

Все задания упорядочены по главам и параграфам, а внутри параграфов пронумерованы. Так, 7.3.4 обозначает 4-е задание в параграфе 7.3.

Надстрочный индекс «к», как, например, в $2.2.4^{к}$, означает, что это задание стоит использовать для коллективного решения

Звездочкой помечены задания повышенной трудности.

Индекс В, как в 16В4, соответствует заданию для внеаудиторной работы.

А индексом С, как в 14С1, помечены те задания, которые непосредственно связаны с другими главами.

Иллюстрации для упражнений — например, илл. 1 для упр. 1.5.3 — приводятся рядом с самими заданиями. Кроме того, у нас есть еще несколько массивов данных в разделе, озаглавленном «Приложение для упражнений», который следует сразу за данным разделом.

ГЛАВА 1

1.1.1. Выделите четыре стадии в организации анализа данных и сформулируйте некоторые «за» и «против» прохождения этих стадий в противоположных направлениях.

1.1.2.а. Как Стьюдент построил прямой путь вместо бесконечной лестницы?

б. Какие данные он при этом использовал?

1.1.3. Всегда ли, а если нет, то когда, прямой путь Стьюдента — это хорошая идея?

1.1.4. Дайте три разных примера t Стьюдента, причем так, чтобы учитывалась структура каждого конкретного приложения. В каждом случае предложите возможные «подходящие значения». (Выясните, что такое подходящее значение.)

1.1.5*. Проведите различие между «непараметрическими» и «свободными от распределения» методами.

1.2.1. Что Вы скажете о числе степеней свободы для критерия Стьюдента, задающего такие 95%-ные доверительные границы, которые в 2,29 раза шире, чем 2/3 доверительные границы?

1.2.2. Пользуясь критерием Стьюдента, найдите минимум отношения

$$\frac{\text{ширина } 95\%\text{-ных доверительных границ}}{\text{ширина 2/3 доверительных границ}}.$$

Какие степени свободы максимизируют это выражение? Приведите три самых больших значения (для целочисленных степеней свободы).

1.2.3. Найдите нормированные доверительные границы, основанные на критерии Стьюдента для а) 15 степеней свободы, б) 60 степеней свободы (см. примечание 2 к илл.1 2.1, с. 38).

1.2.4. При сколь больших значениях критерий Стьюдента не срабатывает? Особенно интересно, как при этом меняются критические значения.

1.2.5. Рассмотрите прошлые, настоящие и будущие недостатки стьюдентова «прорыва».

1.3.1. «Нормально» ли гауссово распределение? Почему да? Почему нет?

1.3.2. Все ли нормальные распределения имеют одну и ту же «форму»?

1.3.3. Сделайте набросок кривой плотности нормального распределения для $\mu = 1$, $\sigma = 2$. Не забудьте обозначить оси. (Вам может помочь илл. 1.3.1, с. 38)

1.3.4. Назовите какое-нибудь семейство распределений, в котором встречаются распределения самых разнообразных форм.

1.3.5. Когда о «хвостах» распределения говорят, что они «слишком растянуты», а когда — что «очень разбросаны»? На чем основан критерий сравнения этих выражений?

1.3.6. Можете ли Вы, опираясь на центр распределения, описать его «хвосты» и, наоборот, опираясь на «хвосты», описать центр? Почему? Почему нет? Когда да? Когда нет?

1.4.1. Что означают Q_1 и Q_3 в нормальном распределении? А что это будет для равномерного распределения, заданного на отрезке между нулем и единицей?

1.4.2. Уточните, что такое асимметрия и что такое эксцесс?

1.4.3. Приблизительно какая доля наблюдений лежит внутри интервала $0{,}25\,s$ относительно $\bar{x}$ для больших выборок? А за границами $3{,}1s$?

1.4.4. *Контрольная классная работа.* Отыщите какую-нибудь большую выборку и испытайте на ней некоторые из методов, примененных Уилсоном и Хилферти к данным Пирса. (В качестве источников воспользуйтесь, например, работой [Michelson A. A., Paese A. A. and Pearson F. (1935). Measurement of the velocity of light in a partial vacuum. — Astrophysical Journal, Vol. 82, p. 26—61 (табл. VI. 51—54)].

1.5.1. Как Вы думаете, важна ли нерегулярность функции на илл. 1.5.1 (с. 40)? Почему? Почему нет? В каких случаях да, а в каких — нет?

1.5.2. Что происходит с концами кривых на илл. 1.5.2 (с. 41)? Ответьте на тот же самый вопрос для илл. 1 5.3 (с. 41).

1.5.3. Почему отклонения от нормальной формы могут оказаться важными именно на «хвостах»?

1.5.4. Что Вы предпочитаете надежности статистического анализа?

1.5.5. Назовите и опишите не менее двух видов робастности.

1.5.6.[к] Что такое «случайный выстрел?»

1.5.7. Как поведение не на «хвостах» влияет на поведение «хвостов»? Важно ли это влияние? Почему? Почему нет?

1.5.8.[к] Как Вы думаете, велико ли различие в эффективности на 1%?.

1.5.9. Психопат расшвыривает случайным образом половину из множества карточек с числами, которые ему (или ей) предъявляют. Какова эффективность

среднего по оставшимся значениям относительно среднего по отброшенным? А какова она относительно среднего по всем значениям?

1.6.1. Перечислите и обсудите три концепции статистической неопределенности, причем как минимум одна из них не должна фигурировать в параграфах 1.6 или 1.7.

1.7.1^{к}. Перечислите «за» и «против» выборочной медианы как статистической свертки.

1.7.2. Выполните задание предыдущего упражнения, заменив выборочную медиану на выборочное среднее.

1.7.3. Дайте три примера индикаций, два из которых не обсуждаются в параграфе 1.7.

1.8.1. Обсудите ценность анализа данных, дающего только индикацию.

1.8.2. Можно ли χ-квадрат, используемый как индикатор плохой подгонки, превратить в оцениватель?

1.8.3. Проведите различие между индикацией, подсчетом и выводом.

ГЛАВА 2

2.1.1. Почему рецепт из второго абзаца параграфа 2.1 «полностью оторван от действительности»?

2.1.2. Приведите три примера индикации с подробным объяснением, однако, таких, что не представлены прямо или не связаны явно с параграфом 2.1.

2.2.1. Найдите в сегодняшней газете три разных примера индикации. В каждом случае выясните, сколь важно не ограничиться только индикацией.

2.2.2. Найдите и обсудите два примера задач множественности, которые Вам интересны.

2.2.3. Дотошный исследователь проверял каждый из 137 препаратов (на мышах) в надежде обнаружить среди них противораковые. Каждый препарат отдельно сравнивался с инертным материалом, так что понадобилось $2 \cdot 137$ групп мышей, т. е. 274. В итоге 15 препаратов оказались значимыми не менее чем на 5%-ном, 2 препарата — не менее чем на 1%-ном и 1 — на 0,1%-ном уровнях значимости. Обсудите результаты этого исследователя. На что они указывают?

2.2.4^{к}. Один агроном сравнивал 11273 образца растений, подвергнутых гибридизации. После упорядочения образцов по значимости оказалось, что наилучший соответствует 0,03%, 5-й из лучших — 0,17%, 25-й — 0,72%, 125-й — 2,8%, 625-й — 10,3%. Другой агроном работал с 1492 образцами совсем иного растения, но получил после упорядочения образцов по значимости точно такие же результаты. Если бы Вы взялись решать, (а) стоит ли кому-нибудь из агрономов перепроверять некоторые из его растений и (б) если стоит, то сколько, как бы Вы к этому подступились?

2.2.5. Дано одно-единственное множество данных, которое приходится использовать и для исследования, и для проверки. Если бы Вы выбрали треть данных для исследования и поиска «подсказок», а две трети — для проверки, то была бы пропорция сбалансирована? Почему? Почему нет?

2.2.6. Десять исследователей работают над одной и той же задачей. Ни одному из них не удалось получить значимых результатов, но 9 из 10 склоняются к одному направлению. Что Вы могли бы из этого заключить?

2.3.1. Отыщите в прессе не менее трех примеров завуалированных «неформальных» выводов. Выпишите соответствующие высказывания и обсудите, почему Вы находите или не находите эти выводы «явными».

2.4.1. Служит ли среднее генеральной совокупности параметром? В каком смысле?

2.4.2. Можно ли параметризовать семейство гауссовых (нормальных) распределений с помощью μe^{σ^2} и $\log \mu$, а с помощью μe^{σ^2} и $\sqrt{\mu^2 + \sigma^2}$? Почему? Почему нет?

2.4.3. Если бы Вы точно знали интересующий Вас вопрос и то, что любой из двух оценивателей приводит к подходящим оценкам, что еще Вы хотели бы узнать для выбора между этими двумя оценивателями? Могли бы Вы указать относительный объем счета, требуемого для обеспечения Вашего решения? Почему? Почему нет? Когда да? Когда нет?

2.4.4. Какая из приведенных ниже функций от χ-квадрат служит мерой качества подгонки модели (при объеме выборки n) для равного числа $k = \nu + 1$ наблюдений в ячейках χ^2, χ^2/k, $\sqrt{n}\,(\chi^2-\nu)$, $(\chi^2-\nu)/\sqrt{n}$, $(\chi^2-\nu)/n\nu$? Почему? Почему нет? Каково оцениваемое?

2.4.5. Возьмем оцениватель, который (1) в выборках из двух элементов всегда дает ответ либо «0», либо «∞», а (2) в довольно больших выборках (скажем, $\geqslant 100$) он обеспечивает ответы, накрывающие значения $\mu\, e^{\sigma^2}$. Что же будет тогда оцениваемым в выборках объема 2? В выборках объема 2000? Объясните Ваши ответы.

2.5.1. Рассмотрим срединное среднее (усеченное среднее при 25%-ном усечении с каждого конца) вместо среднего, усеченного на 10%, как было в параграфе 2.5. Чего можно ожидать от гауссовой эффективности? Каково отношение эффективностей срединного среднего и среднего при 10%-ном усечении в гауссовском случае?

2.5.2. (Продолжение упр. 2.5.1.) Если распределение начнет отклоняться от гауссовской формы в сторону более растянутых «хвостов», то что, по вашему мнению, случится с распределением эффективностей?

2.5.3^{к}. Найдите какую-нибудь таблицу (равномерных) случайных чисел и воспользуйтесь илл. 1 и 2 к данной задаче для получения выборок объемом 20 из распределений с *крайне* растянутыми «хвостами» и *просто* растянутыми «хвостами». (Причем каждый элемент надо согласовывать с некоторым числом выборок, не меньшим, чем три, взятых из разных частей таблицы случайных чисел.) Для каждой выборки подсчитайте (а) медиану, (б) срединное среднее (см. упр. 2.5.1), (в) среднее, усеченное на 10%. Систематизируйте результаты по каждой из шести комбинаций растянутости «хвостов» для каждой из трех оценок. Обсудите эти результаты и проведите сравнение трех оценок как индикаторов для каждой степени растянутости «хвостов». (Сохраните обе выборки и оценки всех вариантов для дальнейшего использования.)

Иллюстрация 1 к упражнению 2.5.3

Таблица для построения выборок из распределений с крайне растянутыми «хвостами» с помощью последовательностей случайных чисел.

А. СПОСОБ ПРИМЕНЕНИЯ

Извлекайте из таблицы равномерно распределенных случайных чисел цифры до тех пор, пока в последовательности не появятся либо все девятки, либо все нули. Пример (не совсем случайный): 1, 7, 2, 4, 90, 001, 4, 6, 94, 98, 07, 0000001, 999992. Ниже приведена таблица, где 1 соответствует -220, 7 дает 0,24, на 2 приходится $-1{,}45$, далее идет 4, потом 90, что дает $2{,}2\times 10^2$, а из 001 получается L и т. д.

Б. ТАБЛИЦА соответствий для значений от 001 до 998 (числа, обозначенные L, меньше -10^{100}, а те, что обозначены H, больше $+10^{100}$).

Вход		Вход		ВХОД		Вход		Вход	
000?	L	←		←		90	$2{,}20\times 10^2$	990	$2{,}69\times 10^{41}$
001	L	01	$-2{,}69\times 10^{41}$	1	$-220{,}0$	91	$6{,}69\times 10^2$	991	$1{,}80\times 10^{44}$
002	L	02	$-5{,}18\times 10^{19}$	2	$-1{,}45$	92	$2{,}68\times 10^3$	992	$1{,}94\times 10^{52}$
003	L	03	$-3{,}00\times 10^{12}$	3	$-0{,}24$	93	$1{,}60\times 10^4$	993	$1{,}10\times 10^{60}$
004	L	04	$-7{,}20\times 10^8$	4	$-0{,}07$	94	$1{,}73\times 10^5$	994	$2{,}41\times 10^{70}$
005	$-7{,}33\times 10^{84}$	05	$-4{,}85\times 10^7$	5	0,00	95	$4{,}85\times 10^7$	995	$7{,}33\times 10^{84}$
006	$-2{,}41\times 10^{70}$	06	$-1{,}73\times 10^5$	6	0,07	96	$7{,}20\times 10^8$	996	H
007	$-1{,}10\times 10^{60}$	07	$-1{,}60\times 10^4$	7	0,24	97	$3{,}00\times 10^{12}$	997	H
008	$-1{,}94\times 10^{52}$	08	$-2{,}68\times 10^3$	8	1,45	98	$5{,}18\times 10^{19}$	998	H
009	$-1{,}80\times 10^{44}$	09	$-6{,}69\times 10^2$	→		→		999?	H

В. КРАЙНИЕ ЗНАЧЕНИЯ редко бывают нужными.

Если требуются численные значения для случаев, обозначенных «L» или «H» в таблице выше, то последовательность цифр рассматривают как дробь u и принимают

$$\text{значение} \doteq -e^{1/u}/100 = -10^{(0,434/u)-2} \text{ для } u \leqslant 0{,}004;$$

$$\text{значение} \doteq e^{1/(1-u)}/100 = 10^{(0,434/(1-u))-2} \text{ для } u \geqslant 0{,}996,$$

где экспоненты надо брать до наименьшего целого, обеспечивающего подходящую точность.

Так,

Последовательность	Дробь u	0,434/ [u или $(1-u)$]	Округление МИНУС 2
001	0,001	434,0	432
0000001	0,0000001	434000,0	433998
99999	0,99999	4340,0	4338
003	0,003	144,7	143

Г. Формула, лежащая в основе использования дроби u:

$$\text{значение} = \frac{1}{100}\left(e^{1/(1-u)} - e^{1/u}\right).$$

Иллюстрация 2 к упражнению 2.5.3

Таблица для построения выборок из распределений с растянутыми «хвостами» с помощью последовательностей случайных чисел.

А. СПОСОБ ПРИМЕНЕНИЯ — см. илл. 1 к упр. 2.5.3.

Б. ТАБЛИЦА СООТВЕТСТВИЙ (на с. 270) для значений от 0001 до 9998 (для крайних значений смотри пункт В ниже).

В. КРАЙНИЕ ЗНАЧЕНИЯ редко бывают нужными.

Последовательность 0000...: умножьте значения в левом столбце таблицы из В на 100 столько раз, сколько дополнительных нулей будет сверх трех. Последовательность 9999...: умножьте значения в правом столбце таблицы из Б на 100 столько раз, сколько дополнительных девяток будет сверх трех.

Г. ФОРМУЛА:

Для u, соответствующего правильной дроби (между 0 и 1), имеем

$$\text{значение} = \frac{1}{100}\left(\frac{1}{(1-u)^2} - \frac{1}{u^2}\right).$$

2.5.4^{К}. Постройте таблицы такие же, как в илл. 1 и 2 из упр. 2.5.3, но без крайних значений.

2.5.5^{К}. (Продолжение упр. 2.5.4.) Воспользуйтесь таблицами из предыдущего упражнения так же, как таблицами из упр. 2.5.3 (сохраните результаты).

2.6.1. Некий спортивный репортер утверждает, что для предсказания разности в числе перебежек в какой-то игре двух команд Вы должны взять среднюю разность всех остальных результатов встреч этих команд в данном сезоне. Можем ли мы устроить перепроверку этого утверждения, имея полный отчет за сезон? Если да, то как именно надо действовать? Если нет, то почему? (Скорее всего речь идет об игре в бейсбол. — *Ю. А.*)

2.6.2. Другой репортер предлагает делать то же самое при предсказании разности забитых мячей в игре двух футбольных команд. Ответьте на те же вопросы.

Вход		Вход		Вход		ВХОД		Вход		Вход		Вход	
0000?	L	←		←		←		90	0,98	990	100	9990	$1{,}00\times10^{4}$
0001	$-1{,}00\times10^{6}$	001	—10000	01	—100,0	1	—0,98	91	1,22	991	123	9991	$1{,}23\times10^{4}$
0002	$-2{,}50\times10^{5}$	002	—2500	02	—25,0	2	—0,23	92	1,55	992	156	9992	$1{,}56\times10^{4}$
0003	$-1{,}11\times10^{5}$	003	—1111	03	—11,1	3	—0,09	93	2,03	993	204	9993	$2{,}04\times10^{4}$
0004	$-6{,}25\times10^{4}$	004	—625	04	—6,25	4	—0,03	94	2,77	994	278	9994	$2{,}78\times10^{4}$
0005	$-4{,}00\times10^{4}$	005	—400	05	—3,99	5	0,00	95	3,99	995	400	9995	$4{,}00\times10^{4}$
0006	$-2{,}78\times10^{4}$	006	—278	06	—2,77	6	0,03	96	6,24	996	625	9996	$6{,}25\times10^{4}$
0007	$-2{,}04\times10^{4}$	007	—204	07	—2,03	7	0,09	97	11,1	997	1111	9997	$1{,}11\times10^{5}$
0008	$-1{,}56\times10^{4}$	008	—156	08	—1,55	8	0,23	98	25,0	998	2500	9998	$2{,}50\times10^{5}$
0009	$-1{,}23\times10^{4}$	009	—123	09	—1,22	→		→		→		9999?	H

2.6.3. Какие разности получались бы в упр. 2.6.1 или 2.6.2, если бы для предсказания брались только уже сыгранные игры?

2.6.4[к]. Что представляла бы в примерах из упр. 2.6.1 и 2.6.2 перепроверка?

2.6.5. Возьмите 10 любых пар чисел (x, y), не укладывающихся строго на прямую. Подберите для них уравнение прямой $y = a + bx$ самым лучшим методом, какой вы знаете. Теперь случайным образом разделите ваши 10 пар пополам, постройте прямые для каждой половины и проверьте их по оставшимся точкам. Насколько будет меняться остаточная сумма квадратов $\Sigma(y_{\text{набл}} - y_{\text{предск}})^2$ при переходе от одной половины к другой? Как вы могли бы это объяснить?

2.6.6. (Продолжение 2.6.5, но с более громоздкими вычислениями.) Подберите прямые $y = a + bx$ по каждым 9 из 10 пар предыдущего упражнения. Проверьте по оставшейся паре. Сравните с тем, что было раньше.

2.6.7. Проделайте все то же самое, что и в упр. 2.6.5, но для пар (x, y), задаваемых соотношением $(i, (i - 5{,}5)^2)$, где $i = 1, 2, \ldots, 10$.

2.6.8. (Продолжение упр. 2.6.7, но с более сложными вычислениями.) Проделайте то же, что и в упр. 2.6.6, но с данными из упр. 2.6.7.

2.6.9[к]. Пусть мы имеем 8 чисел $a, b, c, \ldots, h$; сколькими различными путями, каждый из которых симметричен соответствующему пути, вы можете разделить их пополам (4 : 4)?

2.6.10. Каковы преимущества всех объективных формализованных методов подбора и нет ли методов, теряющих свою «объективность»? Обсудите это.

ГЛАВА 3

3.1.1. В «опоре и консоли»

2	01	что есть «опора»? А где «консоль»? Каково множество представленных значений?
3		
4	67	

3.1.2, 3.1.3, 3.1.4. Постройте «опоры и консоли» для населения 50 штатов США плюс округ Колумбия соответственно для 1960 г. (упр. 3.1.2), 1965 г. (упр. 3.1.3) и 1970 г. (упр. 3.1.4) (данные возьмите в табл. 3 приложения для упражнений).

3.1.5. (Продолжение упр. 3.1.2—3.1.4.) Сравните «опоры и консоли» в упр. 3.1.2, 3.1.3 и 3.1.4.

3.1.6. Постройте «опоры и консоли» соответственно для рождений и смертей в 50 штатах США за 1960—1970 гг. (данные возьмите в табл. 3 приложения для упражнений).

3.1.7. Сравните «опоры и консоли» из упр. 3.1.6 с каждым из представлений в упр. 3.1.2—3.1.4 и, если сможете, выразите их сходства и различия словами.

3.1.8. На илл. 1 к упр. 3.1.8 показано население ряда городов США, где на 1 июля 1973 г. оно превышало 250000. Постройте соответствующее представление в виде «опоры и консоли» с мощным стволом переменной толщины (как в илл. 3.2.1, с. 75).

Иллюстрация 1 к упражнению 3.1.8

Население городов США, превышавшее 250000 в 1973 г.

