Статистика для "чайников"™

Statistics FOR DUMMIES®

by Deborah Rumsey, Ph.D.

Wiley Publishing, Inc.

Статистика для "чайников"™

Дебора Рамси

ДИАЛЕКТИКА

Москва • Санкт-Петербург • Киев
2008

ББК (С)60.6
Р21
УДК 311

Компьютерное издательство "Диалектика"
Главный редактор *С.Н. Тригуб*
Зав. редакцией *А.В. Назаренко*
Перевод с английского *А.Н. Свирид*
Под редакцией *А.В. Назаренко*

По общим вопросам обращайтесь в издательство "Диалектика" по адресу:
info@dialektika.com, http://www.dialektika.com
115419, Москва, а/я 783; 03150, Киев, а/я 152

Рамси, Дебора.

Р21 Статистика для "чайников". : Пер. с англ. — М. : ООО "И.Д. Вильямс", 2008. — 320 с. : ил. — Парал. тит. англ.

ISBN 978-5-8459-1369-2 (рус.)

Цель этой книги заключается в том, чтобы научить вас понимать и критически оценивать невероятное количество статистической информации, с которой вам приходится сталкиваться ежедневно (диаграммы, графики, таблицы, а также газетные заголовки, посвященные результатам последних опросов, экспериментов или других научных исследований). Благодаря этой книге вы разовьете способность разбираться в статистических результатах и принимать на их основе важные решения (например, о результатах новейших медицинских исследований). Не забывайте о том, что с помощью статистических данных вас могут попытаться ввести в заблуждение, поэтому учитесь справляться с такими проблемами.

В данной книге приводится масса примеров из реальных источников, имеющих отношение к повседневной жизни: от последних открытий в медицине, исследований преступности и тенденций этнического состава жителей страны до опросов на тему знакомств в Интернете, использования сотовых телефонов и худшего автомобиля тысячелетия. Читая главы книги, вы начнете понимать, как пользоваться диаграммами, графиками и таблицами, а также научитесь оценивать результаты последних опросов, экспериментов и других исследований. Вы даже узнаете, как с помощью сверчков измерить температуру воздуха и как сорвать джек-пот в лотерее.

ББК (С)60.6

ISBN 978-5-8459-1369-2 (рус.)

ISBN 0-7645-5423-9 (англ.)

Оглавление

Содержание

Об авторе

Дебора Рамси получила степень доктора в сфере статистики в Государственном университете Огайо (ГУО) в 1993 году. После окончания учебы она начала работать на факультете статистики Государственного университета Канзаса, где в 1998 году получила престижную Президентскую премию в области образования, а затем и более высокую должность. В 2000 году она вернулась в ГУО в качестве директора Центра по изучению математики и статистики, где и работает по сей день. Дебора — редактор одной из рубрик издания *Journal of Statistics Education*, кроме того, она пишет статьи и выступает с лекциями по теме обучения статистике, в которых особое внимание уделяется статистической грамотности (умению понимать статистические данные в повседневной жизни и в работе) и обучению в ходе погружения в среду (ситуации, в которых студенты учатся развивать собственные идеи). Она обожает рыбалку, любит наблюдать за птицами и болеет за футбольную команду университета Огайо.

Посвящение

Моему мужу Эрику и сыну Клинту Эрику — вы мои самые любимые учителя.

Благодарности автора

Я бы хотела поблагодарить людей, которые помогли этой книге выйти в свет: Кэти Кокс за идею написать книгу, о которой я давно мечтала; Тери Дрент, благодаря которой книга была напечатана вовремя и в правильном формате; Дженет Данн за ее вдумчивое редактирование; Джона Габросека из Государственного университета Гранд-Велли за техническую корректуру. Также благодарю отдел верстки издательства *Wiley Publishing*, сотрудники которого прекрасно справились со сложными оформительскими задачами.

Спасибо Пэг Стайгервальд, Майку О'Лири и моему коллеге Джиму Хиггинсу: частные беседы с ними о статистике сформировали мое видение этой дисциплины. Я благодарна вам за поддержку. Спасибо Тони Баркаускасу, первому и самому лучшему преподавателю в моей жизни, за вдохновение. Я очень признательная своим друзьям с факультета статистики и в Центре по изучению математики и статистики в Государственном университете Огайо за их поддержку и понимание. А также огромное спасибо моей семье за любовь и веру в меня.

Введение

Цель этой книги заключается в том, чтобы научить вас понимать и критически оценивать невероятное количество статистической информации, с которой вам приходится сталкиваться ежедневно. (Вы понимаете, о чем я говорю: диаграммы, графики, таблицы, а также газетные заголовки, посвященные результатам последних опросов, экспериментов или других научных исследований.) Благодаря этой книге вы разовьете способность разбираться в статистических результатах и принимать на их основе важные решения (например, о результатах новейших медицинских исследований). Не забывайте о том, что с помощью статистических данных вас могут попытаться ввести в заблуждение, поэтому учитесь справляться с такими проблемами.

В данной книге приводится масса примеров из реальных источников, имеющих отношение к повседневной жизни: от последних открытий в медицине, исследований преступности и тенденций этнического состава жителей страны до опросов на тему знакомств в Интернете, использования сотовых телефонов и худшего автомобиля тысячелетия. Читая главы книги, вы начнете понимать, как пользоваться диаграммами, графиками и таблицами, а также научитесь оценивать результаты последних опросов, экспериментов и других исследований. Вы даже узнаете, как с помощью сверчков измерить температуру и как сорвать джек-пот в лотерее.

Еще вам будет полезно немного посмеяться над статистиками (которые подчас воспринимают себя слишком серьезно). И все это потому, что для понимания статистики вам совсем не обязательно быть одним из них.

Об этой книге

Эта книга коренным образом отличается от привычных учебников, справочников и пособий по статистике, потому что обладает следующими особенностями.

- ✓ Полезные и доступные объяснения статистических понятий, идей, методов, формул и вычислений.
- ✓ Ясное и точное поэтапное изложение метода, благодаря чему вы легко поймете, как решать статистические задачи.
- ✓ Интересные примеры из реальной жизни, основанные на ситуациях, которые могут произойти с вами в быту и на работе.
- ✓ Честные и откровенные ответы на ваши вопросы, подобные: “Что это вообще значит?” и “Когда и как мне этим пользоваться?”

Условные обозначения, принятые в книге

Работая с книгой, вы должны помнить, что в ней были использованы следующие условные обозначения.

- ✓ Определение размера выборки (n): говоря о размере выборки, я, как правило, имею в виду количество элементов, отобранных для участия в опросе, исследовании или эксперименте. (Размер выборки обозначается буквой n.) Но представим себе, что для участия в опросе было *выбрано*

100 человек, но только 80 из них стали *респондентами*. Тогда какое из этих двух чисел нужно обозначить *n*: 100 или 80? Давайте условимся, что в таком случае мы будем использовать число 80, т.е. количество человек, действительно ответивших на вопросы, причем это число может быть меньше количества тех, кого приглашали к участию в опросе. Значит, всякий раз, когда вы встретите фразу "размер выборки", помните, что это окончательное количество элементов, участвовавших в исследовании и предоставивших информацию для него.

- ✓ Два значения слова "статистика": в одних ситуациях я понимаю под ним предмет исследования, отрасль знания, например: "Статистика — это действительно очень интересная дисциплина". В других случаях статистикой я называю статистические показатели, к примеру: "Самые распространенные статистические показатели — это среднее значение и среднеквадратичное отклонение".

Предположения автора

Я не рассчитываю, что у вас уже был опыт работы со статистикой помимо "неявного" — вы ведь являетесь представителем широкой общественности, которая ежедневно оказывается под напором статистики в виде чисел, процентов, диаграмм, графиков, "статистически значимых" результатов, "научных" исследований, опросов, экспериментов и т.п.

Я все же предполагаю, что вы способны выполнить простейшие математические действия и понять кое-что из основных условных обозначений, принятых в алгебре, например, переменные *x* и *y*, знак суммы, квадратного корня, возведение числа в квадрат и т.д. Если же вам нужно освежить ваши знания по алгебре, обратитесь к соответствующему учебнику.

Однако не забывайте, что в действительности статистика совсем не похожа на математику. Она в большей степени связана с научным методом, т.е. формулировкой вопросов для исследования, разработкой исследований и экспериментов, сбором данных, организацией, обработкой и анализом данных, интерпретацией результатов и подведением итогов. Короче говоря, с помощью данных вы формулируете доказательство ответов на интересующие вас вопросы о мире. Математика нужна при этом только для вычисления основных статистических показателей и проведения некоторых видов анализа, но это всего лишь крошечная часть того, чем занимается статистика.

Не хочу вас обманывать: в этой книге вы действительно найдете много формул, потому что для статистики нужно решать кое-какие задачи. Но пусть это вас не волнует. Я медленно и осторожно проведу вас по всем этапам вычислений, которые вам придется делать. Кроме того, я приведу примеры, чтобы вы могли познакомиться с вычислениями, почувствовать себя увереннее и научиться делать их самостоятельно.

Как построена эта книга

Книга состоит из семи больших частей, в которых изложен основной материал, а также заключительной, восьмой части, где вы найдете десять самых нужных советов и подсказок. Каждая часть разделена на главы, в которых основной материал представлен в понятной форме.

Часть I. Важнейшие сведения о статистике

В этой части объясняется, с каким количеством и качеством статистической информации вы сталкиваетесь каждый день на работе и в быту. Кроме того, вы узнаете, что огромное количество этих статистических сведений ошибочно, причем ошибки делаются либо случайно, либо намеренно. Вы сделаете первый шаг к статистической грамотности, познакомившись с некоторыми профессиональными инструментами, представив себе статистику как процесс получения и интерпретации информации, а еще выучив кое-какие понятия из профессионального жаргона.

Часть II. Основы решения задач

Эта часть научит вас увереннее работать с разными видами *визуального представления данных* (известными также как диаграммы, графики и т.п.). Здесь вы найдете советы по интерпретации этих диаграмм и графиков, а также научитесь замечать ошибки в них. Еще вы узнаете, как подытоживать данные с помощью самых распространенных статистических показателей.

Часть III. Определение шансов

В этой части рассказывается об основах вероятности: как ею пользоваться, что нужно знать и с чем вам придется столкнуться в азартных играх. В чем суть? Вероятность и интуиция — не всегда одно и то же!

Вы узнаете, как вероятность влияет на вашу повседневную жизнь, и выучите некоторые основные правила вероятности. Кроме того, вы узнаете о подводных течениях в азартных играх, т.е. о том, как работают казино и почему эти заведения никогда не проигрывают.

Часть IV. Разбираемся в результатах

Здесь вы познакомитесь с тонкостями, благодаря которым существует статистика, в том числе выборочным распределением, точностью, пределом погрешности, процентилями и стандартными показателями. Вы научитесь вычислять и интерпретировать два критерия относительного положения — стандартные показатели и процентили. Помимо этого вы узнаете, что статистики называют “королевой статистики” (центральную предельную теорему) и насколько проще с ее помощью толковать статистическую информацию; а также каким образом статистики оценивают изменчивость от выборки к выборке и почему это так важно. Из этой части вы узнаете, что же все-таки означает широко используемый термин *предел погрешности*.

Часть V. Уверенные предположения

В этой части рассказывается о том, как довольно точно оценивать среднее или пропорцию совокупности, когда правильный ответ не известен. (Например, среднее количество часов в неделю, которое взрослые люди проводят перед телевизором, или процентное отношение людей в США, у которых на стекле автомобиля есть хотя бы одна наклейка.) Кроме этого вы прочтете, как сделать довольно точную оценку, имея в своем распоряжении сравнительно небольшую выборку (по сравнению с размером совокупности). В этой части описаны основные характеристики доверительных интервалов, указано их назначение и определены основные элементы доверительных интервалов (предположение плюс/минус предел погрешности). Помимо этого, вы выясните, какие факторы

влияют на размер доверительного интервала (например, размер выборки), и познакомитесь с формулами, поэтапными вычислениями и примерами самых распространенных доверительных интервалов.

Часть VI. Переходим к проверке гипотез

В этой части рассказывается о процессе принятия решений и о том, какую огромную роль играет при этом статистика. Вы узнаете, как исследователи формулируют (или должны это делать) и проверяют свои утверждения, а также поймете, как нужно оценивать полученные ими результаты, чтобы убедиться, что они все сделали правильно и пришли к убедительным выводам. Здесь же вы получите подробные инструкции для выполнения вычислений при проведении самых распространенных видов проверки гипотез и для правильного толкования результатов.

Часть VII. Статистические исследования: взгляд изнутри

В этой части говорится об основных характеристиках опросов, экспериментов, исследований по результатам наблюдений и контроля качества. Вы узнаете, чем занимаются все эти исследования, как они проводятся, каковы их недостатки и как их оценить, чтобы понять, стоит ли доверять полученным результатам.

Часть VIII. Великолепные десятки

В этой последней части вы найдете десять критериев хорошего опроса и десять самых распространенных ошибок в статистике, которые допускают исследователи, средства массовой информации и общественность.

Приложение

Одна из основных задач книги — подтолкнуть вас к тому, чтобы копать глубже и находить истинную информацию, необходимую для принятия взвешенных решений относительно определенных статистических показателей. В приложении указаны все источники, которыми я пользовалась для того, чтобы привести примеры в книге. Возможно, вы захотите подробнее узнать о каких-либо из них.

Пиктограммы в этой книге

В книге используются некоторые пиктограммы, привлекающие внимание читателей к некоторым регулярно встречающимся особенностям. Вот что они означают.

Это полезные рекомендации, идеи и подсказки, которыми вы можете воспользоваться для экономии времени. Здесь же предлагаются альтернативные варианты видения определенных понятий.

Рядом с этой пиктограммой указана информация, которую будет полезно запомнить.

Здесь говорится о специфических способах, которыми исследователи или средства массовой информации, используя статистику, могут вводить вас в заблуждение, а также объясняется, как с этим бороться.

На эти абзацы вы должны обратить особое внимание, если захотите больше узнать о различных технических аспектах статистических вопросов. Однако их можно и пропустить, если у вас нет желания вдаваться в детали.

Куда двигаться дальше

Эта книга написана так, чтобы дать вам толчок к движению в любом направлении, но чтобы при этом вы не теряли нить происходящего. Поэтому изучите содержание или предметный указатель, найдите интересующую вас информацию и откройте книгу на нужной странице.

Если вы не знаете, с чего начать, тогда начните с главы 1 и читайте всю книгу, не отрываясь.

От издательства

Вы, читатель этой книги, и есть главный ее критик. Мы ценим ваше мнение и хотим знать, что было сделано нами правильно, что можно было сделать лучше и что еще вы хотели бы увидеть изданным нами. Нам интересны любые ваши замечания в наш адрес.

Мы ждем ваших комментариев и надеемся на них. Вы можете прислать нам бумажное или электронное письмо, либо просто посетить наш Web-сервер и оставить свои замечания там. Одним словом, любым удобным для вас способом дайте нам знать, нравится или нет вам эта книга, а также выскажите свое мнение о том, как сделать наши книги более интересными для вас.

Отправляя письмо или сообщение, не забудьте указать название книги и ее авторов, а также свой обратный адрес. Мы внимательно ознакомимся с вашим мнением и обязательно учтем его при отборе и подготовке к изданию новых книг.

Наши электронные адреса:

E-mail: `info@dialektika.com`

WWW: `http://www.dialektika.com`

Наши почтовые адреса:

в России: 115419, Москва, а/я 783

в Украине: 03150, Киев, а/я 152

Часть I

Важнейшие сведения о статистике

The 5th Wave

Рич Теннант

В этой части...

Включая телевизор или открывая газету, вы попадаете в окружение цифр, графиков, схем и статистических результатов. Начиная с последнего опроса и заканчивая новейшими достижениями в медицине, цифры все прибывают и прибывают. И все же очень многое в той статистической информации, которую вас призывают воспринять, на самом деле неверно — случайно или даже намеренно. Чему же верить? Давайте проведем кропотливое детективное расследование.

Благодаря данной части книги вы сможете пробудить скрытого в себе сыщика. Вы поймете, как статистика влияет на вашу повседневную жизнь и работу, сможете оценить качество большей части получаемой информации и научитесь принимать правильные решения. Кроме того, эта часть поможет вам овладеть некоторыми полезными статистическими терминами.

Глава 1

Статистика в повседневной жизни

В этой главе...

- Встреча со статистикой в обычной жизни: что и как часто вы видите
- Объяснение, как статистика используется в работе

Современное общество полностью покорено цифрами. Цифры есть везде, куда бы вы ни посмотрели: от рекламных щитов, которые сообщают последние данные о количестве абортов, до спортивных передач, в которых речь идет о ставках в Лас-Вегасе на предстоящий футбольный матч, и вечернего выпуска новостей, сюжеты в котором рассказывают об уровне преступности, предположительной продолжительности жизни человека, который употребляет нездоровую пищу, и последнем рейтинге популярности президента. Обычно вы пять, десять, а то и двадцать раз сталкиваетесь с различными статистическими данными (а в ночь выборов это случается еще чаще). Прочитав воскресную газету от первой страницы до последней, вы найдете сотни статистических данных в статьях, рекламных объявлениях и заметках, рассказывающих о чем угодно — от супа (какое количество съедает за год обычный человек?) до орехов (сколько орехов нужно съесть, чтобы повысить свой уровень IQ?).

Цель данной главы — показать важность статистики в нашей повседневной жизни и работе, а также обратить ваше внимание на то, как именно статистические данные предлагаются публике. На сегодняшний день средства массовой информации буквально атакуют нас разными цифрами, и не всегда удается понять, что все эти цифры значат. Следует признать, что статистика занимает в нашей жизни важное место, поэтому полезно будет научиться разбираться в тонкостях этой непростой науки.

Статистика и атака СМИ: больше вопросов, чем ответов?

Откройте любую газету и попытайтесь найти примеры статей и заметок, в которых задействованы цифры. Пройдет совсем немного времени, и цифры начнут то и дело попадаться вам на глаза. Читателей засыпают результатами исследований, сообщениями о новых открытиях, статистическими отчетами, прогнозами, предположениями, диаграммами, графиками и сводками. То, насколько активно статистические данные используются в СМИ, просто поражает. Вы можете даже не осознавать, сколько раз сталкиваетесь с цифрами в наш информационный век. Ниже приведено всего несколько примеров из одной воскресной газеты. Читателей могут охватить сомнения: чему можно верить, а чему нет? Расслабьтесь! Эта книга как раз и призвана помочь вам научиться отличать хорошую информацию от плохой. (Этот вопрос подробно рассматривается в главах 2–5).

Проблемы с попкорном

Первая статья с цифрами, на которую я натолкнулась, называется “Санитарно-эпидемиологическая проверка на фабрике по производству попкорна”. Подзаголовок гласит: “Заболевшие работники утверждают, что ароматические добавки вызывают болезни легких”. В статье говорится, что Центр контроля над заболеваниями (ЦКЗ) выражает обеспокоенность в связи с возможной взаимосвязью между использованием химических веществ в качестве ароматических добавок для попкорна и несколькими зафиксированными случаями обструкционных легочных болезней. Восемь работников одной из фабрик по производству попкорна подверглись этому заболеванию, а четыре из них ожидают теперь пересадки легких. Согласно статье, о подобных случаях сообщается и на других фабриках. И вот вы можете задать вопрос: “А как же будут чувствовать себя люди, которые едят такой попкорн?” В статье сказано, что ЦКЗ “не видит никаких оснований полагать, что людям, которые едят попкорн, есть чего бояться”. Сотрудники ЦКЗ намерены тщательнее обследовать работников, в том числе провести исследование их здоровья и возможной подверженности указанным химическим веществам, проверить состояние легких и качество воздуха. Вопрос звучит так: сколько случаев этого легочного заболевания можно считать закономерностью, а не простой случайностью или статистической аномалией? (Подробнее об этом читайте в главе 4.)

Обратимся к вирусам

Во второй статье воскресного номера речь шла об атаке в киберпространстве — о новом вирусе, появившемся в Интернете и замедляющем работу браузеров, а также отправку электронных сообщений во всем мире. Сколько компьютеров были им поражены? Эксперты, слова которых приводятся в статье, утверждают, что зараженными оказались 39 тыс. компьютеров, что повлияло еще на сотни тысяч других систем. Но откуда они взяли это число? Разве его так просто определить? Неужели были проверены все имеющиеся компьютеры, чтобы установить, не заражены ли они? Тот факт, что статья была написана менее чем через сутки после атаки, наталкивает на мысль, то приводимая цифра — это предположение. Тогда почему же не сказать не 39, а 40 тыс.? Чтобы больше узнать о надежных приближенных оценках (а также о том, как оценивать данные других людей), см. главу 11.

Разберемся с авариями

Затем в газете идет статья о возрастающем количестве аварий на мотоциклах. Специалисты говорят, что с 1997 года это число увеличилось более чем на 50%, и никто не может объяснить причину. Статистика приводит следующий интересный факт: в 1997 году погибло 2 116 мотоциклистов, в 2001 году погибших было 3 181, как показывают данные Национальной администрации безопасности дорожного движения (НАБДД). В статье рассматриваются многие возможные причины увеличения количества смертей, в том числе и тот факт, что сегодня мотоциклисты стали старше (средний возраст погибших мотоциклистов увеличился с 29,3 года в 1990 году до 36,3 лет в 2001 году).

Еще одно из возможных объяснений — увеличение размеров мотоциклов. Размер двигателя среднего мотоцикла вырос почти на 25% — с 769 см^3 в 1990 году до 959 см^3 в 2001 году. Дополнительный вариант — это тот факт, что некоторые штаты США делают послабления в законе относительно ношения шлема. Специалисты, слова которых цитируются в статье, говорят, что необходимо более обширное изучение причин, но

оно, вероятно, так и не будет проведено, потому что затраты на него составят от 2 до 3 млн. долл. При этом в статье ничего не говорится о количестве людей, которые ездят на мотоциклах, в 2001 и 1997 году. Естественно, что большее число людей на дорогах означает больше аварий, даже если все остальные факторы остаются прежними. Однако в статье приведен еще и график, отображающий количество смертей мотоциклистов на 100 млн. миль, которые были преодолены в США с 1997 по 2001 год. Касается ли это увеличения количества людей на дрогах? Здесь же приводится и столбиковая диаграмма, в которой сравниваются число смертей мотоциклистов с количеством людей, погибших в авариях на других видах транспорта. Из этой диаграммы видно, что уровень смертности мотоциклистов составляет 33,4 смерти на 100 млн. преодоленных миль по сравнению с показателем всего лишь в 1,7 на то же количество миль, преодоленных на машине. В этой статье множество цифр и самых разных статистических данных, но что все это значит? Объем и разнообразие статистических данных очень скоро может сбить с толку. Глава 4 поможет вам разобраться с графиками, диаграммами и данными, которые к ним относятся.

Размышления о врачебных ошибках

Дальше в газете следует статья о последних исследованиях в сфере страхования врачей на случай судебного преследования. Итак, насколько серьезна данная проблема? В статье сказано, что один из пяти врачей в штате Джорджия отказался от проведения опасных процедур (например, принятие родов) из-за постоянно растущих страховых ставок от судебного преследования в этом штате. Это описывается как “национальная эпидемия” и “кризис здравоохранения” в стране. Приводятся некоторые сведения об исследовании проблемы; в статье утверждается, что из 2200 врачей штата Джорджия, принявших участие в опросе, 2800 (которые, как говорится, составляют 18% от общего числа участников) скорее всего, откажутся от проведения рискованных процедур. Минуточку! Разве это возможно? Из 2200 врачей 2800 человек не участвуют в рискованных процедурах, и это должно составлять 18%? Но это же невозможно! В числителе дроби не может быть большее число, если эта дробь меньше 100%, правильно? Это один из многих примеров ошибок в статистике, которые приводятся в средствах массовой информации. Так каков же настоящий процент? Можно только догадываться. В главе 5 подробно описываются особенности подсчета статистических данных, чтобы вы знали, что искать, и сразу же могли заметить ошибку.

Бьемся над потерей земли

В той же воскресной газете находим статью об уровне освоения земель и торговле земельными участками в стране. Учитывая количество зданий, которые, скорее всего, будут построены в данной местности, это очень важный вопрос. Приводятся статистические данные, касающиеся акров пахотной земли, которая ежегодно теряется из-за застройки, и все это превращается в квадратные мили. В качестве дополнительной иллюстрации того, как много земли теряется, эта площадь представлена также в соответствующем количестве футбольных полей. В этом конкретном случае эксперты отмечают, что в центре штата Огайо в год теряется 150 тыс. акров земли, что составляет 234 квадратные мили или 115 385 футбольных полей (включая зону защиты). Но как были получены эти цифры и насколько они точны? И неужели проще представить количество потерянной земли с помощью футбольных полей?

Хотим сорвать джек-пот

Вы когда-нибудь мечтали выиграть в Суперлотерее, в среднем имея для этого шансы 1 к 89 миллионам?

Не паникуйте! Чтобы представить себе этот 1 шанс из 89 миллионов, вообразите, что в одной огромной куче находятся 89 миллионов лотерейных билетов, и где-то среди них есть и ваш. Предположим, я скажу, что у вас есть один шанс залезть в эту кучу и вытянуть свой билет — как вы думаете, это вам удастся? Вот именно таковы ваши шансы выиграть в одной из крупных лотерей. Но, имея определенную дополнительную информацию, вы можете увеличить свой джек-пот на случай выигрыша. (А я бы хотела получить свою долю от этого выигрыша, если мой совет вам поможет.) Подробнее об этом и других подсказках, касающихся азартных игр, см. главу 7.

Изучение самых разных исследований

Исследования и опросы — это, наверное, самое популярное средство, которым пользуются современные средства массовой информации, чтобы привлечь ваше внимание. Создается впечатление, что провести опрос оказывается по плечу маркетологам, страховым компаниям, телеканалам, общественным группам и даже старшеклассникам. Вот всего несколько примеров того, какие результаты опросов попадают в выпуски новостей.

Учитывая старение трудоспособных американцев, компании планируют, какие лидеры им нужны в будущем. (Но откуда они знают, что трудоспособное население стареет, и если это так, то насколько именно оно стареет?) Одно из исследований показывает, что почти 67% опрошенных менеджеров по кадрам заявили, что планирование преемственности за последние пять лет стало намного важнее, чем было прежде. Но если вы хотите бросить свою теперешнюю работу и претендовать на должность генерального директора, не спешите. Из исследования также видно, что 88% из 210 респондентов сказали, что они обычно или часто принимают на должности руководителей высшего звена уже имеющихся сотрудников компании. (Но сколько менеджеров *не* участвовали в опросе, и разве 210 респондентов достаточно, чтобы результаты такого исследования попали на первую полосу раздела о бизнесе?) Верьте или нет, но только начните искать, и в новостях вы сразу же наткнетесь на многочисленные примеры исследований, в которых принимало участие еще меньше человек.

Некоторые исследования основаны на принципах попроще. Например, что американцы считают сегодня самым важным для себя — зубную щетку, хлеборезку, компьютер, автомобиль или мобильный телефон? По результатам опроса 1042 взрослых и 400 подростков (почему они взяли именно эти числа?) 42% взрослых и 34% подростков назвали самым важным предметом зубную щетку, а совсем не компьютер, автомобиль или сотовый. Действительно ли это так важно? Зачем нужно сравнивать такой важный предмет личной гигиены как зубная щетка, с сотовыми телефонами и хлеборезками? (На втором месте оказался автомобиль. Но разве без специального исследования до этого нельзя было бы додуматься?). Подробнее об исследованиях см. главу 16.

Рассматриваем школу

Следующая тема в газете — это школьная подготовка, а именно вопрос о том, помогают ли дополнительные занятия повысить успеваемость учащихся. В статье утверждается, что 81,3% учеников конкретно взятого округа, которые посещали дополнительные занятия, сдали письменный тест, а среди тех, кто не занимался дополнительно, этот тест сдали всего 71,7%. Но слишком ли существенна эта разница, чтобы с ее помощью можно было оправдать 386 тыс. долл., выделяемых ежегодно? И что именно происходит на этих занятиях, благодаря чему повышается успеваемость? Может, ученики уделяют боль-

ше времени только подготовке к экзаменам, а не изучению нового материала в целом? А вот и самый важный вопрос: не были ли слушатели этих дополнительных занятий добровольцами, что давало бы им дополнительный стимул по сравнению с обычными учениками, пытающимися улучшить свой результат? Неизвестно. Подобные исследования проводятся постоянно, и единственный способ узнать, чему же можно верить, — это понимать, какие вопросы нужно задавать, а также уметь критически подходить к качеству исследования. Все это элементы статистики! Хорошая новость состоит в том, что с помощью нескольких дополнительных вопросов вы сможете быстро оценить статистические исследования и их результаты. В этом вам поможет глава 17.

Изучаем спорт

Спортивная страничка — это, наверное, тот раздел газеты, в котором встречается больше всего цифр. Помимо результатов последних матчей, процентного соотношения проигрышей и выигрышей каждой команды лиги и их сравнительного изучения, специализированные статистические данные, которые используются в мире спорта, так сложны, что в них сходу не разобраться. Например, баскетбольная статистика подсчитывается по команде, по четверкам и даже по отдельному игроку. И чтобы понять все это, нужно быть помешанным на баскетболе, и к тому же разбираться в многочисленных сокращениях (которые нигде не поясняются).

- ✓ MIN: Сыграно минут
- ✓ FG: Игровые попадания
- ✓ FT: Штрафные броски
- ✓ RB: Броски при отскоке
- ✓ A: Результативные передачи
- ✓ PF: Нарушения правил
- ✓ TO: Потери мяча
- ✓ B: Блоки
- ✓ S: Перехваты
- ✓ TP: Общий счет

Кому все это надо, кроме мам самих игроков? Статистика — это то, чего спортивным болельщикам никогда не хватает, а игроки ее терпеть не могут. Статистика становится темой для обсуждения у автомата газированной воды и источником кухонных перебранок во всем мире.

Деловые новости

В разделе газеты, посвященном бизнесу, вы найдете статистические данные о фондовом рынке. Прошлая неделя была неудачной, цена акций упала на 455 пунктов. Это много или мало? Чтобы понять это, вам придется подсчитать проценты. В том же разделе вы найдете отчеты о самых высоких доходах по депозитам в стране. (Кстати говоря, откуда известно, что они самые высокие?) Здесь же можно обнаружить и сведения о процентных ставках по кредитам: ставки по 30-летней долгосрочной ссуде, 15-летней долгосрочной ссуде, годовому займу с переменным процентом, займам на покупку новых автомобилей, подержанных машин, жилищным кредитам и ссудам от вашей бабуш-

ки (что в общем-то вряд ли возможно, хотя если бы бабушка разбиралась во всей этой статистике, она бы могла подумать об увеличении той мелочи, которую вы можете получать с ее денег!). Наконец, вы увидите множество рекламных объявлений для тех, кто обожает кредитные карточки — в этих объявлениях указываются процентные ставки, годовая плата и количество дней в цикле выставления счетов по кредитке. Как сравнить всю эту информацию об инвестициях, ссудах и кредитных карточках, чтобы принять правильное решение? Какие данные важнее всего? И самое главное: представленные в газете сведения дают полную картину или же все-таки нужно провести дополнительное расследование, чтобы докопаться до правды? Глава 3 поможет вам научиться понимать эти цифры и с их помощью принимать верные решения.

Собираемся в путешествие

Вам не удастся скрыться от атакующих цифр даже на странице о путешествиях. В этом разделе я обнаружила, что самый часто задаваемый вопрос, который поступает на горячую линию Управления безопасности перевозок (в среднем в неделю поступает 2000 телефонных звонков, 2500 электронных сообщений и 200 писем — вы лично хотите все это подсчитывать?), звучит так: “Можно взять *это* в самолет?”. В этой фразе словом “это” может обозначаться что угодно, от животного до огромной упаковки попкорна. (Я бы не советовала брать с собой такую упаковку. Вам придется положить ее на верхнюю полку горизонтально, а поскольку во время полета вещи двигаются, крышка может открыться, и когда вы, покидая самолет, пойдете забирать свой попкорн, вас с приятелями осыплет с ног до головы. Да, я сама такое видела.)

После всего вышесказанного возникает интересный статистический вопрос: сколько потребуется операторов, которые бы отвечали на все эти звонки? Прежде всего, нужно определить количество входящих звонков, и ошибка ударит вам по карману (если вы переоцените число) или сделает антирекламу (если вы недооцените его).

Поговорим о сексе (и статистике) с доктором Рут

На отдельной странице воскресной газеты вниманию читателей предлагаются результаты новейших исследований доктора Рут, касающихся сексуальной жизни человека. Она утверждает, что секс не прекращается в 60 лет и даже в 70. Знать это приятно, но как доктор определяет то, как часто люди в таком возрасте занимаются сексом? Она не говорит (может быть, о некоторой статистике лучше промолчать?). И тем не менее, доктор Рут советует людям в таком возрасте не обращать внимания на опросы, сообщающие, сколько раз в неделю, месяц или год пара занимается сексом. На ее взгляд, большинство опрашиваемых просто хвалятся. Возможно, в этом доктор права. Представьте, что кто-то проводит опрос — звонит людям по телефону, просит их уделить ему несколько минут, чтобы обсудить их сексуальную жизнь... Захотят ли люди говорить об этом? И что они скажут в ответ на вопрос: “Сколько раз в неделю вы занимаетесь сексом?”. Они скажут чистую правду или все же немного преувеличат? Результаты исследований, о которых сообщает сам ученый, могут стать источником предубеждений и привести к созданию ошибочной статистики. Поэтому не слишком критикуйте доктора Рут (кстати, автора книги «Секс для “чайников”»). Что бы вы могли ей посоветовать о том, как побольше узнать об этом очень интимном вопросе? Иногда исследование намного сложнее, чем кажется. В главе 2 вы найдете еще много примеров того, как статистика может ошибаться и чего нужно опасаться.

Ставки в Лас-Вегасе

Если проследить, как цифры используются (и фальсифицируются) в повседневной жизни, то нельзя не обратить внимание на мир спортивных пари — бизнес с оборотом в несколько миллиардов долларов в год. К этому бизнесу имеют отношение как обычные люди, делающие ставку, так и профессиональные и заядлые игроки. На что можно заключать пари? На любую ситуацию, у которой есть два исхода. Сумасшедшие пари, которые можно заключить в Лас-Вегасе, не имеют пределов (простите за каламбур). Ниже приведены примеры некоторых самых насущных вопросов, касающихся Суперкубка, на которые можно сделать ставку в букмекерской конторе (месте, где заключают пари) Лас-Вегаса.

- ✓ Какая команда получит наибольшее число пенальти?
- ✓ Какая команда забьет последний гол в первом периоде?
- ✓ Что будет сначала — гол или пант?
- ✓ Сколько ярдов пройдут обе команды (больше или меньше 675)?
- ✓ Сделают ли обе команды бросок с 33 ярдов или дальше?

Хмм... Почему бы не добавить сюда еще и количество фунтов соуса гуакамоле, который потребляют телезрители Суперкубка, и количество травинок на футбольном поле? Игроки, начинайте считать.

Интерес людей к явлениям природы

В статьях о погоде тоже имеется масса статистических данных, если учесть прогнозируемую высокую или низкую температуру (интересно, почему синоптики называют не 15, а именно 16 градусов?) или сведения об уровне ультрафиолета, содержании пыльцы в воздухе, индексе загрязнения, а также качестве и количестве воды. (Откуда у них эта информация? Они берут образцы? Тогда сколько образцов исследуется и где они их берут?). Вы можете даже получить прогноз на три дня, неделю, месяц или даже год вперед! Но насколько точны современные прогнозы погоды? Учитывая то, сколько раз вы попадали под дождь, когда вам говорили, что день будет солнечным, можно сказать, что синоптикам еще нужно работать и работать над своими прогнозами!

Тем не менее, вероятность и компьютерное моделирование действительно играют важную роль в современном прогнозировании погоды. Особенно они полезны в том, что касается заметных явлений, таких как ураганы, землетрясения и извержения вулканов. Конечно, компьютеры не умнее людей, которые их программируют, поэтому ученым по-прежнему есть над чем поработать, прежде чем они смогут предсказывать торнадо до того, как те появятся. Подробнее о моделировании и статистике см. главу 6.

Немного о кино

Переходим к следующей странице газеты, посвященной искусству... Здесь читателям предлагается несколько анонсов фильмов. В каждом объявлении помещены цитаты из отзывов популярных кинокритиков: “За — обеими руками!”, “Удивительное приключение наших дней”, “Просто великолепно!” или “Один из десяти лучших фильмов года!”. Вы обращаете внимание на слова критиков? Как вы решаете, какой фильм посмотреть? Эксперты утверждают, что, хотя оценка критиков (хорошая или плохая) может повлиять на популярность фильма, когда он только вышел в прокат, в долгосрочной перспективе намного более надежным критерием успешности фильма будут отзывы зрителей.

Кроме того, исследования показывают, что чем напряженнее фильм, тем больше попкорна продается. Да, индустрия развлечений подсчитывает даже то, как часто вы жуете во время просмотра. Но как происходит сбор такой информации и как она влияет на то, какие фильмы снимаются? Вот еще одна из целей статистики: разработка и проведение исследований, которые должны определить аудиторию, установить, что ей нравится, а затем использовать эти сведения при создании продукта. Поэтому когда в следующий раз человек с папкой в руках спросит, есть ли у вас минутка, чтобы поучаствовать в опросе, вы можете остановиться и тоже выразить свое мнение.

Подробнее о гороскопах

Ох уж эти гороскопы: вы читаете их, но верите ли им? Стоит ли им доверять? Можно ли предсказывать то, что происходит не просто случайно? Статистики могут определить это, используя так называемый критерий проверки гипотезы (см. главу 14). Пока им не удалось найти человека, способного читать мысли, но поиски продолжаются!

Использование статистики в работе

Оторвемся от воскресной газеты, которую вы читаете в уютной обстановке у себя дома, и перейдем к повседневной рутине на рабочем месте. Если вы работаете в бухгалтерской фирме, тогда, конечно же, цифры — это неотъемлемый атрибут вашей жизни. Ну а как же медсестры, фотографы, продавцы, газетные репортеры, офисные работники и даже строительные рабочие? Важны ли числа в таком случае? Даже не сомневайтесь. В этом разделе вы найдете несколько примеров, которые демонстрируют тесную связь статистики с профессией человека.

Не нужно слишком далеко ходить, чтобы проследить путь статистики, который она прокладывает в вашу жизнь и работу. Секрет заключается в том, чтобы понимать, что все это значит, чему можно верить, а также уметь принимать разумные решения, основанные на истине, скрытой за многочисленными цифрами, с тем, чтобы научиться управлять и постепенно привыкнуть к статистике в повседневной жизни.

Рождение ребенка — и информация

Сью работает медсестрой ночной смены в родильном отделении университетской больницы. В ее обязанности входит заботиться о своих пациентах, и она делает все возможное, чтобы им помочь. Каждый раз, приходя на смену, Сью должна представиться, написать свое имя на доске в палате и поинтересоваться у пациентки, есть ли у нее какие-то вопросы. Дело в том, что по возвращении домой из больницы женщинам уже через несколько дней звонят и спрашивают о качестве обслуживания, интересуются их мнением по поводу того, что больница может сделать для улучшения качества обслуживания, а также что может сделать персонал, чтобы в этот роддом обращались чаще, чем в другие больницы города. Качество обслуживания очень важно, и молодым мамам, которые остаются в больнице, когда медсестры сменяются каждые восемь часов, нужно знать их имена, потому что благодаря этому они могут своевременно получить ответы на свои вопросы. Повышение Сью зависит от ее способности удовлетворять потребности молодых мам.

Позирование для фото

Кэрол недавно начала работать фотографом в фотостудии универмага. Одна из сильных ее сторон — это работа с маленькими детьми. Исходя из количества фотографий, купленных клиентами за многие годы, в универмаге пришли к выводу, что люди будут покупать больше профессионально поставленных снимков, чем естественных. В результате менеджеры магазина поощряют своих фотографов делать именно такие фотографии.

В студию приходит мама с малышом и высказывает особое пожелание: “Нельзя ли сделать так, чтобы поза моего ребенка не казалась слишком надуманной? Мне бы хотелось, чтобы его фотография была естественной”. Что на это скажет Кэрол? “Извините, не могу. Моя зарплата зависит от умения заставить ребенка позировать”. Ух-ты! Можно не сомневаться, что мама, высказавшая такую просьбу, обязательно примет участие в опросе относительно качества обслуживания после подобного сеанса — и не только для того, чтобы сэкономить пару долларов во время следующего визита (если она вообще когда-нибудь вернется в эту студию).

Роемся в данных о пицце

Терри — менеджер в небольшой пиццерии, которая продает нарезную пиццу. Его обязанность — определять, сколько работников необходимо иметь в штате в определенное время, сколько штук пиццы нужно приготовить заранее, чтобы удовлетворять спрос, а также сколько требуется заранее заказать и натереть сыра, чтобы минимизировать общий расход времени и ингредиентов. В пятницу заполночь в пиццерии нет ни одного клиента. У Терри в распоряжении пять работников. Пять противней с пиццей уже можно ставить в духовку и приготовить около 40 кусков пиццы. Нужно ли отправить двоих работников домой? А может, лучше приготовить большее количество пиццы? Терри знает, что, скорее всего, случится, потому что владелец магазина в течение нескольких недель следил за спросом, и ему уже известно, что в пятницу вечером жизнь замедляется между 10 и 12 часами вечера, но потом около полуночи толпа начинает прибывать, и двери не закрываются до полтретьего ночи. Поэтому Терри не отпускает работников, начиная с полуночи он отправляет пиццу в духовку с интервалом в полчаса, а наградой за это ему будет хорошая ночная выручка, довольные клиенты и счастливый босс. Подробнее о том, как с помощью статистики делать правильные оценки, см. главу 11.

Работа со статистикой в офисе

Возьмем Диджи, помощника руководителя в компьютерной компании. Как статистика связана с ее работой? Очень просто. В каждом офисе есть масса людей, которые хотят получить ответы на вопросы, поэтому им нужен человек, который бы “щелкал задачки”, “объяснил, что это значит”, “выяснил, есть ли у кого-то такие данные” или просто ответил на вопрос: “Это число имеет смысл?” Им нужно знать все — от количества довольных клиентов до изменений в товарно-материальных запасах в течение года, от процентного выражения того, сколько времени сотрудники тратят на электронную почту, до стоимости поставок за последние три года. В каждом офисе очень много статистики, поэтому успех Диджи на рынке и ценность ее как работника может вырасти, если она окажется именно тем человеком, к которому все обращаются за помощью. Каждому офису нужен собственный статистик — так почему бы вам не стать им?

Глава 2

Когда со статистикой возникают проблемы

В этой главе...

- Важность проблем со статистикой
- Отказ от традиционных табу в статистике
- Осознание того, что со статистикой возникли проблемы

От такого количества цифр может закружиться голова (но с помощью нашей книги вы научитесь понимать бо́льшую часть тех статистических данных, которые встречаются вам в жизни)! Задача этой главы — вынести на поверхность еще одно чувство — скептицизм! Но не такой, как в предложении: "Я больше ни во что не верю", а тот скептицизм, что звучит во фразах: "Хм, интересно, откуда взялись эти цифры?", "Это действительно так?", "Прежде чем верить таким результатам, надо бы побольше узнать об этом исследовании". В средствах массовой информации вы найдете множество примеров ошибочной статистки. И когда вы научитесь замечать подобные проблемы, то очень скоро почувствуете себя намного увереннее во всем, что касается статистики, и будете готовы принять вызов цифр!

Берем ситуацию под контроль: так много цифр и так мало времени

Статистические данные, которые вы встречаете на экране телевизора или в газете, — это результат длительного процесса. Прежде всего, исследователям, изучающим какой-либо вопрос, нужно получить определенные результаты. Этими людьми могут быть специалисты по опросам общественного мнения, врачи, исследователи рынка, правительственные исследователи и другие ученые. Они называются *исходным источником* статистической информации. Получив результат, эти исследователи хотят рассказать о нем другим людям, поэтому обычно выпускают пресс-релиз или публикуют статью в журнале. Тут появляются журналисты, которых называют *медийным источником* информации. Журналисты ищут интересные пресс-релизы, просматривают отраслевые журналы, т.е. по сути, творят новые сенсационные заголовки. Когда репортеры заканчивают статьи, статистические данные представляются общественности. Это можно сделать через самые разные СМИ: телевидение, газеты, журналы, Web-сайты, проспекты и т.д. Теперь информация готова для подачи третьей группе — *потребителям* информации (вам!). Вы и другие потребители информации — именно те люди, перед которыми стоит задача прослушивать и читать информацию, копаться в ней и принимать на ее основе реше-

ния. И, как вы уже, наверное, догадались, на любом этапе этого долгого процесса — во время исследования, представления результатов или потребления информации — могут произойти ошибки либо случайно, либо преднамеренно.

Обнаружение ошибок, преувеличений и просто откровенного вранья

Проблемы со статистикой могут возникать по разным причинам. Во-первых, это может быть простая, искренняя погрешность. Ведь такое может случиться с каждым, правильно?

В другой раз ошибка представляет собой нечто большее, нежели простая, искренняя погрешность. В кульминационный момент, если человек всецело поддерживает какую-то идею, а цифры не слишком наглядно подтверждают то, что хочет сказать исследователь, статистические данные слегка приукрашиваются, т.е. делаются преувеличения, которые могут касаться как самих цифр, так и того, как эти цифры подаются или обсуждаются.

Наконец, бывают ситуации, когда цифры оказываются полностью сфабрикованными и ими нельзя пользоваться, потому что таких результатов не было в природе. Это худший из возможных вариантов, но и он встречается в реальном мире.

В этом разделе вы найдете советы, которые помогут замечать ошибки, преувеличения и ложь. Здесь же содержатся примеры всех видов ошибок, с которыми вы как потребитель информации можете столкнуться.

Проверка вычислений

При изучении статистических данных или результатов статистического исследования хочется задать вопрос: “Правильны ли эти цифры?”. Не принимайте их на веру! Вы удивитесь, как много обычных математических ошибок совершается при сборе, подытоживании, изложении или интерпретации статистических данных. Помните, что еще один вид ошибок — *ошибка упущений*, т.е. отсутствует информация, наличие которой кардинально изменило бы общую картину, скрытую за цифрами. Именно поэтому непросто решить вопрос о правильности данных, ведь у вас нет всей необходимой информации.

Чтобы найти арифметические ошибки или упущения в статистике, вы должны следовать таким рекомендациям.

- ✓ Проверьте все действия. Действительно ли сумма всех процентов на диаграмме равна 100? Равно ли количество людей во всех категориях заявленному общему количеству?
- ✓ Перепроверьте даже самые простые вычисления.
- ✓ Всегда ищите итоговый результат, чтобы можно было взглянуть на ситуацию с правильной точки зрения. Не обращайте внимания на результаты, основанные на исследованиях крошечных групп.

Ошибки в процентном соотношении

Многие статистики делят результаты на группы, показывая, какой процент людей в каждой группе дал определенный ответ на поставленный вопрос, или отмечая демографический показатель (например, возраст, пол и т.д.). Это один из эффективных способов представить статистические данные, но только если все проценты в сумме дают 100%.

Например, в газете *USA Today* сообщалось о результатах опроса общественного мнения для компании *Tupperware* относительно разогрева пищи в микроволновых печках. В статье было сказано, что 28% опрошенных разогревают еду в микроволновке почти каждый день, 43% сказали, что делают это два–четыре раза в неделю, а 15% утверждают, что пользуются микроволновкой раз в неделю. Если предположить, что любой человек относится к одной из этих групп, то результат должен равняться 100%. Проверим: 28% + 43% + 15% = 86%. Где еще 14%? Кого не учли? К какой группе они относятся? Неизвестно. Данные не суммируются.

Четыре из пяти — да неужели?

Еще один момент, который можно быстро проверить, — это наличие общего числа респондентов. Возьмем, к примеру, рекламу жевательной резинки Trident, где говорилось: "Четыре из пяти опрошенных стоматологов рекомендуют своим пациентам жевать Trident". Этой рекламе уже несколько лет, но недавно о ней снова заговорили благодаря забавной серии новых роликов, в которых задается вопрос: "А что случилось с пятым стоматологом?", а затем показываются разные происшествия, которые мешают пятому стоматологу нажать на кнопку с надписью "Да". И все же: сколько же всего стоматологов было опрошено? Это неизвестно, потому что в исследовании об этом не говорится. Невозможно перепроверить даже информацию, написанную мелким шрифтом, потому что в подобных рекламных роликах она не требуется.

Но что изменилось бы, если бы было известно общее число респондентов? Дело в том, что надежность статистики частично зависит от объема информации, поступившей на обработку (если это хорошая и правильная информация). Если в рекламе говорится: "четыре из пяти стоматологов", то вполне может быть, что опросили всего пять человек. Если же опрошенных было 5000, то в таком случае жевательную резинку рекомендуют 4000 из них. Таким образом, неизвестно, сколько стоматологов действительно советуют эту резинку. Чтобы узнать это, придется провести дополнительное расследование. В большинстве случаев именно потребителю придется искать информацию. Если неизвестно общее число людей, принимавших участие в исследовании, будет весьма трудно понять, насколько надежна представленная информация.

Обнаружение ошибочной статистики

Даже найдя ошибку в статистике, вы не всегда сможете определить, к какому типу она относится — это действительная погрешность или же человек, имеющий доступ к подаче информации, подтасовывал факты. Пока самой распространенной проблемой статистики остается незначительное, но все же эффективное преувеличение правды. Даже если с вычислениями все в порядке, ошибочными могут быть лежащие в основе статистические данные: они могут быть несправедливыми, искажать правду или приукрашать факты. Данные, вводящие в заблуждение, сложнее обнаружить, чем простые незначительные погрешности, но они оказывают огромное влияние на общество и, к сожалению, встречаются постоянно.

Уровень преступности, которого нет

Заметив данные, вводящие в заблуждение, вы хотите проверить тип использованной информации. Справедлива ли она? Применима ли? Имеет ли она вообще практический смысл? Если вас волнует только правильность вычислений, то вы упустите из виду более страшную ошибку, когда статистической обработке подвергаются заведомо ошибочные данные.

Отличным примером того, как статистика используется для представления двух версий истории (хотя правильный из них только один), является уровень преступности. О преступлениях часто говорят во время политических дебатов, когда один кандидат (как правило, уже занимающий определенный пост) утверждает, что за время его работы уровень преступности снизился, а его оппонент обычно заявляет, что этот уровень наоборот вырос (и это дает ему повод покритиковать соперника). Но как оба политика могут говорить о том, что уровень преступности движется в разных направлениях? Предположив, что с вычислениями все в порядке, как это можно объяснить? В зависимости от того, как вы измеряете количество преступлений, можно будет получить либо тот, либо другой результат. В табл. 2.1 показано количество преступлений в Соединенных Штатах Америки с 1987 по 1997 годы по данным ФБР.

Таблица 2.1. Количество преступлений в США (1987–1997)

Год	Количество преступлений
1987	13 508 700
1988	13 923 100
1989	14 251 400
1990	14 475 600
1991	14 872 900
1992	14 438 200
1993	14 144 800
1994	13 989 500
1995	13 862 700
1996	13 493 900
1997	13 175 100

Итак, уровень преступности растет или уменьшается? Кажется, что в общем он падает, но на эти данные можно посмотреть под разными углами и представить их так, что вся тенденция будет казаться иной. Самый главный вопрос здесь: правильно ли эти данные отражают действительность?

Например, сравним годы с 1987 по 1993. В 1987 году в США произошло 13 508 700 преступлений, а в 1993 году их общее число было 14 144 800. Кажется, что за шесть лет преступлений стало больше. Представьте, что вы кандидат, который бросает вызов действующему президенту. На этом кажущемся росте уровня преступности можно построить свою платформу. А если посмотреть на данные за 1996 году, то в этом году было совершено 13 493 900 преступлений, что всего лишь немногим меньше, чем общее число преступлений в 1987 году. Так что же, действительно ли за период с 1987 по 1993 годы многое было сделано для борьбы с преступностью? Кроме того, эти цифры говорят не все. Можно ли назвать общее число совершенных преступлений за определенный год самым подходящим критерием для измерения уровня преступности в США?

Кое-какая важная информация оказалась упущена (и поверьте, такое происходит чаще, чем вы думаете)! За период с 1987 по 1993 годы вырос не только уровень преступности, но и население США. Общая численность населения страны тоже должна играть роль в подсчете статистики преступлений, потому что когда увеличивается количество людей, живущих в стране, то логично предположить, что вырастет также число потенциальных преступников и их потенциальных жертв.

Вся подноготная соотношений, уровней и процентов

В статистике существует множество разных единиц, с помощью которых выражаются статистические данные, и в этом разнообразии можно запутаться.

- ✓ *Отношение* — это дробь, в которой делятся два числа. Например: "Отношение девочек и мальчиков 3 к 2", т.е. на каждых трех девочек приходится два мальчика. Это не значит, что в группе всего три девочки и два мальчика, отношения выражаются в упрощенном виде. Таким образом, может быть 300 девочек и 200 мальчиков, отношение все равно останется 3 к 2.
- ✓ *Уровень* — это соотношение, которое отражает определенное количество, содержащееся в конкретной единице (например, ваша машина едет со скоростью 60 миль в час или уровень грабежей в округе составляет 3 случая на 1000 домов).
- ✓ *Процент* — цифра от 0 до 100, отражающая часть от целого. Например, рубашка продается со скидкой в 10% или 35% населения выступают за четырехдневную рабочую неделю. Чтобы проценты превратить в десятичную дробь, разделите их на 100 или передвиньте запятую в дроби на две цифры влево. Проще это можно запомнить так: 100% равны 1 или 1,00, и чтобы перейти от 100 к 1, вы делите на 100 или переносите запятую в десятичной дроби на два промежутка влево. (Значит, проделав это в обратном порядке, мы превратим десятичную дробь в процент.)

Можно сказать, что проценты используются для определения того, насколько увеличивается или уменьшается значение. Предположим, что в одном городе совершается уже не 50, а 60 преступлений, а в другом число нарушений закона выросло с 500 до 510. В обоих случаях количество преступлений увеличилось на 10, но для первого города эта разница намного больше, если представить ее в процентном отношении от общего числа преступлений. Чтобы определить рост процентного отношения, из количества "после" вычитаем количество "до", а получившийся результат делим на количество "до". В случае с первым городом это означает, что преступлений стало больше на (60 – 50)/50 = 10/50 = 0,20 или 20%. Для второго города изменения составили всего 2%, потому что (510 – 500)/500 = 10/500 = 0,02 или 2%. Для того чтобы определить снижение в процентном отношении, поступать нужно так же. Вы получите отрицательное число, которое и покажет снижение.

Поэтому чтобы правильно изучить ситуацию с преступностью, нужно учесть не только количество преступлений, но и численность населения. Как это делается? ФБР определяет коэффициент преступности, который и представляет собой уровень преступности. *Уровень* — это соотношение, т.е. число людей или событий, которые вас интересуют, деленное на общую численность всей группы.

В табл. 2.2 показана приблизительная численность населения в США за 1987–1997 годы, а также примерное число преступлений и приблизительный *уровень преступности* (количество преступлений на 100 000 человек).