Город	Население на 1 июля 1973 г., тыс.	Город	Население на 1 июля 1973 г., тыс.
1	2	1	2
Акрон, Огайо	677	Бирмингем, Алабама	787
Олбани, Нью-Йорк	800	Бостон, Массачусетс	2898
Альбукерке, Нью-Мексико	376	Бриджпорт, Коннектикут	397
Атланта, Джорджия	1748	Буффало, Нью-Йорк	1345
Остин, Техас	375	Чарлстон, Южн. Каролина	352
Балтимор, Мэриленд	2128	Чарлстон, Зап. Виргиния	256

Продолжение

Город	Население на 1 июля 1973 г., тыс.	Город	Население на 1 июля 1973 г., тыс.
1	2	1	2
Чикаго, Иллинойс	7002	Ланкастер, Пенсильвания	335
Цинцинати, Огайо	1383	Лас-Вегас, Невада	308
Кливленд, Огайо	2006	Лексингтон, Кентукки	282
Колумбия, Южн. Каролина	349	Лос-Анджелес, Калифорния	6924
Колумбус, Огайо	1057	Луисвилл, Кентукки — Индиана	886
Дейтон, Огайо	848		
Денвер, Колорадо	1377	Мадисон, Висконсин	301
Детройт, Мичиган	4446	Майами, Флорида	1370
Эль-Пасо, Техас	390	Милуоки, Висконсин	1417
Эри, Пенсильвания	273	Миннеаполис — Сант-Пол, Миннесота	2000
Форт-Лодердейл, Флорида	756		
Фресно, Калифорния	435	Нашвилл, Теннеси	732
Гаррисберг, Пенсильвания	425	Нассау-Суффолк, Нью-Йорк	2630
Хартфорд, Коннектикут	733	Нью-Хейвен, Коннектикут	415
Гонолулу, Гавайи	686	Новый Орлеан, Луизиана	1083
Хьюстон, Техас	2168	Нью-Йорк, Нью-Йорк — Нью-Джерси	9739
Индианаполис, Индиана	1137		
Джэксон, Миссисипи	275	Ньюарк, Нью-Джерси	2053
Джэксонвилл, Флорида	661	Филадельфия, Пенсильвания — Нью-Джерси	4806
Джерси-Сити, Нью-Джерси	598		
Канзас-Сити, Монтана — Канзас	1299	Финикс, Аризона	1127
		Питсбург, Пенсильвания	2365

3.2.1. Что такое медиана?

3.2.2. Чему равна медиана для 24 чисел? А для 25?

3.2.3. Найдите медианные значения для выборок из илл. 3.1.1Б и 3.2.1Б (с. 74 и 76 соответственно).

3.2.4. Найдите «четвертушки» и «осьмушки» выборок в упр. 3.1.2 и 3.1.4.

3.2.5. Сравните структуру особых точек (как на илл. 3.2.2) для данных илл. 3.1.1 (с. 76 и 74 соответственно).

3.2.6, 3.2.7, 3.2.8. (Продолжение упр. 3.1.4.) Сравните структуру особых точек для данных из упр. 3.1.2 (3.2.6), 3.1.3 (3.2.7) и 3.1.4 (3.2.8).

3.3.1. Найдите особые точки, вплоть до размахов, для населения США в 1965 г. и для рождений и смертей между 1960—1970 гг. (см. выборки в упр. 3.1.2 и 3.1.4).

Об упр. 3.4. В табл. 1 приложения для упражнений (после упражнений) приводятся разнообразные полезные сведения о 152 городах с населением 250000 или больше (по состоянию в 1960 г.) в трех переписных округах США: Север Среднего Запада, Юг Среднего Запада и Горный округ. Каждый, кто собирается решать следующие упражнения, должен получить свои собственные подвыборки объемом примерно 50, 20 и 5 из этой совокупности данных. Однако прежде лучше ответить на несколько вопросов.

3.4.1^{к}. Если бы у Вас была таблица случайных чисел (случайных десятичных цифр), то как бы Вы ею воспользовались для получения упомянутых подвыборок?

3.4.2^{к}. Если бы у Вас был полный комплект пронумерованных карточек и абсолютное (хотя, по-видимому, и необоснованное) доверие к своему умению тасовать карточки до тех пор, пока порядок их не станет случайным, то как бы Вы получили подвыборки, настолько близкие к случайным, насколько позволяет Ваше тасование (тех же объемов порядка 50, 20 и 5)?

3.4.3. Два аналитика, не имевших случайных чисел и не доверявших своим способностям тасования, поступили следующим образом: сначала, упорядочив города в табл. 1 приложения для упражнений, они разделили их все (152) на

три группы объемом 51, 50 и 51. Затем они разделили каждую группу по 51 точно так же на три подгруппы по 17, а для группы объемом в 50 получилось 17, 16 и 17. Наконец, они проделали эту операцию в третий раз, что дало им еще меньшие группы — по 6, 5 и 6 или 6, 5 и 5 единиц.

Один из них настаивал, что взять третью группу объемом 51, пятую объемом 16 и седьмую из 6 единиц — это вполне удовлетворительный способ получения подвыборок. Другая возражала, что тогда в подвыборке объемом 51 получилось бы слишком много городов из штата Техас — 44, а всем известно: Техас — совсем другое дело. Она утверждала, что будет гораздо лучше, если взять 1, 2 и 3-й города из списка первой группы, а дальше выбирать тройки городов «случайно» в каждой из трех первичных групп, пока не наберется 50 или 51.

Теперь Вы должны выступить в роли эксперта, не только сформулировав Ваше предпочтение, но и взвесив «за» и «против» каждого варианта. Тогда, может быть, Вы предпочтете вообще какой-нибудь иной подход.

3.4.4.[к]. Даны две разные монеты, скажем, пени и никль. (Вы можете взять, например, копейку и пятак — *Ю. А.*). Можете ли Вы с помощью их подбрасывания выбрать одну из трех альтернатив с действительно равными вероятностями? (*Факультативное задание*: а если бы были три разные монеты, Вам было бы легче?)

3.4.5[к]. Помогло бы двум аналитикам из упражнения 3.4.3, если бы у них была копия таблиц случайных перестановок Мозеса и Оукфорда ((Moses, Oakford) — см. библиографию к гл. 10)?

3.4.6. Извлеките свои подвыборки объемом около 50, около 20 или около 5 из выборки в 152 города (табл. 1 из приложения для упражнений). Воспользуйтесь при этом наилучшим, с Вашей точки зрения, методом. *Сохраните эти выборки для дальнейшей работы.* Проще всего для этого сделать копию табл. 1.

3.4.7[к]. Некий третий аналитик сказал: «Я хотел бы обеспечить дублирование, так чтобы моя выборка объемом около 5 входила как часть в выборку объемом около 20, а та в свою очередь входила бы в выборку объемом около 50». Взвесьте «за» и «против» допущения такой последовательности типа «матрешка».

3.5.1.[к]. (Надо назначить каждому, кто берется решать эту задачу, какую-нибудь переменную, но не медианный семейный доход). Для 152 городов, представленных в табл. 1 из приложения для упражнений, $k = \sqrt{152} = 12^{+}$ (это означает, что 12 — целое значение корня с избытком. — *Ю. А.*). Сосредоточьте Ваше внимание на отыскании: (а) 12 городов с наибольшими значениями заданной Вам переменной, (б) и 12 городов — с наименьшими. Оставьте по 10 значений с каждой стороны после отбрасывания 5-го и 9-го значений среди наиболее крайних. Для двух множеств по 10 оставшихся значений постройте графики наподобие тех, что представлены на илл. 3.5.2 (с. 78). Нанесите Вашу переменную против двух других, одна из которых — медианный семейный доход. Соберите (в классе) все графики и обсудите, о чем говорит их сопоставление.

3.5.2. (Продолжение упр. 3.5.1 с той же самой заданной переменной.) Возьмите выборку порядка 50 городов, которую Вы получили в упр. 3.4.6. Нанесите на график Вашу переменную против медианного семейного дохода для этих «50». Хорошо ли согласуются график 10 — и — 10 точек из упр. 3.5.1 и допустимый прогноз по 50? А наоборот?

3.5.3. Повторите предыдущее упражнение, но на Вашей выборке порядка 20. Ответьте на те же вопросы.

3.6.1. (Назначение переменных решающим эту задачу осуществляется так же, как в упр. 3.5.1, кроме медианного семейного дохода.) Возьмите Вашу выборку в 50 городов и постройте для нее «опору и консоль» медианных семейных доходов, а затем с помощью этого представления выпишите эти города в порядке убывания медианного семейного дохода. Если вдруг, среди них окажутся совпадающие,— отбросьте их. Теперь вернитесь к заданной Вам переменной и выпишите снизу ее значения в качестве y (сохраняя тот же порядок, и заменяя значения для отброшенных городов их медианами). Теперь у Вас есть либо 50 (если не было совпадений), либо несколько меньше (если совпадения были) значений y. Сгладьте их текущими медианами по трем соседним точкам так, как это показано на илл. 3.6.1 (с. 79). Нанесите результаты на график. Что бы это, по Вашему мнению, значило?

3.6.2. Повторите предыдущее упражнение с другой переменной в качестве y.

3.6.3. Проделайте упр. 3.6.1, взяв из табл. 1 приложения для упражнений *иную переменную*, чем медианный семейный доход или заданная Вам переменная (или, наконец, переменная из предыдущего упражнения), и упорядочите Ваши 50 городов по ее значениям, но оставьте те же y. Рассмотрите сходства и различия (как внешние, так и по существу) между тем, что получилось, и результатом упр. 3.6.1.

3.6.4. Объедините условия упр. 3.6.2 и 3.6.3, а затем повторите 3.6.3.

3.6.5. «*Возрастные сгустки*». Это широкоизвестный демографический феномен, состоящий в том, что люди склонны называть свой возраст округленно ближайшим числом, кратным пяти годам. В табл. 1 к упр. 3.6.5 представлено распределение женщин по возрастам в Мексике на 1960 г. в диапазоне от 0 до 75 лет. [Mexico, Direction General de Estadistica, VIII censo general de poblacion, Cuadro 7, 1962]. Проанализируйте эти данные с помощью скользящих медиан. Будет ли какая-нибудь разница, если брать скользящие медианы по 3, 5 или 7 точкам? О чем говорят структуры остатков? Подтверждают ли они гипотезу о «возрастных сгустках»?

Иллюстрация 1 к упражнению 3.6.5

Распределение сообщаемых возрастов женщин в Мексике в 1960 г., тыс.

Возраст	Число индивидов	Возраст	Число индивидов	Возраст	Число индивидов
0	558				
1	513	26	243	51	42
2	582	27	220	52	86
3	604	28	283	53	61
4	584	29	182	54	66
5	566	30	412	55	148
6	562	31	113	56	69
7	524	32	208	57	46
8	529	33	163	58	83
9	430	34	146	59	48
10	497	35	310	60	245
11	369	36	168	61	20
12	455	37	130	62	43
13	398	38	224	63	32
14	404	39	129	64	32
15	382	40	331	65	103
16	366	41	51	66	29
17	346	42	136	67	23
18	403	43	91	68	40
19	300	44	77	69	16
20	409	45	231	70	109
21	226	46	90	71	9
22	325	47	77	72	25
23	294	48	148	73	15
24	289	49	78	74	14
25	380	50	281	75	48

3.6.6. Повторите упр. 3.6.5, работая с квадратными корнями чисел, соответствующих каждому возрасту.

3.7.1. (Продолжение упр. 3.6.1.) Проведите «на глаз» прямую

$$y \approx a + b \times (\text{медианный семейный доход})$$

по сглаженным результатам из 3.6.1. Найдите для каждого из 50 индивидуальных значений отклонение от этой прямой. Сгладьте получившиеся остатки последовательными медианами по трем точкам. Нанесите их на график. Похоже ли,

что эта прямая охватывает все неслучайные связи между y и медианным семейным доходом? Почему? Почему нет? Постройте такую же прямую, но для сглаженных остатков и нанесите на аналогичный график вместе с данными из 3.6.1. Как сравнить эти два множества результатов? Почему Вы думаете, что это возможно?

3.7.2. (Продолжение упр. 3.6.2.) Сделайте то же самое для данных из 3.6.2 (вместо 3.6.1).

3.7.3. На илл. 3.8.3 (с. 87) опущены (по сравнению с илл. 3.8.2 с. 86) 60 точек, соответствующих четным числам, делящимся на 6. Выпишите под ними (в упорядоченной последовательности) их числа Гольдбаха и сгладьте их последовательными скользящими медианами по трем точкам. Найдите разности между результатами и сплошной «кривой» с илл.3.8.3 и нанесите их на график.

3.7.4. (Продолжение упр. 3.7.3.) Возьмите полученные выше разности и снова сгладьте их последовательными скользящими медианами по трем точкам. Нанесите на график. Можете ли вы предложить простое приближение?

3.8.1. (Упр. 3.8.1—3.8.3 также могут быть полезны для классной работы.) Постройте график, аналогичный илл. 3.8.2 (с. 86), используя числа Гольдбаха от 252 от 430 (вместо 2—180).

3.8.2. Повторите предыдущее упражнение для чисел от 9502 до 9680.

3.8.3. Снова повторите, но теперь для чисел от 9752 до 9930.

3.8.4 — 3.8.6. (Продолжение предыдущих задач.) Сделайте то же самое по илл. 3.8.3, вплоть до построения ломаной линии сглаженных результатов (как в параграфах 3.6 и 3.7) для чисел Гольдбаха от четных чисел, делящихся на три (как в упр. 3.7.3).

3.9.1. (Классная работа.) В таблице, из которой извлечена илл. 3.9.3 (с. 88.) [World Almanac, 1973, p. 248], за некоторый 30-летний период представлены: (а) обычные минимальные температуры января; (б) обычные максимальные температуры июля; (в) обычные минимальные температуры июля; (г) наивысшая зарегистрированная температура; (д) наинизшая зарегистрированная температура; (е) обычные годовые осадки. Эти шесть переменных распределяются по одной на шестую часть группы. Теперь надо построить графики, аналогичные илл. 3.6.2 (с. 79).

3.9.2. (Продолжение упр. 3.9.1.) Определите «на глаз» наклон только что построенного графика и графика остатков, в зависимости от долготы, как на илл. 3.9.2 (с. 88) (те, кому нравится располагать запад слева, могут направить горизонтальную ось справа налево). Обсудите Ваши результаты.

3.9.3. (Продолжение упр. 3.9.2.) Выделите все города, для которых остатки из упр. 3.9.2 выглядят явно необычно, подыщите самую лучшую — наиболее полезную, наиболее удобную для регрессии — переменную, какую Вы только сможете, и постройте для нее график зависимости от этих остатков. Обсудите полученные результаты.

3.10.1. Для Вашей выборки объемом около 50 городов соотнесите медианный семейный доход и долю живущих в отдельных квартирах: (а) в терминах графика «10 плюс 10» и, (б) в терминах сглаженных медиан, медианами, объединенными в блоки (примерно) по 3 города (16 блоков). Постройте аналог илл. 3.10.2 (с. 91) и обсудите результаты (1) сами по себе, (2) в сравнении с илл. 3.10.2.

3.10.2. Повторите предыдущее упражнение (кроме сравнения (2)) для медианного семейного дохода и заданной Вам переменной.

ГЛАВА 4

4.1.1. Когда стоит подбирать преобразование $z = t^{1/3}$ на множестве степенных преобразований? И стоит ли это делать вообще?

4.1.2. Когда стоит подбирать преобразование $z = \sqrt{t + 4}$ на множестве степенных преобразований? И стоит ли вообще?

4.1.3. На каких ступеньках «лестницы степеней» имеет смысл останавливаться при $t < 0$?

4.2.1. Рассмотрите кривую, определенную (хотя и не полно, но вполне достаточно для наших целей) следующей таблицей:

x :	0,73	1,27	1,73	2,31	2,50	2,72	2,91
y :	0,05	0,26	0,65	1,54	1,95	2,52	3,08

Какое преобразование y спрямит эту кривую? (Проведите исследование, как в начале параграфа 4.2.)

4.2.2. Как найти спрямляющее преобразование x для той же кривой из упр. 4.2.1? (Проведите исследование аналогично второй части параграфа 4.2.)

4.2.3. Для данных из упр. 4.2.1 испытайте $\log y$ со «сдвигом», как в третьей части параграфа 4.2).

4.2.4. Найдите спрямляющее преобразование y для кривой, задаваемой следующей таблицей:

x :	0,73	1,27	1,73	2,31	2,50	2,72	2,91
y :	1,11	1,33	1,48	1,63	1,67	1,72	1,76

4.2.5. Преобразуйте x для спрямления кривой из упр. 4.2.4.

4.2.6. Испытайте $\log y$ со «сдвигом» для спрямления кривой из упр. 4.2.4.

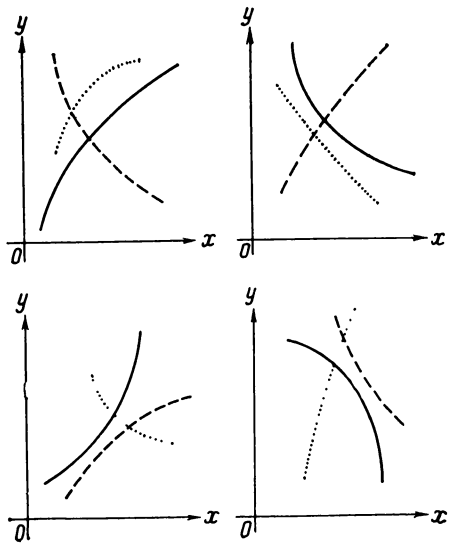
4.2.7. Преобразуйте x для спрямления кривой, задаваемой следующей таблицей:

x :	0,73	1,27	1,73	2,31	2,50	2,72	2,91
y :	2,22	2,34	2,44	2,55	2,59	2,63	2,67

4.2.8к. Найдите самое лучшее преобразование, на какое Вы только способны, для данных таблицы из упр. 4.2.7.

4.3.1. На илл. 1 к упр. 4.3.1 представлены различные кривые. Расскажите, как бы Вы стали действовать, если бы Вам пришлось выбирать степень спрямляющего преобразования (хотя бы приближенного) для y и для x сплошной кривой на левом верхнем рисунке.

Иллюстрация 1 для упражнения 4.3.1



4.3.2. То же самое для пунктирной кривой этого рисунка.

4.3.3. А как быть с точечной кривой на этом рисунке?

4.3.4 — 4.3.6. Аналогичные задания для трех остальных рисунков.

4.3.7. Ниже приведены значения q_x для США в 1970 г. Это вероятности того, что некий человек, обозначенный индексом x, должен умереть в течение следующих пяти лет. Как преобразовать x, q_x или обе переменные, чтобы получше спрямить эту кривую? (x = возраст, q_x = вероятность смерти)

x	q_x	x	q_x
5	0,0021	45	0,029
10	0,0020	50	0,043
15	0,0056	55	0,066
20	0,0074	60	0,096
25	0,0072	65	0,138
30	0,0097	70	0,198
35	0,012	75	0,290
40	0,019	80	0,407
		85	0,555

4.4.1. В следующей таблице приведены числа простых множителей (N), меньших, чем n, для отдельных значений n:

n :	4	16	64	256	1024	4096	16384
N :	2	6	18	54	172	564	1900

Постройте график зависимости $\log N$ от $\log n$. Что подсказывает этот график насчет спрямления зависимости N от n?

4.4.2. Ниже представлены значения $\operatorname{ctg}\theta$ для некоторых θ:

θ:	0,1°	0,2°	0,5°	1°	2°	5°	10°	20°	50°
$\operatorname{ctg}\theta$:	1145,9	286,3	114,59	57,29	28,63	11,43	5,67	2,75	0,84

Нанесите на график $\log \operatorname{ctg}\theta$ и θ. Что этот график говорит о спрямлении исходной зависимости?

4.4.3. Опираясь на таблицу из упр. 4.4.1, постройте график N/n от n. Что он подсказывает о спрямлении зависимости? Теперь сравните $\log(N/n)$ от n с зависимостью N/n от $\log n$. Что бы Вы предпочли для спрямления?

4.4.4. Опираясь на таблицу из упр. 4.4.2, постройте зависимость $\theta \operatorname{ctg}\theta$ от θ. Подсказывает ли она Вам что-нибудь? Подберите преобразование θ, спрямляющее его зависимость от $\theta \operatorname{ctg}\theta$. Что приводит к более полезному результату? Какую зависимость построили бы Вы теперь?

4.5.1к. (Группа делится на 8 частей и каждой части дается одна из переменных, иная, чем медианный семейный доход.) Вернитесь к илл. 3.5.1 (с. 77) и 3.5.2 (с. 78) и добавьте 10 новых городков, не входящих в объединения и имеющих медианный семейный доход, близкий к медианам для всех 88 городков. (Данные возьмите из нижеследующей таблицы.)

Нанесите данные по всем десяти городкам на график (медианы данной Вам переменной и медианы семейных доходов). Найдите преобразование, которое вытянет эти точки вдоль прямой. Посмотрите, что получилось. (Соберите все результаты по восьми переменным вместе и обсудите их.)