Таблица 2.2. Количество преступлений, приблизительная численность населения и уровень преступности в США (1987–1997)

Год	Количество преступлений	Примерная численность населения	Уровень преступности (на 100 000 чел.)
1987	13 508 700	243 400 000	5 550,0
1988	13 923 100	245 807 000	5 664,2
1989	14 151 400	248 239 000	5 741,0
1990	14 475 600	248 710 000	5 820,3
1991	14 872 900	252 177 000	5 897,8

Окончание табл. 2.2

Год	Количество преступлений	Примерная численность населения	Уровень преступности (на 100 000 чел.)
1992	14 438 200	255 082 000	5 660,2
1993	14 144 800	257 908 000	5 484,4
1994	13 989 500	260 341 000	5 373,5
1995	13 862 700	262 755 000	5 275,9
1996	13 493 900	265 284 000	5 086,6
1997	13 175 100	267 637 000	4 922,7

Если снова сравнить 1987 и 1993 годы, то вы увидите, что количество преступлений выросло с 13 508 700 в 1987 году до 14 144 800 в 1993 году. (Обратите внимание, что это составит прирост в 4,7%, потому что 14 144 800 – 13 508 700 равно 636 100, а если это число разделить на исходное значение, 13 508 700, то вы получите 0,047, т.е. 4,7%.) С такой точки зрения можно сказать, что за период с 1987 по 1993 годы уровень преступности вырос на 4,7%. Но эти 4,7% представляют собой прирост в *общем числе* преступлений, а не в количестве преступлений *на человека*, т.е. на 100 000 человек. Чтобы определить, как с течением времени изменилось количество преступлений на 100 000 человек, нужно подсчитать и сравнить уровни преступности в 1987 и 1993 году. Вот как это делается: (5 484,4 – 5 550,0)/5 550,0 = 65,6/5 550,0 = –0,012 = –1,2%. Итак, на самом деле преступлений на 100 000 человек (уровень преступности) стало меньше на 1,2%.

В зависимости от того, как вы интерпретируете числа, результат может подтверждать разные точки зрения: уровень преступности за период с 1987 по 1993 год либо вырос, либо снизился. Но теперь, когда известна разница между количеством преступлений и уровнем преступности, становится ясно, что в некоторых отчетах нельзя было бы просто указывать общее число случаев, а следовало бы говорить об уровнях (т.е. количестве случаев, деленном на численность всей группы).

Не доверяйте использованным данным, пока не разберетесь с результатом. Этот критерий справедливый и уместный? Можно ли назвать его надежным способом отразить действительность, скрытую за данными, или же есть лучший вариант?

Шкала подскажет, как выиграть в лотерею!

Графики и диаграммы — это хороший способ наглядно и быстро проиллюстрировать свою мысль, если такие рисунки сделаны правильно и точно. В чем же разница между графиком и диаграммой? Небольшая. Статистики, говоря о наглядном отображении статистической информации, пользуются этими понятиями как синонимами. Хотя на самом деле, если на рисунке показаны столбики, окружности или другие фигуры, то это диаграмма. А график — это координатная сетка (x, y), на которой точками представлено изменение показателей с течением времени. (Подробнее об этом см. главу 18.)

К сожалению, очень часто графики и диаграммы, которыми сопровождается привычная статистика, выполнены с ошибками, поэтому всегда нужно помнить о возможных погрешностях. Одна из самых важных проблем, о которой нельзя забывать, — это то, как именно график или диаграмма промаркированы. Всегда учитывайте *цену деления* графика — какое количество обозначает каждая отметка на осях координат. С каждым

шагом количество возрастает на 10, 20, 100 или 1000? Шкала может иметь огромное значение в том, что касается наглядности графика или диаграммы.

Например, лотерея Kansas Lottery регулярно показывает последние результаты розыгрыша Pick 3 Lottery. Один из приводимых критериев — количество раз, когда на три выигрышных номера выпадает каждая цифра (от 0 до 9). В табл. 2.3 показано, сколько раз каждая цифра выпала 15 марта 1997 года (в течение 1 613 розыгрышей Pick 3, общее количество — 4 839 номеров). В зависимости от того, под каким углом посмотреть на эти результаты, всю статистику опять-таки можно представить совершенно по-разному.

Таблица 2.3. Количество выпадений каждой цифры (Kansas Pick 3 Lottery, 15 марта 1997 года)

Цифра	Количество выпадений
0	485
1	468
2	513
3	491
4	484
5	480
6	487
7	482
8	475
9	474

Результаты лотерей, подобные приведенным в табл. 2.3, обычно показывают так, как представлено на рис. 2.1. Так, из диаграммы кажется, будто цифра 1 выпадала совсем не так часто (всего 468 раз) по сравнению с цифрой 2 (513 раз). Складывается впечатление, что разница между этими столбиками очень значительная, что преувеличивает реальную разницу между количеством выпадений этих двух цифр. Но в данном случае настоящая разница составляет 513 – 468 = 45 из общего числа выпадений в 4 839. В том, что касается общего числа выпадений каждой цифры, разница между цифрами 1 и 2 составляет 45/4 839 = 0,009 или всего девять десятых процента.

Почему в этой диаграмме настолько преувеличена разница? Здесь проявляются два момента, влияющие на внешний вид диаграммы. Во-первых, обратите внимание, что на вертикальной оси показано количество выпадений (или частота) каждой цифры, при этом цена деления равна 5. Поэтому складывается впечатление, что разница в 5 раз из 4 839 действительно важна. Это распространенная уловка, которой пользуются, чтобы преувеличить результаты — растянуть шкалу так, чтобы разница казалась большей, чем есть на самом деле. Во-вторых, отсчет на диаграмме начинается не с нуля, а с 465, поэтому на самом деле здесь представлена только верхушка каждого столбика, где и проявляется различие. Такой ход тоже помогает приукрасить результаты.

В табл. 2.4 показана более реалистичная картина того, сколько раз во время розыгрыша Pick 3 Lottery вытягивали каждую цифру, поскольку здесь представлены процентные отношения всех выпадений.

Рис. 2.2 представляет собой столбиковую диаграмму, на которой показано не количество выпадения каждой цифры, а их процентное отношение. Заметьте, что на этой диаграмме использована более реалистичная шкала, чем на рис. 2.1, а также то, что она начинается с нуля, благодаря чему разница между столбиками воспринимается такой, какой она и есть на самом деле — не слишком значительной.

Таблица 2.4. Процент выпадений каждой цифры

Цифра	Количество выпадений	Процент выпадений
0	485	10,0% = 485/4 839
1	468	9,7% = 486/4 839
2	513	10,6% = 513/4 839
3	491	10,1% = 491/4 839
4	484	10,0% = 484/4 839
5	480	9,9% = 480/4 839
6	487	10,0% = 487/4 839
7	482	10,0% = 482/4 839
8	475	9,8% = 475/4 839
9	474	9,8% = 474/4 839

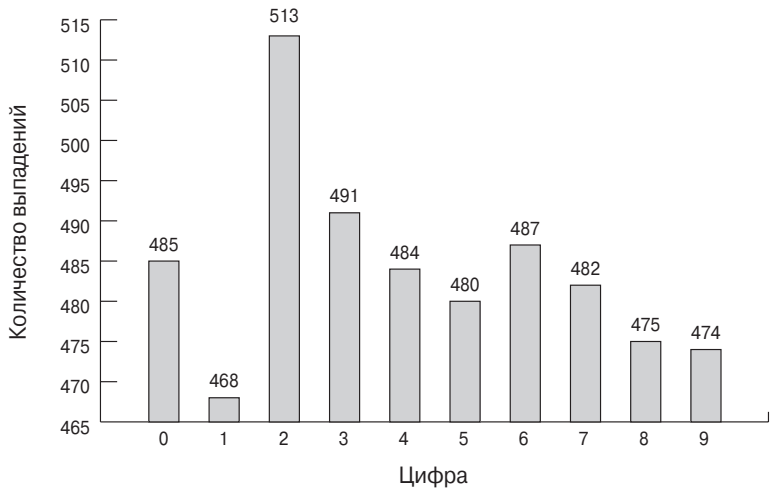


Рис. 2.1. Столбиковая диаграмма, показывающая количество выпадений каждой цифры

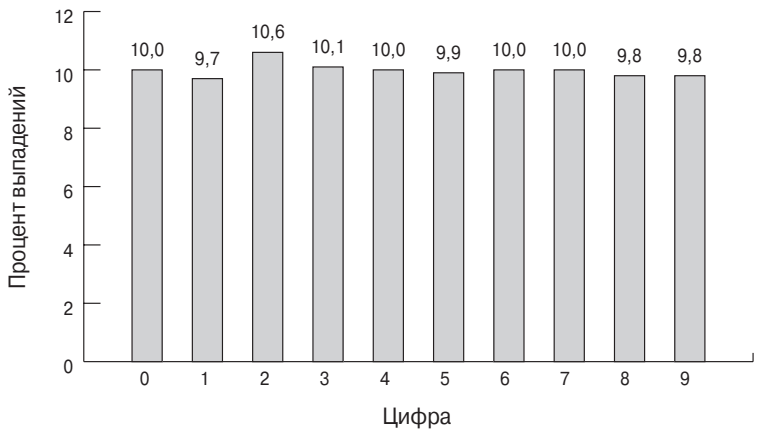


Рис. 2.2. Столбиковая диаграмма, показывающая процентное отношение количества выпадений каждой цифры

В таком случае спрашивается, зачем организаторы лотереи так поступают? Возможно, они хотят, чтобы вы поверили, что получаете некую секретную информацию. И если полагать, что цифру 1 вытягивали слишком редко, то вы купите лотерейный билет и выберете 1, потому что это "обязательно" случится (кстати, это *не*правда; подробнее об этом см. главу 7). Или же вы предпочтете цифру 2, потому что ее вытягивали очень часто, и она "в ходу" (и опять никаких гарантий). Но что бы вы ни решили, устроители лотереи хотят, чтобы вы думали, что с цифрами связана какая-то "магия", и не винили их. В этом и состоит их работа.

Графики, вводящие в заблуждение, приводятся в СМИ ежедневно! Репортеры и другие люди могут растягивать шкалу (т.е. делать слишком маленькой цену деления) и/или начинать отсчет не с нуля, а с другого числа, чтобы разница казалась большей, чем она есть на самом деле. Кроме того, шкалу можно сжать (сделать цену деления очень большой), чтобы складывалось впечатление, что "нет никакой разницы". Таким образом мы получаем множество вариантов вводящего в заблуждения изложения истины. (Подробнее о подобных графиках см. главу 4.)

Изучение шкалы графика или диаграммы помогает правильно осознать результаты.

Проверка источников

Проверьте источник информации. Лучшие результаты часто публикуются в издании, которое хорошо известно специалистам в конкретной области знаний. Например, в медицине это *Journal of the American Medical Association* (JAMA), *New England Journal of Medicine*, *The Lancet* и *British Medical Journal,* относящиеся к уважаемым изданиям, в которых врачи публикуют результаты своих исследований и читают о новых открытиях.

Изучая результаты любого исследования, обратите внимание на источник и учтите все проведенные изыскания, а не только те, результаты которых опубликованы в журналах или появились в рекламе. Столкновение интересов между учеными может привести к появлению неверной информации.

Учитываем размер выборки

Количество опрашиваемых — это еще не все, но все же оно очень многое значит в том, что касается опросов и исследований. Если исследование разработано и проведено правильно и если участники были выбраны наугад (т.е. без смещения; подробнее об этом см. главу 3), то при определении точности и надежности результатов важным фактором является размер выборки. (Подробнее о разработке и проведении исследований см. главы 16 и 17.)

Складывается впечатление, что во всех исследованиях принимает участие множество людей. Действительно, в большинстве случаев это так, но это бывает не всегда, особенно если речь идет об исследованиях, которые связаны с тщательно контролируемыми экспериментами. На проведение эксперимента может потребоваться очень много времени, иногда на воссоздание ряда ситуаций уходят месяцы и даже годы. Кроме того, экспериментальные исследования могут быть очень дорогостоящими. Некоторые эксперименты связаны с изучением не только людей, но и продукции, например, компьютерных чипов

или военного оборудования, которые стоят тысячи и даже миллионы долларов. Если для эксперимента при тестировании продукта его нужно уничтожить, то стоимость такого опыта может оказаться довольно высокой. Из-за дороговизны определенных исследований некоторые из них основываются на привлечении небольшой группы участников или использовании небольшого количества продукции. Однако меньший размер выборки (или меньшее число протестированных продуктов) означает меньше информации в целом, поэтому исследования с небольшим количеством участников (или продукции) в общем менее точны, чем такие же эксперименты, численность участников в которых была больше.

Большинство исследователей пытаются довести число участников до максимума и уравновешивают стоимость эксперимента с точностью. Однако иногда люди просто слишком ленивы и не хотят возиться с большой группой респондентов. Подчас такие исследователи вообще не понимают, какие последствия имеет небольшая численность участников эксперимента. А кое-кто просто надеется, что вы-то ничего не знаете об этих последствиях. Хотя теперь вам это уже известно.

Наихудший пример совершенно неадекватно подобранного размера выборки — это телереклама, в которой участвует всего один человек. Как правило, в таких роликах представлена видимость эксперимента, чтобы убедить зрителей, что один продукт лучше другого. Возможно, вы видели рекламные ролики, в которых один вид бумажных салфеток сравнивается с другим, когда ими вытирают одно и то же количество красного сока. Такие примеры могут показаться глупостью, но любой человек легко попадется в ловушку и сделает вывод, основанный на эксперименте всего с одним участником. (Вы когда-нибудь говорили человеку не покупать определенный продукт, потому что вам с этой маркой не повезло?). Не забывайте, что анекдот (или анекдотическая история) — это, на самом деле, эксперимент с одним участником.

Проверяйте численность участников, чтобы убедиться, что исследователи собрали достаточно информации для подведения итогов. Предел погрешности (см. главу 10) также поможет понять, что такое размер выборки, потому что низкий предел погрешности, скорее всего, значит, что размер выборки был большим.

Драгоценное время врача: количество или качество?

Газетные заголовки — это хлеб всех СМИ, но и они могут вводить в заблуждение. Очень часто заголовки в газетах грандиознее, чем "реальная" информация, особенно когда история связана со статистическими данными и исследованиями, в ходе которых эти данные были получены. Более того, в таких историях вы зачастую можете заметить настоящее несоответствие между заголовком и телом статьи.

Во время одного исследования, проведенного несколько лет назад, оценивались видеозаписи приемов у врача 1 265 пациентов, которые обратились к 59 врачам первичной медико-санитарной помощи и 6 хирургам в штатах Колорадо и Орегон. В результате было обнаружено, что врачи, на которых *не* подавали в суд за плохое качество обслуживания, в среднем уделяли каждому пациенту 18 минут по сравнению с 16 минутами, которые тратили врачи, прежде обвиненные в небрежности. Неужели эти две минуты действительно так важны? Когда сообщения об исследовании появились в СМИ под за-

головком “Врачебный такт спасает от судебных исков”, то складывалось впечатление, будто в этом исследовании утверждается, что если вы доктор и на вас подали в суд, то все, что вам нужно сделать, — это больше времени уделять своим пациентам, тогда все будет в порядке.

Но что происходит на самом деле? Неужели же врач, на которого подавали в суд, должен уделять каждому пациенту на пару минут больше, чтобы избежать нового иска? Подумайте о других возможных объяснениях. Может быть, врачи, против которых не выдвигались иски, профессионалы в своем деле — задают больше вопросов, внимательнее выслушивают пациентов и больше рассказывают им о том, чего ожидать, на что и уходит больше времени? Если это так, то здесь важны именно действия врача во время приема, а не то, сколько времени он тратит на осмотр пациента. Однако есть и другое объяснение: возможно, врачи, на которых подавали в суд, выполняют более сложные операции, или же они имеют определенную специализацию. К сожалению, в статье вы не найдете такой информации. Еще одно возможное объяснение заключается в том, что у врачей, на которых в суд не подавали, меньше пациентов, поэтому они могут больше времени уделять каждому, а значит, и лучше лечить их. Как бы там ни было, содержание статьи не совсем соответствует броскому заголовку, и когда вы слышите или читаете подобные заголовки, обращайте внимание на расхождения между заглавиями и тем, что действительно было установлено в ходе исследования.

Отчеты без шкалы

Интересно, как политики могут предугадать настроение своих избирателей? Они проводят опросы и исследования. Многие исследования организовывают независимые группы, например, *Gallup Organization*, а другие проводят представители самих кандидатов. Поэтому их методы могут во многом отличаться у разных кандидатов и в ходе проведения различных опросов.

Во время президентских выборов 1992 года в США ярким явлением на политической арене стал Росс Перо. Его группа под названием United We Stand America получила известность и, в конечном счете, Росс Перо и его сторонники повлияли на результаты выборов (по мнению многих аналитиков он отобрал голоса у Дж. Буша-старшего и принес победу Биллу Клинтону). Часто во время дебатов и программных выступлений Перо приводил статистические данные и делал выводы о том, как американцы относятся к определенным вопросам. Но всегда ли точно господин Перо чувствовал настроение “американцев”? Может, он хорошо знал, что ощущают его сторонники? Одно из средств, которое Росс Перо использовал для того, чтобы влиять на мнения американцев, была публикация анкеты в выпуске *TV Guide* от 21 марта 1992 года с просьбой к читателям заполнить ее и отослать по указанному адресу. После этого он подвел итоги опроса и сделал их частью своей предвыборной платформы. На основе этих результатов он пришел к выводу, что свыше 80% американцев разделяют его точку зрения по данным вопросам. (Но обратите внимание, что во время выборов он получил всего 18,91% голосов.)

Одна из проблем с утверждениями господина Перо состоит в том, как проводился опрос. Чтобы принять в нем участие, вам нужно было купить этот номер газеты, иметь горячее желание поучаствовать в опросе, чтобы заполнить анкету, и отправить ответ по почте за свои деньги. Кто, скорее всего, сделает это? Те, кто имеет твердые убеждения. Кроме того, формулировка вопросов в анкете, возможно, подталкивала людей, согласных с Россом Перо, участвовать в опросе и отсылать свои ответы. Те же, кто не разделял его мнения, вероятнее всего, просто проигнорировал опрос.

Если, исходя из формулировки вопроса, вы можете понять, что от вас хочет услышать исследователь, знайте, что перед вами — *наводящий вопрос*. (Подробнее о том, как замечать эту и похожие проблемы в опросах, см. главу 16.)

Вот пример некоторых вопросов, которые господин Перо использовал в своей анкете. Я перефразировала их, но исходный смысл остался неизменным. (И это сделано не для того, чтобы придраться к господину Перо, многие политики и их сторонники поступают так же.)

- ✓ Должен ли президент иметь право использовать постатейное вето для того, чтобы сократить расходы?
- ✓ Должен ли Конгресс самоустраниться от законов, которые передает нам?
- ✓ Обязательно ли все новые крупные программы сначала в деталях представлять на рассмотрение американского народа?

Мнения людей, знавших об опросе и пожелавших принять в нем участие, скорее всего, совпали с точкой зрения господина Перо. Это один из примеров того, как выводы, сделанные в ходе опроса, вышли из зоны компетенции самого исследования, поскольку результаты не представляли мнения "всех американцев", как полагали некоторые избиратели. Как можно узнать мнение всех американцев? Нужно правильно провести хорошо разработанное исследование с применением случайной выборки респондентов. (Подробнее о проведении исследований см. главу 16.)

Рассматривая результаты любого исследования, внимательно изучите как группу, которая действительно была в центре опроса (или группу, которая действительно принимала в нем участие), так и более широкий круг людей (или лабораторных мышей, блох, в зависимости от исследования), которых, как предполагается, изучаемая группа представляла. Затем взгляните на сделанные выводы. Проверьте, совпадают ли они. Если нет, постарайтесь понять, каковы правильные выводы, и реально воспринимайте утверждения, которые делались до того, как вы самостоятельно приняли какое-либо решение.

Ищем места, где скрывается ложь

Вы уже видели примеры искренних погрешностей, приводящих к возникновению проблем, а также осознали, каким образом преувеличенная правда может вызвать неприятности. Иногда можно наблюдать ситуации, в которых статистические данные просто выдумываются, фабрикуются или подделываются. Благодаря отраслевым изданиям, которые читают многие специалисты в определенной отрасли, наблюдательным советам и правительственным указам, такое происходит не очень часто. Однако подчас вы все же слышите о том, как кто-то подделал те или иные дынные или "подчистил цифры". Пожалуй, самая распространенная ложь, связанная со статистикой и данными, — это когда люди отбрасывают информацию, которая не подтверждает их гипотезу, не укладывается в схему или кажется "несущественной". Когда была явно допущена ошибка (например, написано, что человеку — 200 лет), стоит попытаться исправить данные, либо устранив ошибочную информацию, либо попробовав исправить ошибку. Но отбрасывать информацию только потому, что она не соответствует вашей идее, нельзя. Исключение данных (помимо тех случаев, когда это явная ошибка) — этически неправильно. К сожалению, такое все же случается.

Если речь зашла о нехватке данных относительно эксперимента, очень часто можно слышать фразу: “Среди тех, кто участвовал в исследовании до конца...” А как же те, кто не закончил исследование, особенно если это медицинский эксперимент? Они что, умерли? Устали от побочных эффектов экспериментального лекарства и бросили всю затею? Может, их вынуждали давать определенные ответы или подтверждать гипотезы исследователей? Или же они совершенно разочаровались, потому что, несмотря на продолжительность эксперимента, им не становилось лучше, поэтому они и решили отказаться от дальнейшего участия?

В опросах участвуют не все, и даже те люди, кто обычно стремится принять в них участие, иногда замечают, что у них нет ни желания, ни времени отвечать на вопросы каждой встреченной анкеты. Современное общество “сдвинулось” на опросах, и практически каждый месяц вас приглашают к участию в опросе по телефону, через Интернет или по почте, причем темы бывают самые разные, от ваших любимых товаров до мнения по поводу нового постановления насчет лая собак в округе. Результаты опросов сообщаются только заказавшим их людям, а мнения участников могут абсолютно не совпадать с точкой зрения тех, кто отказался от участия. Но сообщают ли вам об этом исследователи — это уже другой вопрос.

Например, можно сказать, что вы разослали 5 000 анкет, получили 1000 ответов и основываете на них свои выводы. Тогда вы могли бы подумать: “Тысяча ответов! Это масса информации, должно быть, это очень точное исследование”. Ошибаетесь. Проблема в том, что 4 000 человек, выбранных для участия в опросе, отказались от него, и вы не имеете никакого представления о том, что бы они ответили, если бы все же участвовали в исследовании. Нет никаких гарантий того, что мнение этих 4 000 человек представлено теми, кто отвечал на вопросы. Более того, скорее всего, справедливо противоположное утверждение.

Что обеспечивает высокую *долю ответивших* (т.е. число респондентов, деленное на количество разосланных анкет)? Иногда опрос считается заслуживающим доверия, если доля ответивших составляет не менее чем 70%, но, как говорит телеведущий доктор Фил, статистикам нужно “спуститься с небес на землю”. Подобная доля ответивших на опрос бывает крайне редко. Однако в целом чем ниже доля ответивших, тем ненадежнее результаты и тем больше эти результаты будут склоняться в сторону точки зрения участников опроса. (При этом не забывайте, что респонденты склонны иметь твердые убеждения по вопросам, которые они хотят обсудить.)

Чтобы заметить подделку или пропущенную информацию, ищите сведения об опросе, в том числе то, сколько человек было отобрано для участия, сколько из них дошли до конца и что случилось со всеми участниками, а не только с теми, кто показал положительные результаты.

Влияние ошибочной статистики

Как неправильные статистические данные влияют на вашу жизнь? Воздействие может оказаться значительным или не очень, в зависимости от типа данных, с которыми вы столкнулись, и того, что вы намерены делать с полученной информацией. Самое серьезное влияние статистики (положительное или отрицательное) на вас — это принятие решения в повседневной жизни.

Вспомните примеры, описанные в этой главе, и подумайте, как бы они повлияли на ваши решения. Скорее всего, вы не станете засиживаться за полночь, размышляя о тех 14% опрошенных, которые непонятно что делают с микроволновкой. Однако статистика проявляется и в более серьезных вещах, ее влияние может быть огромным, поэтому вы должны быть к этому готовы и уметь видеть подвох. Рассмотрим несколько примеров.

- ✓ Вам доказывают, что четыре из пяти опрошенных согласны: налоги нужно увеличить, значит, вы тоже должны с этим согласиться! Вы ощутите давление или попробуете сначала найти более подробную информацию? (Помните, как в детстве вам говорили: “Так все делают”?)
- ✓ Кандидат во время предвыборной кампании присылает вам брошюру, в которой содержатся статистические данные о выборах. Поверите ли вы тому, что там написано?
- ✓ Если вас когда-либо выбирали для участия в суде присяжных, вполне может быть, что в какой-то момент вы станете свидетелем того, как адвокат использует статистику в качестве довода. Вы должны обдумать всю информацию и решить, убедительны ли и “бесспорны” эти доказательства. Другими словами, есть ли шанс, что обвиняемый виновен? (Подробнее о толковании вероятности см. главы 7 и 8.)
- ✓ В выпуске новостей по радио в час пик сообщают, что сотовые телефоны вызывают опухоль мозга. Ваш супруг постоянно пользуется телефоном. Будете ли вы беспокоиться?
- ✓ А как насчет всех этих бесконечных рекламных роликов фармацевтических компаний? Представьте, какое давление должны испытывать врачи со стороны пациентов, которые заранее убеждены благодаря рекламе, что им *сейчас* же нужно принять определенное лекарство. Быть информированным — это одно, но считать себя информированным благодаря рекламному ролику, проспонсированному производителем продукции, — это совсем другое.
- ✓ Если у вас или вашего знакомого проблема со здоровьем, возможно, вы ищете новые лекарства или способы лечения, которые могут помочь. В мире медицинских экспериментов масса статистики, которая может легко сбить с толку.

В реальной жизни встречается все: от простых арифметических ошибок до преувеличений, подделки данных, *выуживания данных* (подбор подходящих результатов) и отчетов, которые просто упускают из виду информацию или сообщают только о тех итогах, которые исследователь хочет до вас донести. Хотя все же следует подчеркнуть, что не все статистические данные вводят в заблуждение и не каждый исследователь хочет вас “подловить”, но все же нужно не терять бдительность. Отделяйте хорошую информацию от плохой, и тогда вы научитесь находить ошибки в статистике.

Статистика проникает в повседневную жизнь

Каждый день вы принимаете решения, основанные на статистике и статистических исследованиях, о которых вы слышали или читали, и очень часто сами о том не подозреваете. Вот некоторые примеры решений, которые вам приходится принимать в течение дня.

- ✓ “Стоит ли сегодня обуть ботинки? Какой вчера был прогноз погоды? Ах да, 30% вероятность снега”.
- ✓ “Сколько воды стоит пить? Раньше я думал, что 8 стаканов по 200 мл — это идеальный вариант, но теперь говорят, что слишком много воды может быть вредно! “
- ✓ “Может, сегодня стоит сходить за витаминами? Мэри говорит, они ей помогают, но мой желудок витамины плохо переносит”. (Кстати, *когда* лучше всего принимать витамины?)
- ✓ “Начинает болеть голова, наверное, нужно выпить аспирин. Может, попробовать больше времени проводить на солнце? Слышал, что это помогает при мигрени”.
- ✓ “Ой, надеюсь, Рекс не сгрызет опять половик, пока я на работе. Где-то я слышал, что собаки, которым дают антидепрессант Прозак, легче переживают одиночество. Собака на антидепрессантах? И как они подбирают дозировку? А что я скажу друзьям?”
- ✓ “Может, опять заскочить в Макдональдс на ленч? Что-то я слышал о “вредном холестерине”. Но, наверное, вся еда в фастфуде вредна для здоровья.”
- ✓ “Интересно, напустится ли босс на тех, кто общается по электронной почте? Исследование показало, что люди проводят в среднем два часа в день, проверяя и отправляя с рабочего места электронные письма. Хотя у меня, конечно же, на это уходит намного меньше времени!”
- ✓ “Вот еще один водитель, который говорит по мобильному за рулем. Когда же примут закон, запрещающий это безобразие? Несомненно, из-за этого происходит множество аварий!”

Не во всех примерах есть цифры, но все они касаются того, что называется статистикой. На самом деле статистика связана с процессом принятия решений, проверкой теорий, сравнением групп или методов и постановкой вопросов. “Перемалывание” чисел происходит за кулисами, производя на людей глубокое впечатление, что в конечном счете сказывается на их повседневных решениях.

Глава 3

Инструменты статистики

В этой главе...

- Статистика как процесс
- Знакомство с некоторыми профессиональными терминами

В наши дни, когда балом правят цифры, у всех на языке слово “данные”, как, например, во фразах “У вас есть данные, подтверждающие эту мысль?”, “Какие данные у вас есть по этому вопросу?”, “Данные подтверждают первоначальное предположение о том, что...”, “Статистические данные свидетельствуют, что...” и “Как видно из данных...”. Но статистика касается не только данных, это целый процесс, связанный со сбором доказательств, которые бы помогли ответить на вопросы о мире в том случае, если такими доказательствами оказываются числовые данные.

В этой главе статистика показана как процесс, в котором свою роль играют цифры. Кроме того, вы познакомитесь с самыми распространенными понятиями в статистике, которые являются частью этого процесса. Поэтому когда в следующий раз вы услышите, как кто-то говорит: “Предел погрешности в этом опросе — плюс/минус 3%”, то уже будете иметь общее представление о том, что это означает.

Статистика: больше чем просто цифры

Большинство статистиков не хотят, чтобы об этой науке думали как о “всего лишь статистических данных”. Хотя все остальные и придерживаются именно такого мнения, сами статистики не считают себя обычными вычислительными устройствами, скорее, они склонны называть себя хранителями научного метода. (Конечно, в своей работе статистики зависят от специалистов из других областей, которые задают интересные вопросы, ведь не статистикой единой жив человек.) *Научный метод* (постановка вопросов, проведение исследований, сбор доказательств и их анализ, формулировка выводов) — возможно, с таким понятием вы уже сталкивались, а может быть, не понимаете, какое отношение он имеет к статистике.

Любое исследование начинается с вопроса, например такого, как указано ниже.

- ✓ Можно ли выпить так много воды?
- ✓ Каков уровень жизни в Сан-Франциско?
- ✓ Кто победит в следующих президентских выборах?
- ✓ Действительно ли травы помогают сохранить здоровье?
- ✓ Получит ли моя любимая телепередача награду в следующем году?

Ни один из этих вопросов напрямую не связан с цифрами. И все же для поиска ответа на любой из них необходимо использовать данные и статистические процессы.

Предположим, исследователь хочет установить, кто победит в следующих президентских выборах в США. Чтобы уверенно ответить на этот вопрос, ему нужно проделать следующие шаги.

1. **Установить группу респондентов.**

   В этом случае он опрашивает зарегистрированных избирателей, которые планировали голосовать на следующих выборах.

2. **Собрать данные.**

   Это сложный этап, потому что невозможно опросить всех жителей Соединенных Штатов Америки, собираются ли они голосовать, а если да, то за кого. Помимо этого, представьте, что кто-то говорит: "Да, я планирую голосовать". Но действительно ли этот человек придет на избирательный участок в день выборов? А также скажет ли он вам, за кого действительно проголосует? Что, если этот человек передумает чуть позже или проголосует за другого кандидата?

3. **Систематизировать, подытожить и проанализировать данные.**

   После того как исследователь собрал необходимые данные, их систематизация и анализ помогут ответить на вопрос. Именно это большинство людей считают основной задачей статистики.

4. **На основе собранных данных, диаграмм и графиков сделать выводы, чтобы попытаться ответить на исходный вопрос исследования.**

   Конечно, исследователь не может быть на 100% уверен, что ответ будет правильным, ведь опросили не всех жителей США. Но он может получить ответ, который *почти* на 100% будет правильным. Более того, имея выборку в 2500 человек, сделанную честно и *объективно* (т.е. когда у любого члена общества есть шанс попасть в эту группу), исследователь может получить точные результаты с погрешностью плюс/минус 2,5% (если на всех этапах исследовательский процесс проводился правильно).

Делая выводы, исследователь должен понимать, что у любого исследования есть свои границы, и поскольку всегда существует вероятность ошибки, результаты могут оказаться неверными. То, насколько исследователь уверен в своих результатах и насколько точными они кажутся, можно выразить в виде числового значения. (Подробнее о пределе погрешности см. главу 10.)

После того как исследование завершено и ответ найден, результаты, как правило, приводят к появлению еще большего числа вопросов и подталкивают к новым исследованиям. Например, если мужчины отдают предпочтение Мисс Калькуляции, а женщины поддерживают ее оппонента, то следующим будет вопрос: "Кто чаще голосует на выборах — мужчины или женщины — и какие факторы влияют на их предпочтения?".

На самом деле область статистики — это использование научного метода для того, чтобы ответить на вопросы исследователя о мире. Статистические методы используются на любом этапе хорошего исследования, как в разработке, так и при сборе данных, как при систематизации данных, так и при их анализе, во время подведения итогов, обсуждения недостатков и, наконец, при разработке следующего исследования, которое должно ответить на новые возникшие вопросы. Статистика — это больше, чем просто число, это процесс!

Знакомство с основными терминами

В каждой профессии есть свои инструменты, и статистика не исключение. Если вы представите себе статистический процесс как ряд этапов, через которые нужно пройти, чтобы ответить на вопрос, то сможете догадаться, что на каждом из этих этапов используются свои инструменты и термины (или статистический жаргон). Не волнуйтесь, никто не заставляет вас становиться экспертом в статистике и погружаться в дебри статистических терминов. Кроме того, вам не придется носить с собой тяжелый калькулятор, как это делают статистики.

Но поскольку мир все больше заполняют цифры, статистические термины все чаще появляются в средствах массовой информации и на вашем рабочем месте, поэтому, разбираясь в этих понятиях, вы почувствуете себя намного увереннее. Помимо этого, если вы читаете эту книгу, потому что хотите больше узнать о том, как подсчитывать простейшие статистические данные, прежде всего, нужно выучить основные термины. Итак, в этом разделе вы познакомитесь со статистическим жаргоном (здесь также будет указано, в каких главах можно подробнее почитать о том или ином понятии).

Совокупность

Для ответа буквально на любой вопрос о мире, который вас интересует, придется обращать внимание на определенную группу элементов (скажем, определенную группу лиц, городов, животных, образцов скальной породы, результатов экзаменов и т.д.).

- ✓ Что американцы думают о внешней политике президента?
- ✓ Какой процент зерновых в штате Висконсин был поврежден оленями в прошлом году?
- ✓ Какие шансы есть у больных раком груди, испытывающих новое экспериментальное лекарство?
- ✓ Какой процент всех тюбиков зубной пасты наполняется в соответствии с тем, что на них написано?

Во всех этих примерах задается вопрос. И в каждом случае вы можете установить конкретную группу изучаемых элементов: американцы, все зерновые в Висконсине, все пациенты с раком груди и все заполняемые тюбики зубной пасты соответственно. Группа элементов, которую вы хотите изучать, чтобы ответить на вопрос исследования, называется *генеральной совокупностью*. Но определить совокупность бывает очень сложно. В хорошем исследовании ученые дают четкое определение, а в плохом — крайне небрежное.

Вопрос о том, действительно ли младенцы лучше спят под музыку — вот хороший пример того, насколько сложно бывает установить генеральную совокупность. Как именно вы будете определять младенца? До трех месяцев? До года? Вы собираетесь изучать детей только в США или по всему миру? Результаты для детей постарше и помладше могут отличаться. То же касается американских и европейских детей и т.д.

Очень часто исследователи планируют изучать и делать выводы относительно большой совокупности, но, в конце концов, чтобы сэкономить время и деньги или просто потому, что они не видят другого выхода, они рассматривают только узкую совокупность. Это может привести к большим неприятностям при

подведении итогов. Например, представим, что преподаватель университета хочет определить, как телереклама заставляет потребителей покупать продукцию. Это исследование основывается на группе его собственных студентов, которые участвуют в опросе, чтобы получить пять дополнительных баллов к зачету. Возможно, это удобная выборка, но полученные результаты нельзя относить ни к какой другой совокупности помимо этих студентов, потому что в исследовании другие совокупности не были представлены.

Выборка

Когда вы пробуете суп, то что делаете? Помешиваете жидкость в кастрюле, набираете немного в ложку и пробуете. После этого вы делаете вывод обо всей кастрюле с супом, попробовав лишь одну ложку. Если проба сделана правильно (например, вы выбираете не только самые лучшие кусочки), то вы сможете определить истинный вкус супа. Так происходит и в статистике. Исследователи хотят выяснить что-то о совокупности, но у них нет времени или денег, чтобы изучить каждый элемент этой совокупности. Тогда что они делают? Они отбирают небольшое количество элементов совокупности, изучают их и используют эту информацию, чтобы сделать вывод обо всей совокупности. Это называется *выборкой*.

Звучит вроде бы понятно, но, к сожалению, не все так просто. Обратите внимание, что я говорю *отбирают* выборку. Похоже, ничего сложного, но так ли это на самом деле? То, как делается выборка из совокупности, имеет большое значение для результатов исследования, которые соответственно могут оказаться правильными или ни на что не годными. Для примера предположим, что вы хотите узнать мнение подростков о том, не слишком ли много времени они проводят в Интернете. Если вы рассылаете анкету по электронной почте, то полученные результаты не будут представлять мнение *всех подростков*, которые должны были быть вашей целевой совокупностью. Это будет точка зрения только тех из них, у кого есть доступ к Сети. Часто ли происходит такая путаница в статистике? Можете не сомневаться.

Одна из грубейших ошибок в статистике, связанная с неправильной выборкой, — это опросы, проводимые в Интернете. Можно назвать тысячи случаев, когда для участия в подобных опросах необходимо зайти на определенный сайт и высказать свое мнение. Даже если 50 тыс. человек в США ответят на вопросы такой анкеты, они все равно не будут представлять всю совокупность американцев. Эта группа представляет только тех людей, у которых есть доступ в Интернет, которые зашли на конкретный сайт и заинтересовались настолько, что приняли участие в опросе (что, как правило, означает, что у них уже давно сложилось твердое мнение по исследуемому вопросу).

Когда в следующий раз вы натолкнетесь на результаты исследования, поинтересуйтесь сделанной выборкой участников и спросите себя, представляет ли эта выборка всю совокупность. Не доверяйте никаким выводам, сделанным относительно более широкой совокупности, чем той, которая действительно была изучена.

Случайность

Случайная выборка — хорошая штука, благодаря ей у любого члена совокупности появляется шанс быть выбранным, при этом для отбора используется определенный случайный механизм. Другими словами, это значит, что люди не сами выказывают желание участвовать в опросе, и никому в совокупности не отдается предпочтение перед другими.

Чтобы понять, как специалисты делают такую выборку, рассмотрим пример компании *Gallup Organization*. Процесс случайной выборки начинается с компьютерного перечня всех телефонных коммутаторов в Америке и приблизительного подсчета домов, в которых есть такие коммутаторы. Компьютер использует процедуру, которая называется *случайный цифровой набор* (СЦН), чтобы наугад создавать телефонные номера, а затем из получившихся номеров отбирает определенное количество. Значит, компьютер создает перечень *всех возможных* домашних телефонных номеров в Америке, после чего выбирает ряд телефонов, по которым будут звонить представители *Gallup*. (Обратите внимание, что некоторые из этих номеров могут пока не существовать, что создает дополнительные проблемы с логистикой.)

Еще один пример случайной выборки связан со сферой производства и понятием контроля качества. Большинство производителей придерживаются жестких правил изготовления своей продукции, и ошибки в этом процессе могут стоить им денег, времени и доверия потребителей. Многие компании стараются предотвратить возникновение проблем, осуществляя контроль над всеми процессами и используя статистку, чтобы решить, правильно идет процесс или его нужно остановить. Подробнее о контроле качества и статистике см. главу 19.

К примерам *неслучайной* (другими словами, *плохой*) выборки можно отнести телефонные опросы, при которых вы высказываете свое мнение. Это не может быть случайной выборкой в полном смысле слова, потому что не у каждого представителя совокупности имеется шанс принять участие в опросе. (Если вам нужно купить газету или посмотреть телепрограмму, а потом изъявить желание написать или позвонить, можете не сомневаться — процесс выборки не случаен.) Подробнее о выборке и опросах см. главу 16.

Всякий раз, когда вы видите результаты исследования, основанного на выборке элементов, прочтите все детали, написанные мелким шрифтом, и посмотрите, встречается ли там сочетание “случайная выборка”. Если такой термин есть, читайте дальше, чтобы понять, как именно проводилась выборка, и воспользуйтесь описанным определением, чтобы установить, была ли выборка действительно случайной.

Смещение

Это слово вы слышите постоянно и, наверное, осознаете, что оно означает что-то нехорошее. Но в чем именно состоит смещение? *Смещение* — это *систематическая ошибка*, которая проявляется в процессе сбора данных и приводит к необъективным, сбивающим с толку результатам.

Смещение может произойти самыми разными способами.

- ✓ **В том, как делается выборка.** Например, вы хотите оценить, сколько подарков жители вашего округа собираются купить на Рождество. Для этого вы берете опросный лист и идете в торговый район сразу после Дня благодарения, чтобы поинтересоваться у покупателей их планами.

Таким образом, в процессе выборки вы делаете смещение. В такой выборке, скорее всего, будут представлены те заядлые покупатели конкретного торгового района, которые в отдельно взятый день оказались в толпе.

- ✓ **В том, как собираются данные.** Вопросы анкет — один из основных источников смещения. Поскольку исследователей зачастую интересуют конкретные результаты, то вопросы, которые они задают, уже изначально отражают ожидаемый ответ. Например, вопрос налогообложения в поддержку местных школ, который периодически возникает перед любым избирателем. Вопрос анкеты, звучащий: "Не думаете ли вы, что поддержка местных школ была бы прекрасной инвестицией в будущее?" содержит определенную степень смещения. С другой стороны, то же касается и вопроса: "Не надоело ли вам платить деньги из своего кармана за то, чтобы чьи-то чужие дети получали образование?". Формулировка вопроса может оказать огромное влияние на результаты. Подробнее о разработке анкет и опросов см. главу 16.

Изучая результаты опроса, которые имеют для вас большое значение или в которых вы особенно заинтересованы, узнайте, какие вопросы задавались и как именно они были сформулированы, прежде чем делать какие-либо выводы.

Данные

Данные — вот настоящий критерий, который вы получаете в ходе исследования. Большинство данных можно отнести к одной из двух групп: числовые или категорийные (дискретные) данные (подробнее см. главу 5).

- ✓ *Числовые данные* — данные, которые имеют смысл в качестве единицы измерения, например, рост человека, вес, уровень IQ или артериальное давление, количество акций, которыми он владеет, количество зубов у его собаки и все, что угодно, что можно посчитать. (Статистики называют числовые данные также *количественными данными* или *данными измерений*.)
- ✓ *Категорийные данные* представляют собой характеристики, например, пол человека, мнение, раса и даже форма пупка (с глубокой ямкой или выпуклый — не осталось уже ничего святого?). Хотя подобные характеристики тоже можно выразить цифрами (например, 1 — мужчина, 2 — женщина), но у этих цифр не будет конкретного значения. Например, их нельзя будет сложить. (В статистике такие данные еще называются *дискретными данными*.)

Не все данные равны. Если установить, как именно они собирались, это во многом поможет определить, стоит ли доверять результатам и какие выводы можно на их основе сделать.

Набор данных

Набор данных — это совокупность всех данных, полученных из выборки. Например, если вы измеряете вес пяти упаковок, которые весят 12, 15, 22, 68 и 3 фунта соответст-

венно, то эти пять чисел (12, 15, 22, 68, 3) и представляют набор данных. Но чаще наборы данных намного больше, чем в этом примере.

Статистика

Статистика — это число, которое резюмирует данные, собранные из выборки. Для резюмирования данных используются разные статистики. Например, это можно сделать в виде процентного отношения (у 60% семей, выбранных в США, есть больше двух автомобилей), среднего (средняя цена дома в этой выборке составляет...), медианы (медиана зарплат 1000 специалистов по компьютерам в этой выборке равна...) или процентиля (вес вашего малыша в этом месяце соответствует 90-му процентилю, исходя из данных осмотра 10 000 детей...).

Конечно же, не все статистические показатели честны или справедливы. Сам по себе факт, что вам называют статистический показатель, не гарантирует того, что он будет научным или законным! Возможно, вы слышали, как говорят: “Цифры не лгут, но лжецы выдумывают цифры”.

Статистика использует не данные генеральной совокупности, а данные выборки. Если вы собираете данные со всей совокупности, этот процесс называется *перепись*. Если затем вы выразите всю информацию, полученную в ходе переписи, одним числом, то это число будет *параметром*, а не статистикой. Чаще всего исследователи наоборот пытаются оценить параметры с помощью статистики. Например, задача Бюро переписей США состоит в подсчете общего числа жителей Соединенных Штатов Америки, для чего эта организация и проводит перепись. Но из-за проблем с перемещениями при выполнении такой изнурительной задачи (например, общение с бездомными), в конечном итоге данные переписи можно назвать только приближенными, и они округляются в сторону большего значения, чтобы учесть тех людей, которые не попали в перепись. Длинную анкету переписи заполняет только случайная выборка семей, затем Бюро переписей США использует эту информацию, чтобы сделать выводы относительно всего населения (не заставляя каждого жителя заполнять такую анкету).

Среднее значение

Среднее значение — это самый популярный статистический показатель, который используется для измерения центра или середины в наборе числовых данных. Среднее значение — сумма всех чисел, деленная на общее количество чисел. Подробнее об этом см. главу 5.

Среднее значение может быть не совсем объективным отражением данных, потому что на него очень просто повлиять с помощью *выбросов* (очень большого или очень маленького значения в наборе данных, которое не является типичным).

Медиана

Медиана — еще один способ измерить центр набора числовых данных (помимо старого доброго среднего значения). В статистике медиана очень похожа на разделяющую

полосу между частями магистральной автодороги. На шоссе это середина дороги, по обе стороны такой разделяющей линии лежат равные половины дороги. В наборе числовых данных *медиана* — это точка, от которой вверх и вниз находится половина данных. Таким образом, медиана — это середина набора данных. Подробнее об этом см. главу 5.

Когда вы в следующий раз наткнетесь на сообщение о среднем значении, проверьте, указана ли там и медиана. Среднее значение и медиана — это два разных выражения середины в наборе данных, поэтому они очень часто могут характеризовать данные совершенно по-разному.

Стандартное отклонение

Слышали ли вы когда-нибудь, как об определенном результате говорили, что он представляет собой "два стандартных отклонения от среднего значения"? Все чаще люди хотят доказать, насколько важны полученные результаты, и количество стандартных отклонений от среднего — один из способов сделать это. Но что *именно* называется стандартным отклонением?

Стандартное отклонение — это способ, с помощью которого статистики измеряют степень изменчивости (разброса или рассеяния) чисел в наборе данных. Как видно из самого термина, это среднее (или типичное, стандартное) отклонение (т.е. расстояние) от среднего значения. Значит, грубо говоря, стандартное значение — это среднее расстояние от среднего значения. Подробнее о вычислениях см. главу 5.

Стандартное отклонение используется также для обозначения того, в какой области должно быть, так сказать, наибольшее количество данных по сравнению со средним значением. Например, очень часто около 95% данных будут находиться в пределах двух стандартных отклонений от среднего значения. (Этот результат называется эмпирическим правилом. Подробнее об этом см. главу 8.)

Формула стандартного отклонения (s) выглядит так:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}},$$

где

n = количество значений в наборе данных (объем выборки);
$\bar{x}$ = среднее всех значений;
x = каждое значение в наборе данных.

Подробнее о подсчете стандартного отклонения см. главу 5.

Стандартное отклонение — важный статистический показатель, но когда сообщаются статистические результаты, о нем часто забывают. Без этого показателя вы видите только часть информации относительно данных. Статистики любят рассказывать историю о человеке, одной ногой стоящем в ведре с ледяной водой, а второй — в ведре с кипятком. В среднем несчастный должен чувствовать себя отлично! Но вспомните о разнице двух температур для каждой его ноги. Например, средняя цена жилого дома ничего не скажет вам о разбросе цен, с которым вы столкнетесь, когда действительно будете подыскивать себе жилье. Средняя зарплата может не в полной мере отражать реальное положение дел в вашей компании, если разброс окладов очень большой.

Не довольствуйтесь наличием только среднего значения — обязательно также выясняйте стандартное отклонение. Без него вы не сможете узнать, каким может быть разброс значений. (Например, если речь идет о начальном окладе, то это может быть очень важно!)

Процентиль

Возможно, вы уже слышали такое слово. Если вам приходилось выполнять какой-нибудь стандартизованный тест, то вы знаете, что ваш результат представляется в сравнении с другими людьми, которые проходили этот тест. Скорее всего, такой сравнительный критерий был выражен в виде процентиля. *Процентиль* по отношению к определенному результату — это процент значений в наборе данных, которые находятся ниже этого результата. Например, если вам говорят, что ваш результат соответствует 90-му процентилю, это значит, что 90% других участников теста набрали меньше баллов, чем вы (а 10% — больше вас). Подробнее о процентилях см. главу 5.

Процентили используются самым разным образом для сравнения и определения *сравнительного положения* (т.е. того, как определенные данные соотносятся с остальными в группе). Например, вес младенцев часто показывают в процентилях. Кроме того, процентили используются компаниями для определения своего положения среди других компаний в том, что касается продаж, прибыли, удовлетворения запросов потребителей и т.д.

Нормированная переменная

Нормированная переменная — отличный способ представить результаты в перспективе, при этом не приводя слишком много подробностей, — как раз то, что любят средства массовой информации. *Нормированный параметр* представляет собой количество стандартных отклонений от среднего значения (независимо от того, каким на самом деле является стандартное отклонение или среднее значение).

Для примера представим себе, что Боб недавно сдал стандартный тест для десятого класса и набрал 400 баллов. Что это значит? Для вас, возможно, ничего, потому что вы ни с чем не можете соотнести эту цифру. Но зная, что нормированный параметр Боба в этом тесте равен +2, все становится ясным. (Браво, Боб!) Теперь предположим, что его нормированный параметр –2. В таком случае это не очень хорошо (для Боба), потому что это значит, что показатель Боба на два стандартных отклонения *ниже* среднего значения.

Формула нормированной переменной такова:

$$\text{нормированная переменная} = \frac{\left(\text{исходный показатель} - \bar{x}\right)}{s},$$

где

$\bar{x}$ — среднее значение всех переменных;

s — стандартное отклонение всех переменных.

Подробнее о подсчете и интерпретации нормированных переменных см. главу 8.

Гауссово (нормальное) распределение (или колоколообразная кривая)

Когда числовые данные систематизированы, они часто располагаются от меньшего к большему, делятся на группы и представляются в виде графиков и диаграмм, чтобы можно было изучить форму или распределение данных. Самый распространенный вид распределения данных называется *колоколообразной кривой*, при которой большая часть данных концентрируется возле среднего значения, а чем дальше от середины, тем меньше наблюдается данных. На рис. 3.1 показана колоколообразная кривая (обратите внимание, что форма кривой напоминает старинный колокол).

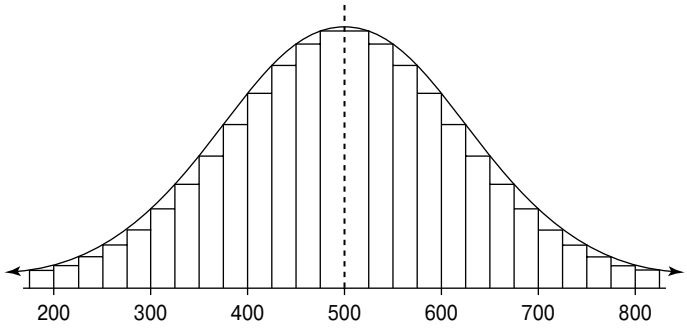


Рис. 3.1. Колоколообразная кривая

В статистике принято еще одно название такой кривой, когда существует множество возможных значений данных — *гауссово (нормальное) распределение*. Такое распределение используется для описания данных, которые придерживаются колоколообразной модели, в том числе и то, какой диапазон значений ожидается или где располагаются отдельные значения по отношению к другим. Например, если данные имеют нормальное распределение, то можно предположить, что бо́льшая часть данных находится в пределах двух стандартных отклонений от среднего значения. Поскольку у каждой отдельной совокупности данных имеется собственное среднее значение и стандартное отклонение, то существует бесконечное число различных нормальных распределений, каждое со своим средним значением и собственным стандартным отклонением. Подробнее о нормальном распределении см. главу 8.

Гауссово распределение используется также для того, чтобы определить точность многих статистических данных, в том числе среднего значения, при помощи важного понятия в статистике, которое называется *центральной предельной теоремой*. Эта теорема позволяет установить, насколько будет варьироваться среднее значение вашей выборки, без необходимости сравнивать его со средними значениями других выборок (слава Богу!). По сути, она гласит, что среднее значение выборки имеет нормальное распределение независимо от того, на что похоже распределение исходных данных (при условии, что размер выборки достаточно велик). Подробнее о центральной предельной теореме (известной в статистике как “королева всех статистических данных”) см. главу 9.

Если набор данных обладает нормальных распределением и вы нормируете все данные, чтобы получить стандартные показатели, то такие нормированные величины называются *Z*-значениями. У *Z*-значений наблюдается так называемое стандартное нормальное распределение (или *Z*-распределение). *Стандартное нормальное распределение* — это особое гауссово распределение, при котором среднее значение равно 0, а стандартное отклонение равно 1. Стандартное нормальное распределение полезно для изучения данных и определения таких показателей, как процентили или процентное отношение данных, которые находятся в пределах двух значений. Поэтому если исследователи устанавливают, что данным присуще гауссово распределение, то они, как правило, прежде всего нормируют данные (преобразовав все значения в *Z*-величины), а затем используют стандартное нормальное распределение для более подробного изучения данных.

Эксперименты

Эксперимент — это исследование, при котором испытуемые и их окружение находятся под определенным контролем (например, соблюдают определенную диету, принимают установленную порцию лекарства или плацебо или бодрствуют в течение указанного периода времени). Цель большинства экспериментов состоит в том, чтобы установить причинно-следственную взаимосвязь между двумя переменными (такими как потребление алкоголя и нарушение зрения). Вот вопросы, ответы на которые пытаются найти исследователи в ходе экспериментов.

- ✓ Помогает ли употребление цинка сократить продолжительность простуды? Некоторые исследования утверждают это.
- ✓ Влияет ли форма и положение подушки на качество ночного сна? Центр *Emory Spine* в Атланте говорит: “Да”.
- ✓ Влияет ли высота каблука на удобство при ходьбе? Результаты исследования, проведенного в *UCLA*, свидетельствуют, что каблук высотой в дюйм полезнее, чем плоская подошва.

В этом разделе вы найдете более подробную информацию о том, как проводятся (или должны проводиться) экспериментальные исследования, а глава 17 полностью посвящена этой теме. Пока же давайте просто сосредоточимся на основных понятиях, связанных с экспериментами.