Иллюстрация 1 к упражнению 4.5.1

Десять не входящих в объединения городков, население которых имело в 1959 г. медианный доход, близкий к медиане (подробности см. в илл. 3.5.1, с. 77)

Городок	Медианный семейный доход	212	229	244	246	248	249	256	264
Брейнтри, Массачусетс	7474	31,0	60,7	52,2	7,7	5,8	88,8	19,4	6,7
Росс, Пенсильвания	7475	31,1	52,6	4,0	12,9	5,7	86,1	28,6	6,5
Илмонт, Нью-Йорк	7494	31,1	68,2	44,2	33,2	5,6	88,5	16,6	19,0
Фреймингхэм, Массачусетс	7495	29,1	44,5	53,0	5,9	5,6	79,8	32,7	21,0
Арлингтон, Массачусетс	7538	34,8	61,8	60,4	27,7	5,8	56,8	22,3	7,8
Нейтик, Массачусетс	7550	28,7	57,8	55,7	9,1	5,9	81,6	23,7	5,4
Юинг, Нью-Джерси	7597	31,1	56,0	48,7	6,8	5,6	91,5	23,0	24,0
Миддлтон, Пенсильвания	7656	22,6	19,2	56,3	7,9	6,1	99,2	35,8	24,3
Кейтонсвилл, Мэриленд	7662	32,0	50,5	64,5	14,9	6,0	82,3	25,9	14,8
Хамден, Коннектикут	7741	35,5	59,4	55,3	13,7	5,7	84,5	21,4	10,6

ГЛАВА 5

5.1.1[к]. Соотнесите каждое из следующих ниже измерений с одной из категорий «итог», счет, «баланс», счетная доля, ранг, «ярлык»: концентрация соли в морской воде; содержание углекислого газа в воздухе; доля лиц испанского происхождения на Манхаттане; количество (тонн) соли в море; число молекул углекислого газа в воздухе; среднее (ожидаемое) число лет жизни для лиц женского пола, родившихся в ФРГ в этом году; часть прибыли, идущей на выплату налогов в крупной корпорации; налоги индивида; класс бейсбольной лиги; соотношение мячей, с которым выигран футбольный матч; место футбольной команды в итоговой таблице (сколько впереди и сколько наравне); буквенные обозначения разделов в учебнике математики; насколько далеко «к востоку от Бостона» некий город в штате Вермонт; как глубоко ушел под воду вулкан или вулканический остров; сколько пальцев осталось у покалеченного человека; сколько покрышек износит старый автомобиль на плохой дороге.

5.1.2. Если бы Вы захотели преобразовать каждый пример из предыдущего упражнения, то что бы вы сделали прежде всего?

5.2.1. Исследуйте подробно, как с помощью илл. 5.2.1 (с. 115) получить log от числа 321, 98 с двумя десятичными знаками? А от числа 0,000098321?

5.2.2. Два специалиста дебатировали вопрос о том, чтоит ли использовать $\log(y + 0{,}01)$ или $\log(y + 0{,}02)$. Для каких значений y их пререкания действительно имеют смысл?

5.2.3. Другая пара специалистов выбирала между $\log(y + 1)$ и $\log(y + 3)$. А когда разумна эта дискуссия? Какой из данных логарифмов легче построить?

5.2.4. Вычислите log 17 1/6; log 23 1/6.

5.3.1. Выясните, как с помощью илл. 5.2.3 (с. 116) получить значения квадратных корней из чисел 321,98 и 0,000098321 с двумя и тремя знаками.

5.3.2[к]. Опишите достоинства и недостатки работы с каждым из следующих преобразований: $1000/y$, $1/y$, $-1/y$, $-1000/y$.

5.3.3. Еще одна пара специалистов спорила насчет выбора между $\sqrt{y + 0{,}05}$ и $\sqrt{y + 0{,}1}$. Для каких y этот выбор имел бы значение?

5.3.4. Четвертая пара специалистов колебалась между $\sqrt{y + 1}$ и $\sqrt{y + 2}$. Для каких y эти терзания имели бы смысл?

5.3.5. Опишите, как с помощью илл. 5.3.1 (с. 118) получить отрицательные обратные числа для 321,98 и 0,000098321 с двумя и тремя знаками.

5.3.6. Чему равны $\sqrt{17\frac{1}{6}}$, $\sqrt{23\frac{1}{6}}$, $\sqrt{47\frac{1}{6}}$ и $\sqrt{289\frac{1}{6}}$?

5.3.7. Чему равны $-1000/17\frac{1}{6}$, $-1000/23\frac{1}{6}$, $-1000/47\frac{1}{6}$ и $-1000/289\frac{1}{6}$?

5.4.1. Как выбрать согласованные преобразования для 86,5%? Для 13,5%? Для 98, 9%? А для 1,1%?

5.4.2. Подробно опишите, как проще всего найти

$$\log(\text{одна сумма}) - \log(\text{другая сумма}),$$

если обе суммы имеют одинаковые «сдвиги» на $\frac{1}{6}$, а исходные суммы были $17+133 = 150$.

5.4.3. Проделайте то же, что и в упр. 5.4.2, но для

$$\sqrt{\text{одна сумма}} - \sqrt{\text{другая сумма}}$$

при тех же сдвигах и исходных данных.

5.4.4^{к}. Рассмотрите преимущества и недостатки различных преобразований долей, которые согласованы в окрестностях 50%.

5.5.1. Почему, когда согласуются степени и логарифмы, наблюдается наибольшее значение, а для преобразований долей ничего подобного нет?

5.5.2. Если известно, что согласованное преобразование y не отличается от самих y больше, чем на ± 1 в диапазоне от 280 до 320, то можно ли было это обеспечить степенным преобразованием?

5.5.3. Тот же вопрос, что и в предыдущем упражнении, но при $\pm$ 10 и для $2800 \leqslant y \leqslant 3200$.

Предупреждение. Упражнения к параграфу 5.6 требуют значительного объема счета.

5.6.1. Хартман (Hartman), на которого ссылается Мейтер в книге [Mather K. A. Statistical Analysis in Biology. 1949, p. 196], приводит данные для чисел мужчин и женщин, не ощущающих следы фенилтиокарбамида при различных разведениях. Вот эти данные:

Мужчины:	15	А	35	Б	46	В	31	Г	23	Д	13	Е	9	Ж	7	З	10	И	13	К	25	Л	63	Σ	290
Женщины:	42	А	52	Б	38	В	30	Г	19	Д	17	Е	6	Ж	5	З	10	И	19	К	33	Л	43	Σ	314

Здесь A — самый концентрированный раствор, в котором следы не улавливают 15 мужчин, тогда как 35 реагируют на A, но не чувствительны к $Б$. И так далее до 63 мужчин, реагирующих на самый слабый раствор. Руководствуясь илл. 5.6.1 (с. 123), и с учетом илл. 5.6.2 (с. 124) подберите два преобразования $A, Б, \ldots, Л$ — одно для мужчин, второе для женщин.

5.6.2. Дэвис и Ричардс [Davis and Richards. Journal of Ecology, 21, 250—384 and 22, 106—155] сообщают данные о числе деревьев семи пород (причем одна из пород на самом деле состоит из двух, но трудно различимых) из шести упорядоченных по величине диаметра ствола классов в смешанном тропическом лесу (джунглях) Морабалли Крик, Британская Гвиана (с 1966 г. Кооперативная Республика Гайана — *Примеч. ред.*). Классы были промерены, но Вам забыли сказать, где проходят их границы. Вот эти данные по возрастанию классов: Eschweilera (decolorans или pallida) (128, 19, 14, 7, 1, 1); Eschweilera sagotiana (48, 5, 6, 14, 13, 1); Licania laxiflora (24, 18, 13, 5, 6, 0); Licania heteromorpha (176, 29, 12, 0, 0, 0); Licania venosa (128, 26, 16, 9, 10, 0); Ocotea rodiaei (16, 8, 2, 5, 10, 3); Pentaclethre macroloba (80, 31, 20, 16, 6, 0). (Восстановление точных русских названий исследуемых в этом упражнении растений оказалось за пределами возможностей редактора, однако для компенсации можем все-таки сообщить, что род Eschweilera относится к семейству миртовых (Myrtaceae), славящихся корабельной древесиной. Licania принадлежит семейству розовых (Rosaceae), древесина которых идет на строительные конструкции и разные поделки. Ocotea из семейства лавровых (Lauraceae), которые также широко используются в кораблестроении. Наконец, Pentaclethre — из семейства бобо-

вых (Leguminosaceae) — это масличное растение. Так что исследователи руководствовались не одним только любопытством. — *Ю. А.*) Найдите преобразования для каждого размера класса (опуская пустые классы) по очереди для всех пород. Эти семь преобразований выглядят достаточно хорошо согласованными?

5.6.3. (Продолжение упр. 5.6.2.) Найдите среднее преобразование для класса каждого размера из упр. 5.6.2 (охватывающее все породы), ограниченное тремя меньшими классами, и вычтите полученный результат из всех значений. Теперь у Вас есть трижды семь таких скорректированных преобразований. Воспользуйтесь их медианой как первым приближением к общему преобразованию для класса этого размера. Постройте зависимости каждого исходного преобразования от этого приближения. Разумно ли считать, что эти зависимости линейны?

5.6.4. (Продолжение упр. 5.6.3.) Проведите «на глазок» прямые на каждом из семи графиков из упр. 5.6.3. Подсчитайте значения

$$\text{вторая поправка} = \frac{\text{остатки от прямой}}{\text{угловой коэффициент}}$$

для каждого сочетания породы и диаметра ствола (класса). Возьмите медианы по породам. Кажутся ли они достаточно большими? Образуйте улучшенное общее преобразование для каждого класса по формуле

приближенное общее ПЛЮС медиана второй поправки.

Насколько близкими к равным по объему кажутся наши классы?

5.7.1. Преобразования рангов иногда полезны и для численных данных. На илл. 1 к упр. 5.7.1 приведены данные за 1960 г. о населении (РОР) и медианном семейном доходе (МСД) для округов штата Аризона, для каждого избирательного округа (ИО) по выборам в конгресс штата Арканзас (за 1964 г.) и для всех трех естественных «зон» штата Калифорния, упорядоченные по населению в 1960 г. Для штата или его части, которая Вам выделена, найдите преобразование $\log(j - 1/3)$ МИНУС $\log(n + 1 - j - 1/3)$ и для (РОР), и для (МСД). Постройте графики зависимости (МСД) от (РОР) до преобразования и после него. Какой из графиков кажется Вам более полезным?

Иллюстрация 1 к упражнению 5.7.1

Население (РОР) и медианный семейный доход (МСД) в округах некоторых регионов США

Аризона, 1960			Первый ИО, Арканзас, 1960		
Округ	РОР	МСД	Округ	РОР	МСД
Марикопа	663510	5896	Миссисипи	70174	2725
Пайма	265660	5690	Криттенден	47564	2506
Пинол	62673	4412	Крейгхид	47303	3408
Кочиз	55039	5107	Филлипс	43997	2360
Юма	46235	5360	Сент-Франсиз	33303	1973
Коконино	41857	5398	Поинсетт	30834	2591
Навайо	37994	4237	Грини	25198	2654
Аппачи	30438	2832	Клей	21258	2633
Явапаи	28912	5197	Ли	21001	1710
Джила	25245	5087	Кросс	19551	2480
Грахэм	14045	4593			
Гринли	11059	5168			
Санта-Круз	10808	4620			
Мохейв	7736	5111			

Второй ИО Арканзас 1960		
Округ	РОР	МСД
Пуласки	242980	4935
Уайт	35795	2893
Лоноки	24551	2708
Фолкнер	24303	2968
Арканзас	23355	3348
Джексон	22843	2995
Индепенденс	20048	2502
Монро	17327	2162
Лауренс	17267	2255
Конвей	15430	2751
Вудрафф	13954	1902
Рэндолф	12520	2497
Прейри	10515	2853
Клиберн	9059	2137
Изард	6766	2699
Фултон	6657	1886
Шарп	6319	1902
Стоун	6294	1740
Перри	4927	2217

Третий ИО Арканзас 1960		
Округ	РОР	МСД
Себастьян	66685	3089
Вашингтон	55797	3683
Бентон	36272	3180
Кроуфорд	21318	3122
Поуп	21177	3046
Буни	16116	2837
Логан	15957	2376
Джонсон	12421	2484
Йелл	11940	2600
Каролл	11284	2555
Франклин	10213	2611
Бэкстер	9943	2800
Медисон	9068	1928
Сирси	8124	2066
Скотт	7297	2168
Ван Бурен	7228	1968
Марион	6041	2210
Ньютон	5963	1666

Четвертый ИО Арканзас, 1960		
Округ	РОР	МСД
Джефферсон	81373	3671
Юнион	49518	4361
Гарленд	46697	3511
Миллер	31686	3372
Учита	31641	3686
Солайн	28956	4483
Колумбия	26400	3438
Эшли	24220	3432
Хот-Спрингс	21893	3881
Кларк	20950	3127
Дишей	20770	2430
Хемпстид	19661	2676
Чикот	18990	2013
Дрю	15213	2614
Линкольн	14447	1911
Бредли	14029	3069
Полк	11981	2694
Лафайетт	11030	2245
Хауард	10878	3033
Невада	10700	2538
Даллас	10522	2809
Севиер	10156	3089
Литтл-Ривер	9211	2725
Грант	8294	2985
Райк	7864	2614
Кливленд	6944	2363
Колхаун	5991	2394
Монтгомери	5370	2572

Побережье Калифорнии, 1960		
Округ	РОР	МСД
Лос-Анджелес	6038771	7046
Сан-Диего	1033011	6545
Аламеда	908209	6786
Сан-Франциско	740316	6717
Санта-Клара	642315	7417
Сакраменто	502775	7100
Сан-Матео	444387	8103
Контра-Коста	409030	7327
Сан-Джоакин	249919	5889
Сентура	199138	6466
Монтерей	198351	5770
Санта-Барбара	168962	6833
Солано	134597	6190
Гумбольдт	104892	6282
Санта-Круз	84219	5325
Напа	65890	6524
Мендосино	51058	5803
Дел-Норт	17771	6277

Округа Калифорнии, граничащие с прибрежными, 1960			Остальная Калифорния, 1960		
Округ	РОР	МСД	Округ	РОР	МСД
Орандж	703925	7219	Тула	166403	4815
Сан-Бернардино	503591	5998	Батт	82030	5408
Фресно	365945	6603	Шаста	59468	5989
Риверсайд	306191	5693	Мадера	40466	4596
Керн	291981	5933	Юба	33859	5031
Станиславс	157294	5260	Невада	20911	5419
Сонома	147325	5725	Туолумн	14404	5602
Марин	146820	8110	Лассен	13597	5861
Мерсед	90448	4806	Колуза	12075	5604
Сан-Луис-Обиспо	81011	5659	Инёу	11689	5837
Империал	72105	5507	Плюмас	11260	5834
Еуло	65727	6240	Модок	8308	5709
Плейсер	56468	5989	Марипоза	5064	4704
Кингз	49954	4957	Сиерра	2247	5863
Суттер	33380	5670	Моно	2213	6321
Сискуок	32885	5558	Олпин	397	—
Эльдорадо	29309	6603			
Техама	25305	5589			
Гленн	17245	5290			
Сан-Бенито	15396	5538			
Лейк	13786	4438			
Калаверас	10289	5824			
Амадор	9900	5636			
Тринити	9706	6210			

Источник. County and City Data Book, 1962.

5.7.2. В каждой из 8 групп округов из илл. 1 к упр. 5.7.1 возьмите по 3 точки: вторую (в малых группах) или третью (в больших) с каждого конца и среднюю точки. Теперь для полученной выборки выясните, ведет ли себя степенное (или логарифмическое) преобразование достаточно линейно относительно $\log(j - 1/3)$ МИНУС $\log(n + 1 - j - 1/3)$, построенного в предыдущем упражнении.

5.7.3. (Продолжение упр. 5.7.2.) Для выделенных Вам округов постройте график зависимости некоторых степенных преобразований для населения 1960 г. от преобразования рангового типа.

5.7.4. (Для тех, кто хочет еще поупражняться и может заглянуть в справочник County and City Data Book, 1962). Вот дополнительные столбцы, которые могут оказаться интересными порознь или вместе: (4) плотность населения, (5) процент прироста населения, (13) медианный возраст и отношение (107) числа продовольственных магазинов к (86) числу промышленных предприятий.

5.8.1. Можно ли обойтись только параграфом 5.8 и пропустить все предыдущие в этой главе?

5.8.2. У Вас есть много треугольников, причем каждый имеет одну сторону длиной в 10 единиц, а вторую — в 5. Как по правилам «первой помощи» преобразовать длину третьей стороны?

5.8.3^{к}. Браковщик качества резисторов отбирает из партии такие изделия, сопротивление которых оказывается между 99 и 101% номинала. Как бы Вы преобразовали измеренные сопротивления для 500-омных резисторов? Объясните Ваш ответ.

5.8.4. «*Нормы*» *смертности.* Уильям Брасс утверждал в работе [Brass William. On the Scale of Mortality, p. 69—110, в Biological Aspects of Demography, London: Taylor and Francis, Ltd., 1971], что существуют «нормы» смертности, ко-

торые определяются всего двумя параметрами и при этом вполне удовлетворительно описывают почти весь опыт, накопленный человечеством. На илл. 1 к упр. 5.8.4 приведены три современные «таблицы жизни», содержащие доли людей в каждой популяции, которые доживут до указанного возраста. Вместе с ними даны и «нормы» Брасса. Он отмечал, что зависимость между логитами любой такой таблицы и его «нормами» близка к линейной. Работая с данными для каждой страны отдельно, выясните, действительно ли логит — наилучшее преобразование? Какие еще стоит испытать? Рассмотрите теперь все данные вместе. Хорошо ли подходит логит-преобразование? Верите ли Вы, что действительно существуют «нормы» смертности, основанные на этих данных?

Иллюстрация 1 к упражнению 5.8.4

Доля переживших указанный возраст

Возраст	«Нормы» Брасса	Швеция жен., 1959	Италия жен., 1901—1911	Япония муж., 1959
1	0,850	0,987	0,848	0,964
5	0,769	0,984	0,739	0,952
10	0,750	0,982	0,721	0,947
20	0,713	0,979	0,699	0,938
30	0,652	0,974	0,641	0,917
40	0,590	0,965	0,592	0,892
50	0,511	0,943	0,541	0,846
60	0,396	0,891	0,466	0,738
70	0,238	0,757	0,317	0,524
80	0,076	0,447	0,107	0,214

(Данные воспроизведены с разрешения автора и издателя.)

5.9.1. Два специалиста анализируют данные о числе дорожных происшествий час за часом для трех соседних больших городов на протяжении целой недели (168 часов). Один из них считает нужным рассматривать логарифмы часовых итогов, но другой возражает, поскольку в данных оказалось слишком много нулей. Какими двумя совершенно разными путями мог бы пойти первый аналитик, чтобы преодолеть все трудности с логарифмами и нулями?

5.9.2. Один исследователь, изучая наклоны кривых, полученных на «самописце», собирался сделать логарифмическое преобразование. Он считал это нужным отчасти потому, что проскальзывание ленты должно по-разному сказываться на наклонах в зависимости от скорости лентопротяжки. Спрашивается, как ему преодолеть трудности, подстерегающие его и по горизонтали, и по вертикали и связанные с нулевыми и бесконечными наклонами? Что Вы на это скажете?

5.9.3. В условиях упр. 5.9.2 будет ли зависеть Ваш ответ от того, наблюдает ли исследователь два нуля и три бесконечности на 1000 отсчетов или же ему встретилось 97 нулей и 43 бесконечности на те же 1000 отсчетов? Почему? Почему нет?

ГЛАВА 6

6—1. Если Вы всегда будете получать от преобразований какие-нибудь дивиденды, то всегда ли «овчинка» будет стоить «выделки»? Почему? Почему нет?

6—2. Если Вы располагаете наблюдениями: 1, 1,5, 3,5,..., 63, 89, 121, и верите, что логарифмы дадут лучшее выражение, то вероятен ли успех данного предприятия в этом случае? Почему? Почему нет? Когда «да»? Когда «нет»?

6—3. Ответьте на вопросы предыдущего упражнения для значений в диапазонах от 23 до 29 и от 365 до 513.

6—4. Ответьте на вопросы из упр. 6—2 для значений в диапазонах от 51 до 54 и от 473 до 551.

6—5. Ответьте на те же вопросы для значений от 1,04 до 1,09 и от 1,73 до 1,81.