Испытуемая группа и контрольная группа

В ходе большинства экспериментов исследователи стремятся установить, оказывает ли определенное условие (или важный фактор) какое-либо воздействие на результат. Например, помогает ли цинк сократить продолжительность простуды? Испытуемые, отобранные для участия в эксперименте, как правило, делятся на две группы — испытуемую и контрольную. *Испытуемая группа* состоит из тех, кто подвергается воздействию, которое, как ожидается, влияет на результат (в данном случае, они принимают цинк). В *контрольную группу* входят люди, не подвергающиеся такому воздействию или получающие стандартное, хорошо известное лечение, результаты которого будут сравниваться с результатами нового лекарства (например, в случае с исследованием цинка это может быть витамин С).

Плацебо

Плацебо — ненастоящее лекарство, например, сахарная пилюля. Его часто дают членам контрольной группы, при этом испытуемые не знают, принимают они какое-то лекарство (например, цинк) или не принимают ничего. Плацебо дают контрольной группе, чтобы контролировать явление, известное как эффект плацебо, когда пациенты, получающее ожидаемое лечение в виде таблетки (даже если она будет целиком из сахара), сообщают о результате, будь то положительном ("Да, мне уже лучше") или отрицательном ("Ой, у меня начинает кружиться голова"). Это связано с психологическими причинами. Без плацебо исследователи не могли бы наверняка сказать, что результаты были получены благодаря действительному эффекту, ведь некоторые (или все) испытуемые могли попасть под влияние эффекта плацебо.

Слепой или двойной слепой эксперимент

Слепым называется эксперимент, при котором участники не знают, к какой группе они относятся — испытуемой или контрольной. В случае с цинком было бы использовано плацебо, по внешнему виду напоминающее таблетку с цинком, и пациентам не сообщили бы, какое именно лекарство они принимают. С помощью слепого эксперимента исследователи пытаются устранить любое смещение в свидетельствах испытуемых.

Двойной слепой эксперимент призван контролировать потенциальное смещение со стороны как пациентов, так и самих исследователей. Ни пациенты, ни исследователи, собирающие данные, не знают, кто из испытуемых получает лекарство, а кто — нет. Двойной слепой эксперимент — самый лучший вариант, потому что хотя сами исследователи и заявляют о своей абсолютной объективности, но зачастую они заинтересованы в конкретных результатах — в противном случае им вообще не нужно было бы проводить исследование!

Опросы

Опрос — это метод оценивания, который чаще всего используется для выяснения общественного мнения наряду с некоторой демографической информацией, имеющей отношение к делу. Поскольку многие политики, маркетологи и специалисты в других областях хотят "держать руку на пульсе американского общества" и пытаются выяснить, что думает и чувствует средний американец, то многим теперь кажется, что им уже просто негде скрыться от настойчивых просьб принять участие в различных опросах. Более того, вы сами тоже, наверное, получали множество подобных просьб, оставались глухи к ним, просто выбрасывая присланные анкеты в мусорное ведро или говоря "нет", когда к вам обращались по телефону.

При правильной организации опрос может оказаться действительно информативным. Опросы используются, чтобы узнать, какие телепрограммы нравятся американцам (и другим людям), что потребители думают о покупках в Интернете и нужна ли Соединенным Штатам Америки система противоядерной обороны. Многие компании используют опросы, чтобы оценить, насколько довольны их продукцией потребители, чтобы выяснить, какая продукция им нужна, и установить, кто именно покупает их продукцию. Телеканалы используют опросы, чтобы узнать непосредственную реакцию на новые программы и события, а производители кинофильмов — чтобы определить, какой финал лучше подойдет для их новых работ.

Если бы мне нужно было описать общее положение опросов в современных средствах массовой информации всего одним словом, я бы сказала скорее не *качество*, а

количество. Другими словами, плохих опросов очень много. Чтобы выяснить, правильно ли был проведен опрос, достаточно задать несколько основных вопросов. Подробнее об этом речь пойдет в главе 16.

Оценивание

Одна из самых популярных сфер использования статистики — это приближенная оценка чего-либо (статистики используют термин *оценивание*), как это видно на следующих примерах.

- ✓ Каков доход средней семьи в Америке?
- ✓ Каково процентное отношение семей, которые в этом году смотрели церемонию вручения награды "Оскар"?
- ✓ Какова предполагаемая средняя продолжительность жизни ребенка, который рождается в наши дни?
- ✓ Насколько эффективно новое лекарство?
- ✓ Чем отличается состав воздуха на сегодняшний день по сравнению с воздухом, которым человек дышал десять лет назад?

Чтобы ответить на любой из этих вопросов, потребуется определенная числовая оценка, но получение точной и объективной оценки может оказаться довольно сложной задачей. Ниже рассказывается об основных составляющих этого процесса. Подробнее о проведении и интерпретации оценивания см. главу 11.

Предел погрешности

Возможно, вы уже слышали такую фразу: "Предел погрешности данного опроса составляет плюс/минус 3 процента". Что это значит? Все опросы основаны на информации, полученной от выборки, а не из целой совокупности. Поэтому непременно будет наблюдаться определенная погрешность — не в смысле арифметической ошибки (хотя и такое возможно), а в смысле *ошибки выборки* или ошибки, которая непременно произойдет, потому что исследователи опрашивают не всех. Предел погрешности должен определять максимальное значение, на которое результаты выборки, как ожидается, отличаются от результатов опроса всей совокупности. Поскольку результаты большинства опросов можно представить в виде процентов, то и предел погрешности чаще всего показывается таким же образом.

Но как понимать предел погрешности? Предположим, вам известно, что 51% опрошенных говорят, что на предстоящих выборах собираются голосовать за Мисс Калькуляцию. Проецируя эти результаты на всю совокупность людей, обладающих правом голоса, вам пришлось бы добавить и отнять предел погрешности и установить диапазон возможных результатов, чтобы уверенно говорить о том, что между вашей выборкой и всей совокупностью нет существенной разницы. В таком случае (предположим, что предел погрешности составляет плюс/минус 3 процента) вы могли бы быть практически уверены, что, исходя из результатов опроса выборки, от 48 до 54% населения проголосуют на выборах за Мисс Калькуляцию. Значит, Мисс Калькуляция может получить немного больше или немного меньше большинства голосов, т.е. может либо выиграть, либо проиграть на выборах. В последние годы такая ситуация повторяется довольно часто, когда СМИ хотят сообщить о результатах в Ночь выборов, но, исходя из результатов опроса, исход выборов "пока еще трудно предсказать". Подробнее о пределе погрешности см. главу 10.

Предел погрешности измеряет точность, а не степень возможного смещения. Результаты, которые кажутся научными и точными в числовом выражении, ничего не означают, если были собраны необъективно.

Доверительный интервал

Соединив полученную оценку с пределом погрешности, вы получите *доверительный интервал*. Например, представим, что в среднем на дорогу до работы каждый день у вас уходит 35 минут, а предел погрешности составляет плюс/минус 5 минут. Вы предполагаете, что в среднем дорога до работы займет от 30 до 40 минут. Это и есть доверительный интервал. В нем учитывается тот факт, что результаты опроса выборки могут отличаться, и указывается, какое именно различие ожидается. Подробнее о доверительном интервале см. главу 11.

Одни доверительные интервалы шире других (и это плохо, потому что это означает меньшую точность). На ширину доверительного интервала влияют несколько факторов, например, размер выборки, степень изменчивости исследуемой совокупности и то, насколько вы можете быть уверены в своих результатах. (Большинство исследователей довольствуются степенью уверенности 95%.) Подробнее о факторах, влияющих на доверительный интервал, см. главу 12.

В научных исследованиях используются различные доверительные интервалы, в том числе доверительный интервал для средних значений, пропорций, разницы двух средних значений или пропорций или парных разниц. Подробнее о самых распространенных критериях для проверки гипотезы см. главу 13.

Вероятность или шансы

Вероятность — это степень возможности того, что событие произойдет. Другими словами, вероятность — это шанс того, что что-то случится. Например, если шансы того, что завтра будет дождь, равны 30%, то, скорее всего, дождя завтра не будет, но все же шанс остается 3 из 10. (Учитывая такую вероятность, возьмете ли вы завтра с собой зонтик?) Шанс дождя в 30% также означает, что за многие и многие дни с теми же погодными условиями, что и завтра, дождь шел в 30% случаев.

Вероятность подсчитывается самыми разными способами.

- ✓ Для выведения чисел используется математика (например, определение ваших шансов на победу в лотерее или установление иерархии карт на руках игрока в покер).
- ✓ Собираются данные, и вероятность определяется на их основе (например, предсказание погоды).
- ✓ Сложные вычисления и компьютерное моделирование используются для того, чтобы предсказать поведение и возможность появления в будущем природных явлений (например, ураганов и землетрясений).

Законы вероятности часто действуют вопреки вашей интуиции и убеждениям относительно того, что, как вам кажется, произойдет (именно поэтому казино по-прежнему процветают). Подробнее о вероятности см. главу 6.

Шансы и вероятность — немного разные понятия. Лучше всего объяснить эту разницу можно на примере. Предположим, что вероятность того, что определенная лошадь выиграет в скачках, составляет 1 из 10. Это значит, что вероятность для нее победить равна 1 на 10 или 1/10 или 0,10. Вероятность показывает возможность победы. Теперь каковы же шансы у этой лошади? 9 к 1. Это объясняется тем, что, по сути, шансы — это отношение возможности проиграть к возможности выиграть. У этой лошади 9 из 10 вероятность проиграть и 1 из 10 вероятность победить в скачках. Возьмем 9/10 и 1/10, отбросим десятые в знаменателе, и у вас останется 9/1, что на языке шансов читается как "9 к 1". Подробнее об азартных играх см. главу 7.

Закон средних чисел

Возможно, вы уже слышали о законе средних чисел. Может быть, это был местный баскетбольный обозреватель, который сокрушался, что команда бросила вызов судьбе, выиграв 50 игр и проиграв только 12 за первые три месяца сезона, а теперь начала проигрывать, уступая действию закона средних чисел. А может, такое выражение прозвучало в контексте азартных игр ("Наверняка закон средних чисел скоро мне отомстит — пока что я постоянно выигрываю!"). Но что такое закон средних чисел и правильно ли используют этот термин люди?

Закон средних чисел — это правило вероятности. Он гласит, что в долговременной перспективе результаты будут усреднены до своих ожидаемых значений, но в краткосрочной перспективе никто не знает, что случится. Например, в казино все игры устроены так, что у заведения всегда остается чуть больше шансов на выигрыш. Это значит, что в долгосрочной перспективе пока люди продолжают играть, казино в среднем всегда будут немного впереди. Конечно, иногда будут победители, именно благодаря этому люди играют дальше, надеясь оказаться в числе счастливчиков. Но в общем количество проигравших перевешивает число победителей (не говоря уже о том, что очень часто те, кто выигрывают большие суммы, зачастую просто возвращают свои деньги в игру и снова их проигрывают). Подробнее о законе средних чисел см. главу 7.

Проверка гипотезы

Проверка гипотезы — это термин, с которым вы, вероятно, не так часто сталкивались в процессе изучения чисел и статистических данных. И все же очевидно, что проверка гипотезы играет в вашей жизни и работе важнейшую роль просто потому, что она крайне важна в промышленности, медицине, сельском хозяйстве, государственной деятельности и множестве других областей. Всякий раз, когда кто-то говорит о том, что полученные им результаты показывают "статистически значимую разницу", перед вами — результаты проверки гипотезы. По сути, проверка гипотезы — это статистическая процедура, при которой данные собираются и оцениваются относительно определенного утверждения о совокупности. Например, если сеть доставки пиццы утверждает, что пицца будет доставлена в течение 30 минут после заказа, вы можете проверить, правда ли это, сделав случайную выборку доставок за определенный период и проверив продолжительность доставки.

Поскольку вы основываете решение не на всей генеральной совокупности, а на выборке из нее, проверка гипотезы может иногда подтолкнуть вас к неправильным выводам. Но у вас есть только статистические данные, которые, при правильной организации, могут подвести вас максимально близко к истине, даже если она достоверно и неизвестна. Подробнее об основе проверки гипотезы см. главу 14.

В научных исследованиях проводятся самые разные проверки гипотез, в том числе *t*-тесты, парные *t*-тесты и проверки пропорций и средних значений для одной или нескольких совокупностей. Подробнее о самых распространенных вариантах проверки гипотезы см. главу 15.

P-значение

Проверка гипотезы используется для того, чтобы подтвердить или опровергнуть утверждение, сделанное относительно совокупности. По сути, рассматриваемое утверждение называется *нулевой гипотезой*. Доказательства в этом судебном разбирательстве — это ваши данные и связанная с ними статистика. Любая проверка гипотезы в конечном итоге использует *p*-значение для того, чтобы взвесить значимость доказательств (т.е. то, что вам говорят о совокупности представленные данные). *P-значение* — это число от 0 до 1, которое отражает значимость данных, используемых для оценки нулевой гипотезы. Если *p*-значение невелико, то доказательства против нулевой гипотезы убедительны. Большое *p*-значение указывает на слабые доказательства. Например, если пиццерия обязуется доставить пиццу менее чем за 30 минут (это нулевая гипотеза), а в вашей случайной выборке из 100 доставок в среднем доставка происходила за 40 минут (что составляет более 2 стандартных отклонений от предполагаемого среднего времени доставки), то *p*-значение для этого теста невелико, и вы могли бы сказать, что у вас есть убедительные доказательства против утверждения пиццерии.

Статистически значимый

Всякий раз, когда для проверки гипотезы собираются данные, исследователь, как правило, стремится получить значимый результат. Обычно это значит, что он нашел что-то необычное. (Исследование, просто подтверждающее то, что уже было известно, к сожалению, редко попадает на первые страницы газет.) *Статистически значимый* результат — это результат, у которого очень мала вероятность случайного повторения. Эта вероятность отражена в *p*-значении.

Например, если установлено, что некое лекарство эффективнее борется с раком груди, чем применяемое в настоящее время, то исследователи говорят, что новое лекарство показывает статистически значимое улучшение уровня выживаемости пациентов, страдающих раком груди (или что-то подобное). Это значит, что, исходя из полученных данных, разница в результатах между пациентами, использующими новое лекарство, и теми, кто проходил привычный курс лечения, настолько велика, что ее нельзя назвать простым совпадением.

Иногда выборка не представляет всю совокупность (просто случайно), и такие результаты приводят к неверным выводам. Например, положительный эффект, который засвидетельствовали участники выборки, принимающие новое лекарство, может оказаться всего лишь неожиданным везением. (Представим на минутку, что вам наверняка известно, что данные не были сфабрикованы, подделаны или преувеличены.) Прелесть медицинского исследования состоит в том, что как только появляется пресс-релиз, в котором сказано, что результаты имеют большое значение, все сразу же хотят повторить эти результаты. А если повторить их невозможно, это может означать, что исходные результаты были по какой-то причине неверными. К сожалению, пресс-релизы, объявляющие о “крупном изобретении”, все чаще встречаются в средствах массовой информации, а вот последующие исследования, опровергающие эти результаты, никогда не появляются на первой полосе.

Один статистически значимый результат не должен подтолкнуть к скоропалительным выводам. В науке важно не отдельное замечательное исследование, а ряд доказательств, которые собираются постепенно, наряду с разнообразными последующими, хорошо разработанными исследованиями. Возьмите любое крупное изобретение, о котором вам говорят, и подождите, пока не будет проведена проверка результатов, прежде чем использовать информацию, полученную в ходе единственного исследования, для принятия важных решений в жизни.

Корреляция и причинность

Из всех неправильно понимаемых вопросов в статистике самая большая проблема — это ошибочное использование понятий корреляции и причинности.

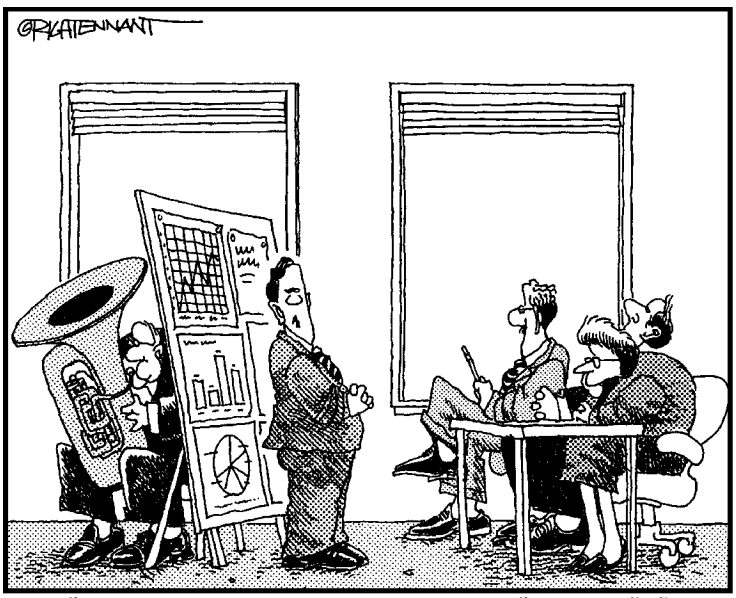
Корреляция означает, что между двумя числовыми переменными наблюдается определенная линейная взаимосвязь. Например, частота трелей сверчка зависит от температуры: когда на улице холодно, сверчки поют не так часто. (И это действительно так!) Еще один пример корреляции — набор кадров в полицию. Часто оказывается, что количество преступлений (на душу населения) связано с количеством полицейских на данной территории. Если территорию патрулирует больше полицейских, преступлений совершается меньше, и наоборот. Но между на первый взгляд не связанными между собой событиями тоже может наблюдаться корреляция. Один из примеров — потребление мороженого (пинты на человека) и количеством убийств в определенном районе. Если большее число полицейских влияет на уровень преступности, это понятно, но какое отношение к преступлениям имеет мороженое? В чем разница? А разница в том, что в случае с корреляцией между двумя переменными x и y существует связь. Если речь идет о причинности, то закономерна фраза: “Изменение в x повлечет за собой изменение в y”. Слишком часто во время исследований, в СМИ или в понимании статистики общественностью подобная взаимосвязь делается там, где ее на самом деле нет. Когда это можно сделать? Когда проводится хорошо разработанный эксперимент, в котором отсутствуют любые факторы, которые могут повлиять на результат. Подробнее о корреляции и причинности см. главу 18.

Часть II

Основы решения задач

The 5th Wave

Рич Теннант



"Приготовься, по-моему, они уже "поплыли"."

В этой части...

Решение задач — это грязная работа, но кто-то же должен ее делать. Почему бы не вы? Даже если вы не слишком дружите с числами, а вычисления — не ваш конек, поэтапный подход, изложенный в этой части, возможно, окажется именно тем, что вам нужно для того, чтобы в дальнейшем вы уверенно и с пониманием работали со статистическими данными.

В этой части вы познакомитесь с основами "перемалывания" чисел, от составления и интерпретации диаграмм до определения и понимания средних значений, медиан, стандартных отклонений и многого другого. Кроме того, вы научитесь критически смотреть на статистическую информацию других и докапываться до истины, скрытой за данными.

Глава 4

Получаем картину: диаграммы и графики

В этой главе...

- Варианты представления данных, с которыми вы можете столкнуться
- Понимание цели, скрытой за картинкой
- Интерпретация и критика диаграмм и графиков
- Внимание: неправильное представление данных

Кто-то однажды сказал, что картина стоит тысячи слов. В статистике изображение может стоить тысячи единиц данных — конечно, до тех пор, пока это изображение сделано правильно. Визуальное представление данных, например, диаграммы и графики, очень часто используется в повседневной жизни. Они могут рассказывать обо всем, начиная с результатов выборов, классифицированных по всем возможным характеристикам, и заканчивая тем, как изменилось состояние фондового рынка за прошедшие годы. Современное общество — это общество заведений быстрого питания и быстрой информации, все хотят знать итог и не терять время на детали. Главная сфера применения статистики — сжатие информации до максимальных пределов, а визуальное представление данных — естественный способ сделать это. Но действительно ли благодаря ему вы видите полную картину того, что происходит с данными? Все зависит от качества представления и его изначальной цели. Изображения могут сбивать с толку (иногда намеренно, а иногда случайно), и не все варианты визуального представления данных, которые вы встречаете, будут правильными. Эта глава поможет вам лучше понять использование диаграмм и графиков в средствах массовой информации и на работе, а также объяснит, как нужно читать и понимать эти варианты представления данных. В этой главе вы также найдете некоторые советы относительно того, как лучше оценивать визуальное представление информации и замечать те случаи (столь многочисленные!), которые запросто могут ввести в заблуждение.

Связь изображения со статистикой

Основная задача визуального представления данных — донести определенную идею, сделать это четко и эффективно, а также правильно. Диаграмма или график, например, используются для того, чтобы придать особое значение определенной характеристике данных, подчеркнуть произошедшие за некоторый период времени изменения, сравнить и противопоставить мнения или демографические данные, или же продемонстрировать взаимосвязь между отдельными блоками информации. Визуализация данных “прерывает” статистический рассказ, который автор хочет донести до вас с помощью этих данных, с тем, чтобы читатель быстрее вник в суть дела и пришел к определенному выводу.

Для этого визуальное представление данных крайне эффективно: при правильном использовании оно может быть информативным и действенным, а при неправильном оно становится деструктивным и сбивает с толку.

Визуальное представление данных может повлиять на вашу жизнь в большом и малом, поэтому критическое восприятие таких изображений и понимание того, что они говорят, а о чем умалчивают, поможет вам стать искушенным потребителем статистических данных. Вы захотите познакомиться с разными вариантами визуального представления данных, с которыми можете столкнуться, и понять, как они используются в средствах массовой информации и на работе.

Исследователи и журналисты по-разному представляют две основных разновидности данных — категорийные, которые отражают качества или характеристики (например, пол или политическая партия), и числовые, отражающие измеряемые количества (как высота или доход).

Самые распространенные варианты визуального представления данных таковы.

- ✓ Круговые диаграммы (см. раздел "Кусок круговой диаграммы").
- ✓ Столбиковые диаграммы (см. раздел "Поднимаем планку столбиковой диаграммы").
- ✓ Таблицы (см. раздел "Вносим данные в таблицу").
- ✓ Временны*е* диаграммы, известные также как *линейные графики* (см. раздел "В ногу со временными диаграммами").

Для визуального представления числовых данных чаще всего используются таблицы. Кроме того, для отображения числовых данных можно широко использовать и гистограммы (хотя это делается нечасто), поэтому я включаю их в раздел "Отображение данных на гистограмме".

В этой главе приводятся примеры каждого варианта визуального представления данных, изложены некоторые мысли автора относительно их интерпретации, здесь же вы найдете советы для критического рассмотрения каждого типа.

Кусок круговой диаграммы

Круговая диаграмма — один из самых широко используемых вариантов визуального представления данных, потому что его легко читать и понимать. Скорее всего, вы уже встречали такие диаграммы — они кажутся такими простыми. Разве может быть какая-то ошибка в безобидной круговой диаграмме? Ответ — "Может".

Круговая диаграмма берет категорийные данные и делит их на группы, показывая процентное отношение единиц в каждой группе. Поскольку круговая диаграмма представлена в виде окружности, или пирога, то "куски" этого пирога, означающие каждую группу, можно легко сравнить и противопоставить друг другу. А поскольку каждый элемент в группе относится только к одной категории, то сумма всех кусков пирога должна составлять 100% или около того (если учесть небольшую погрешность при округлении).

Подсчет личных затрат

Когда вы тратите деньги, то на что именно? Какие три основных статьи ваших расходов? Согласно данным, полученным Бюро трудовой статистки США, в 1994 году тремя основными источниками потребительских расходов были коммунальные услуги (32%), транспорт (19%) и питание (14%). На рис. 4.1 эти результаты показаны в виде круговой

диаграммы. (Обратите внимание, что на этой диаграмме категория “Другое” изображена немного большей. Но в данном случае определить, какие именно статьи расходов отнести к этой категории, было бы очень трудно, потому что для разных людей возможны самые разные варианты.)

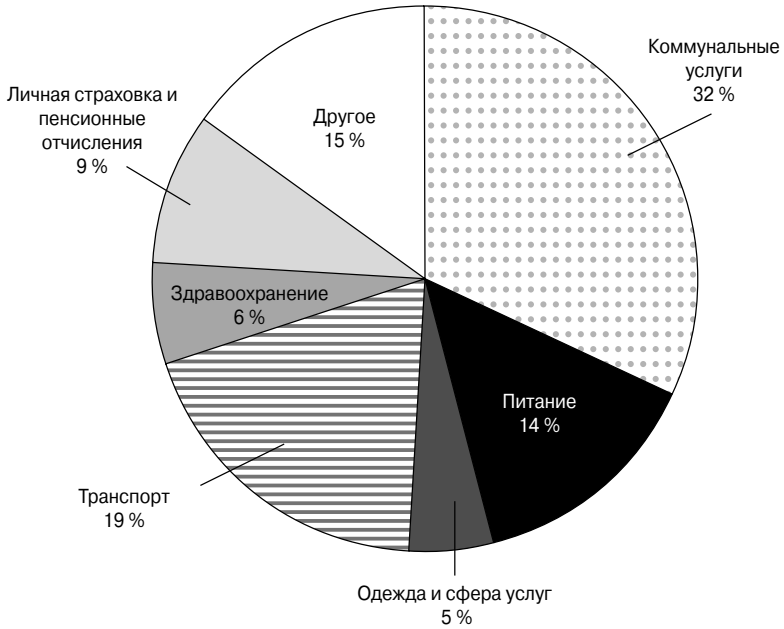


Рис. 4.1. Общие потребительские расходы в США в 1994 году

Каким же образом американское правительство получило такие данные? Благодаря так называемому Опросу потребительских расходов. Многие федеральные агентства обязаны собирать данные (часто с помощью опросов) и распространять результаты в виде письменных отчетов. (Правительство США — хороший источник информации о многих аспектах повседневной жизни в Соединенных Штатах Америки.)

Оценка лотереи

Государственные лотереи приносят огромный доход, а также возвращают бо́льшую часть полученных денег. При этом часть дохода используется на призы, а часть выделяется на проведение государственных программ, например, поддержку образования. Но откуда приходят эти деньги? На рис. 4.2 представлена секторная диаграмма, показывающая разновидности розыгрышей и процентное отношение дохода, который поступает от лотереи в Огайо.

На этой диаграмме вы видите, что большая часть доходов от продажи лотерейных билетов в Огайо (49,25%) поступает от мгновенных розыгрышей (в которых на билете нужно стереть защитный слой). Остальной доход идет от различных видов розыгрышей, в которых игроки выбирают ряд цифр и выигрывают, если их цифры совпадают с теми, которые будут выбраны лотереей. Но почему мгновенные розыгрыши приносят такой большой доход? Одна из возможных причин состоит в том, что выплаты по мгновенным розыгрышам проводятся часто, а выигрыши не слишком велики. Кроме того, при таком

виде лотереи вы видите результат сразу же, а в других видах приходится ждать непосредственно самого розыгрыша. С другой стороны, возможно, людям просто нравится стирать защитный слой в нужных ячейках!

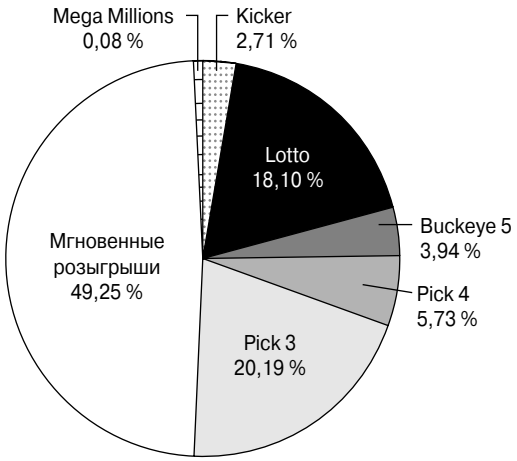


Рис. 4.2. Анализ доходов лотереи Огайо

Обратите внимание, что эта круговая диаграмма не показывает, *сколько* денег поступает, а только *процентное отношение* средств, полученных от каждого вида розыгрыша. Другими словами, вы знаете, как поделен пирог, но самое главное — неизвестно, насколько он велик. Именно это может вас заинтересовать как потребителя этой информации. Примерно половина денег (49,25%) поступает от мгновенных розыгрышей. Означает ли это сумму в миллион долларов, два миллиона, десять или больше? Круговая диаграмма на рис. 4.2 не сообщает этой информации, и вы не можете ответить на такой вопрос самостоятельно, не зная общего дохода в долларах. Но мне удалось выяснить эту информацию по другой диаграмме, представленной лотереей в Огайо: общий доход в 2002 году от продажи лотерейных билетов был равен “1 983,1 млн долларов” — или другим словами 1,983 млрд долларов. Поскольку 49,25% поступает от мгновенных розыгрышей, это соответствует доходу в 976 676 750 долларов за более чем 10-летний период. Долго же нужно вытирать билеты!

Круговые диаграммы часто показывают деление или процентное отношение от общего, приходящееся на каждую группу или категорию. Но они не говорят, каково общее число в каждой группе в том, что касается исходных единиц (количества долларов, людей и т.д.). В результате такого подхода теряется информация, за данными вы уже не можете увидеть полной картины и думаете, каким же было общее, которое потом делилось на части. Всегда можно перейти от количества к процентам, но, не зная общей суммы, нельзя перейти от процентов назад к исходному количеству. В случае с результатами опроса, отсутствие такой информации может оказаться серьезной проблемой. Очень часто на круговых диаграммах показано процентное отношение людей, давших определенный ответ на поставленный вопрос, но неизвестно, сколько всего человек принимало участие в опросе — а это важнейшая информация, необходимая для того, чтобы оценить точность результатов. (Подробнее о точности и пределе погрешности в опросах см. главу 10.)

Изучая любое визуальное представление данных, всегда ищите общее количество участвующих элементов.

Устроители лотереи во Флориде используют круговую диаграмму, чтобы сообщить, на что расходуются деньги, которые вы отдаете, покупая их лотерейные билеты (см. рис. 4.3). Можно заметить, что половина дохода лотереи Флориды (50 центов из каждого потраченного доллара) идет на призы, а 38 центов от каждого доллара — на образование. В этой круговой диаграмме показано, как тратится каждый доллар дохода, но вас, возможно, заинтересует также вопрос, *сколько* именно долларов расходуется на участие в лотерее. Продажи лотерейных билетов лотереи Флориды в 2001 году составили 2 360,6 млн долларов (или 2,36 млрд долларов), что составляет 147,70 долларов на душу населения (т.е. на человека), как показано в табл. 4.1.

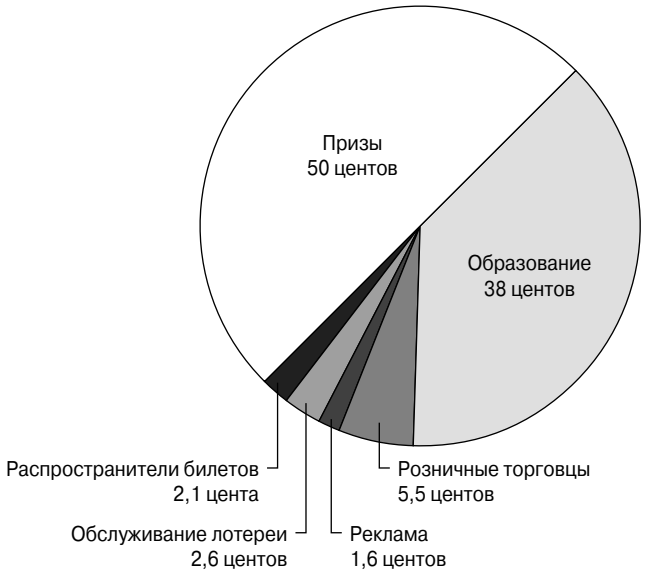


Рис. 4.3. Расходы лотереи Флориды (2001–2002 финансовый год)

Интересно, что на Web-сайте мичиганской лотереи указана сумма в долларах, которая ежегодно выделяется на образование, но не ее процентное отношение от всего дохода лотереи. Например, утверждается, что в 2001 году лотерея в Мичигане выделила на образование 587 млн долларов. Поскольку из табл. 4.1 известно, что общий доход от продажи лотерейных билетов в Мичигане составил 1 615 млн долларов (т.е. 1,6 млрд), можно подсчитать процентное отношение средств, выделенных на образование в этом штате. В Мичигане около 36% (587 млн долларов/1 615 млн долларов $\times$ 100%) пошли на образование.

Круговые диаграммы легко использовать для сравнения размеров кусков в самой диаграмме, но их можно применить также и для сравнения отдельных диаграмм. Например, лотерея Нью-Йорка сообщает о своих расходах с помощью круговой диаграммы (рис. 4.4).

Таблица 4.1. Десять ведущих лотерей (2001 год)

Место	Лотерея	Население (млн)	Билетов продано (млн)	Призы (млн)	Чистый доход (млн)	Призы (% от доходов)	Продажи на человека ($)
1	Нью-Йорк	18,976	4 178	2 274	1 447	54,4%	220,16
2	Массачусетс	6,349	3 923	2 774	865	70,7%	617,85
3	Калифорния	33,872	2 896	1 492	1 048	51,5%	85,49
4	Техас	20,852	2 826	1 639	865	58,0%	135,50
5	Флорида	15,982	2 361	1 180	862	50,0%	147,70
6	Джорджия	8,186	2 194	1 142	692	52,0%	267,98
7	Огайо	11,353	1 920	1 113	637	58,0%	169,11
8	Нью-Джерси	8,414	1 807	991	695	54,8%	214,72
9	Пенсильвания	12,281	1 780	996	627	55,9%	144,93
10	Мичиган	9,938	1 615	874	586	54,1%	162,49

Призы	56%	$ 2,664 млрд
Помощь образованию	33%	$ 1,58 млрд
Комиссионные розничных торговцев	6%	$ 284 млн
Зарплата контракторов	3%	$ 119 млн
Административные расходы	2%	$ 107 млн

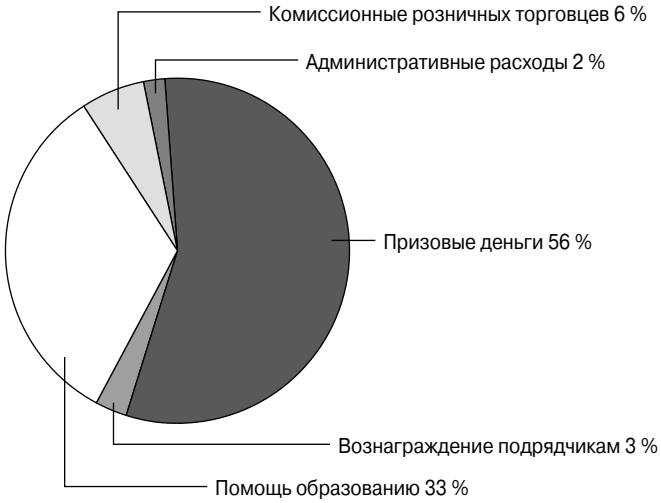


Рис. 4.4. Расходы лотереи Нью-Йорка (2001–2002 годы)

Сравнив рис. 4.3 и 4.4, вы увидите, что в случае с лотереей Нью-Йорка 56% денег идет на призы (в процентном отношении немногим больше, чем во Флориде), а 33% — на образование (в процентном отношении немногим меньше, чем во Флориде). Наряду с каждой круговой диаграммой Нью-йоркской лотереи составляется таблица, где показаны реальные долларовые эквиваленты этих процентов, благодаря чему вы можете лучше понять общую картину. (Но учредители лотереи Нью-Йорка заставляют вас самостоятельно делать сложение. Общая сумма превышает 4,5 млрд долларов.)

Штат Нью-Йорк хочет также, чтобы вы осознали, как много денег вкладывается в образование, для этого составляется специальная круговая диаграмма, показывающая доходы школ Нью-Йорка (очень умный ход с политической точки зрения). На рис. 4.5 показано, что если в 2001–2002 годах 4% дохода школы штата получали из федеральных источников, то 5% было выделено Нью-йоркской лотереей. И снова круговая диаграмма сопровождается таблицей, в которой представлены реальные суммы в долларах. (В действительности единственная сумма, которая нужна, — это общая сумма в долларах, потому что, имея ее и проценты в секторной диаграмме, вы можете сами подсчитать числа в таблице. Хотя если не нужно делать дополнительную роботу, это тоже неплохо.)

Распределение налоговых отчислений

Налоговое управление США (Internal Revenue Service, IRS) предпочитает, чтобы налогоплательщики знали, на что тратятся их деньги, — если вы сообщите, сколько именно налогов заплатили в прошлом году, вам покажут, как эти деньги были распределены.

Местный доход	45%	$ 15,168 млрд
Другая помощь образованию от государства	42%	$ 14,148 млрд
Нью-йоркская лотерея	5%	$ 1,58 млрд
Федеральная помощь	4%	$ 1,48 млрд
Другие источники	4%	$ 1,43 млрд

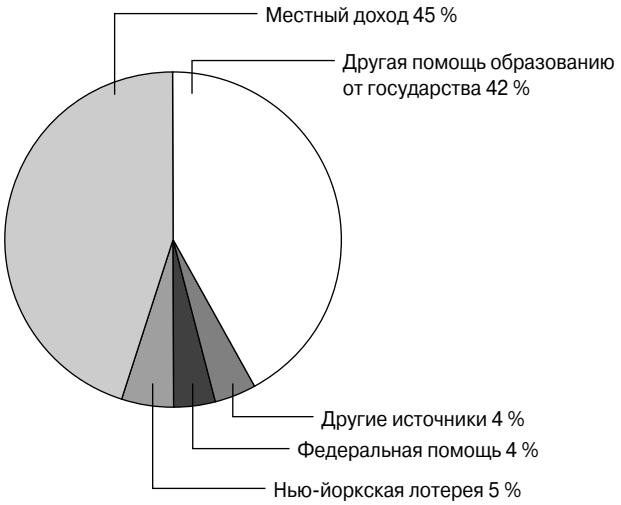


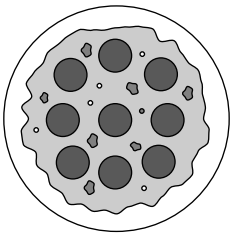
Рис. 4.5. Доход школ штата Нью-Йорк (2001-2002 годы)

На рис. 4.6 представлены результаты, которые гражданин США получает из IRS, если в прошлом году он заплатил налогов на 10 000 долларов.

Это визуальное представление данных выполнено творчески, но слегка непривычно (я не рискну сказать "странно", чтобы не нарваться на неприятности). Во-первых, эта диаграмма похожа скорее на пиццу, чем на пирог. При чем же здесь пицца, если она к тому же не была порезана так, чтобы показать, куда ушли ваши деньги? Процентные соотношения представлены в таблице рядом с пиццей, значит, они все же есть. Эта диаграмма оказалась бы эффективнее, если бы IRS указало реальные "куски" пиццы, соответствующие процентам в таблице. Однако хорошо то, что в подобном представлении данных показано как количество денег, так и их процентное соотношение, израсходованное на каждое направление. (Кстати, независимо о того, какова общая сумма налоговых отчислений, процентные отношения, показывающие, куда были распределены деньги, не меняются. Меняются только суммы в долларах.)

Изучив рис. 4.6, вы заметите, что наибольший кусок налоговых отчислений идет на социальное страхование (23%), на втором месте — национальная оборона (17%). Хотя странно, что налоговое управление выделяет определенные категории, на которые идет однозначное число отчислений (например, 7% на медицинскую помощь бедным), а третий по величине кусок пирога (в данном случае пиццы) соответствует категории "другие расходы" (16%).

Федеральные расходы



На что были потрачены деньги	Ваша доля ($)	Процентное отношение (%)
Национальная оборона	1 700,00	17
Бесплатная медицинская помощь неимущим	700,00	7
Медицинское страхование	1 200,00	12
Безработица, инвалидность и другие выплаты	1 400,00	14
Социальное обеспечение	2 300,00	23
Выплата процентов	1 100,00	11
Другие расходы	1 600,00	16
Всего выплачено	**10 000,00**	**100**

Рис. 4.6. Как распределяются налоги (2002 год)

В идеале секторная диаграмма делится на не очень много кусков, потому что если их много, это отвлекает читателя от более важных вопросов, которые такая диаграмма пытается осветить. Однако если после объединения таких менее значительных категорий в одну под названием "Другое" этот кусок становится одним из самых крупных во всей секторной диаграмме, читатели недоумевают: что же относится к этому куску пирога?

Возможно, вам тоже интересно, что входит в эти "другие расходы" из диаграммы IRS. Если тщательнее поискать на сайте Налогового управления, то вы увидите, что "другие расходы" означают "пенсионные выплаты по федеральной программе занятости, выплаты фермерам и другие направления деятельности". Здесь содержится не так уж много дополнительной информации. Нужно отдать должное IRS: очевидно, подробности изложены в каком-нибудь детально продуманном правительственном докладе.

Прогнозирование расового состава населения

Бюро переписей США снабжает свои доклады массой визуальных представлений данных относительно населения Соединенных Штатов Америки. На рис. 4.7 представлены две секторные диаграммы, сравнивающие расовый состав населения США в 1995 году (реальные цифры) с прогнозируемым расовым составом в 2050 году, если сохранятся существующие тенденции. Итак, в 1995 году около 73,6% населения составляли белые, а второй по величине была группа афроамериканцев, составивших 12,0%, за ними почти вплотную шли латиноамериканцы, к которым относилось 10,2% населения.

(Обратите внимание, что хотя латиноамериканцы, как правило, относятся к белой или черной группе, они выделены в отдельную категорию независимо от расовой принадлежности.) Бюро переписей прогнозирует, что в будущем доля белого населения США будет сокращаться, а количество латиноамериканцев будет расти быстрее, чем число не относящихся к ним черных. Это наглядно видно на двух секторных диаграммах, в отличие от таблиц, где содержатся только процентные соотношения.

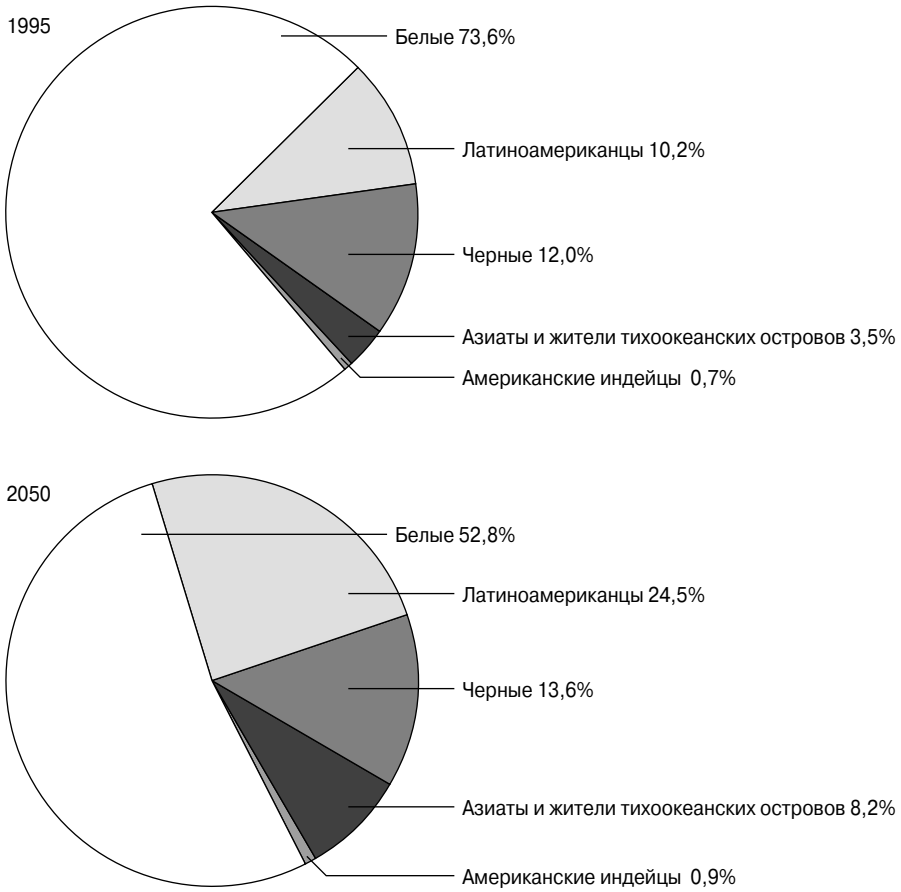


Рис. 4.7. Этнические тенденции в Соединенных Штатах Америки

Оценка секторной диаграммы

Чтобы определить, насколько точны статистические данные в секторной диаграмме, следуйте таким рекомендациям:

- ✓ Проверьте, точно ли сумма всех процентных соотношений будет равна 100% или около того (любая погрешность при округлении должна быть очень маленькой).

- ✓ Проследите, нет ли в пироге кусков под названием “Другое”, которые больше многих других секторов.
- ✓ Ищите указания на общее количество единиц, с помощью которого можно определить размер пирога до деления на куски.

Поднимаем планку столбиковой диаграммы

Столбиковая диаграмма — это, наверное, самый распространенный вид визуального представления данных, который используется в средствах массовой информации. Подобно секторной диаграмме в случае со столбиковой диаграммой дискретные данные делятся на группы, при этом указывается количество элементов в каждой группе. На столбиковой диаграмме эти группы показаны в виде столбиков разной высоты, а не кусков пирога разного размера. И если на секторной диаграмме количество элементов в каждой группе чаще всего показано в процентных отношениях, то в случае со столбиковой диаграммой используется либо число элементов в каждой группе, либо процентное отношение от общего количества. При этом высота столбика передает количество или проценты в каждой группе.

Отслеживаем расходы на транспорт

Какую часть своих доходов люди тратят на транспорт? Все зависит от того, сколько они зарабатывают. Бюро транспортной статистики в 1994 году провело исследование транспорта в США, и многие находки были представлены в виде столбиковых диаграмм, подобных показанной на рис. 4.8.

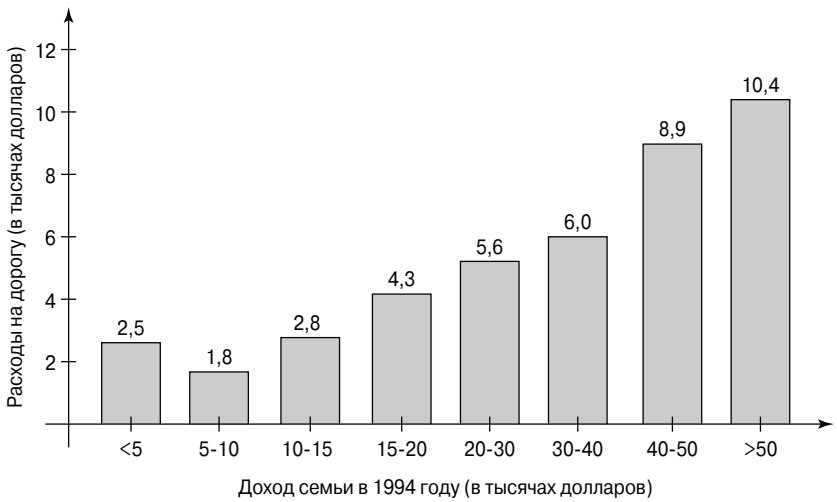


Рис. 4.8. Транспортные расходы семьи в 1994 году

На этой столбиковой диаграмме показано, сколько денег тратят на дорогу люди из семей разного достатка. Кажется, что с увеличением доходов семьи возрастают также и расходы на транспорт. Возможно, так и должно быть, ведь чем больше у людей денег, тем больше они могут потратить. Но изменилась бы диаграмма, если бы мы рассмотрели транспортные расходы не в виде общих сумм в долларах, а в качестве процентных отношений от всего

дохода семьи? Семьи в первой группе зарабатывают менее 5 000 долларов в год и должны тратить за этот же период 2 500 долларов на дорогу. (Обратите внимание, что надпись гласит: “2,5”, но поскольку все числа даны в тысячах долларов, то 2,5 нужно понимать как 2 500 долларов). Эти 2 500 долларов составляют 50% годового дохода семей, получающих 5 000 долларов в год. Для тех же, кто зарабатывает менее 5000 в год, этот процент будет еще выше. Семьи, зарабатывающие 30–40 тыс. долларов в год, тратят на дорогу 6 000 долларов ежегодно, что составляет от 15% до 20% их дохода. Значит, хотя люди, которые больше зарабатывают, больше тратят на проезд, но в процентном отношении эти расходы будут меньше по сравнению с их общим доходом. В зависимости от того, как посмотреть на расходы, столбиковая диаграмма может рассказать совершенно разные истории.

У этой столбиковой диаграммы есть еще одна особенность. Категории семейного дохода показаны неравными. Например, четыре первых столбика представляют семейные доходы с интервалом в 5 000 долларов, но следующие три группы возрастают каждая на 10 000 долларов, а к последней относятся семьи, зарабатывающие более 50 000 долларов в год, что составляет довольно большой процент, даже для 1994 года. Столбики диаграммы с разными размерами категорий, как это представлено на рис. 4.8, только затрудняют сравнение групп.

Доля работающих матерей

Столбиковые диаграммы часто используются для сравнения двух групп, при этом каждая группа делится на категории, которые затем представляются в виде соседних столбиков. Например, диаграмма на рис. 4.9 отвечает на вопрос: “Изменилось ли со временем процентное отношение работающих матерей?” положительно. На рис. 4.9 видно, что общий процент работающих матерей за период с 1975 по 1998 год вырос с 47 до 72%. Учитывая возраст детей, все меньше матерей выходят на работу до того, как их дети идут в школу, но разница за период с 1975 по 1998 год по-прежнему составляет 25%, что и видно на рис. 4.9.

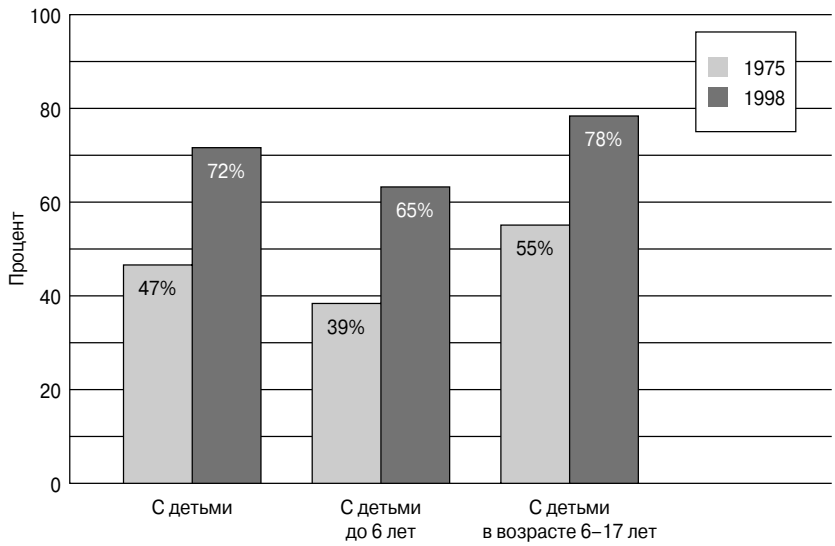


Рис. 4.9. Процентное отношение работающих матерей в зависимости от возраста ребенка (1975 и 1998 годы)

И снова лотерея в Огайо

Устроители лотереи Огайо показывают свои продажи и расходы за 2002 год с помощью столбиковой диаграммы (рис. 4.10). Чтобы понять всю заложенную здесь информацию, придется провести дополнительную работу. Первая проблема, связанная с данной диаграммой, заключается в том, что столбики показывают разные типы единиц. Первый столбик представляет продажи (разновидность дохода), а остальные демонстрируют расходы. Эта диаграмма была бы более понятной, если бы не был включен первый столбик. Например, общие продажи можно было бы показать в примечаниях. Кроме того, расходы можно было бы изобразить в виде секторной диаграммы, как это делают устроители лотерей в других штатах (см. рис. 4.3. и 4.4). Следующая проблема в том, что сумма всех расходов (2 013,2 млн. долларов — другими словами, 2,0132 млрд. долларов) превышает продажи (1,9831 млрд. долларов), значит, какой-то дополнительный доход не был показан на этой диаграмме (или же лотерее Огайо вскоре придется уйти из игорного бизнеса!). Пытаясь разыскать более подробную информацию на Web-сайте этой лотереи, я выяснила, что помимо продаж был заявлен дополнительный доход в 124,1 млн. долларов, который был получен благодаря "выплатам по процентам и из других источников".

Таким образом, общий доход в 2002 году составляет

$$\$1{,}9831 \text{ млрд.} + \$124{,}1 \text{ млн.} = \$2{,}1072 \text{ млрд.},$$

значит, прибыль в этом году была получена в размере

$$\$2{,}1072 \text{ млрд.} - \$2{,}0132 \text{ млрд.} = \$0{,}094 \text{ млрд.}, \text{ т.е. } 94\,000\,000 \text{ долларов.}$$

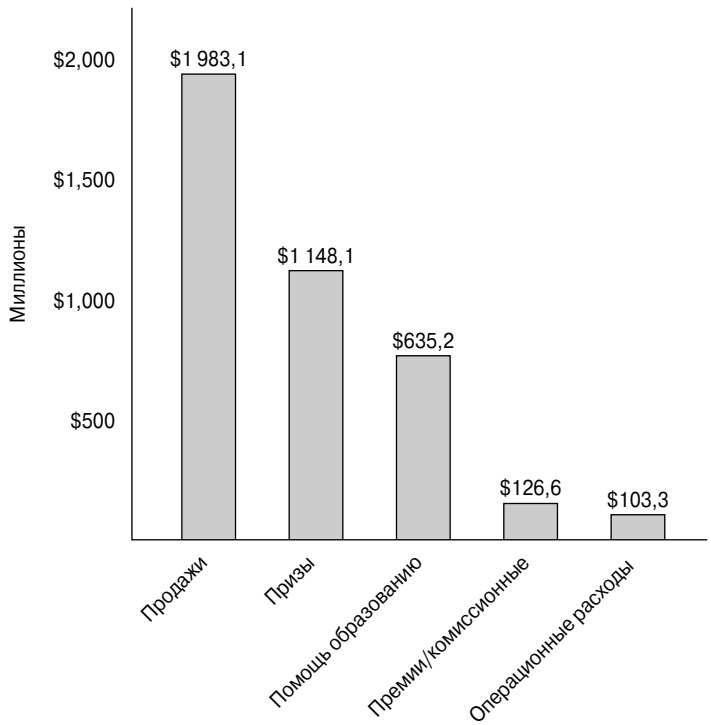


Рис. 4.10. Продажи и расходы лотереи в Огайо за 2002 год

Обратите внимание, что в примере с лотерей штата Огайо все единицы показаны в миллионах, поэтому то и дело встречаются необычные числа, например, $1 983,1 млн., что на самом деле означает $1,9831 млрд. Зачем лотерея выбрала такие единицы измерения? Может быть, чтобы складывалось впечатление, что устроители зарабатывают меньше, чем есть на самом деле. На рис. 4.10 числа кажутся слегка приукрашенными. Не думайте, что лотереи не обращают внимания на такие тонкости — они мастера хитростей. (Вспомните: чтобы подсчитать прибыль, вы вынуждены были самостоятельно делать вычисления, кроме того, искать числа для них пришлось в двух разных местах. Не все так просто в лотерее!)