6—6. *Места и голоса* (домашнее задание). При двухпартийном парламенте или конгрессе не приходится ожидать, что партия, собравшая $x\%$ голосов избирателей, получит и $x\%$ мест. Для описания этой зависимости предлагались различные модели. Простейшая из них известна как «кубический закон», а именно

$$\frac{S}{1-S}=\left(\frac{V}{1-V}\right)^3,$$

где S — доля мест, а V — доля голосов. А вот более сложная модель для «логитов»:

$$\log\left(\frac{S}{1-S}\right)=\beta_0+\beta_1\log\left(\frac{V}{1-V}\right).$$

Тафт (E. R. Tufte) предложил более простую модель

$$S-0{,}5=\beta_0+\beta_1(V-0{,}5),$$

где β_0 элементарно интерпретируется как «смещение» системы, а β описывается как «ножницы», т. е. процент мест, выигранных на 1% превышения числа голосов. Каковы теоретические положения, на которые опираются первая и третья модели? Как соотносятся между собой кубический закон и логит-модель? На илл. 1 к упр. 6—6 приведены доли мест и голосов для 36 выборов в конгресс США, относящихся к XX в. Какая модель кажется Вам наиболее подходящей, если основываться на этих данных и учитывать теоретические соображения? В свете данной простейшей интерпретации, «достаточно ли точна» линейная модель в большинстве случаев? Можете ли Вы предложить лучшую модель?

Иллюстрация к упражнению 6—6

Доли мест в Конгрессе и голосов избирателей (для демократической партии США), 1900—1972

Год	% голосов, отданных за демократов	% мест в конгрессе	Год	% голосов, отданных за демократов	% мест в конгрессе
1900	46,60	43,59	1936	58,48	78,91
1902	48,68	46,23	1938	50,82	60,79
1904	43,66	35,23	1940	52,97	62,24
1906	46,55	42,49	1942	47,66	51,51
1908	48,11	43,99	1944	51,71	56,12
1910	50,50	58,46	1946	45,27	43,32
1912	57,11	69,54	1948	53,24	60,60
1914	50,34	54,48	1950	50,04	54,04
1916	48,88	49,30	1952	49,94	49,08
1918	45,10	44,63	1954	52,54	53,33
1920	37,67	30,56	1956	50,97	53,79
1922	46,40	47,92	1958	56,10	64,91
1924	42,09	42,56	1960	54,97	59,95
1926	41,57	45,14	1962	52,42	59,45
1928	42,84	37,91	1964	57,50	67,82
1930	45,87	49,77	1966	51,33	57,01
1932	56,87	72,79	1968	50,92	55,86
1934	56,18	75,76	1970	54,32	58,62

Источник. Tufte E. R. The relationship between seats and votes in two-party systems.—American Political Science Review, 68, June 1974, 540—554. (Данные воспроизводятся с разрешения автора и журнала.)

ГЛАВА 7

7.1.1. Вернитесь к илл. 1.4.1 (с. 40) и для дней со 2-го по 24-й выпишите кажущуюся дисперсию дневного среднего: (1) основываясь на разностях между дневными средними и (2) как среднее интервальных оценок ($s_{\bar{x}}^2$) дисперсий дневных средних. Если

$$\text{«истинная» дисперсия дневного среднего} = \frac{s^2}{n_{\text{эфф}}},$$

то

$$\frac{n}{n_{\text{эфф}}} = \frac{\text{дисперсия, основанная на разностях (внешняя)}}{\text{дисперсия, основанная на } \left(s_{\bar{x}}^2\right) \text{ (внутренняя)}}.$$

Найдите оценку $n/n_{\text{эфф}}$. При n порядка 500, какова будет приблизительная величина $n_{\text{эфф}}$. Что это означает?

7.1.2.[к] Почему невозможно найти подходящую формулу для оценки обоснованной меры неопределенности?

7.2.1. Палумбо и Стругала в статье [P a l u m b o F. A. and S t r u g a l a E. S. Industrial Quality Control, November 1945, 6—8] приводят данные о долях дефектов (при вулканизации) прокладок батарей, используемых в портативных дуплексных радиостанциях, для 32 последовательных партий. В каждой паре чисел первое — объем выборки, второе — число дефектов. Вот данные: (140, 77), (140, 19), (140, 24), (140, 20), (140, 27), (155,0), (155,0), (210,0), (155, 0), (210,0), (50,50), (50,4), (50,17), (90,0), (105,0), (105,4), (155, 8), (155,2), (155, 0), (210, 4), (155,5), (155,7), (105,3), (210,12), (190, 9), (125, 7), (125,5), (125,2), (75,0), (75,4), (125,1), (125,2). Всего было 9 партий объемом от 50 до 105 единиц, 10 — от 125 до 140 и 13 выборок объемом 155 и более. Подсчитайте для каждой из этих трех групп партий внутреннюю оценку (среднее есть pq/n) дисперсии партии p и внешнюю оценку (обычное s^2, основанное на p значениях для этих партий). Сравните их и обсудите полученное сравнение.

7.3.1. При изучении широты различий в классе команд Национальной футбольной лиги (НФЛ) были собраны различные физические характеристики для дальнейшего соотнесения их с классом команд. На этот раз мы хотим изучить различия в классе между четырьмя ассоциациями, на которые делится НФЛ. Наши ресурсы ограничены, и мы вынуждены взять только по две команды из каждой ассоциации. Из некоторых соображений решено включить в исследование самую сильную и самую слабую команды в каждой ассоциации. Как лучше всего найти ошибку сравнения ассоциаций? Почему ? Если этот член, характеризующий ошибку, имеет смещение, то в какую сторону оно скорее всего направлено?

7.3.2. Одним из аспектов изучения пяти поваренных книг, содержащих рецепты, характерные для Новой Англии, была оценка простоты приготовления блюд. Для практической проверки в специально оборудованных кухнях выборки их получали двумя способами. По первому способу из каждой книги случайным образом отбиралось 20 рецептов. По второму рецепты были сначала расклассифицированы на 5—10 категорий (таких, как мясные блюда, рыбные, овощные, соусы и подливы и т. п.), а затем из каждой категории случайно отбиралось заданное число рецептов. Как бы Вы оценили член, содержащий ошибку, применительно к условиям каждого из этих планов отбора?

7.3.3. (Продолжение упр. 7.3.2.) Если бы пришлось использовать несколько сочетаний блюд с кухнями, то как бы Вы их сбалансировали с поваренными книгами?

7.3.4. Сравнение шести крупных универсамов по ценам продуктов питания, без учета распродаж уцененных продуктов, включало деление всех продающихся продуктов на 20 категорий и 10 более или менее постоянных наименований в каждой категории. Затем в течение года дважды в неделю (в систематически чередующиеся дни) определялись цены всех отобранных продуктов во всех шести магазинах, так что всего удалось собрать 124800 данных. Что бы Вы рекомендовали как основу для оценки члена, содержащего ошибку? Как зависит Ваш ответ

от того, предназначаются ли результаты для описания годовых наблюдений или для руководства к тому, где в будущем следует покупать продукты?

7.4.1. Два специалиста ведут некое исследование по двум различным планам. Реализация их планов требует различных денежных затрат. Первый собирает больше данных в меньшее число независимых групп — всего 1000 наблюдений в пяти группах. Второй собирает меньше данных, но в большее число групп — всего 700 наблюдений в 10 группах. Если число групп сказывается только на соответствующем числе степеней свободы, а не на межгрупповой вариации, то какой из этих планов даст в среднем более короткий 95%-ный доверительный интервал?

7.4.2. (Продолжение упр. 7.4.1.) А что можно сказать о плане, включающем (900 наблюдений в 9 группах?

7.5.1. Прошан приводит в работе [P r o s c h a n F. Technometrics, 5,376 (1963)] данные об интервалах времени между отказами систем кондиционеров 12 реактивных самолетов 720-й модели. Вот они (мы приводим сначала бортовой номер машины, затем, после точки с запятой, следуют периоды безотказной работы (в часах), разделяемые капитальными ремонтами по любой причине): (7907; 194, 15, 41, 29, 33, 81), (7908; 413, 14, 58, 37, 100, 65, 9, 169, 447, 184, 36, 201, 118), (7909; 90, 10, 60, 186, 61, 49, 14, 24, 56, 20, 70, 84, 44, 59, 29, 118, 25, 156, 310, 76, 26, 44, 23, 62), (7910; 74, 57, 48, 29, 502, 12, 79, 21, 29, 386, 59, 27), (7911; 55, 320, 56, 104, 220, 239, 47, 246, 176, 182, 33), (7912; 23, 261, 87, 7, 120, 14, 62, 47, 225, 71, 246, 21, 42, 20, 5, 12, 120, 11, 3, 14, 71, 11, 16, 90, 1, 16, 52, 95), (7913; 97, 51, 11, 4, 141, 18, 142, 68, 77, 80, 1, 16, 106, 206, 82, 54, 31, 216, 46, 111, 39, 63, 18, 191, 18, 163, 24), (7914; 50, 44, 102, 72, 22, 39, 3, 15, 197, 188, 79, 88, 46, 5, 5, 36, 22, 139, 210, 97, 30, 23, 13, 14), (7915; 359, 9, 12, 270, 603, 3, 104, 2, 438); (7916; 50, 254, 5, 283, 35, 12); (8044; 487, 18, 100, 7, 98, 5, 85, 91, 43, 230, 3, 130), (8045; 102, 209, 14, 57, 54, 32, 67, 59, 134, 152, 27, 14, 230, 66, 61, 34). Для каждого самолета найдите: (а) медианное время между отказами; (б) долю отказов с интервалами меньше 20; (в) медиану длин интервалов, меньших, чем 20; (г) медиану длин интервалов, равных или больших, чем 20. Воспользуйтесь t-критерием Стьюдента для построения доверительных границ средних каждого из этих четырех периодов времени для каждого самолета.

7.5.2. Андерсон ([A n d e r s o n R. L. Journal of the American Statistical Association, 42, 612—634 (1947)]) исследовал логарифмы и отношения цен на поросят в Цинциннати и в Луисвилле, полученные для двух весовых категорий на протяжении 5 лет (1937—1941) все 12 месяцев по 5 дней в неделю. Вот его обобщенные данные а) по месяцам (от января к декабрю): (197,4; 194,3; 190,5; 208,9; 215,6; 200,9; 192,1; 185,9; 146,6; 137,6; 144,9; 162,8); б) по дням (от понедельника к пятнице): 175,2; 174,0; 180,0; 173,6; 183,8); в) по годам (от 1937 к 1951): (169,5; 195,3; 220,4; 168,7; 132,6). Обработайте эти данные в отдельности по месяцам, дням и годам и воспользуйтесь ими как основой для прямой оценки ошибки с помощью t-критерия Стьюдента и доверительных границ общего среднего. На каких теоретических принципах должен основываться наиболее подходящий выбор члена, содержащего ошибку? Исследуйте ваш ответ. После рассмотрения этих данных могли ли бы Вы внести в Ваш ответ какие-либо изменения? Почему? Почему нет?

7.6.1. (Головоломка; средства для ее разгадки в тексте отсутствуют.) Данные Андерсона о ценах на поросят (см. упр. 7.5.2 выше) включают следующие средние по дням недели, скомбинированные по годам и приведенные здесь округленно (с точностью до единиц) в виде (год; понедельник, вторник, среда, четверг, пятница): (1897; 156, 168, 173, 158, 173), (1938; 202, 192, 200, 195, 201), (1939; 207, 210, 211, 214, 219); (1940; 144, 118, 132, 131, 151), (1941; 125, 130, 131, 123, 132). Примем, раз уж это необходимо, что и годы, и дни недели должны вносить свой вклад в дисперсию, связанную с общим средним. Продемонстрируйте подробные вычисления, в которых оцениваемая дисперсия учитывает оба частных источника вариации, а заодно и остаточную вариацию. Сравните полученные 95%-ные доверительные границы с теми, которые получились бы, если бы каждый из источников вариации учитывался сам по себе.

7.7.1. О некотором химическом анализе на основе многолетних исследований известно, что вариация между днями недели оценивается дисперсией $\sigma^2 = 2 \times 10^{-5}$, между неделями месяца — дисперсией $\sigma^2 = 1 \times 10^{-5}$ а между

месяцами года — величиной $\sigma^2 = 3 \times 10^{-5}$. Однажды за день было сделано 16 определений, которые дали величину $s_x^2 = 0{,}0000173$. Какую дисперсию надо было бы приписать среднему этих 16 анализов для учета (а) среднего внутри месяца или (б) среднего внутри года?

7.7.2. (Продолжение упр. 7.7.1.) В другой раз за день снова сделали 16 анализов, которые дали s^2 (не $s_{\bar{g}}^2$) $= 0{,}000279$. Какую дисперсию надо было бы приписать общему среднему двух множеств по 16 измерений (с учетом как (а), так и (б) в предыдущем упражнении), если (1) эти два дня были на самом деле одним и тем же днем, (2) эти два дня были различны, но принадлежали одной неделе, (3) они относились к разным неделям одного месяца, (4) — к разным месяцам одного года?

7.7.3. (Продолжение 7.2.2.) Должен ли зависеть Ваш ответ в упр. 7.7.2 от цели, ради которой получается оценка? Например, если она получается для недельного, месячного, годового среднего или среднего многолетнего? Проиллюстрируйте Ваш ответ численно для случая (2) из упр. 7.7.2.

ГЛАВА 8

8.1.1. Назовите два важных использования «складного ножа». Почему этот метод так называется?

8.1.2. (а) Что такое $y_{\text{общ}}$? (б) Что такое $y_{(j)}$? (в) Что такое $y*_j$? (г) Что такое $y*$?

8.1.3. Проверьте уравнение (1) из параграфа 8.1 и выясните, почему псевдозначения можно рассматривать как аналоги экспериментальных данных y?

8.1.4. Пусть ожидаемое значение какой-то статистики, основанное на выборке объема n, равно $\mu + (a/n)$. Такая статистика будет смещенной оценкой μ. Для $n = k$ найдите ожидаемое значение $y*_j$ в уравнении (1). (Учтите, что $E\,(y_{(j)}) = \mu + [a/(k-1)]$.) Объясните, как этот результат подтверждает второе замечание в первой части параграфа 8.1.

8.1.5. Когда статистика представляет собой выборочное среднее. Получите $y*$(т. е. когда $y_{\text{общ}} = \Sigma y_j/k$, а

$$y_{(j)} = \frac{(\Sigma y_j) - y_j}{k-1}\Big).$$

8.1.6. Приведите оценку дисперсии нормального распределения с неизвестным средним. Покажите, что смещение этого оценивателя $y_{\text{общ}} = \Sigma\,(x - \bar{x})^2/n$ имеет порядок $1/n$. Будет ли смещение y_j больше, чем смещение $y_{\text{общ}}$? Покажите, что $y*$ — несмещенная оценка.

8.2.1. Среднее геометрическое k измерений дается формулой $g_k = \sqrt[k]{y_1 y_2 \ldots y_k}$. Имеем 4 измерения: 1, 2, 2, 4. С помощью «складного ножа» оцените геометрическое среднее генеральной совокупности и постройте для него $\frac{2}{3}$-доверительные интервалы.

8.2.2. Воспользуйтесь логарифмами в упр. 8.2.1 для получения доверительных границ $\log g_k$, а затем и для g_k.

8.2.3. Для подгонки прямой, проходящей через начало координат, исследователь получил 6 точек (x_j, y_j). Вот они:

(2,1) (2,2) (4,2)
(2,2) (3,2) (5,4)

Для угловых коэффициентов он взял оценку $m = \dfrac{\Sigma y_j}{\Sigma x_j}$. Воспользуйтесь методом «складного ножа» для построения 2/3-доверительных границ для «истинного» наклона.

8.2.4. (Продолжение упр. 8.2.3.) В качестве альтернативы своему методу оценки угловых коэффициентов исследователь рассматривал

$$m_1 = \frac{1}{k}\left(\frac{y_1}{x_1} + \frac{y_2}{x_2} + \ldots + \frac{y_k}{x_k}\right).$$

Найдите оценку «складного ножа» для данных из упр. 8.2.3 и сравните s_*^2 двух методов. Из этих двух оценок какая, на Ваш взгляд, предпочтительнее?

8.3.1. Проверьте столбец $j = 3$ из илл. 8.3.2. (с. 165).

8.3.2. Зачем в илл. 8.3.2 понадобилась строка «1000 (округл. $z*_j$ — 0,100)»?

8.3.3. Используйте отдельные группы из илл. 8.2.3 как основу для получения усеченных на 20% средних методом «складного ножа», а затем постройте доверительные границы этой величины. Вспомните, что эти средние получаются путем отделения 20% наибольших значений и 20% — наименьших. Затем от того, что осталось, берется среднее арифметическое.

8.3.4. Что же оценивает такое усеченное среднее?

8.3.5. Должно ли измениться ожидаемое значение $y*$ в упр. 8.3.3, если Вы продублируете наблюдения всех объектов, но число этих объектов не измените? Почему? Почему нет?

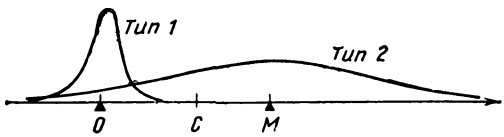
8.4.1. Зачем нужна дискриминантная функция в задаче о «The Federalist» из параграфа 8.4?

8.4.2. Как эта дискриминантная функция связана с обычным уравнением регрессии? Специально выясните роль y.

8.4.3. Почему для дискриминантной функции не встает вопрос о конкретных значениях параметров A и B ($A \neq 0$)? Что случится, если A удвоить?

8.4.4. Как правило, в задачах дискриминации есть разделяющая точка. Что фактически играет здесь ее роль?

8.4.5. Бывает, что распределения, связанные с двумя группами наблюдений, в задаче дискриминации выглядят так:



Можете ли Вы считать разделяющей точкой полпути между средними? Что Вы от этого выиграете?

8.4.6. Если средними типов 1 и 2 положить соответственно 0 и M, а стандартными отклонениями σ_1 и σ_2, да еще принять, что распределения дискриминируемых типов приближенно нормальны, то как найти формулу для такой разделяющей точки (C), которая обеспечивает равное число ошибок классификации для обоих типов?

8.4.7. Обычно в методе «складного ножа» мы отбрасываем наблюдения по одному, но в задаче о «The Federalist» из статей Гамильтона составляли случайным образом пары со статьями Мэдисона и отбрасывали уже пары. Почему мы так поступили? Была ли в этом необходимость? А чего можно пожелать? Теряется ли при этом что-нибудь?

8.4.8. Воспользуйтесь частотами слова «of» (илл. 8.4.1, с. 166) для разделения авторства между Гамильтоном и Мэдисоном. Для этого постройте дискриминантную функцию $y = Ab_3X_3 + B$ и оцените по ней 0 или 1 для Мэдисона и Гамильтона соответственно.

8.4.9. (Продолжение упр. 8.4.8.) Соедините ваш метод дискриминации со «складным ножом», действуя так же, как и в илл. 8.4.3 (с. 168).

8.4.10. Воспользуйтесь словами «of» и «and» для построения дискриминантной функции для задачи о Гамильтоне и Мэдисоне. В основном придерживайтесь илл. 8.4.2 (с. 168), но для этих двух слов.

8.4.11. (Продолжение упр. 8.4.10.) Воспроизведите илл. 8.4.3 (с. 168) для слов «of» и «and».

8.5.1. (Продолжение упр. 8.4.9.) Как перепроверить Ваши оценки?

8.5.2. (Продолжение упр. 8.5.1.) Сравните работу слова «of» отдельно от всех 5 слов в тексте.

8.5.3. (Продолжение упр. 8.4.11.) Воспроизведите илл. 8.5.1 (с. 169) для «of» и «and».

8.5.4. (Продолжение упр. 8.5.3.) Сравните работу слов «of» и «and» со всеми 5 словами и отдельно со словом «of».

8.5.5. Найдите из илл. 8.5.1 (с. 169) оцениваемое значение вероятности ошибки классификации статей Гамильтона и Мэдисона.

8.6.1. (Проект) (Продолжение упр. 8.4.8, 8.4.9, 8.5.1, 8.5.2.) Воспроизведите два аналогичных способа «отбрасывания по одному» из параграфа 8.6 для дискриминант, основанных на слове «of» из обсуждаемого примера.

ГЛАВА 9

Предупреждение. Упражнения этой главы требуют громоздкого счета и тщательных проверок. Имеет смысл ограничиться при выборе упражнений, например, 1, 2, 5, 6, 9 или 10 и провести их через всю главу. Упр. 9.1.1 и 9.1.2 имеют продолжения.

9.1.1. Бен привел в [Ben U. Annalen der Physik, (4), 1, 257—269 (1900)] средние значения «грамм-атомной теплоемкости» многих элементов для трех температурных интервалов. Мы приведем здесь некоторые из них (в виде: (элемент; интервал от — 180° до — 79°, от — 79° до + 18°, от + 18° до + 100° C)) в порядке снижения атомного веса: (свинец, Pb; 6,0, 6,2, 6,4), (платина, Pt; 5,4, 6,1, 6,3), (сурьма, Sb; 5,5, 5,8, 6,0), (олово, Sn; 5,8, 6,1 6,5), (кадмий, Cd; 5,6, 6,0, 6,3), (серебро, Ag; 5,4, 5,9, 6,0), (палладий, Pd; 5,2, 6,0, 6,3), (цинк, Zn; 5,2, 5,8, 6,1), (медь, Cu; 4,5, 5,6, 6,0), (никель, Ni; 4,3, 5,8, 6,4), (железо, Fe; 4,0, 5,6, 6,3), (алюминий, Al; 4,2, 5,3, 6,0), (магний, Mg; 4,6, 5,7, 6,1). Проанализируйте результаты таблицы 3 × 13 построчно, как предлагается в параграфе 9.1. Приведите их к стандартной форме. Что же вы видите?