Информация, показанная на диаграмме, содержит отнюдь не все, что вам нужно знать. Приготовьтесь копать глубже, если придется заполнить пробелы (которыми славятся диаграммы и графики в средствах массовой информации!). Обычно для поиска всего необходимого (или, по крайней мере, для того, чтобы дискредитировать представленную информацию, если вы обнаружите неточности или смещение) не требуется много времени.

Составители столбиковых диаграмм все в душе поэты — человек, составляющий диаграмму, сам определяет, какую шкалу он хочет использовать, т.е. информацию можно представить так, что она будет вводить в заблуждение. Используя более мелкую шкалу (например, если каждый сантиметр высоты столбика соответствует 10 единицам вместо 50), вы можете растянуть правду, сделать так, что разница станет более броской, или преувеличить значения. Воспользовавшись более крупной шкалой (если сантиметр высоты показывает не 10 единиц, а 50), вы приуменьшите разницу, результаты станут менее наглядными, чем есть на самом деле, а незначительные расхождения вообще сотрутся. (Примеры такой ситуации см. в главе 2.)

Заметьте, что в случае с секторной диаграммой шкалу нельзя изменить так, чтобы она как-либо искажала результаты. Как бы вы ни поделили этот пирог, вы в любом случае делите окружность, поэтому пропорции всего пирога и каждого отдельного куска не изменятся, даже если окружность будет больше или меньше.

Оценка столбиковой диаграммы

Чтобы столбиковая диаграмма не вводила вас в заблуждение, обратите внимание на такие аспекты.

- ✓ Столбики, которые делят значения числовой переменной (например, доход), должны быть одинаковой ширины, поскольку это облегчает сравнение.
- ✓ Обратите внимание на шкалу диаграммы (единицы измерения, в которых измеряется информация) и установите, подходит ли она для адекватного представления информации.
- ✓ Не верьте, что информация, показанная на столбиковой диаграмме, содержит все, что вам нужно знать. Будьте готовы при необходимости копать глубже.

Вносим данные в таблицу

Таблица — это способ визуального представления данных, при котором итоговая информация о наборе данных показана в виде столбцов и строк. Некоторые таблицы понять легко, другие оставляют желать большего. Если, как правило, секторная и столбиковая диаграммы имеют дело с одним или максимум двумя вопросами, то в таблице можно одновременно представить несколько вопросов (что тоже может быть или хорошо, или плохо, в зависимости от эффекта, который таблица производит на читателя).

Исследователи собирают статистическую информацию не только для собственных отчетов, но и для того, чтобы другие могли воспользоваться ею для своей работы и ответить на свои вопросы. В такой ситуации зачастую используются таблицы.

Изучаем статистику рождаемости

Департамент общественного здравоохранения и защиты окружающей среды штата Колорадо составляет таблицы со статистическими данными рождаемости для жителей Колорадо. В табл. 4.2 показано количество родившихся детей в зависимости от пола, а также количественная характеристика родов (один ребенок или двойня, тройня и т.д.) в разные годы за период с 1975 по 2000 год. Вот вопросы, на которые можно ответить с помощью этой таблицы: “Каков уровень рождаемости мальчиков по сравнению с девочками в штате Колорадо?”, “Меняется ли уровень многоплодных родов?”. Изучив эту таблицу, вы можете увидеть, что за 25 лет процентное отношение родившихся девочек осталось неизменным на уровне 49%, процентное отношение родившихся мальчиков тоже не изменилось и составило 51%. (Может показаться странным, почему обе эти цифры так близки к 50%. Это вопрос к демографам, т.е. к ученым, изучающим тенденции человеческого населения, и биологам, а не статистикам.). Также видно, что уровень многоплодных родов (по сравнению с одноплодными), похоже, за эти годы изменился. Кажется, что процент многоплодных родов возрастает, но на какой столбик вы смотрите — количество многоплодных родов или процентное отношение? Разве это важно? Да, конечно!

Проценты или общее количество

Как сделать вывод о тенденциях в количестве многоплодных родов, используя статистические данные из таблицы? Если посмотреть только на число многоплодных родов в 1975 году по сравнению с 2000 годом, то их стало больше — с 763 до 1 982. Кое-кто может заявить, что это означает прирост в 160%, т.е. что за 25 лет многоплодных родов стало в 1,6 раза больше ([1 982 – 763]/763). В 2000 году было больше многоплодных родов, чем в 1975 году, но за тот же период стало больше и одноплодных родов. В связи с этим единственный точный способ сравнить эти данные — это подсчитать процентное отношение одноплодных и многоплодных родов, затем сравнить полученные результаты. Из табл. 4.2 известно, что процент многоплодных родов в 1975 году составлял 1,9%, а в 2000 году он был равен 3,0%. Значит, можно сделать вывод, что за указанный период процентное отношение многоплодных родов действительно увеличилось, даже если учесть возросшее число всех родов. Однако этот прирост составляет не 160%, а около 58% ([3,0 – 1,9]/1,9) × 100%.

Таблица 4.2. Рождаемость в штате Колорадо с учетом пола ребенка и количественной характеристики родов

Год	Общее число родов	Количество родившихся девочек	Процент родившихся девочек	Количество родившихся мальчиков	Процент родившихся мальчиков	Число одноплодных родов	Процент одноплодных родов	Число многоплодных родов	Процент многоплодных родов
1975	40 148	19 447	48,4	20 701	51,6	39 385	98,1	763	1,9
1980	49 716	24 282	48,8	25 434	51,2	48 771	98,1	945	1,9
1985	55 115	26 925	48,9	28 190	51,1	53 949	97,9	1 166	2,1
1990	53 491	26 097	48,8	27 394	51,2	52 245	97,7	1 246	2,3
1995	54 310	26 431	48,7	27 879	51,3	52 669	97,0	1 641	3,0
2000	65 429	31 953	48,8	33 476	51,2	63 447	97,0	1 982	3,0

Не делайте выводов, используя представление данных, на котором *количество* единиц сравнивается с *процентным отношением*. Процентное отношение — это относительное сравнение количеств (часто за определенный период времени), обычно это точный способ сравнить разные количества, особенно если со временем меняется и общее число единиц или событий. Рассматривая изменения в процентах, вы учитываете тот факт, что изменилось и общее количество. (Конечно, если вас действительно интересует, как изменилось *количество* единиц в каждой категории, тогда нужно смотреть не на проценты, а на количество.)

В табл. 4.3 показано количество родов в Колорадо с учетом возраста матерей за определенные годы (с 1975 по 2000 год). Числовые переменные возраста поделены на непересекающиеся категории с одинаковой продолжительностью (5 лет). Благодаря этому можно точно и объективно сравнить возрастные группы. Однако в таблице представлено только число родов для каждой категории, поэтому нельзя взглянуть на нее и понять, какие же тенденции наблюдались с течением времени в том, что касается возраста рожениц. Такую проблему можно решить, если в скобках добавить к общему числу в каждой категории еще и проценты, чтобы читатель мог легко произвести сравнение. Еще один способ представить информацию — включить секторную диаграмму для каждого года, в которой бы были показаны проценты от общего числа родов для каждой из непересекающихся возрастных групп рожениц.

Таблица 4.3. Рождаемость в штате Колорадо по возрасту рожениц

Год	Общее число родов*	Возраст рожениц (лет)							
		10–14	15–19	20-24	25–29	30–34	35–39	40–44	45–49
1975	40 148	88	6627	14533	12565	4 885	1211	222	16
1980	49 716	57	6530	16642	16081	8 349	1842	198	12
1985	55 115	90	5634	16242	18065	11231	3464	370	13
1990	53 491	91	5975	13118	16353	12444	4772	717	15
1995	54 310	134	6462	12935	14286	13186	6184	1071	38
2000	65 429	117	7546	15865	17408	15275	7546	1545	93

* Примечание: Общее число родов может не совпадать с суммой родов из-за неизвестного или необычного (от 50 и выше) возраста рожениц.

Обратите внимание, что поскольку в таблице указано общее количество родов, вы можете, если захотите, самостоятельно подсчитать процентные отношения. (Если бы в таблице содержались только проценты, вам было бы проще их сравнить. Но тогда вы были бы ограничены в возможностях при подведении итогов, потому что не знали бы общего количества.) Просто чтобы сэкономить вам время, я уже подсчитала эти проценты для объединенной группы рожениц в возрасте от 40 до 49 лет. Из этой таблицы вы видите, что, похоже, возраст рожениц постепенно увеличивается. Все больше женщин рожают после сорока, и это количество постоянно растет.

Примечание к табл. 4.3 (слегка измененный вариант оригинала, который приводит департамент общественного здравоохранения и защиты окружающей среды штата Колорадо) говорит о том, что роженицы старше 50 не попали в этот набор данных.

Новейшие исследования позволяют предположить, что все больший процент (хотя в целом и не очень большой) женщин рожает в возрасте за 50, поэтому, возможно, эти данные придется со временем подправить, чтобы включить и указанную возрастную группу.

Таблица 4.4. Процент родов у женщин в возрасте 40–49 лет в штате Колорадо

Год	Количество родов	Число родов у женщин 40–49 лет	Процент родов у женщин 40–49 лет
1975	40 148	238	0,59
1980	49 716	210	0,42
1985	55 115	383	0,69
1990	53 491	732	1,4
1995	54 310	1 109	2,0
2000	65 429	1 638	2,5

Проценты под правильным углом

Не пребывайте в заблуждении, что только потому, что определенные процентные отношения невелики, они не имеют особого значения или их нельзя ни с чем сравнить. Все проценты из табл. 4.4 малы (меньше или равны 2,5%), но процентное отношение для 2000 года (2,5%) в четыре раз больше, чем процентное отношение для 1975 года (0,59%), что означает, так сказать, очень большой прирост. Точно так же не стоит думать, что если сообщается о больших процентах, то ситуация касается большого числа людей. Предположим, что кто-то заявляет, что за прошедшие несколько лет уровень заболеваемости определенной болезнью вырос в четыре раза. Это не значит, что болеет больший процент людей, а означает, что процентное отношение увеличилось в четыре раза. Начнем с того, что процентное отношение людей, пораженных этим заболеванием, может быть крайне мало. Рост остается ростом, но в некоторых ситуациях информация только о процентных отношениях может сбивать с толку. Рост заболеваемости нужно соотнести с общим количеством заболевших.

Процентное отношение — это относительная величина. Но чтобы правильно понять общую картину, нужна информация и об общем количестве.

Следим за единицами

Иногда в таблицах можно запутаться, особенно если вы будете невнимательны. Например, налоговое управление IRS на своем сайте приводит “Краткую информацию о налогах”, и некоторые данные (в том же виде, что и в докладе IRS) показаны в табл. 4.5.

С первого взгляда в глаза бросаются две особенности этой таблицы. Во-первых, в одной таблице налоговое управление сводит данные за разные отчетные годы, например, 2001 ф.г. (что означает финансовый год, т.е. период в 12 месяцев с 1 июля 2000 года по 31 июня 2001 года), 1999 н.г. (налоговый или календарный год 1999) и 2000 н.г. Обратите внимание, что налоговый и финансовый год пересекаются, и, например, налоговые декларации за 2000 н.г. должны поступить в IRS в апреле 2001 года (т.е. в 2001 финансовом году). Еще не запутались? Также заметьте, что нельзя сравнивать медиану скорректированного

валового дохода (медиана СВД) в этой таблице с 1% или 10% СВД из той же таблицы, потому что эти данные указаны для разных лет (2000 н.г. и 1999 н.г. соответственно).

Таблица 4.5. Данные о налоговых декларациях граждан

Количество деклараций (2001 финансовый год)	129 783 221
Валовой сбор (в млн долларов, 2001 ф.г.)	1 178 210
1% граждан с наибольшим скорректированным валовым доходом (СВД) (1999 налоговый год)	293 415 долл.
10% граждан с наибольшим СВД (1999 н.г.)	87 682 долл.
10% граждан с наименьшим СВД (1999 н.г.)	4 718 долл.
Медиана СВД (2000 н.г.)	27 355 долл.
Процент претендующих на стандартные налоговые вычеты (2000 н.г.)	66,2%
Процент претендующих на постатейные вычеты (2000 н.г.)	32,9%
Процент использующих помощь специалистов при заполнении налоговой декларации (2000 н.г.)	53,4%
Процент отправивших декларацию по электронной почте (2000 н.г.) до 03.05 2002	38,3%
Количество деклараций с СВД больше 1 млн. долларов (2000 н.г.)	241 068
Количество возмещенных выплат (в млн. долларов за 2000 н.г.)	93,0
Сумма возмещенных выплат гражданам (в млрд. долларов, 2000 н.г.)	167,6

Во-вторых, то, как IRS подает долларовые единицы, может вызвать путаницу. Например, валовые сборы по налоговым декларациям в 2001 финансовом году (указаны в млн. долларов) составляют 1 178 210. Это значит, что было собрано 1 178 210 млн. долларов. Но ведь вы обычно не указываете суммы в долларах таким образом. Правильнее было бы сказать, что 1 178 210 млн. долларов — это на самом деле 1 178 210 000 000 долларов, т.е. 1,178 триллионов долларов. Не удивительно, что налоговое управление не показывает доходы таким образом — это слишком невообразимое число!

Изучая таблицу, досконально рассмотрите все единицы, которые в ней представлены, и следите за изменениями единиц (например, разные годы) в пределах одной таблицы.

Предполагалось, что табл. 4.5 высветит несколько вопросов, и вот некоторые из них. В 2000 налоговом году отношение граждан, претендовавших на стандартные вычеты, к тем, кто претендует на постатейные вычеты, составило примерно 2 к 1. (Это можно было бы очень хорошо показать на секторной диаграмме.) Около половины человек, подавших налоговые декларации за 2000 год, воспользовались услугами специалистов по заполнению таких документов, а процент тех, кто отправил свою декларацию по электронной почте, в 2001 налоговом году был равен 38% (это тот показатель, рост которого с течением времени очень порадовал бы налоговое управление). Средний размер возмещенных выплат в 2000 налоговом году составил 1 802,15 долларов (общее количество средств возмещения (в млрд. долл.), деленное на общее число возмещенных выплат (в млн. долл.). (Но разве не лучше было бы указать общую сумму средств возмещения в противовес среднему числу возмещенных выплат?). Кроме того, обратите внимание, что налоговое управление не указывает общее число налоговых деклараций, заполненных в 2000 году или процентное отношение граждан, получивших возмещение за этот нало-

говый год. Знать этот процент было бы намного полезнее, чем иметь представление об общем количестве возмещенных выплат за данный налоговый год. Ведь как бы там ни было, но не все получают возмещение — многим приходится платить точно указанную сумму.

Таблицы составляются для того, чтобы особо подчеркнуть определенные моменты, а другие сделать менее заметными. Но иногда самой важной бывает информация, отодвинутая на задний план или вообще упущенная!

Оценка таблицы

Чтобы выяснить, надежна ли информация в таблице, придерживайтесь следующих рекомендаций.

- ✓ Помните о разнице между процентами и общим количеством, а также о том, как эти два вида данных используются для интерпретации результатов. Зачастую для сравнения разных результатов разумнее всего использовать именно процентные отношения.
- ✓ В случае с числовыми данными следите за тем, чтобы категории в таблице не пересекались и были поделены равномерно для их правильного сопоставления.
- ✓ Внимательно следите за единицами и тем, как они представлены в таблице.
- ✓ Проанализируйте то, как представлена информация. Очень часто таблицы составляются так, чтобы оттеснить определенные детали и подчеркнуть только то, что исследователи или докладчики хотят довести до вашего сведения.

В ногу со временными диаграммами

Временная диаграмма — это вариант визуального представления данных, основная задача которого — проследить за тенденциями, сложившимися за определенный период. Второе название временно*й* диаграммы — *линейный график*. Обычно на горизонтальной оси отмечаются единицы времени (такие как год, день, месяц и т.д.), а на вертикальной — измеряемая величина (например, средний доход семьи, уровень рождаемости, общий уровень продаж, процентное отношение людей, поддерживающих президента и т.п.). На каждом отрезке времени количество обозначается точкой, после чего эти точки соединяются, и получается временн*ая* диаграмма.

Анализируем изменения в заработной плате

В 1999 году Бюро трудовой статистики США выпустило доклад о тенденциях рабочей силы в Соединенных Штатах Америки, сопроводив его прогнозами на будущее. В этом докладе содержится множество временн*ых* диаграмм, в том числе и те, что показаны на рис. 4.11 и 4.12. На рис. 4.11 мы видим изменения в почасовой оплате работников сферы производства, произошедшие за период с 1947 по 1998 год. (В связи с инфляцией было бы бессмысленно просто показать уровень почасовой оплаты за это время. Вас ведь интересует информация, касающаяся “реальной зарплаты” — другими словами, то,

что можно сравнить для отдельных отрезков времени. Здесь бюро показывает все сведения по отношению к доллару в 1998 году). Из рис. 4.11 понятно, что с 1947 года до начала 1970-х годов оплата труда работников сферы производства росла, в 1970-х годах наблюдался спад, и до конца 1990-х годов заработная плата оставалась, по сути, на одном уровне, затем началось небольшое повышение.

На рис. 4.12 показано, что разница в оплате труда квалифицированных и неквалифицированных рабочих за период с 1979 по 1997 год увеличилась.

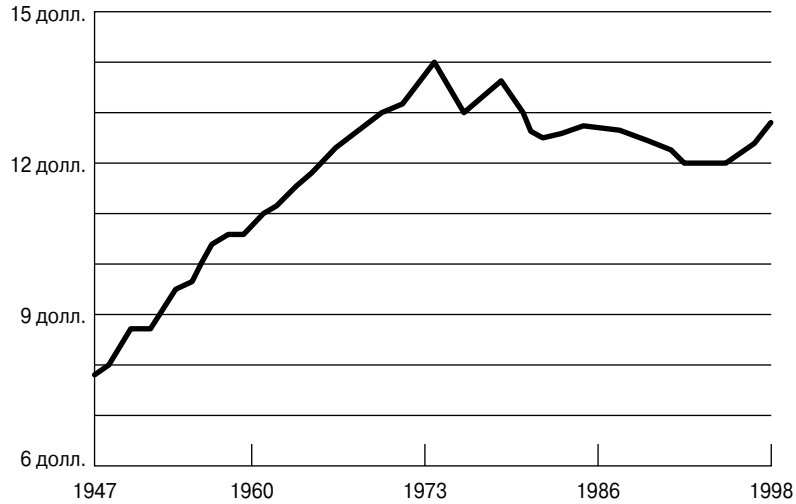


Рис. 4.11. Средняя почасовая оплата работников сферы производства, 1947–1998 годы (в долларах 1998 года)

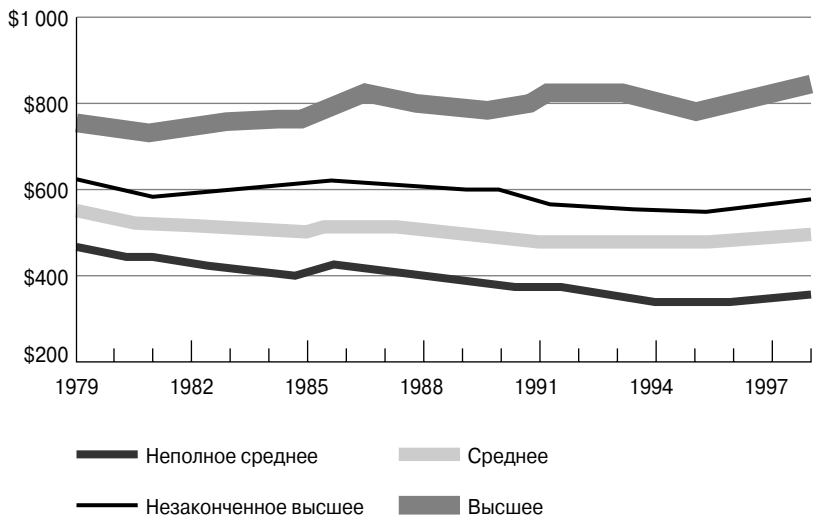


Рис. 4.12. Средний еженедельный заработок в зависимости от уровня образования, 1979–1997 годы (в долларах 1998 года)

Статистика сообщает факты — другими словами, отмечает, что происходит в мире. Но статистические данные не объясняют, почему события развиваются именно так. В докладе, выпущенном Бюро трудовой статистики, не только изложены сведения об изменениях в заработной плате, но делается и шаг вперед, потому что в этом документе описаны также некоторые из возможных причин, *почему* средняя зарплата в сфере производства не менялась с конца 1970-х до начала 1990-х годов и *почему* разница в оплате труда между более образованными и менее образованными рабочими все увеличивается. Ответить на вопрос "почему" намного сложнее, чем на вопрос "что". Хотя у Бюро трудовой статистки, несомненно, есть и другие данные в доказательство своих выводов относительно замеченных тенденций в зарплате) этим может похвалиться далеко не каждый, кто занимается статистикой).

Многие пытаются с помощью простого способа визуального представления данных показать, что происходит, а также объяснить причины таких событий. Не располагая достаточными данными, исследователи могут приходить к ложным выводам. Если вы подозреваете, что кто-то слишком далеко заходит в своих выводах, необходимо проверить, оправданы ли такие утверждения.

Фиксация многоплодных родов

При работе с данными рождаемости в штате Колорадо временную диаграмму можно использовать, чтобы проследить тенденции и количестве многоплодных родов за определенное время. Такая временная диаграмма будет напоминать представленную на рис. 4.13. Складывается впечатление, что процентное отношение многоплодных родов со временем растет, особенно если сравнить 1975 и 2000 годы.

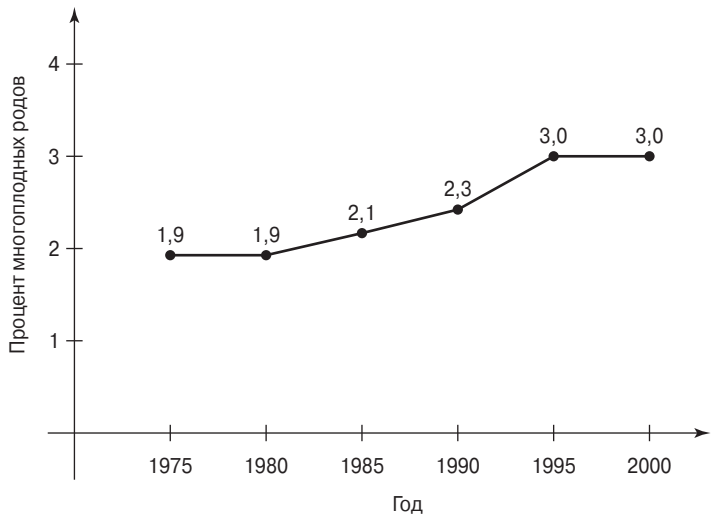


Рис. 4.13. Процент многоплодных родов для жителей штата Колорадо, 1975–2000 годы

Как и в случае со столбиковой диаграммой, разницу, показанную на временной диаграмме, можно приуменьшить или преувеличить, если изменить шкалу на вертикальной оси, поэтому всегда обращайте внимание на шкалу, когда будете изучать результаты с помощью временной шкалы.

В примере с многоплодными родами в Колорадо увеличение их количества со временем можно сделать явственнее, если поменять цену деления на вертикальной оси с 1 на 0,2%, в результате чего график вытянется вертикально. И наоборот, тенденция будет практически незаметной, если цену деления на вертикальной оси выбрать не 1, а 5%. Самая разумная цена деления — что-то между 0,2 и 5%. В своем варианте графика я выбираю 1%.

Еще один фактор, на который нужно обращать внимание в работе с временными диаграммами, — это ось времени или то, насколько близко или далеко друг от друга проставлены единицы данных. Данные, отмеченные вскоре друг за другом, но значительно отличающиеся по своей величине, приведут к тому, что временная диаграмма будет слишком угловатой. Это характерно для диаграмм, на которых представлены изменчивые данные, например, цены на фондовом рынке в течение дня, что часто показывают в выпусках деловых новостей. На других временных диаграммах могут быть представлены более плавные, долгосрочные изменения, например, на рис. 4.13, где отмечено лишь процентное отношение многоплодных родов за каждые пять лет, а не за каждый год. И снова все зависит от того, какую информацию составитель диаграммы хочет донести до вас. Чтобы прояснить все недоразумения, обязательно нужно задуматься над тем, как была составлена диаграмма и, при необходимости, задавать дополнительные вопросы.

Иногда возникает еще одна проблема — когда на временной диаграмме информация представлена необъективно, например, указывается количество преступлений, а не уровень преступности (количество преступлений на душу населения). Старайтесь до конца разобраться в данных, представленных на диаграмме, а затем проверьте их на объективность и соответствие (подробнее об этом см. главу 2).

Оцениваем временную диаграмму

Чтобы понять, все ли правильно в такой диаграмме, придерживайтесь следующих рекомендаций:

- ✓ Обратите внимание на шкалу на вертикальной (количество) и горизонтальной (отрезки времени) осях. Если всего лишь немного изменить шкалу, результаты могут выглядеть более или менее внушительно, чем есть на самом деле.
- ✓ Учитывайте единицы, использованные в диаграмме, и следите за тем, чтобы они подходили для сравнения величин за разные отрезки времени (например, если на временной диаграмме представлено изменение цен на какой-то товар, учитывается ли при этом уровень инфляции?).
- ✓ Помните, что люди могут попытаться объяснить, почему наблюдаются определенные явления, не имея дополнительных данных, подтверждающих их утверждения. По сути, временная диаграмма показывает, что происходит. А вот почему это происходит — это совсем другая история!

Отображение данных на гистограмме

Числовые данные в сыром, необработанном виде трудно переварить. Например, посмотрите на табл. 4.6, в которой показана предполагаемая численность населения в каждом из 50 штатов (а также в округе Колумбия) в 2000 году.

Таблица 4.6. Прогнозируемая численность населения по штатам (перепись 2000 года)

Штат	Перепись населения за 2000 год
Айдахо	1 293 953
Айова	2 926 324
Алабама	4 447 100
Аляска	626 932
Аризона	5 130 632
Арканзас	2 673 400
Вайоминг	493 782
Вашингтон	5 894 121
Вермонт	608 827
Виржиния	7 078 515
Висконсин	5 363 675
Гавайи	1 211 537
Делавэр	783 600
Джорджия	8 186 453
Западная Виржиния	1 808 344
Иллинойс	12 419 293
Индиана	6 080 485
Калифорния	33 871 648
Канзас	2 688 418
Кентукки	4 041 769
Колорадо	4 301 261
Коннектикут	3 405 565
Луизиана	4 468 976
Массачусетс	6 349 097
Миннесота	4 919 479
Миссисипи	2 844 658
Миссури	5 595 211
Мичиган	9 938 444
Монтана	902 195
Мэн	1 274 923
Мэриленд	5 296 486
Небраска	1 711 263
Невада	1 998 257
Нью-Гемпшир	1 235 786
Нью-Джерси	8 414 350
Нью-Йорк	18 976 457
Нью-Мексико	1 819 046
Огайо	11 353 140
Оклахома	3 450 654
Округ Колумбия	572 059
Орегон	3 421 399
Пенсильвания	12 281 054
Род-Айленд	1 048 319
Северная Дакота	642 200
Северная Каролина	8 049 313
Теннеси	5 689 283
Техас	20 851 820
Флорида	15 982 378
Южная Дакота	754 844
Южная Каролина	4 012 012
Юта	2 233 169
Всего в США	**281 421 906**

Таблица составлена Бюро переписей США. Внимательно рассмотрите ее секунд 30, после чего попытайтесь быстро ответить на приведенные ниже вопросы.