9.1.2^{k}. Стэм, Кратц и Уайт привели [Stam P. B., Kratz R. F. and White H. J. Textile Research Journal, 22, 448—465 (1952)] данные об увеличении диаметра сухих человеческих волос под воздействием относительной влажности воздуха, когда она сначала растет, а потом падает. Мы приводим часть из них ниже в таком порядке: (N выборки; изменение диаметра в процентах при изменении влажности на 10, 40, 60, 90, 100, 90, 60, 40, 10, 0%, где «на x%» соответствует увеличению диаметра (в тысячных долях исходного) после стабилизации процента относительной влажности (N 9; 25, 57, 63, 89, 142, 83, 62, 50, 20, 9), (N 9A; 25, 43, 57, 84, 115, 87, 61, — 21, 6), (N 11; 54, 72, 110, 133, 171, 120, 79, 69, 42, 0), (N 12; 10, 29, 44, 80, 116, 100, 54, 38, 16, 1), (N 14; 30, 75, 101, 142, 164, 141, 109, 84, 40, 3). Проанализируйте результаты, сведенные в таблицу 5 × 10, как в параграфе 8.1. Приведите их к стандартной форме. Видите ли Вы что-нибудь интересное?

9.1.3. Повторите упр. 9.1.2, используя квадратные корни приведенных данных.

9.1.4. Повторите упр. 9.1.2, используя логарифмы данных, предварительно увеличенных на единицу. (Так, например, 25 соответствует $\log 26 = 1{,}41$.)

9.1.5. Браун и Грейхэм представили [Brown H. M. and Graham J. S. Textile Research Journal, 20, 418—425 (1950)] таблицу данных, показывающих зависимость качества хлопка (в мг/дм) от сорта и процента зрелых волокон, вот некоторые из этих данных в форме (сорт; 65%; 70%, 75%, 80%): («Половина на половину»; 5,10, 5,40, 6,00, 6,81), («Альпийская дельта 11»; 3, 95, 4, 25, 4, 60, 5, 07), («Стоунвилл 5»; 4,35, 4,55, 4,68, 4,78), («Стоунвилл 2B (8275)»; 3,40, 3,82, 34,22, 4,65), («Гибрид»; 4,33, 4,73, 5,14, 5,60), («Кволла»; 4,63, 4,88, 5,10, 5,28), («Звезда текстиля»; 4,45, 4,82, 5,30, 6,05), (Мексиканский «Большая коробочка»; 4,10, 4,32, 4,59, 4,88), («Мечта фермера»; 4,68, 4,95, 5,15, 5,32), («Кливлендский»; 4,90, 5,16, 5,30, 5,36), («Арканзас»; 3,85, 4,08, 4,34, 4,67), («Смеска-912»; 4,25, 4,52, 4,81, 5,10), («Роджерс акала»; 3,72, 3,95, 4,13, 4,30), («Триумф (759)»; 4,56, 4,85, 5,15, 5,48), («Триумф (44)»; 4,47, 4,83, 5,23, 5,70). Проанализируйте таблицу 5×4 для первых пяти сортов в соответствии с параграфом 9.1. Представьте результаты в стандартной форме. И что Вы видите?

9.1.6. Проанализируйте полную таблицу 15 × 4 для данных из упр. 9.1.5 и сообщите, что Вы видите?

9.1.7. Повторите упр. 9.1.6 с логарифмами вместо исходных значений.

9.1.8. (Продолжение упр. 9.1.6 и 9.1.7.) Можете ли Вы сравнить и соотнести друг с другом результаты упр. 9.1.6 и 9.1.7?

9.1.9. Юден привел в [Jouden W. J. Analytical Chemistry, 19, 946—950] окончательные убедительные данные, полученные в прецизионных измерениях соотношения между йодом и серебром (служащие основанием для вычисле-

ний атомного веса). В этих опытах два соединения йода исследовались во многих, но не во всех сочетаниях с пятью соединениями серебра. Мы приводим медианы сравнений в виде (используемое соединение йода; для A, для B, для C, для D, для E) в единицах значимого седьмого знака после запятой, где, A, B, C, D, E — соединения серебра (йодид-I; 24*, 41*, 29**, 50*, 55), (йодид-II; 18**, 18*, —, 61, —), причем «*» означает медиану по двум наблюдениям, «**» — медиану по трем наблюдениям. [Остается думать, что отсутствие звездочек говорит о единственном наблюдении, а тире — о том, что наблюдений нет вовсе. — *Ю. А.*] Проведите анализ таблицы с двумя входами 2 × 5 при двух пустых ячейках как можно лучше. Воспользуйтесь стандартной формой. Обсудите результаты.

9.1.10. В отчете «The Report of the Royal Society, IGY Antarctic Expedition to Halley Bay. etc.» (Sir David Brunt, ed., 1962) об Антарктической экспедиции в рамках Международного геофизического года приводятся (в табл. 14, p. 191) среднемесячные температуры на различных высотах над уровнем моря (более точно, на высотах, где давление воздуха падает до удельного давления) по данным с шаров-зондов, запускавшихся в полдень по Гринвичу. На илл. 1 к упр. 9.1.10 представлено извлечение из этой таблицы в виде таблицы температур размером 14 × 12. Проанализируйте эти данные в соответствии с рекомендациями параграфа 9.1 и приведите результаты к стандартной форме. Как вы думаете, будет ли этот анализ успешным? Сможет ли он стать основой для дальнейшего исследования?

Иллюстрация 1 к упражнению 9.1.10

Среднемесячные температуры (все отрицательные) в десятых долях градуса Цельсия (за 1958 г.)

Уровень давления, мбар	Месяцы года											
	Я	Ф	М	А	М	И	И	А	С	О	Н	Д
30	343	423	529	687	787	870	917	885	827	651	394	326
30	354	425	530	665	779	849	901	891	852	704	443	348
50	368	425	517	640	778	837	888	886	850	720	464	350
60	375	428	509	636	761	822	870	877	839	728	483	371
80	385	431	498	610	719	789	843	851	829	731	520	392
100	396	435	496	585	696	761	811	831	820	740	597	403
150	412	435	472	557	644	717	768	792	789	740	601	446
200	423	435	468	551	630	701	750	762	759	720	626	461
250	453	456	492	584	652	666	701	719	708	677	628	540
300	502	496	548	571	614	616	639	654	644	616	586	556
400	430	427	463	472	509	500	521	525	512	499	465	459
500	334	329	361	373	411	396	417	417	399	385	354	355
700	179	186	212	227	270	254	257	272	248	225	204	209
850	93	123	139	177	208	206	207	218	211	154	128	128

9.2.1. (Продолжение упр. 9.1.1.) Постройте график, подобный илл. 9.2.1 (с. 189) для результатов анализа упр. 9.1.1.

9.2.i. (Продолжение упр. 9.1.1 для $i = 2, \ldots, 10$.) Сделайте то же, что и в упр. 9.2.1, для той из задач от 9.1.2 до 9.1.10, которую Вам предложат.

9.2.(10 + i). (Продолжение упр. 9.2.i; $i = 1, 2, \ldots, 10$.) Постройте графики ранее найденных остатков, аналогичные илл. 9.2.2 и 9.2.3 (с. 190), для выделенной Вам задачи. Что они говорят Вам?

9.3.i. ($i = 1, 2, \ldots, 10$.) Поверните график из упр. 9.2.i (для Вашего i) на 45° и нанесите горизонтальные риски, как на илл. 9.3.3 (с. 191).

9.3.(10 + i). (Продолжение упр. 9.1.i; $i = 1, 2, \ldots, 10$.) Постройте по Вашим данным график, аналогичный илл. 9.3.4 (с. 192), пользуясь остатками из упр. 9.1.i. Что Вы видите?

9.4.i. (Продолжение упр. 9.1.i; $i = 1, 2, \ldots, 10$.) Уточните анализ упр. 9.1.i, как показано в параграфе 9.4. Возьмите медианы. Насколько это улучшит результаты?

9.4.(10 + i). (Продолжение упр. 9.1.i; $i = 1, 2, \ldots, 10$.) То же, что и упр. 9.4.i, но не для медиан, а для чего-нибудь другого.

9.4.21. На илл. 1 к упр. 9.4. 21 приводятся доли детских смертей, приходящихся на 1000 рождений для некоторых регионов США, классифицированные по образовательному цензу отцов. Изучите эти данные с помощью аддитивной модели и анализа остатков.

Иллюстрация 1 к упражнению 9.4.21

Детская смертность в США, 1964—1966 гг.

Регион	Образовательный ценз, отцов в годах обучения				
	⩽8	9—11	12	13—15	⩾16
Северо-Восток	25,3	25,3	18,2	18,3	16,3
Север Центра	32,1	29,0	18,8	24,3	19,0
Юг	38,8	31,0	19,3	15,7	16,8
Запад	25,4	21,0	20,3	24,0	17,5

Источник. Infant Mortality Rates: Socioeconomic Factors, United States. U. S. Department of HEW, NCHS, Vital and Health Statistics, Series 22, Number 14.

9.5.i. (Продолжение упр. 9.1.i или 9.4.i; $i = 1, 2, \ldots, 10$.). Постройте графики, подобные илл. 9.5.2 (с. 198), опираясь на анализ упр. 9.1.i или 9.4.i. Каковы Ваши выводы?

9.5.(10 + i). (Продолжение 9.5.i; $i = 1, 2, \ldots, 10$.) Если на графике из упр. 9.5.i появится какой-нибудь наклон, то проведите дальнейший анализ, как в илл. 9.5.3 (с. 199).

9.6.1. Проведите упр. 9.1.3 через всю главу, если Вы этого еще не сделали.

9.B.1. *Рост движения за равноправие* (проект). Рост движения за равноправие можно определить долей женщин, имеющих равное положение с мужчинами (число рожденных ранее в иных условиях), которые действительно пришли к новой жизни. Бин и Вуд (Bean, Wood) оценили эти доли для трех этнических групп населения Юго-Запада США в 1960 и в 1970 гг. Они приведены в илл. 1 к упр. 9.B.1.

Эти данные можно обрабатывать двояко. Во-первых, можно скомпоновать шесть строк как три разные совокупности и применить аддитивный анализ таблицы с двумя входами. С другой стороны, мы можем каждый год анализировать отдельно. Как сравнить эффекты столбцов (или доли достигших равноправия) в этих двух способах анализа? А как насчет эффектов строк (или групп)? Действительно ли второй путь дает настолько более ясную картину остатков, что это оправдывает его дополнительные трудности?

Что говорят остатки и диагностические графики о преобразованиях? Говорят ли они то же самое, что и полная таблица?

Иллюстрация 1 к упражнению 9.В.1

Рост движения за равноправие

Равноправие / 1960	0	1	2	3	4	5
Англо-американцы	0,879	0,797	0,558	0,484	0,480	0,531
Мексиканцы	0,939	0,920	0,818	0,796	0,789	0,774
Негры	0,748	0,778	0,787	0,789	0,774	0,759
1970						
Англо-американцы	0,929	0,882	0,646	0,531	0,476	0,496
Мексиканцы	0,961	0,943	0,839	0,803	0,748	0,713
Негры	0,866	0,877	0,794	0,761	0,723	0,725

Источник: Bean F. D. and Wood C. H. Ethnic variations in the relationship between income and fertility.—Demography, Nov., 1974, 629—640. (Данные воспроизведены с разрешения авторов и журнала).

9.В.2. *Детская смертность* (проект). На илл. 1 к упр. 9.В.2 приведены данные о детской смертности (число умерших, не дожив до года, на 1000 родившихся), расклассифицированные по возрасту матерей и порядку рождения (числу более старших детей в семье + 1) ребенка. Какие преобразования кажутся Вам полезными при описании этих данных? Не осталось ли в остатках еще каких-нибудь закономерностей?

Иллюстрация 1 к упражнению 9.В.2

(Умершие в первый год жизни)/(1000 родившихся)

Возраст матери	Порядковый номер ребенка					
	1	2	3	4	5	6
15—19	26,1	42,7	54,7	63,4	96,9	140,0
20—24	17,2	21,8	27,3	35,1	45,2	58,7
25—29	17,5	17,3	18,9	22,4	28,5	39,7
30—34	24,1	19,3	18,2	20,2	23,6	33,5
35—39	27,7	22,8	21,0	20,9	22,0	32,0
40—44	33,4	31,2	26,8	24,1	25,0	35,3

Источник. A study of infant mortality from liked records by age of mother, total birth order, and other variables. United States, 1960 Live Birth Cohort, National Center of Health Statistics, Vital and Health Statistics, Series 20, Number 14, 1973.

9.В.3. (Головоломка — может подойти для классной работы). Маек, Балог и Джордон в [Maek P. B., Balog J. A., Jordon M. N. Textile Research Journal, 22, 30 — 42] представили данные об изменениях в процентах прочности на разрыв для 6 тканей, 2 режимов стирки и 5 последовательных номеров стирок. Эти результаты для изменения прочности влажной ткани можно представить в десятых долях процента в виде: (ткань, режим; основа после 1-й стирки, 5-, 10-, 15-, 20-й; уток после 1-, 5-, 10-, 15-, 20-й), (солайна, ультразвук; — 40, — 1 —104, —100, — 108; — 54, — 34, — 75, — 109, — 128), (солайна, ручной; — 71, — 99, — 129, — 109, — 129; — 143, — 110, — 177, — 51, — 38), (габардин-I, ультразвук; 158, 106, — 49, — 24, 69; — 36, — 23, 13, 95, 10), (габардин-I, ручной; — 213, — 291, — 286, — 420, — 379; — 5, — 50, — 69, — 104, — 114), (поплин, ультразвук; 17, — 34,—72, — 78, — 115; — 24, — 123,

— 11, 55, — 2), (поплин, ручной; — 202, — 345, — 438, — 523, — 523; 65, 173, 32, 162, 144), (габардин-II, ультразвук; 185, 212, 224, 166, 310, 200, 200, 276, 187, 250), (габардин-II, ручной; — 130, —78, — 146, — 138, — 246; 169, 127, 122, 344, 74), (тафта, ультразвук; 0, — 58, 10, — 48, — 58; 0, 260, 150, 30, 120), (тафта, ручной; — 11, — 112, — 135, — 124, — 180; — 32, — 24, 71, — 48, 16), (сатин, ультразвук; 115, 135, 62, 94, 83; 78, 117, 109, 86, 78), (сатин, ультразвук; 115, 135, 62, 94, 83; 78, 117, 109, 86, 78), (сатин, ручной; — 197, — 193, — 160, — 181, — 354; — 190, — 95, — 48, — 114, 152). Заметьте, что изменения простираются от — 523 (потеря 52,3% исходной прочности) до 344 (приобретение 34,4% дополнительной прочности). Продвиньтесь как можно дальше в анализе конфигурации $6 \times 2 \times 5 \times 2$, представьте Ваши результаты с наибольшей наглядностью и обсудите их следствия.

ГЛАВА 10

10.1.1. Подсчитайте среднее число дней выживания для контрольной группы с учетом и без учета «выбросов». Сделайте то же и для экспериментальной группы по данным из илл. 1 к упр. 10.1.1.

10.1.2. Подсчитайте двумя способами, по медианам и по бивес-средним (для $c = 6$), общую сопротивляемость для контрольной и экспериментальной групп при изучении действия витамина *C* на свиней гвинейской породы.

Иллюстрация к упражнению 10.1.1. (используемая также и в упр. 10.1.2, 10.4.1, 10.5.1, 10.5.6)

Время выживания (сверх 10 дней) поросят гвинейской породы, родившихся от свиноматок экспериментальной и контрольной групп

Норкус и Руссо провели наблюдения над гвинейской свиньей, чтобы установить, не вызовет ли повышенное потребление беременными женщинами витамина *C* (являющегося источником аскорбиновой кислоты) в конечном счете цингу у потомства. Оплодотворенные свиноматки были разделены на две группы. В экспериментальной группе (4 животных) обеспечивался уровень аскорбиновой кислоты, эквивалентный 1500 мг для 70-килограммового человека. А в контрольной группе (5 животных) давалась лишь десятая часть этого количества. Восемь родившихся поросят в экспериментальной группе и 14 в контрольной получали те же дозы, что и их матери в течение 10 дней. На 11 день все поросята были существенно ограничены в получении витамина *C*.

Экспериментаторы сравнивали числа дней выживания после 10 дней в двух группах с помощью *отдельных* стандартных ошибок в каждой группе. Здесь дни выживания представлены в виде «опоры и консоли»:

Контрольная группа (низкий уровень витамина *C*)		Экспериментальная группа (высокий уровень витамина *C*)	
1		1	4466
2	1345	2	0446888
3	133	3	012
4		4	
5	4	5	

Они пришли к выводу, что когда столь важное питательное вещество изымается из рациона, поросята от матерей с высоким уровнем витамина *C* умирают чаще, чем при контрольной диете.

Источник. Norkus E. P. and Russo P. (1975). Changes in ascorbic acid metabolism of the of spring following high maternal intake of this vitamin in the pregnat guinea pig. — Annals of the New York Academy of Sciences, 258, Second Conference on Vitamin *C*, 401—409.

10.1.3. Кларк [Clarke. The data geochemistry. — Bulletin 3304, The U. S. Geological Survey, 1908, p. 608] приводит результаты анализов (в %) 8 образцов самородной платины. Они приведены ниже для 5 наиболее важных ин-

гредиентов после округления. Для выделенного Вам ингредиента найдите (а) среднее, (б) медиану, (в) срединное среднее и (г) бивес-среднее. Сравните результататы.

Ингредиент	Образец							
	1	2	3	4	5	6	7	8
Платина	86,2	85,5	82,8	76,4	49,0	68,2	78,4	73,0
Палладий	0,5	0,6	3,1	1,4	0,2	0,3	0,1	21,8
Родий	1,4	1,0	0,3	0,3	3,3	3,1	1,7	?
Медь	0,6	1,4	0,4	4,1	1,6	3,1	3,9	0
Железо	7,8	6,8	11,0	11,7	18,9	7,9	9,8	0

10.1.4. Кларк в той же работе (р. 523) привел результаты анализов семи образцов минерала серпентина (змеевика). Мы снова их округлили и привели ниже для основных ингредиентов:

Ингредиент	1	2	3	4	5	6	7
SiO_2	40,4	39,1	41,9	40,5	31,0	44,9	13,1
Al_2O_3	1,9	2,1	0,7	0,8	1,0	5,5	1,6
Fe_2O_3	2,8	4,3	0	4,0	4,9	1,8	1,2
FeO	4,3	2,0	4,2	2,0	2,0	3,5	0,2
MgO	36,0	39.8	38,6	37,4	38,4	25,6	58,4
H_2O (I)	10,7	12,7	14,2	13,8	20,8	5,8	24,8

Повторите то, что Вы сделали в упр. 10.1.3.

10.1.5. Все тот же Кларк (р. 377) проанализировал семь образцов нефелинита (лейцитита). Как и раньше, мы приводим округленные данные для основных ингредиентов:

Ингредиент	1	2	3	4	5	6	7
SiO_2	5,19	44,4	50,2	46,5	46,0	42,6	46,1
Al_2O_3	20,3	20,0	11,2	11,9	17,1	9,1	16,0
Fe_2O_3	3,6	5,2	3,3	7,6	4,2	5,1	3,2
FeO	1,2	2,8	1,8	4,4	5,4	1,1	5,6
MgO	0,2	1,8	7,1	4,7	5,3	10,9	14,7
CaO	1,6	8,5	8,0	7,4	10,5	12,4	10,6
Na_2O	8,5	6,5	1,4	2,4	2,2	0,9	1,3
K_2O	9,8	8,1	9,8	8,7	9,0	8,0	5,1
H_2O (I)	1,1	1,4	2,6	3,6	0,4	4,2	1,4

Задание то же, что и раньше.

10.1.6. И еще раз Кларк (р. 317) проанализировал 6 образцов авгита, минерала переменного состава. Вот основные (округленные) из этих данных:

Ингредиент	1	2	3	4	5	6
SiO_2	45,2	49,1	47,5	48,7	54,9	47,1
TiO_2	4,3	0	3,0	0	0	1,8
Al_2O_3	7,7	8,0	4,1	9,3	6,3	7,8
Fe_2O_3	3,0	0	5,6	3,8	2,9	1,3
FeO	4,1	8,3	6,4	6,4	4,6	8,2
MgO	12,2	12,4	10,0	14,7	14,5	13,5
CaO	23,4	22,6	21,6	16,8	15,9	19,3

Задание то же.

10.1.7. Ландольт — Борнштейн [L a n d o l t , B o r n s t e i n. Physikalisch — Chemicche Tabellen, 1923, Vol. 2, p. 254] опубликовали результаты 9 исследователей для удельной теплоемкости воды при различных температурах. Для диапазона от 5 до 30° С (с шагом 5°) эти результаты представлены на илл. 1 к упр. 10.1.7. Большинство экспериментаторов работали с одинаковыми стандартными термометрами, но трое использовали один из двух других стандартов, как показано в последнем столбце. Для назначенной Вам температуры повторите задание упр. 10.1.3. Рассмотрите результаты в свете дополнительной информации насчет использованных стандартов температуры.

Иллюстрация 1 к упражнению 10.1.7

Удельная теплоемкость воды при разных температурах по данным девяти исследователей

Исследователь	5°C	10°C	15°C	20°C	25°C	30°C
Regnoult	0,9994	0,9997	1,0000	1,0004	1,0008	1,0012 (A)
Liidin	1,0027	1,0010	1,0000	0,9994	0,9993	0,9996
Dieterici	1,0050	1,0021	1,0000	0,9987	0,9983	0,9984
Bonsfreld	1,0039	1,0016	1,0000	0,9991	0,9989	0,9990
Callendor	1,0042	1,0019	1,0000	0,9988	0,9980	0,9975 (T)
Ronland	1,0054	1,0019	1,0000	0,9979	0,9972	0,9969
Bartollis	1,0041	1,0017	1,0000	0,9994	1,0000	1,0016
Janke	1,0040	1,0016	1,0000	0,9991	0,9987	0,9988
Jaeger	1,0030	1,0013	1,0000	0,9990	0,9983	0,9979 (T)

10.1.8. В тех же таблицах Ландольта—Борнштейна (vol. 2, p. 801—802) собрано 14 значений постоянной Планка h. Вот они (значения надо умножить на 10^{29} эрг/с; округленно): 667, 662, 656, 655, 652, 658, 654, 650, 653, 657, 656, 653, 656, 654. Повторите задание для 10.1.3.

10.1.9. Найдите средние, медианы и бивес-средние ($c = 6$) для испытуемых 1 и 2 в условиях Лозанны (илл. 1 к упр. 10.1.9). Постройте три разности. Прокомментируйте их сходства и различия.

Иллюстрация 1 к упражнению 10.1.9 (а также и к упр. 10.1.10, 10.4.2, 10.5.3, 10.5.4).

Рефлекс коленной чашечки на разных высотах [над уровнем моря. — *Ю. А.*]

Линдер привел результаты измерений, проведенных в течение 8 дней подряд над коленночашечным рефлексом неких субъектов в Лозанне (высота 550 м), а затем, после однодневного перерыва, еще 8 дней в Цуоце (1750 м над уровнем моря). Часть данных, выраженных в логарифмических единицах (значения, в 20

раз превышающие десятичные логарифмы и ограниченные всюду двумя знаками), приведена ниже.

Испытуемый 1		Испытуемый 2	
Лозанна	Цуоц	Лозанна	Цуоц
77	58	79	59
82	63	80	59
78	81	65	56
75	63	78	56
76	76	78	56
75	71	83	65
74	95	84	66
73	57	79	84

Источник. Linder A. (1950). Statistical analysis of some physiological experiments. Sankhyā, 10, 1—12. (Цитируется по: Bliss C. I., Calhoun D. W. (1954). An Outline of Biometry, Yale Cooperative Corporation, p. 100.) Данные воспроизведены с разрешения автора и редакторов журнала.

10.1.10. Сделайте то же самое для испытуемого 1 в Лозанне и в Цуоце (илл. 1 к упр. 10.1.9).

10.1.11. Повторите построение бивес-среднего в упр. 10.1.3 используя $c = 9$. Прокомментируйте влияние величины c.

10.1.12. Повторите построение бивес-среднего в упр. 10.1.3 используя $c = 4$. Прокомментируйте влияние величины c.

10.1.13. Подсчитайте средние, медианы и бивес-средние ($c = 6$) для полученных в 1939 г. данных о весе маток крыс при разных дозах стильбестрола: 0,20 грана и 0,28 грана на крысу из илл. 1 к упр. 10.1.13 (1 гран = 64,8 миллиграмма.— *Ю. А.*). Найдите три разности и выясните, в чем они похожи, а в чем различны.

Иллюстрация 1 к упражнению 10.1.13 (а также и к упр. 10.1.14, 10.4.3, 10.4.4, 10.5.3, 10.5.4, 10.5.8, 10.5.9)

Веса маток недоразвитых крыс после введения стильбестрола (в 1000 (log вес — 1,4))

Ли, Роббинс и Чен давали недоразвитым крысам различные дозы стильбестрола, а затем взвешивали их матки. Среди прочего, они получили следующие результаты:

Декабрь 1939 г.		Апрель 1940 г.
0,20 грана	0,28 грана	0,28 грана
—38	83	—2
100	138	122
122	173	130
139	197	191
144	197	321
149	232	
214	251	

Источник. Lee H. M., Robbins E. B. and Chen K. K. (1942). The potency of stilbestrol in the immature female rat. Endocrinology, 30, 469—473. (Цитируется по: Bliss C. I. and Calhoun D. W (1954). An Outline of Biometry. Yale Cooperative Corporation, New Haven, p. 103.) Данные воспроизводятся с разрешения журнала и автора.

10.1.14. Сделайте то же самое, что и выше, но для данных за 1939 и 1940 гг. и навески 0,28 грана на крысу.

10.1.15. И еще раз сделайте то же самое, но теперь с данными о модуле упругости деревьев с 5 и 7 делянок из илл. 1 к упр. 10.1.15.

Иллюстрация 1 к упражнению 10.1.15 (используется еще в упр. 10.4.5, 10.4.6, 10.5.5, 10.5.10)

Модуль упругости деревьев красной сосны с коннектикутской плантации

Креймер измерял модуль упругости деревьев красной сосны с наружных участков нескольких делянок 25—30-летних насаждений в Коннектикуте. Вот некоторые из его результатов (выраженные в десятых долях модулей):

Делянка 1	Делянка 5	Делянка 7
136	68	95
138	114	117
140	120	118
143	132	124
146	147	126
149	150	132
150	159	133
157	163	146
159	164	150
	178	152
	197	161

Источник. Kraemer J. H. Dissertation, Yale University, 1943. (Цитируется по: Bliss C. I. and Calhoun D. W. (1954). An Outline of Biometry, Yale Cooperative Corporation, New Haven, p. 59.)

10.1.16^{к}. Мартин измерял длительность жизни мышей после заражения их туберкулезными бациллами. Блисс и Колхаун скомбинировали эксперименты Мартина в две группы А и Б, показанные на илл. 1 к упр. 10.1.16. Воспринимая 14+ как 14,5 дня и т. д., найдите все, какие сможете средние, медианы, бивесредние, срединные средние и усеченные средние во всех группах. Сравните их все между собой. Обсудите то, что Вы наблюдаете.

Иллюстрация 1 к упражнению 10.1.16

Выживание мышей (в днях после прививки)

Группа	Выживание (дни)								
	14^{+}	15^{+}	16^{+}	17^{+}	18^{+}	19^{+}	20^{+}	21^{+}	22^{+}
А	7	6	12	23	23	43	37	23	17
Б	1	—	—	3	2	7	12	14	16
Группа	23^{+}	24^{+}	25^{+}	26^{+}	27^{+}	28^{+}	29^{+}	30^{+}	$\geqslant 31$
А	6	7	8	5	1	2	1	—	3
Б	16	24	12	4	0	0	4	4	19

Источник. Martin A. R. (1946). The use of mice in the examination of drugs for chemotherapeutic activity against mycobacterium tuberculosis.—J. Pathol. Bacteriol., 58, 500—585. (Цитируется по: Bliss C. I. and Calhoun D. W. An Outline of Biometry. Yale Cooperative Corporation, New Haven, p. 62.)

10.1.17. Могли бы Вы найти еще какие-нибудь дополнительные меры положения, подходящие для упр. 10.1.16, если бы Вы работали с 1/(время выживания), а **не** с самим временем выживания?

10.1.18. Повторите вычисления в упр. 10.1.16 для 100/(дни выживания).

10.2.1. Дайте определение робастности к эффективности. (Вернитесь, если надо, к параграфу 1.5.)

10.2.2^{К}. Некая оценка имеет эффективность 92% относительно какой-то другой оценки. Пусть у обеих известны дисперсии. Что могут дать вычисления и будет ли объем выборки столь велик, что мы сможем пренебречь изменчивостью этих оценок и работать с ними так, как будто в них нет ошибок? Каково должно быть отношение оцениваемых дисперсий? А каково будет отношение длин соответствующих доверительных интервалов? Достаточно ли мы об этом заботимся? А стоит ли вообще беспокоиться? Почему? Почему нет?

10.2.3. (10.2.4, 10.2.5). Повторите упр. 10.2.2 для относительных эффективностей 50%; 80%; 10%.

10.3.1. Что бы Вы предпочли из среднего арифметического, медианы и бивес-среднего для оценки по малой выборке из приблизительно нормального распределения без растянутых «хвостов»?

10.3.2. Чтобы обеспечить счет для больших выборок, выберите одну из трех оценок по таблице из илл. 10.3.1 (с. 214).

10.3.3. Для обеспечения «круговой обороны» снова возьмите растянутые «хвосты» в больших выборках. Какую из оценок положения Вы предпочтете теперь?

10.3.4^{К}. При приеме на работу в городскую полицию поступающие должны были проходить между верхней и нижней границами роста. Проводится сравнение роста у уже работающих полисменов — регулировщиков уличного движения и детективов. Какую меру положения Вы бы рекомендовали здесь использовать?

10.3.5^{К}. Гольф-клуб ежегодно проводит турнир «до первой лунки», в котором участвует много специалистов и несколько новичков-оптимистов. Место остановки каждого мяча тщательно отмечается ежегодно. Желательно подытоживать ежегодный опыт (как по расстоянию, так и по отклонению вправо-влево), чтобы можно было приводить в соответствие междугодичную вариацию действительного центра удара и такую же вариацию положения лунки на лужайке. Какую меру положения Вы советуете использовать?

10.3.6. Как часть кампании в поддержку принятия муниципалитетом среднего по размерам города постановления о борьбе с шумом после 11 часов вечера проводились измерения средней силы шума за 10-минутный отрезок времени в 27 точках, разбросанных по всей территории города. Наблюдения велись непрерывно с 11 часов вечера до 5 часов утра 53 дня подряд. Как Вы предложили бы обобщить 53 измерения для какого-нибудь одного 10-минутного интервала в заданном месте?

10.3.7. (Продолжение упр. 10.3.6.) Изменился бы Ваш ответ в упр. 10.3.6, если бы (1) Вы были консультантом комиссии по борьбе с шумом, (2) Вы консультировали бы тех, кто борется против контроля над шумом, (3) Вы знали, что в некоторых местах маршрута было сильно влияние машин скорой помощи и полицейских автомобилей? Как? Почему?

10.4.1. Воспользуйтесь медианой как устойчивой мерой положения для робастной оценки разброса, МАО, и заполните следующую табличку для данных о витамине *C* из илл. 1 к упр. 10.1.1. Обозначение s_{bi}^2 здесь указывает на оценку разброса, модифицированную Лаксом, как показано в параграфе 10.4.

	Контроль	*Эксперимент*
Бивес		
Медиана		
МАО		
s_{bi}^2		

10.4.2. Заполните табличку, аналогичную приведенной выше, для испытуемых 1 и 2 в Лозанне (данные о коленном рефлексе см. в упр. 10.1.9).

10.4.3. Сделайте то же для 0,20 и 0,28 грана на крысу в декабре 1939 г. (данные о весе маток см. в упр. 10.1.13).

10.4.4. Опять то же задание, но для 0,28 грана на крысу в декабре 1939 г. и в апреле 1940 г. (упр. 10.1.13).

10.4.5. Снова то же, но на этот раз для модуля упругости древесины красной сосны с 5-й и 7-й делянок (упр. 10.1.15).

10.4.6. И еще раз аналогичное задание для 1-й и 5-й делянок (упр. 10.1.15).

10.4.7^{к}. Отыщите все интерквартильные размахи (~ разность между квартилями), s^2 и s^2_{bi} для двух групп из упр. 10.1.10, какие Вы сможете.

10.4.8. Могли бы Вы сосчитать еще меры разброса, подходящие для упр. 10.4.7, если бы Вы работали с 1/(время выживания) вместо самого времени выживания?

10.4.9. Проделайте вычисления из упр. 10.4.7 для — 100/(дни выживания).

10.5.1. (Продолжение упр. 10.4.1.) Воспользуйтесь результатами упр. 10.4.1, чтобы сравнить (устойчиво) значения бивес-средних для данных о витамине *C* из илл. 1 к упр. 10.1.1. Постройте 95%-ный доверительный интервал для разности.

10.5.2, 10.5.3, 10.5.4, 10.5.5. (Продолжение 10.1.2, 10.4.3, 10.4.4, 10.4.5). То же, что в 10.5.1, но для 10.4.2 (10.4.3, 10.4.4, 10.4.5.)

10.5.6. (Продолжение 10.5.1.) Проведите сравнение средних (неустойчивое) для данных о витамине *C* с помощью *t*-критерия Стьюдента для построения 95%-ного доверительного интервала. Сравните с результатами упр 10.5.1 и обсудите результаты.

10.5.7/8/9/10. Продолжение упр. 10.5.2 (10.5.3, 10.5.4, 10.5.5), как в упр. 10.5.6.

10.5.11. Для тех же ингредиентов, что и в упр. 10.1.3, используйте *t*-критерий Стьюдента, основанный (а) на $\bar{y}$ и s^2 и (б) на бивесе $л\, s^2$ при построении 95%-ных доверительных границ к его среднему значению. Рассмотрите различие в результатах.

10.5.12 (10.5.13, 10.5.14). Повторите упр. 10.5.11 для выделенного Вам ингредиента из данных упр. 10.1.4 (10.1.5, 10.1.6).

10.5.15. (10.5.16). Повторите упр. 10.5.11 для заданной Вам температуры с данными упр. 10.1.7 (10 1.8).

Об упражнениях параграфа 10.7. Упражнения к этому параграфу смотри в упр. 14.С.1—14.С.4.

10.В.1. *Государственная оценка развития образования* (контрольная работа). Иногда эмпирические исследования могут подсказать нам, какие оценки положения предпочтительны в данной ситуации. Это особенно характерно для больших массивов данных, анализ которых периодически повторяется. Одним из примеров такого исследования для средних, медиан и бивес-средних служат данные государственной оценки развития образования, для которых используются данные из нескольких различных совокупностей, собираемые эмпирически. Исследователи взяли из одной совокупности 70 наблюдений и сосчитали для этой выборки три меры положения. Для каждой популяции они проделали эти операции 400 раз. На илл. 1 к упр. 10.В.1 мы построили «опоры и консоли» для бивес-средних, медиан и обычных средних. Совокупность составляли здесь *изменения в процентах правильных* ответов на 70 вопросов с многовариантными ответами, предъявляемых 13-летним школьникам по всей стране (США) в качестве экзамена между первым и четвертым годами обучения соответствующему предмету. Используйте графики «опор и консолей» из илл. 1 для описания распределений этих трех мер положения и для решения о том, какая из мер предпочтительнее в данном частном случае.

Иллюстрация 1 к упражнению 10.B.1

Графики «опора и консоль»

Источник. Larson R. and Searls D. National Assessment of Educational Progress (частное сообщение).

A. Для средних

15|4
14|
13|4
12|
11|369
10|002456
9|122248
8|0125677889
7|01124447
6|133567777
5|001122344445678
4|000112444556667899
3|011123444455666778999
2|000001111234455555677778888899
1|001222333334667777788999
0|012222333344455566666777777888899
—0|888388777766554444333322222211
—1|999987776655544433332222211111000
—2|99999988777665554444443222221110000
—3|9999987666665554444333332211100
—4|88777665544333221100
—5|998887766655433321100
—6|88764332100000
—7|954321110
—8|85221000
—9|9863
—10|9641
—11|72
—12|
—13|
—14|
—15|1

Б. Для медиан

21|6
20|379
19|2278
18|
17|
16|4
15|
14|
13|000
12|9999
11|11111111666666677778
10|5555
9|000
8|333333338999
7|
6|00000000000000000000000001111111111111111111111111
5|
4|77777777777777777777
3|00111111111114444444444444
2|
1|777778888888
0|000000000000000000111111111111117
—0|666111111111111
—1|
—2|44444444444444444333
—3|977777666
—4|666666666666666554
—5|
—6|97777222222000000000000000
—7|77777777755555555553333333333333333
—8|888888888888555433333333333111111
—9|4443333222221111000000000000
—10|9831
—11|911111
—12|83
—13|0
—14|954
—15|000
—16|1
—17|
—18|31

И. Замечания к частям А — З.

Столбцы преобразованных данных так же, как и на илл. 9.1.1А, соответствуют месяцам от января до июля, а строки — городам: Карибу (верхний), Вашингтон, Ларедо (нижний).

К. Процентили для остатков (для трех значений d)

	$d=70$		(разброс)
M11	0		
H6	0,1	−0,1	0,2
E3 $\frac{1}{2}$	1,4	−1,2	2,6
2	2,9	−3,2	6,1
1	3,6	−5,6	9,2

	$d=60$		(разброс)
M11	0		
H6	0,1	0	0,1
E3 $\frac{1}{2}$	1,0	−1,3	2,3
2	1,5	−3,7	5,2
1	3,1	−5,3	8,4

	$d=40$		(разброс)
M11	0		
H6	1,3	0	1,3
E3 $\frac{1}{2}$	2,7	−2,0	4,7
2	2,9	−3,3	6,2
1	4,5	−7,5	12,0

Л. Квадратичное преобразование

28,2	28,4	31.8	38,4	48,9	58,5	64,7	
39,4	40,0	46,2	54,4	65,6	76,5	81,8	
57,7	62,3	70.0	79,9	87,4	94,4	97,4	

4,8	4,4	0,1	0	−0,2	0	0	−16,0
0	0	−1,5	0	0,5	2,0	1,1	0
−4,0	0	0	3,2	0	−2,4	−5,6	22,3
−15,0	−14,4	−6,7	0	10,7	20,1	26,3	54,4

М. Замечание к части Л

Квадратичное преобразование проведено по общей формуле параграфа 5.5: $\frac{A}{p}\left(\frac{y}{A}\right)^p+\left(1-\frac{1}{p}\right)A$, которое для точки $A=55$ и $p=2$ дает $\frac{1}{110}y^2+\frac{55}{2}$.

В. Для бивес-средних

```
 16| 2
 15| 9
 14|
 13|
 12| 12
 11| 013499
 10| 002444
  9| 011367
  8| 0223455568
  7| 001123466
  6| 023333444449
  5| 00112223334455567778
  4| 00011223456678999
  3| 012233333556677889
  2| 12222334467889
  1| 000011233444455555666677888899999
  0| 00001233334444555556667888888999
 —0| 99988877777665554444322221111 1
 —1| 99977777666666655554333222222111111 10000
 —2| 98877777766555533332222
 —3| 99998888887666555554433222100
 —4| 99988777666554433331111 0
 —5| 9988865444433221 1000
 —6| 999987665553333111 1000
 —7| 876555310
 —8| 87520
 —9| 86200
—10| 520
—11| 54
—12| 80
—13|
—14| 4
```

ГЛАВА 11.

11—1. Когда имеет смысл при сравнениях пользоваться нормализацией?

11—2. Как обеспечить контроль для рандомизированных опытов?

11.1.1. В чем различие между «приближенными» и «скорректированными» долями?

11.1.2. Что конкретно (в цифрах) представляет собой нормированная совокупность для смеси 50 на 50 «легких» и «трудных» больных из примера 1 в параграфе 11.1 (с. 218)?

11.1.3. Продемонстрируйте вычисления для разницы в процентах (II—I) смеси 50 на 50 «легких» и «трудных» больных.

11.1.4. Что конкретно (в цифрах) представляет собой нормированная совокупность для 45% «легких» и 55% «трудных» больных из примера 1 в параграфе 11.1 (с. 218)?

11.1.5. Пусть вероятности успеха для способа I будут: p_{11} для группы «легких» и p_{21} — для «трудных». Какова будет приближенная доля успехов, когда n применяется способ I к совокупности, в которой смешаны t долей «легких» и $1 - t$ «трудных»?

11.1.6. Пусть вероятности успехов p_{ij} задаются следующей табличкой:

Группы	Способы I	Способы II
«Легкие»	p_{11}	p_{12}
«Трудные»	p_{21}	p_{22}

Причем $p_{11} > p_{12}$ и $p_{21} < p_{22}$. Какая смесь «легких» и «трудных» в совокупности приведет к тому, что мы предпочтем способ I, а не способ II?

11.1.7. Воспользуйтесь табл. 2 из приложения для упражнений (оно идет вслед за упражнениями ко всем главам), где приведены данные о связи курения и здоровья. Найдите приближенные доли смертей для некурящих и для всех трех групп курящих.

11.1.8. (Продолжение упр. 11.1.7.) Используйте совокупность некурящих как стандарт и найдите нормированные доли умерших: $R_{н.к}$ для некурящих, $R_{с.т}$ — для курящих сигары и трубку и $R_{с.}$ — для курящих только сигареты.

11.1.9. Сравните результаты упр. 11.1.7 и 11.1.8.

11.1.10. Из тех же данных выделите курильщиков сигарет и некурящих. Разбейте их возраста на две группы: от 0 до 59 и 60 и старше. Найдите разность — курящие минус некурящие — для смеси 50 на 50 из этих двух групп.

11.1.11. Что такое «прямое нормирование»?

11.1.12. Полагая, что между результатами для «легких» и «трудных» нет корреляции, найдите var $(p_{стд.I} - p_{стд.II})$, если нормированная совокупность такая же, как в вопросе 2 в тексте (с. 219).

11.2.1. Что такое W для прямого нормирования долей в вопросах 1—3 из параграфа 11.1 (с. 221)?

11.2.2. Объясните, почему старший возраст населения в штате Мэн имеет приближенную долю выше, тогда как удельные доли в каждой возрастной группе меньше, чем в Южной Каролине (см. илл. 11.2.1 (с. 242))?

11.2.3. В илл. 11.1.1 (с. 241) приближенная разность в долях успехов (способ I — способ II = 60% — 44%) иная, чем если бы для расчета взяли сравнение этих групп. В илл. 11.2.1 (с. 242) приближенные доли тоже вводят в заблуждение. Как преодолеть эту трудность в каждом из приведенных примеров?

11.2.4. Воспользуйтесь следующими числами в каждой возрастной группе как стандартом на миллион жителей США в 1975 г., чтобы скорректировать прямым методом доли в илл. 11.2.2 (с. 243) для штатов Мэн и Южная Каролина.

Возраст (годы)	Нормировано на миллион для США в 1975 г.	Возраст (годы)	Нормировано на миллион для США в 1975 г.
0—4	74 400	25—34	144 600
5—9	81 000	35—44	107 200
10—14	95 500	45—54	111 400
15—19	98 700	55—64	92 200
20—24	90 300	65—74	64 600
		75+	40 200

11.2.5. (Продолжение упр. 11.2.4.) Сравните то, что получилось в упр. 11.4 для 1975 г. с аналогичными данными из илл. 11.2.2 (с. 243) для 1940.

11.2.6. В части Б илл. 1 к упр. 11.2.6 приводится распределение слов различной длины (с разным числом букв) в текстах объемом от 15 до 18 тыс. слов из публикаций Гамильтона и Мэдисона. В части А содержатся частоты (приведенные к 1000 слов) для некоторых слов каждой длины. Объединяя их, мы видим, что у слов каждой длины есть свой шанс попасть в текст того или другого автора. В части В показано нормированное распределение слов по длинам из романа Мелвилла «Моби Дик». Нормируйте распределение длин слов для всех трех авторов и воспользуйтесь им, чтобы найти шанс появления любого конкретного слова в их текстах.

Иллюстрация 1 к упражнению 11.2.6

Длины слов и их частотные распределения

А. Частотные распределения слов, состоящих из 1—12 букв

Слово длиной k	Доля на 1000 слов		Слово длиной k	Доля на 1000 слов	
	Гамильтон	Мэдисон		Гамильтон	Мэдисон
k = 1 («a»)	22,85	20,22	7 («whether»)	0,49	0,97
2 («of»)	64,85	57,80	8 («language»)	0,04	0,21
3 («our»)	2,27	1,11	9 («direction»)	0,22	0,03
4 («what»)	1,38	1,15	10 («thoughout»)	0,04	0,17
5 («among»)	0,39	0,84	11 («destruction»)	0,13	0,1
6 («second»)	0,18	0,37	12 («consequently»)	0,03	0,48

Б. Распределение длин слов

Слово длиной k	Гамильтон	Мэдисон
1	423,2	396,1
2	3531,6	3834,7
3	2925,0	3644,7
4	2042,4	2204,9
5	1580,0	1720,2
6	1116,1	1396,4
7	1026,7	1298,7
8	824,5	1027,4
9	805,7	888,1
10	617,6	743,4
11	396,6	450,4
$\geqslant 12$	385,6	483,0
	15675,0	18088,0

В. Распределение длин слов

k	Нормировано на 1000 из «Моби Дик»
1	45,0
2	162,0
3	228,0
4	196,5
5	122,3
6	79,5
7	66,0
8	45,3
9	27,9
10	15,1
11	8,1
$\geqslant 12$	4,3

Источники: Mosteller F. and Wallace D. L. (1964). Inference and Disputed Authorship. The Federalist. Addison—Wesley, Reading, MA, p. 244—248, 258, 260. Воспроизведено с разрешения издателя и автора. Для «Моби Дик»: Nowlin A. G. (1973). Statistical analysis of linguistic word—frequency distributions and word—length sequences.—Ph. D. thesis, Princeton University. Напечатано с разрешения автора.

11.2.7. В илл. 1 к упр. 11.2.7 даны числа смертельных исходов в результате хирургических операций для двух регионов страны (США). Они распределены по возрасту и полу. Сравните в каждом из регионов прямым методом доли смертей для мужчин и женщин. В качестве стандарта возьмите совокупность с равным соотношением мужчин и женщин.

Иллюстрация 1 к упражнению 11.2.7

Смерть на операционном столе

Следующие данные собраны в двух регионах США за 5-летний период. В совокупность включались все, кто подвергся хирургическому вмешательству, а к умершим причислялись те пациенты, которые умерли во время операции или после нее, но не покидая больницы.

Возраст	Регион I			
	Совокупность		Умершие	
	мужчины	женщины	мужчины	женщины
0—4	2104	1952	34	22
5—14	4272	3911	9	11
15—24	2835	2989	23	5
25—34	2785	2606	19	8
35—44	1930	1886	16	15
45—54	1497	1524	59	40
55—64	960	1013	101	52
65—75	652	855	185	118
76—83	186	287	97	108
84+	69	125	68	103

Возраст	Регион II			
	Совокупность		Умершие	
	мужчины	женщины	мужчины	женщины
0—4	703	689	12	3
5—14	1739	1758	5	2
15—24	1233	1244	14	1
25—34	989	1004	8	3
35—44	897	922	9	13
45—54	921	961	28	15
55—64	686	739	68	37
65—75	611	784	159	73
76—83	189	290	86	88
84+	52	124	70	119

11.2.8. Сравните прямым методом доли умерших для I и II регионов из илл. 1 к упр. 11.2.7, используя в качестве эталона смесь 50 на 50 (мужчин и женщин в каждом регионе объедините).

11.3.1. С помощью нормированного биномиального распределения найдите дисперсию нормированной разности в долях слов из упр. 11.2.6.

11.3.2. Для улучшения оценки дисперсии в упр. 11.3.1 возьмите стратифицированное бином иальное распределение.

11.3.3. Воспользовавшись улучшенным стратифицированным биномиальным распределением, найдите стандартные ошибки разностей $R_{н.к} - R_{с.т}$, $R_{н.к} - R_{с}$ для данных о курении и здоровье из упр.11.1.8. Что бы Вы теперь, имея стандартные ошибки, могли сказать о других долях?

11.3.4. Что лучше подходит для вычисления стандартной ошибки — приближенная оценка, нормированное биномиальное или стратифицированное биномиальное распределения?

11.4.1. Объясните, апеллируя к илл. 11.4.1 (с. 243), почему важно ревизовать любые нормализационные вычисления. Каковы их скрытые трудности?

11.4.2. Машинная программа выдает нормированные доли. Как Вы хотели бы ее организовать, чтобы обеспечить предостережение в случае «диких» результатов?

11.4.3. Оцениваемая доля смертей среди женщин в возрасте от 5 до 24 лет из илл. 1 к упр. 11.2.7 близка к нулю. Какого сорта проблемы это порождает?

11.5.1. Чем косвенная нормализация отличается от прямой?

11.5.2. Почему в примере 5 из параграфа 11.5 рекомендуется сравнивать долю успехов не со всеми остальными долями, а непосредственно с приближенными долями успехов?

11.5.3. Наблюдается ли при косвенной нормализации иное поведение способов I и II из примера 2 в параграфе 11.1?

11.5.4. Если взвесить итоги вариантов в примере 6 из параграфа 11.5, то станут ли варианты I и II одинаковыми?

11.5.5. Когда можно не интересоваться нормированным числом $n_{стд}$ при грубых прикидках величины дисперсии?

11.5.6. Подсчитайте косвенным методом доли смертей для штатов Мэн и Южная Каролина по данным из илл. 11.2.1 (с. 242).

11.5.7. Сосчитайте по данным табл. 2 из приложения для упражнений косвенным методом нормированные доли больных раком легких для курильщиков и для некурящих.

11.5.8. Опишите две несуразности косвенного подхода.

11.6.1. Почему необходима коррекция?

11.6.2. Можно ли избавиться от трудностей нормирования путем коррекции сравнения групп «легких» и «трудных»?

11.6.3. Почему именно логистическое распределение используется как распределение трудности для нормирования?

11.6.4. Каково различие (способ II — способ I) для скорректированных групп для эталонной совокупности из 55% «легких» и 45% «трудных»?

11.6.5. Проводится ли корректировка совокупностей и для прямого, и для косвенного методов?

11.6.6. Выделите из данных табл. 2 приложения для упражнений курильщиков сигарет и некурящих. Найдите центры тяжести для двух возрастных групп: от 0 до 59 лет и от 60 и выше.

11.6.7. Найдите среднюю разность курящие минус некурящие для упр. 11.6.6.

11.6.8. Сравните результаты упр. 11.6.7 и 11.1.10.

11.6.9. Примите за эталон смесь 45% возрастов от 0 до 59 и 55% 60 лет и старше. Найдите скорректированную разность (курящие минус некурящие).

11.6.10. Что надо сделать для вычисления процентов успеха в точке деления?

11.В.1. *Нормированные индексы детородной способности.* (Контрольная работа.) В илл. 1 к упр. 11.В.1 приводится возрастное распределение женщин в период их способности к деторождению по долям рождений в каждой возрастной группе для трех стран Европы. Объединение этих данных дает распределения и доли, получаемые прямым и косвенным нормированием исходных долей. Демографы получили косвенный стандарт по плодовитости хаттеритов*, одной из религиозных сект, распространенных на Западе США и Канады и известных своей исключительно высокой плодовитостью. Сравните три страны по приведенным ниже данным. На первый взгляд кажется очевидным, что плодовитость выше всего во Франции и ниже всего в Великобритании. Говорят ли то же самое все три метода нормирования? Опираясь на каждый метод в отдельности и на все методы вместе, можете ли Вы сказать, что Норвегия лежит примерно посередине между Францией и Великобританией?

Почему демографы предпочли для сравнения всех совокупностей взять какой-то постоянный стандарт? Почему они выбрали именно хаттеритов, имеющих наивысшую из известных плодовитость?

Иллюстрация 1 к упражнению 11.В.1

Возрастное распределение детородной активности

Возраст	Франция (1968)		Норвегия (1973)		Великобритания (1973)		Хаттериты
	число женщин (тыс.)	доля	число женщин (тыс.)	доля	число женщин (тыс.)	доля	доля
15—19	2070,0	0,0255	150,8	0,0443	1696,0	0,0432	0,300
20—24	1851,0	0,1580	145,7	0,1500	1707,0	0,1309	0,550
25—29	1382,0	0,1636	150,9	0,1378	1799,0	0,1354	0,502
30—34	1514,0	0,0985	111,8	0,0732	1443,0	0,0635	0,447
35—39	1646,0	0,0479	96,5	0,0309	1387,0	0,0246	0,406
40—44	1657,0	0,0146	101,2	0,0073	1424,0	0,0061	0,222
45—49	1561,0	0,0013	113,0	0,0004	1497,0	0,0004	0,061

Источник. U. N. Demographic Yarbooks, 1972 and 1974; данные о хаттеритах: Coale A. Y. Factors associated with the development of low fertility. В: U. N. World Population Conference, 1965, Vol. II, U. N.: New York, 1967.

*Хаттериты — религиозная секта, возникшая в Южной Германии в 1528 г. В 1873—1875 гг. хаттериты эмигрировали в Северную Америку и поселились главным образом в штате Южная Дакота и западных провинциях Канады, образовав сельскохозяйственные коммуны. Для этой секты характерны пацифизм, строгий протестантизм и крайний консерватизм. — *Примеч. ред.*

● ПРИЛОЖЕНИЕ ДЛЯ УПРАЖНЕНИЙ

Перечень упражнений, в которых используются приводимые ниже данные

Таблица 1. Города средней Америки: 3.4.6—3.6.4., 3.7.1., 3.7.2, 14.3.1, 14.3.2.

Таблица 2. Курение и здоровье: 11.1.7—11.1.10, 11.3.3, 11.5.7, 11.6.6—11.6.9.

Таблица 3. Население США в 1960—1970 гг. по штатам: 3.1.2—3.1.7, 3.2.4, 3.2.6—3.2.8, 3.3.1. (Упр. 14.3.1; и 14.3.2 помещены во втором выпуске этой книги. Там же приведено и продолжение приложения для упражнений. — *Ю. А.*)

Таблица 1. Города средней Америки

В этой таблице собраны различные сведения о 152 городах с населением от 325 000 (1960 г.) в трех «средних» переписных округах США: средний северо-запад, средний юго-запад и горный округ. Вот рассматриваемые переменные:
— порядковый номер в списке (список составлен в порядке латинского алфавита);

301 — площадь (кв. мили, округленно);

302 — ранг в списке городов США (1-й наивысший);

303 — население в 1960 г. (сотни);

306 — процент небелого населения в 1960 г.;

313 $\frac{1}{2}$ — процент населения либо родившегося за пределами США, либо имеющего хотя бы одного из родителей, родившихся за пределами США (сумма столбцов 313 и 314);

320 — медианный семейный доход в 1959 г.;

325 — процент лиц (в возрасте 25 лет и старше), окончивших меньше чем 5 классов школы;

327 — процент лиц (в возрасте 25 лет и старше), закончивших колледж;

331 — процент лиц (в возрасте 5 лет и старше), живущих в 1960 г. в тех же домах, что и в 1955 г.;

352 — процент лиц (в возрасте 5 лет и старше), имеющих отдельные квартиры (включая и владельцев нескольких домов);

357 — процент проживающих в квартирах с менее чем одной комнатой на человека;

358 — процент жилых квартир, поменявших квартиросъемщика в период с 1958 по 1960 г.

Все эти данные есть небольшое извлечение из 160 столбцов дополнительной информации, приведенной в справочнике: 1962. County and City Data Book.

Штат и город	#	301	302	303	306	313 1/2	320	325	327	331	352	357	358
Аризона													
Меса	1	15	488	338	2,8	14,4	5598	5,5	8,7	33,7	86,4	16,9	47,3
Финикс	2	187	219	4392	5,8	16,6	6117	6,0	8,9	34,7	87,2	12,9	46,5
Тусон	3	71	54	2129	4,4	22,3	5703	5,7	11,1	33,9	89,3	14,2	48,1
Колорадо													
Аврора	4	9	317	485	1,2	12,6	6627	0,6	13,5	26,4	78,7	13,4	58,7
Боулдер	5	7	429	377	1,2	13,0	6726	1,0	29,9	26,4	70,5	7,0	48,6
Колорадо-Спрингс	6	17	204	702	5,0	13,7	5669	2,4	12,2	33,7	70,3	7,8	45,3
Денвер	7	71	23	4939	7,1	18,7	6361	4,3	12,2	42,0	65,6	8,0	39,3
Энглвуд	8	6	498	331	0,8	12,4	6744	1,8	9,9	43,5	84,5	10,2	38,5
Форт-Коллинс	9	6	675	250	1,0	14,3	5409	4,2	18,4	28,6	78,5	7,1	48,8
Грили	10	5	647	283	0,9	17,3	5351	5,2	14,2	35,4	61,7	7,5	45,2
Пуэбло	11	17	147	912	2,6	17,9	5698	8,5	6,5	49,3	81,5	15,0	33,2
Айдахо													
Бойсе	12	10	473	345	1,1	14,7	5851	2,3	11,2	43,4	71,6	6,9	42,0
Айдахо-Фолс	13	8	506	332	0,9	12,3	6844	2,1	14,7	36,5	79,1	16,4	44,9
Покателло	14	8	595	285	2,6	13,7	6023	2,8	9,3	43,7	68,9	15,2	40,5
Айова													
Эймс	15	8	635	270	1,1	11,6	6191	0,7	32,9	29,5	79,6	8,0	45,6
Берлингтон	16	12	519	324	1,6	11,7	5848	2,5	6,5	50,6	77,0	7,3	30,7
Сидар-Рапидс	17	33	145	920	1,4	14,5	6687	2,9	10,4	46,9	76,0	8,6	33,3
Клинтон	18	11	493	336	1,1	18,7	6146	3,3	6,5	54,6	76,8	8,4	26,7
Каунсил-Блафс	19	16	276	556	1,1	12,6	5967	3,0	4,7	46,2	81,7	13,4	33,6
Давенпорт	20	47	155	890	2,1	16,4	6479	2,3	7,8	47,0	68,4	10,5	33,3
Де-Мойн	21	64	55	200	5,1	12,5	6436	3,1	9,7	47,3	74,9	6,8	34,8
Дубьюк	22	14	271	566	0,3	14,0	6973	2,2	8,5	53,1	67,4	11,6	28,2
Форт-Додж	23	7	598	284	1,1	16,8	6059	2,6	8,1	47,0	69,8	7,7	38,6
Айова-Сити	24	8	495	334	1,5	12,1	5769	1,6	29,4	30,6	60,7	8,5	44,8
Мейсон-Сити	25	13	550	306	0,9	19,3	5979	3,3	7,8	46,2	86,2	8,7	30,9
Оттамуа	26	12	485	339	1,3	0,7	5647	2,9	5,8	52,7	85,2	9,7	28,8
Сок-Сити	27	49	153	891	2,1	17,5	5812	3,8	7,8	49,8	75,0	9,8	31,7
Уотерлу	28	34	200	718	6,9	12,2	6526	2,9	6,5	46,6	81,4	10,3	32,0
Канзас													
Атчинсон	29	11	432	376	3,3	8,2	5469	3,4	7,5	41,3	80,4	8,7	39,1
Канзас-Сити	30	41	98	1219	23,2	11,6	5583	8,0	4,9	53,7	77,1	12,8	30,0

Лоренс	31	8	510	329	9,1	7,4	5427	2,6	23,2	30,5	77,7	6,9	45,4
Прери-Виллидж	32	5	668	254	0,2	11,1	10225	0,5	30,9	43,0	100,0	2,4	31,4
Салайна	33	8	367	432	3,7	10,8	5475	2,1	10,2	34,7	83,6	10,9	48,1
Топика	34	36	100	1194	8,2	9,0	6039	3,1	11,5	42,6	77,6	9,5	37,8
Уичито	35	52	51	2547	8,3	6,4	6121	3,0	10,4	41,2	78,8	10,4	39,8
Луизиана													
Александрия	36	10	394	403	43,4	4,4	3768	19,3	7,8	51,2	84,6	15,7	33,8
Батон-Руж	37	31	80	1524	29,9	4,3	5789	11,8	13,0	51,1	86,1	15,7	33,9
Бессир-Сити	38	11	512	328	11,2	5,4	5043	7,1	7,1	23,6	80,7	16,5	62,7
Лафейетт	39	7	392	404	28,3	2,9	4361	27,7	11,5	50,8	87,5	18,0	35,6
Лейк-Чарльз	40	16	236	634	21,7	4,3	5462	13,4	10,0	42,5	85,0	15,1	42,8
Монро	41	18	295	522	43,8	3,0	3958	18,8	8,6	46,4	83,3	19,2	35,9
Нью-Айбирия	42	7	585	291	23,4	2,4	4663	28,1	6,4	50,4	90,9	22,2	33,7
Новый Орлеан	43	199	15	6275	37,4	8,6	4807	13,7	7,7	50,3	49,5	18,2	32,4
Шривпорт	44	36	76	1644	34,5	3,5	5205	13,2	10,8	45,5	83,2	14,7	37,6
Миннесота													
Остин	45	5	611	279	0,1	18,9	7618	2,1	7,6	51,1	81,1	9,8	28,6
Блумингтон	46	35	308	505	0,3	15,7	7201	0,7	12,8	33,2	98,6	16,1	35,4
Дулут	47	63	122	1069	1,1	36,8	5877	4,7	8,5	53,3	62,3	7,2	29,6
Уодина	48	16	596	285	0,2	20,6	12082	0,4	27,9	39,9	99,0	2,7	31,2
Миннеаполис	49	56	25	4829	3,2	31,2	6401	3,5	9,6	49,5	48,7	6,3	33,6
Миннетонка	50	28	674	250	0,4	19,2	8180	0,9	18,3	39,9	96,7	8,5	36,1
Ричфилд	51	10	372	425	0,3	17,9	7721	0,7	13,6	51,5	94,5	12,6	27,8
Рочестер	52	8	391	407	0,5	19,6	6638	1,9	15,0	41,4	66,7	9,3	40,5
Сент-Клауд	53	9	486	338	0,7	18,7	5592	4,0	8,6	50,0	79,9	15,2	30,9
Сент-Луис-Марк	54	10	365	433	0,5	26,9	7808	1,1	16,2	52,5	88,7	7,3	27,6
Сент-Пол	55	52	40	3134	3,0	27,9	6543	3,7	9,2	53,5	57,4	9,0	30,0
Миссури													
Колумбия	56	11	455	366	7,9	5,2	5616	3,8	25,7	27,5	73,7	9,1	47,0
Флориссент	57	4	419	382	0,1	8,2	7740	1,0	12,9	23,7	98,3	13,3	42,9
Индепенденс	58	14	244	623	1,2	6,8	6535	2,6	6,1	46,0	84,2	8,9	36,2
Джефферсон-Сити	59	10	604	282	10,7	5,8	5876	6,0	9,1	41,0	64,7	7,8	36,5
Джоплин	60	20	409	390	2,2	4,2	4915	5,4	6,9	46,1	85,7	7,9	33,8
Канзас-Сити	61	130	27	4755	17,7	10,9	5906	5,4	7,9	43,2	58,2	8,9	37,6
Керквуд	62	8	574	294	3,2	10,8	8753	1,9	22,8	52,8	94,8	6,1	25,8
Сент-Джозеф	63	28	181	797	3,3	9,6	5522	6,1	5,0	50,7	72,9	9,8	33,0
Сент-Луис	64	61	10	7500	28,8	14,1	5355	9,1	4,5	45,0	35,2	16,4	35,6
Спрингфилд	65	35	137	959	2,6	4,2	4955	5,0	7,8	40,4	82,4	8,9	38,4

Продолжение

Штат и город	#	301	302	303	306	313 1/2	320	325	327	331	352	357	358
Университи-Сити	66	4	302	512	0,4	39,2	8105	6,4	16,3	54,8	63,9	3,6	24,6
Уэбстер-Гловс	67	6	587	290	3,8	12,5	8750	1,9	22,4	62,7	95,4	4,8	19,0
Монтана													
Биллингс	68	9	292	529	1,2	21,6	6638	3,4	13,1	38,3	71,4	9,3	42,7
Бьютт	69	5	615	279	0,9	33,9	5156	4,9	7,0	58,0	65,4	11,4	29,1
Грейт-Фолс	70	11	278	552	1,6	23,5	5257	3,3	8,9	38,5	69,2	13,8	45,4
Мизула	71	6	633	271	0,8	21,9	5847	3,8	13,9	39,1	72,3	11,2	41,4
Небраска													
Гранд-Айленд	72	5	658	257	0,4	17,4	5109	3,2	6,4	46,6	81,1	9,6	33,1
Линкольн	73	25	95	1285	1,9	16,1	5032	1,7	14,6	39,9	68,9	7,3	39,7
Омаха	74	51	42	3016	8,1	21,0	6315	4,1	9,0	47,1	71,3	10,7	34,7
Невада													
Лас-Вегас	75	25	231	644	15,8	17,8	7662	3,8	8,1	27,9	68,8	12,3	53,8
Рино	76	12	301	515	3,2	22,5	7433	2,9	11,6	34,2	70,0	8,5	48,2
Нью Мексико													
Альбукерке	77	56	60	2012	2,9	9,5	6621	4,7	14,7	34,5	86,9	14,0	48,9
Карлсбад	78	7	663	255	2,5	9,2	6293	9,0	8,1	42,0	95,7	12,1	40,3
Хобс	79	11	650	263	8,0	2,9	6229	5,6	8,5	29,9	95,5	19,1	53,5
Лас-Крусес	80	9	577	294	2,9	16,0	5789	11,3	14,5	32,2	85,7	22,2	52,3
Розуэлл	81	13	403	396	4,6	7,1	5543	7,8	10,4	31,5	89,8	18,2	53,5
Санта-Фе	82	26	499	334	1,7	6,3	5502	10,8	13,6	52,3	87,7	20,5	37,2
Северная Дакота													
Бисмарк	83	9	618	277	0,4	29,0	6094	5,8	11,8	36,5	55,3	16,7	45,5
Фарго	84	9	331	467	0,5	26,5	6522	2,6	13,0	41,4	61,6	12,1	39,8
Гранд-Форкс	85	6	475	345	1,0	24,5	5849	3,5	11,9	34,3	61,3	16,7	45,4
Майнот	86	7	551	306	0,8	25,7	5953	3,4	8,9	34,3	68,0	16,6	46,2
Оклахома													
Бартлсвилл	87	6	612	279	4,5	4,3	6606	3,4	18,9	40,4	87,9	7,3	39,0
Инид	88	10	413	389	4,8	7,1	4956	5,0	8,3	43,2	84,9	8,5	37,3
Лотон	89	12	247	617	11,7	8,0	4633	4,7	8,2	29,1	85,3	18,1	58,9
Мидуэст-Сити	90	24	454	360	2,5	4,8	5843	2,3	7,6	34,6	96,1	13,6	46,1
Маскоги	91	13	421	381	22,5	2,9	4449	10,0	9,1	46,5	82,9	8,9	33,5
Норман	92	10	497	334	1,3	4,5	5259	6,4	22,6	29,6	85,0	6,6	50,6
Оклахома-Сити	93	322	37	3243	13,0	4,5	5600	5,2	9,8	42,5	82,6	10,6	39,3

Талса	94	48	50	2617	10,0	4,6	6229	3,9	12,0	42,2	83,2	7,9	37,5
Орегон													
Юджин	95	14	305	510	0,5	14,7	6267	1,7	18,4	33,5	73,7	6,0	47,0
Портленд	96	67	32	3727	3,5	25,8	6335	3,8	3,5	50,5	70,1	4,3	32,4
Сейлем	97	13	315	491	0,5	17,9	5859	6,0	10,0	42,1	81,2	4,1	40,3
Южная Дакота													
Рапид-Сити	98	16	374	424	4,1	13,3	5694	1,2	9,9	29,5	74,6	16,5	51,9
Су-Фолс	99	17	229	655	0,9	18,7	6081	2,0	8,6	43,8	75,9	10,9	35,8
Техас													
Абилин	100	62	149	904	5,1	4,9	5460	7,2	10,9	28,8	86,5	13,8	51,4
Амарилло	101	55	188	1380	5,8	4,4	5877	3,6	9,8	33,3	84,9	12,7	48,4
Арлингтон	102	24	351	448	0,9	3,8	6574	2,4	12,1	31,1	94,9	10,1	45,3
Остин	103	49	67	1865	13,3	10,6	5119	11,7	15,1	39,2	84,3	12,7	42,7
Бейтаун	104	18	606	282	6,7	7,5	6937	7,9	9,9	50,3	92,4	9,0	32,2
Бомонт	105	71	102	1192	29,4	4,8	5577	12,4	8,8	47,6	84,6	13,5	35,5
Биг-Спринг	106	10	542	312	5,2	6,8	5682	8,1	7,4	32,0	89,2	16,4	53,1
Браунсвилл	107	16	321	480	0,1	43,9	3021	36,9	5,9	51,8	90,3	37,2	35,6
Брайан	108	17	622	275	23,4	10,2	4258	15,7	13,1	43,2	92,9	15,7	41,1
Корпус-Кристи	109	38	74	1677	5,6	16,8	5221	17,6	9,4	42,1	90,1	20,7	41,0
Даллас	110	280	14	6797	19,3	6,9	5976	7,4	10,4	39,8	75,6	11,5	41,1
Дентон	111	10	638	268	7,6	3,3	4994	6,2	19,4	30,8	83,0	10,5	44,0
Эль-Пасо	112	115	46	2767	2,7	39,7	5211	16,4	9,1	32,8	73,6	24,4	50,5
Форт-Уэрт	113	140	34	3563	16,0	5,8	5484	4,8	10,0	42,2	82,3	11,7	38,8
Галвестон	114	20	217	672	27,5	18,1	4698	14,3	6,7	47,2	68,8	13,3	35,9
Гарленд	115	19	416	385	3,9	2,3	6792	3,1	9,5	32,0	97,9	10,6	46,3
Гранд-Прери	116	15	555	304	7,0	3,0	5764	4,5	8,8	39,4	91,1	14,2	42,3
Харлинген	117	31	387	412	1,7	35,4	4167	26,8	9,2	36,1	83,5	28,6	47,9
Хьюстон	118	328	7	9382	23,2	9,7	5902	8,9	10,7	41,9	80,1	12,7	39,3
Ирвинг	119	21	339	460	0,1	3,7	6843	3,2	9,5	34,6	96,0	12,3	45,1
Кинговилл	120	5	669	252	4,0	18,2	4366	22,8	11,4	42,7	85,1	23,2	45,5
Ларедо	121	14	252	607	0,4	53,0	2935	38,4	5,0	56,0	90,6	40,1	34,3
Лондвью	122	19	400	400	23,0	1,7	5355	8,1	9,5	42,1	93,0	11,4	40,7
Лаббок	123	75	94	1287	8,1	5,1	5582	8,5	11,5	28,8	86,0	17,5	52,8
Мак-Аллен	124	10	514	327	0,4	44,2	3790	30,0	9,3	45,5	87,3	30,6	40,6
Месквайт	125	21	623	275	0,5	2,9	6241	3,5	6,6	13,5	99,1	13,0	60,8
Мидленд	126	23	243	626	10,0	5,2	7094	5,7	19,5	25,3	87,7	13,9	59,9
Одесса	127	16	179	803	5,8	4,2	6210	6,6	8,1	25,6	91,4	19,8	56,6
Орандж	128	8	660	256	23,0	3,6	5510	10,8	9,3	38,7	86,5	16,4	42,0

Продолжение

Штат и город	#	301	302	303	306	$313\frac{1}{2}$	320	325	327	331	352	357	358
Пасадина	129	22	262	587	0,1	5,3	7176	3,5	9,5	39,9	94,9	10,8	39,3
Порт-Артур	130	46	224	667	30,8	7,9	5659	16,9	5,0	50,1	83,3	15,0	32,8
Сан-Анджело	131	30	260	588	5,4	9,3	4650	11,6	7,8	43,5	89,9	13,9	38,2
Сан-Антонио	132	160	17	5877	7,4	24,0	4691	19,5	6,8	47,3	83,7	21,5	36,0
Темпл	133	19	554	304	18,9	8,5	4509	10,7	6,5	44,7	88,8	12,2	36,0
Тексаркана	134	16	557	302	25,5	2,5	4353	10,6	6,9	48,8	90,1	10,8	31,7
Тексас-Сити	135	45	526	321	19,6	6,1	6101	8,2	8,3	41,6	90,7	14,2	37,1
Тайлер	136	18	303	512	22,3	2,8	5478	5,6	12,4	41,6	90,6	9,4	38,7
Виктория	137	12	508	330	8,3	10,4	5279	16,9	9,3	42,7	89,2	16,2	39,9
Уэйко	138	37	134	978	18,5	7,1	4859	10,2	10,0	42,4	86,5	10,8	38,5
Уигито-Фолс	139	37	127	1017	8,4	6,0	5451	7,0	9,4	33,3	82,0	14,0	48,0
Юта													
Огден	140	19	203	702	3,5	17,4	6145	3,1	8,5	48,9	70,8	13,8	34,9
Прово	141	19	455	360	0,8	14,1	5310	2,8	14,9	37,7	68,1	14,8	44,8
Солт-Лейк-Сити	142	56	65	1895	2,1	23,9	6135	3,2	12,4	48,9	62,4	10,3	35,4
Вашингтон													
Беллингхем	143	22	469	347	1,2	31,3	5735	3,3	8,5	51,1	82,6	5,2	30,3
Бремертон	144	5	589	289	4,3	21,9	6046	8,5	6,4	40,0	77,6	6,6	39,3
Эверетт	145	12	393	403	1,1	29,2	5843	3,1	7,1	46,3	72,8	5,5	34,8
Сиэтл	146	88	19	5571	8,4	31,4	6942	3,4	12,4	46,7	63,3	4,6	36,6
Спокан	147	43	68	1816	2,5	22,4	6044	3,4	9,0	49,2	71,9	6,7	34,0
Такома	148	48	83	1480	5,3	27,6	5993	3,8	6,8	50,4	77,0	5,8	32,6
Ванкувер	149	10	518	325	1,5	18,5	6535	3,1	7,5	45,3	80,7	6,4	34,8
Якима	150	9	366	433	2,8	20,0	5900	4,3	9,0	44,4	79,9	6,8	35,6
Вайоминг													
Каспер	151	7	410	389	1,6	12,6	1157	2,0	13,4	35,5	75,2	10,0	44,7
Шайенн	152	10	364	435	2,5	15,3	6575	3,4	9,8	32,6	62,7	14,3	51,3

Таблица 2. Курение и здоровье

Это исследование проводилось управлением пенсионного обеспечения ветеранов в форме анкетного опроса с учетом дополнительных сведений о смерти пенсионеров, принимавших участие в опросе. Наблюдение велось шесть лет — с 1 июля 1956 г. по 30 июня 1962 г. (Те, кто выкурил за всю свою жизнь не менее чем 100 сигарет, или 10 сигар, или 20 трубок, классифицировались как курильщики) Исследовалась группа ветеранов первой и второй мировых войн, а также войны в Корее, вместе с их взрослыми родственниками-мужчинами (сыновья и отцы).
Возрастные группы соответствуют 1956 г.

Возраст	Некурящие		Курящие только трубки или сигары		Курящие сигареты и что придется		Курящие только сигареты	
	человек	умерло	человек	умерло	человек	умерло	человек	умерло
40—44	656	18	145	2	4531	149	3410	124
45—49	359	22	104	4	3030	169	2239	140
50—54	249	19	98	3	2267	193	1851	187
55—59	632	55	372	38	4682	576	3270	514
60—64	1067	117	846	113	6052	1001	3791	778
65—69	897	170	949	173	3880	901	2421	689
70—74	668	179	824	212	2033	613	1195	432
75—80	361	120	667	233	871	337	436	214
80+	274	120	537	253	345	189	113	63

Источник. Эти данные заимствованы из публикации «A Canadian study of smoking and health—Second report» by Best E. W. R. and Walker C. B., published in the Canadian Yournal of Public Health, 55: 1—1, 1964. Воспроизведено с разрешения автора и журнала. Работа проводилась Управлением национального здравоохранения и Управлением по делам ветеранов Канады вместе с Канадской комиссией по пенсиям.

Таблица 3. Население США в 1960—1970 гг. по штатам и регионам

Регионы и штаты	Население (тыс.)			Изменения в населении от 1960 г. к 1970 г.				
				чистый прирост		родилось	мело	чистая общая миграция
	1960	1965	1970	число	%			
1	2	3	4	5	6	7	8	9
США	179 979	193 526	203 806	23,912	13,3	39 033	18 192	3,070
Северо-Восток*	10 532	11 329	11 883	1 338	12,7	2 169	1 147	316
Мэн	975	997	997	24	2,5	203	109	—69
Нью-Гемпшир	609	676	742	131	21,5	133	71	69
Вермонт	389	404	446	55	14,1	85	45	15
Массачусетс	5 160	5 502	5 706	541	10,5	1 040	574	74
Род-Айленд	855	893	951	90	10,5	171	93	13
Коннектикут	2 544	2 857	3 041	497	19,6	537	255	214
Средняя Атлантика*	34 270	36 122	37 274	3 034	8,9	6 725	3 749	59
Нью-Йорк	16 838	17 734	18 384	1 458	8,7	3 361	1 852	—51
Нью-Джерси	6 103	6 767	7 193	1 101	18,2	1 259	645	488
Пенсильвания	11 329	11 620	11 813	475	4,2	2 105	1 252	—378

Продолжение

1	2	3	4	5	6	7	8	9
Северо-Восток Центра*	36 291	38 406	40 313	4 028	11,1	7 832	3 652	—153
Огайо	9 734	10 201	10 664	946	9,7	2 047	975	—126
Индиана	4 674	4 922	5 202	531	11,4	1 023	475	—16
Иллинойс	10 086	10 693	11 128	1 033	10,2	2 153	1 077	—43
Мичиган	7 834	8 357	8 890	1 052	13,4	1 754	729	27
Висконсин	3 962	4 232	4 429	466	11,8	856	395	4
Северо-Запад Центра*	15 424	15 819	16 518	930	6,0	3 133	1 604	—599
Миннесота	3 425	3 592	3 815	391	11,5	744	327	—25
Айова	2 756	2 742	2 832	68	2,4	541	291	—183
Миссури	4 326	4 467	4 688	358	8,3	857	502	2
Сев. Дакота	634	649	620	—15	—2,3	135	55	—94
Южн. Дакота	683	692	668	—14	—2,1	146	65	—94
Небраска	1 417	1 471	1 488	72	5,1	291	146	—73
Канзас	2 183	2 206	2 249	70	3,2	419	218	—130
Южная Атлантика*	26 091	28 743	30 805	4 700	18,1	5 965	2 598	1 332
Делавэр	449	507	551	102	22,8	109	45	38
Мэриленд	3 113	3 600	3 938	822	26,5	740	303	385
Округ Колумбия	765	797	756	—7	—1,0	182	89	—100
Виргиния	3 986	4 411	4 659	682	17,2	909	369	141
Зап. Виргиния	1 853	1 786	1 751	—116	—6,2	339	190	—265
Сев. Каролина	4 573	4 863	5 098	526	11,5	1 032	412	—94
Южн. Каролина	2 392	2 494	2 597	208	8,7	573	216	—149
Джорджия	3 956	4 332	4 607	646	16,4	975	379	51
Флорида	5 004	5 953	6 848	1 838	37,1	1 107	596	1 326
Юго-Восток Центра*	12 073	12 627	12 839	754	6,3	2 665	1 213	—698
Кентукки	3 041	3 140	3 231	181	6,0	647	313	—153
Теннеси	3 575	3 798	3 937	357	10,0	755	353	—45
Алабама	3 274	3 443	3 451	177	5,4	729	319	—233
Миссисипи	2 182	2 246	2 220	39	1,8	534	228	—267
Юго-Запад Центра*	17 010	18 209	19 388	2 371	14,0	4 012	1 599	—42
Арканзас	1 789	1 894	1 932	137	7,7	401	193	—71
Луизиана	3 260	3 496	3 652	386	11,9	832	316	—130
Оклахома	2 336	2 440	2 567	231	9,9	461	244	13
Техас	9 624	10 378	11 236	1 617	16,9	2 318	847	146
Горный Регион*	6 916	7 740	8 348	1 429	20,8	1 724	602	307
Монтана	679	706	698	20	2,9	144	66	—58
Айдахо	671	686	718	46	6,9	146	58	—42
Вайоминг	331	332	334	2	0,7	70	28	—39
Колорадо	1 769	1 985	2 223	453	25,8	401	163	215
Нью-Мексико	954	1 012	1 023	65	6,8	263	68	—130
Аризона	1 321	1 584	1 792	470	36,1	365	122	228
Юта	900	991	1 066	169	18,9	245	65	—11
Невада	291	444	493	203	71,3	91	31	144
Тихоокеанское побережье *	21 368	24 464	26 600	5 328	25,1	4 808	2 028	2 547
Вашингтон	2 855	2 967	3 413	556	19,5	591	284	249
Орегон	1 772	1 937	2 101	323	18,2	346	182	159
Калифорния	15 870	18 585	20 007	4 236	27,0	3 634	1 511	2 113
Аляска	229	271	304	76	33,6	73	13	16
Гавайи	642	704	774	137	21,7	164	37	11

*Регион, состоящий из следующих штатов.

Источник. Statistical Abstract of the United States, 1975, p. 12—13.

ОГЛАВЛЕНИЕ

ОГЛАВЛЕНИЕ ВЫПУСКА 2

Мостеллер Ф., Тьюки Дж.

М84 Анализ данных и регрессия: В 2-х вып. Вып. 1 / Пер. с англ. Ю. Н. Благовещенского; Под ред. и с предисл. Ю. П. Адлера. — М.: Финансы и статистика, 1982. — 317 с., ил. — (Математико-статистические методы за рубежом).

В пер.: 2 р. 60 к.

В книге исследуются проблемы границ применимости статистических методов к анализу реального мира, проблемы качества статистических выводов — что в них существенно и что несущественно. Под этим углом зрения рассматриваются основные статистические методы, предлагаются новые подходы. Первый выпуск посвящен в основном статистическим данным: их анализу, интерпретации, терминологии, связанной с пониманием данных, представлению о данных, оценке, структуре.

Для статистиков, экономистов, демографов. Полезна студентам старших курсов по этим специальностям.

М $\frac{0702000000-109}{010(01)-82}$ 39—82

ББК 22.172

517.8

Мостеллер Ф., Тьюки Дж.

АНАЛИЗ ДАННЫХ И РЕГРЕССИЯ

Рекомендована к изданию редколлегией серии
5 июня 1979 г.

Зав. редакцией *А. В. Павлюков*
Редактор *К. М. Чижевская*
Мл. редактор *И. Н. Горина*
Техн. редактор *Г. А. Полякова*
Корректоры *Г. В. Хлопцева, Н. П. Сперанская*
Художн. редактор *Э. А. Смирнов*
ИБ № 940

Сдано в набор 27.11.81. Подписано в печать 7.06.82. Формат 60×90¹/₁₆. Бум. тип. № 2. Гарнитура «Литературная». Печать высокая. П. л. 20. Усл. п. л. 20. Уч.-изд. л. 22,66. Тираж 6500 экз. Заказ 615. Цена 2 р. 60 к.
Издательство «Финансы и статистика», Москва, ул. Чернышевского, 7
Московская типография № 4 Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли
129041, Москва, Б. Переяславская, 46.