- ✓ В каких штатах численность населения наибольшая/наименьшая?
- ✓ Сколько жителей насчитывается в большинстве штатов? Укажите приблизительные цифры.
- ✓ Насколько отличается население в разных штатах? (Очень похоже/крайне различно.)

Если эти данные каким-то образом не организовать, вам будет трудно ответить на мои вопросы. Хотя большинство средств массовой информации при работе с числовыми данными отдают предпочтение таблицам, статистики в таких случаях предпочитают гистограмму. А что такое гистограмма, спросите вы?

По сути, *гистограмма* — это столбиковая диаграмма, которая используется для числовых данных. Поскольку данные числовые, то категории располагаются по возрастающей (в отличие от дискретных данных, таких как пол, которые не имеют раз и навсегда установленного порядка). А поскольку вы хотите быть уверенным, что каждое число относится только к одной группе, то столбики гистограммы касаются друг друга, но не пересекаются. Каждый столбик маркируется на оси *x* (горизонтальная ось) значением, соответствующим середине столбика. Например, представим себе гистограмму, на которой показана продолжительность работы определенной детали автомобиля (до поломки) в часах. На такой гистограмме мы видим два соприкасающихся столбика, обозначенные с помощью медиан 1 000 и 2 000 часов, а ширина каждого столбика 500 часов. Это значит, что первый столбик показывает детали автомобиля, прослужившие от 500 до 1500 часов, а второй представляет детали, которые работали от 1500 до 2500 часов. (Числа на границе могут относиться к обеим сторонам, если вы последовательно придерживаетесь этого правила.)

Высота каждого столбика гистограммы показывает либо количество элементов в каждой группе (это также называется *частотой* каждой группы), либо процентное отношение элементов в группах (что также известно как *относительная частота* каждой группы). Например, если 50% деталей автомобиля прослужили от 500 до 1500 часов, то первый столбик из предыдущего примера имеет относительную частоту 50%, и от этого зависит высота столбика.

На рис. 4.14 вы видите гистограмму населения штатов. Бегло взглянув на нее, вы легко сможете ответить почти на все вопросы, предложенные в начале раздела. Во многих ситуациях гистограмма позволяет намного эффективнее представить набор данных, чем таблица.

В большинстве штатов, а также в округе Колумбия (31 из 51, т.е. 60,8%) проживает менее 5 млн. человек. Население 25,5% штатов составляет от 5 до 10 млн. человек. Это значит, что население менее 10 млн. человек имеет 86,3% штатов. Население оставшихся семи штатов очень велико, из-за этого гистограмма кажется скошенной и вытянутой вправо (это называется *скошенной вправо*). За исключением этих нескольких очень больших штатов население в остальных штатах не так сильно варьируется, как может показаться. Из гистограммы не видно, где какой штат, но быстрый просмотр исходных данных поможет определить штаты с наибольшим и наименьшим числом населения. Пять самых густонаселенных штатов — это Калифорния, Техас, Нью-Йорк, Флорида и Иллинойс (за которым впритык идет Пенсильвания). Самый маленький штат — Вайоминг, где проживает около 494 000 человек.

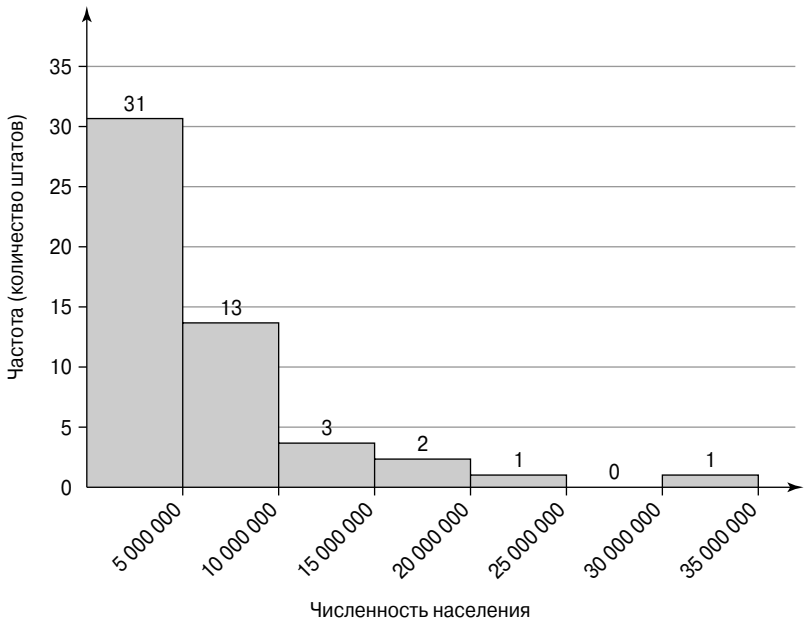


Рис. 4.14. Численность населения по штатам (перепись 2000 года)

Если при изучении изображения у вас возникают вопросы, попытайтесь получить исходный набор данных. Если вы попросите исследователей, они должны предоставить вам необходимую информацию.

Анализируем возраст рожениц

В примере со статистикой рождаемости (см. табл. 4.3) возраст матерей показан в разные годы, с 1975 по 2000. Для любого года в таблице возрастная переменная поделена на группы, количество рожениц представлено в каждой группе. Поскольку дано исходное количество, можно составить гистограмму возраста рожениц, в которой будет показана либо частота, либо относительная частота, в зависимости от того, что лучше подойдет для подтверждения вашей идеи.

Предположим, вы хотите сравнить возраст матерей в 1975 и 2000 году. Можно составить две гистограммы, для каждого года, и сравнить результаты. На рис. 4.15 представлены две такие гистограммы для 1975 (вверху) и 2000 году (внизу). Обратите внимание, что относительные частоты (или процентные отношения) показаны на вертикальной оси, а возрастные группы рожениц — на горизонтальной.

Гистограмма может очень хорошо резюмировать характеристики числовых данных. Одна из характеристик, которую вы можете показать с помощью гистограммы, — это так называемая *форма* данных (т.е. то, как данные распределяются между группами). Проходит ли такое распределение равномерно? *Симметричны* ли данные, т.е. является ли левая половина гистограммы зеркальным отражением правой половины? Имеет ли гистограмма *U-образную форму*, когда большинство данных сосредоточено по концам, а в середине их намного меньше? А может, гистограмма имеет *колоколообразную форму*, т.е. по-

хожа на гору, склоны которой удаляются по обоим направлениям от центра? Или же она имеет *скос*, т.е. напоминает скошенный холм, у которого один длинный склон направлен вправо (значит, данные *скошены вправо*) или влево (тогда данные *скошены влево*)?

Возраст рожениц на рис. 4.14 в 1975 и 2000 годах можно назвать скорее колоколообразным, хотя данные за 1975 год имеют небольшой скос вправо, что говорит о том, что по мере того, как женщины становились старше, все меньше из них рожали детей по сравнению с 2000 годом. То же можно выразить и иначе: в 2000 году больше женщин старшего возраста рожали детей по сравнению с 1975 годом.

Изучив внимательно гистограмму, можно сделать вывод, насколько изменчивы представленные в ней данные. Если гистограмма слегка плоская, то столбики примерно одинаковой высоты, в таком случае расхождения кажутся незначительными, ведь высота столбиков почти одна и та же.

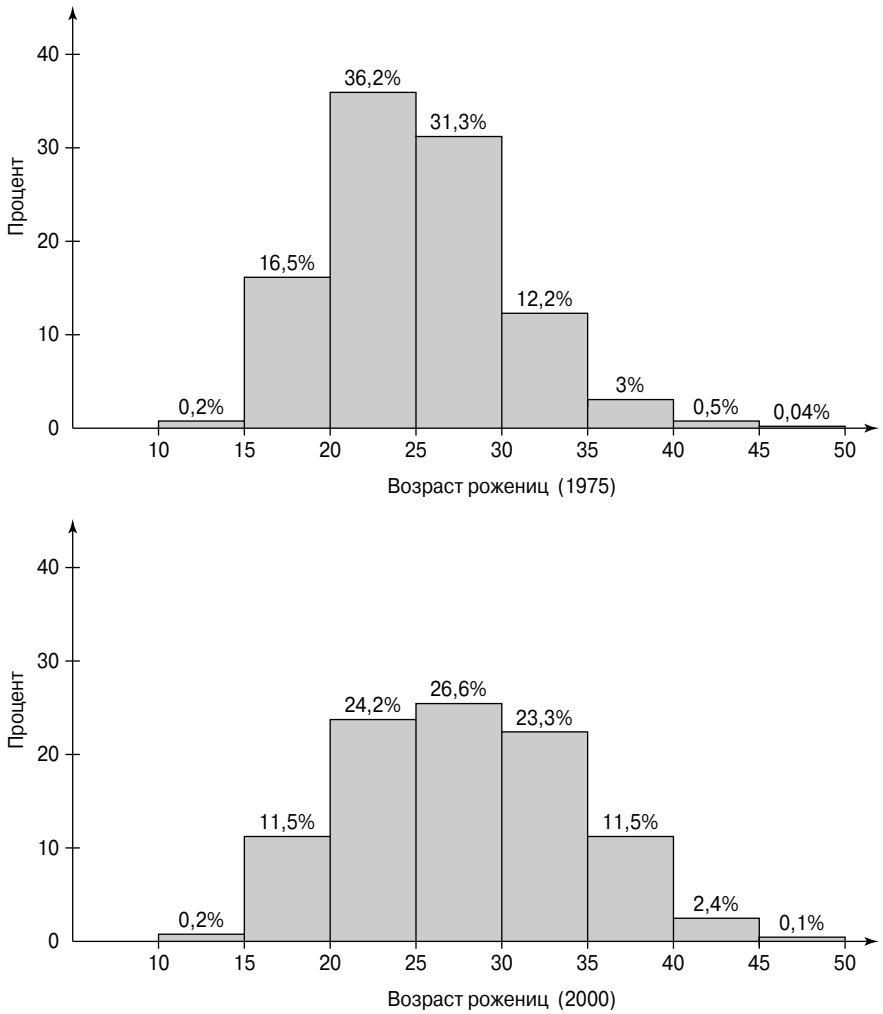


Рис. 4.15. Уровень рождаемости в штате Колорадо (по возрасту рожениц) в 1975 и 2000 годах

Но на самом деле все совсем наоборот. Все объясняется тем, что каждый столбик содержит одинаковое количество элементов, хотя сами столбики показывают разные диапазоны значений, а значит, весь набор данных в целом довольно разнообразен. Если у гистограммы наблюдается большое возвышение в центре, это указывает на то, что в центральных столбиках содержится больше данных, чем во внешних, значит, на самом деле данные ближе друг к другу. Сравнивая возраст рожениц в 1975 и 2000 годах, вы заметите больше изменчивости в 2000 году, чем в 1975. Это опять указывает на изменения во времени. Все больше современных женщин не торопятся с рождением ребенка, тогда как в 1975 году большинство женщин рожало до 30 лет. Варьируется также и продолжительность такой отсрочки. (В главе 5 рассказывается о том, как измерять изменчивость набора данных.)

Изменчивость в гистограмме не нужно путать с изменчивостью во временн*о*й диаграмме (см. раздел “В ногу со временн*ы*ми диаграммами”). Если значения меняются с течением времени, то на временн*о*й диаграмме они показываются в виде подъемов и спусков, а множество переходов от первых ко вторым (со временем) свидетельствует о наличии большой изменчивости. Значит, ровная линия на временн*о*й диаграмме говорит о том, что за определенный период в значениях не наблюдалось изменений. Но если высота столбиков на гистограмме кажется почти одинаковой (единообразной), то это указывает как раз на обратное — значения распределены однообразно среди множества групп, что значит, что в определенный отрезок времени среди данных наблюдается значительная изменчивость.

Гистограмма может также кое-что рассказать о том, где находится центр данных. Центр набора данных определяется разными способами (подробнее об этом см. главу 5). Один из них — представить гистограмму в виде картины, на которой люди сидят на качелях, центр гистограммы — это точка, в которой должно находиться бревно, чтобы оба его конца были уравновешены. Вернемся к рис. 4.15, на котором показан возраст рожениц в штате Колорадо в 1975 и 2000 годах. Складывается впечатление, будто центр находится в районе 25 лет в 1975 году и около 27,5 лет в 2000 году. Отсюда можно сделать вывод, что в 2000 году женщины в Колорадо рожали в среднем позже, чем в 1975.

Гистограммы не так часто используются в средствах массовой информации, как они того заслуживают. Это происходит по непонятным причинам, но анализа числовых данных намного шире применяются таблицы. Хотя и гистограмма может быть информативной, особенно если использовать ее для сравнения одной группы или отрезка времени с другим. Как бы там ни было, если вы хотите представить данные графически, то их всегда можно взять из таблицы и преобразовать в визуальное изображение.

Помните, что в гистограммах могут использоваться необычные шкалы, чтобы вводить читателя в заблуждение. Как и в случае со столбиковыми диаграммами, можно преувеличить разницу, если использовать меньшую шкалу на вертикальной оси гистограммы, или стереть какие-либо различия при помощи большей шкалы. Гистограмма может сбить читателя с толку так, как этого не сделаешь с помощью столбиковой диаграммы. Не забывайте, что гистограмма имеет дело с числовыми, а не категорийными данными. Это значит, что вам нужно определиться, как вы хотите разделить числовые данные на группы, чтобы представить их на горизонтальной оси. От того, каким вы примете деление на группы, будет зависеть внешний вид вашего графика.

Ползем вместе с малышом

Сколько проползает восьмимесячный малыш? На рис. 4.16 показаны две гистограммы, которые представляют один и тот же набор данных (расстояние, которое прополз мой ребенок за шесть часов наблюдений). В каждом случае расстояние было округлено до фута.

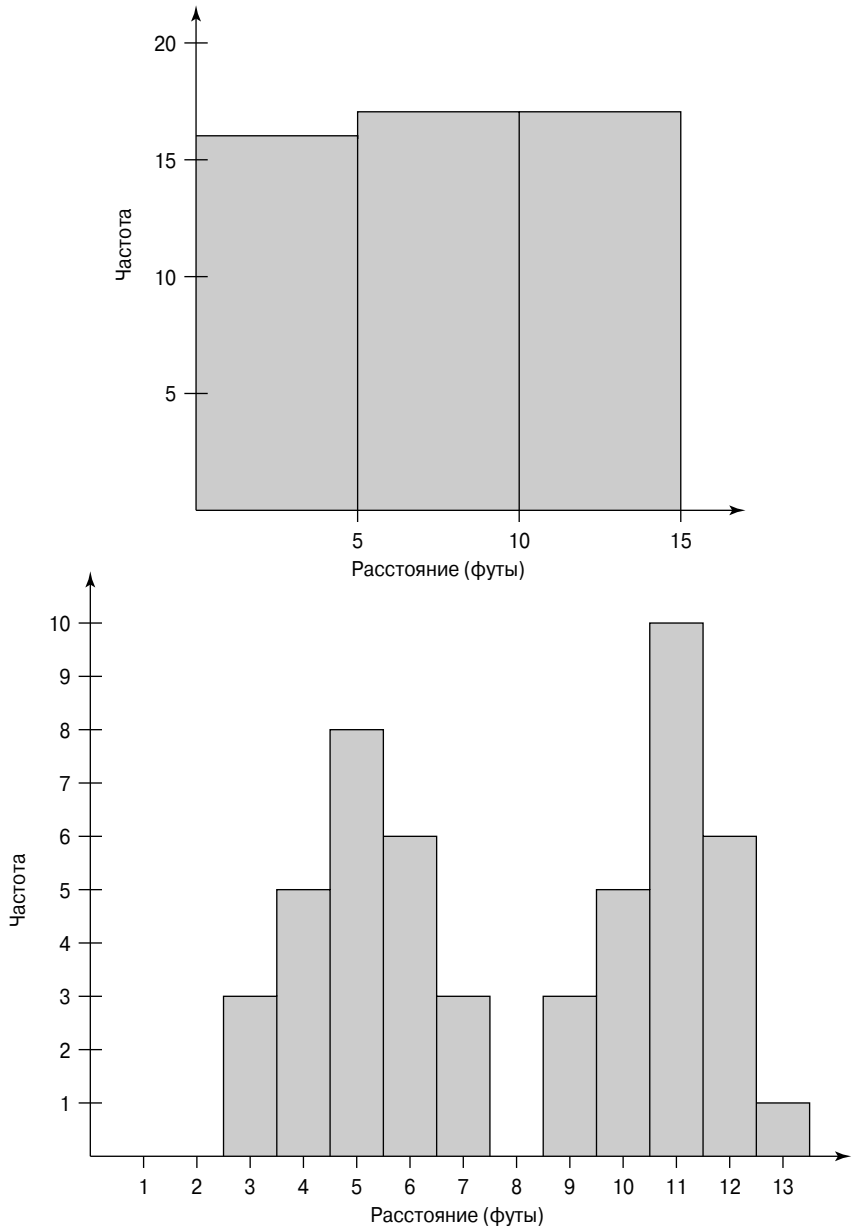


Рис. 4.16. Расстояние, которое проползает ребенок

В верхней части рисунка измерения показаны с ценой деления 5 футов, и кажется, что данные распределены единообразно. Другими словами, количество раз, сколько он прополз каждое расстояние (0–5 футов, 5–10 футов и 10–15 футов[1]), примерно одинаково. Данные кажутся довольно скучными. Но в нижней половине рисунка я выбрала меньшую цену деления, в 1 фут[2], в результате гистограмма совершенно изменилась и стала намного интереснее.

На этой гистограмме вы видите два разных деления расстояний на группы, и становится понятно, что мой ребенок ползал либо на короткие расстояния (около 5 футов), либо на более длинные (около 10 футов). Это имеет свое объяснение, ведь в то время, когда я собирала данные, коляска моего ребенка находилась на расстоянии примерно 5 футов от исходной точки, а пачка газет (еще одна любимая игрушка) — на расстоянии в 10 футов. Вторая гистограмма намного лучше представляет данные — расстояние, которое прополз мой ребенок в предложенных условиях.

Так сколько же он прополз за эти шесть часов? С помощью нижней половины рис. 4.16 вы можете определить общее расстояние, ведь столбики различаются по высоте с ценой деления в 1 фут. Умножьте высоту каждого столбика на расстояние, а затем сложите результаты. За шесть часов наблюдений мой ребенок прополз умопомрачительное расстояние в 398 футов или 121,3 метра — больше, чем длина футбольного поля!

Обратите внимание, что я могла бы выбрать еще меньшую цену деления для расстояния, но из-за этого гистограмма казалась бы слишком загроможденной и не содержала бы никакой дополнительной информации. Золотая середина существует и в том, что касается количества использованных групп и диапазона представленных ими значений. Одинаковых гистограмм нет, но в среднем самое оптимальное количество столбиков — 6–12 штук. Если столбиков на гистограмме слишком мало, данные ни о чем не говорят. Если же столбиков слишком много, то данные получаются слишком разбросанные, и общие закономерности просто теряются.

Изучая результаты исследования, представленные в виде гистограммы, обязательно обращайте внимание на шкалу как на вертикальной, так и на горизонтальной осях. Одни и те же данные могут выглядеть по-разному в зависимости от распределения на группы (например, если групп слишком мало или слишком много) и от шкалы на вертикальной оси, в результате чего столбики могут оказаться выше или короче, чем вы ожидали.

Интерпретация гистограммы

С помощью гистограммы можно выяснить три основные особенности числовых данных.

- ✓ Как данные распределены (симметрично, со скосом вправо, со скосом влево, колоколообразно и т.д.).
- ✓ Степень изменчивости данных.
- ✓ Где находится центр данных (приблизительно).

[1] 0–1,5 м, 1,5–3 м и 3–4,5 м соответственно. — *Примеч. пер.*

[2] 30 см. — *Примеч. пер.*

Оценка гистограммы

Чтобы понять качество гистограммы, вы должны сделать следующее.

- ✓ Изучить шкалу, использованную на вертикальной оси (частота или относительная частота). Помните, что результаты могут показаться преувеличенными или недооцененными, если шкала выбрана неправильно.
- ✓ Проверить единицы на вертикальной шкале, чтобы понять, о чем идет речь — о частоте (количество) или относительной частоте (процентное отношение). Учитывайте это при оценке информации.
- ✓ Изучить шкалу, которая была использована для распределения числовых переменных на группы (на горизонтальной шкале). Если диапазон для каждой группы слишком мал, данные могут показаться чересчур изменчивыми. Если диапазон слишком велик, данные могут выглядеть однороднее, чем есть на самом деле.

Глава 5

Средние значения, медианы, а также многое другое

В этой главе...

- Быстрое подытоживание данных
- Интерпретация самых распространенных статистических показателей
- Что скрывает статистика

Статистика (как параметр) — это число, которое подытоживает какую-либо информацию о наборе данных. Из сотни существующих показателей всего несколько используются настолько широко, что мы уже привыкли встречать их в работе и в самых разных проявлениях повседневной жизни. В этой главе вы узнаете, какие статистические показатели используются чаще всего, как они работают, что означают и какие ошибки в работе с ними допускаются.

У каждого набора данных есть своя история, и при правильном использовании именно благодаря статистическим показателям мы узнаем ее. Неправильно использованные показатели могут рассказать совсем другую историю или только часть реальной истории, поэтому крайне важно знать, как принимать продуманные решения на основе имеющейся информации. В этой главе вы познакомитесь с некоторыми самыми популярными статистическими показателями, больше узнаете о том, что они говорят, а чего и не говорят о данных, которые можно разделить на числовые и дискретные.

Подытоживание данных с помощью статистики

Статистики применяются для того, чтобы подвести итог определенной информации о наборе данных. Это делается по нескольким причинам. Представьте, что к вам подходит начальник и спрашивает: “Как сейчас выглядит ваша клиентская база, и кто покупает нашу продукцию?”. Как бы вы отвечали на этот вопрос — с использованием детального и запутанного потока чисел и статистических данных, от которых взгляд босса, несомненно, потускнеет? Наверное, нет. Вам нужны четкие, понятные и точные числа, которые вкратце расскажут все о клиентской базе, и ваш начальник сразу же поймет, какой вы умный, после чего поручит собрать еще больше данных, чтобы можно было привлечь еще больше клиентов. Значит, статистические данные часто используются для того, чтобы предоставить информацию, которую легко понять и которая отвечает на все существующие вопросы (если ответить на них вообще возможно).

Подытоживание статистических данных делается еще с одной целью. После того, как с помощью опроса или любого другого исследования была собрана определенная информация, следующий шаг, который должен сделать исследователь, — попытаться понять, что же именно было собрано. Как правило, прежде всего исследователи проводят беглое статистическое изучение данных, чтобы приблизительно понять их суть. Затем можно будет провести дополнительный анализ, чтобы сформулировать или проверить утверждения, касающиеся генеральной совокупности, оценить определенные свойства совокупности, найти связи между измеряемыми аспектами и т.д.

Еще один важный элемент исследования — это сообщение результатов не только своим коллегам, но и средствам массовой информации, а также широкой общественности. Если коллеги-исследователи, скорее всего, рассчитывают услышать все о сложных анализах, проведенных с набором данных, то широкая общественность совсем не готова к такому, да ей это и не интересно. Чего же хочет публика? Основную информацию. Значит, статистические показатели, которые ясно и кратко излагают вашу мысль, чаще всего используются для передачи информации СМИ и общественности.

Очень часто с помощью статистик подводится быстрый и грубый итог ситуации, которая на самом деле довольно сложна. В таком случае больше — не всегда лучше, и подчас реальная история может просто затеряться в этой массе. Хотя вам придется смириться с тем, что получение информации в приемлемых количествах — это суровая действительность нашего времени, но следите за тем, чтобы люди, составляющие отчеты, не разбавляли факты несущественными деталями. Помните о том, о каких статистических показателях сообщается, что они на самом деле значат и какой информации не хватает. В данной главе речь пойдет именно об этом.

Подытоживание дискретных данных

Категорийные (дискретные) данные касаются качественных характеристик или свойств элементов, например, цвет глаз человека, пол, политическая партия или мнение по определенному вопросу (т.е. использование таких категорий, как “согласен”, “не согласен” или “не знаю”). Категорийные данные вполне естественно делятся на группы или категории. Например, понятие “политическая партия в США” обычно можно разделить на четыре группы: демократ, республиканец, независимый и другие. Дискретные данные зачастую поступают в ходе опроса, но их также можно собирать и во время эксперимента. Например, в ходе экспериментальной проверки нового лекарственного препарата исследователи могут использовать три категории, чтобы оценить исход эксперимента: состояние пациента улучшилось, ухудшилось или осталось без изменений?

Часто дискретные данные подытоживаются в виде подсчета процентных отношений элементов, относящихся к каждой категории. К примеру, интервьюеры могут сообщить процент сторонников республиканцев, демократов, сторонников независимых партий и других объединений из числа людей, принимавших участие в опросе. Чтобы подсчитать процентное соотношение между группами элементов в определенной категории, количество элементов в каждой группе нужно разделить на общее число участников опроса, а затем умножить результат на 100%.

Например, если среди 2000 опрошенных подростков было 1200 девушек и 800 юношей, то их процентные отношения составят $(1\,200/2000) \times 100\% = 60\%$ девушек и $(800/2000) \times 100\% = 40\%$ юношей.

Дискретные данные можно поделить и с помощью так называемых перекрестных таблиц. *Перекрестные таблицы* (также называемые *таблицами с двумя входами*) — это таблицы со столбцами и строками. Они подводят итог информации о двух дискретных переменных одновременно, например, пол и политическая партия, и таким образом вы можете видеть (или легко подсчитать) процентное отношение элементов в каждой комбинации категорий. Например, если у вас есть данные о поле и политических убеждениях ваших респондентов, вы можете определить процентное отношение женщин, поддерживающих республиканскую партию, мужчин- сторонников этой партии, женщин-демократов, мужчин-демократов и т.д. В этом примере общее количество возможных комбинаций в таблице составит $2 \times 4 = 8$, т.е. общее число категорий пола, умноженное на общее число категорий политических убеждений.

С помощью перекрестных таблиц правительство США подсчитывает и подытоживает огромные массы дискретных данных. Бюро переписей США не просто ведет подсчет численности населения, а также собирает и подытоживает данные о подгруппе, представляющей всех американцев (т.е. тех, кто заполняет длинные анкеты). Сюда относится информация о таких демографических характеристиках, как пол и возраст. Типичные данные о поле и возрасте, собранные Бюро переписей США в ходе опроса 2001 года, показаны в табл. 5.1. (Как правило, возраст относят к числовым данным, но в том виде, как о нем сообщает правительство США, возраст делится на категории, благодаря чему он уже относится к дискретным данным. Подробнее о числовых данных см. следующий раздел.)

Таблица 5.1. Население США, деление по полу и возрасту (2001 год)

Возраст	Всего	Проценты	Количество мужчин	Процент мужчин	Количество женщин	Процент женщин
До 5 лет	19 369 341	6,80	9 905 282	7,08	9464059	6,53
5–9	20 184 052	7,09	10336 16	7,39	9847436	6,79
10–14	20 881 442	7,33	10696244	7,65	10185198	7,03
15–19	20 267 154	7,12	10423173	7,46	9843981	6,79
20–24	19 681 213	6,91	10061983	7,20	9619230	6,63
25–29	18 926 104	6,65	9592895	6,86	9333209	6,44
30–34	20 681 202	7,26	10420677	7,45	10260525	7,08
35–39	22 243 146	7,81	11104822	7,94	11138324	7,68
40–44	22 775 521	8,00	11298089	8,08	11477432	7,92
45–49	20 768 983	7,29	10224864	7,31	10544119	7,27
50–54	18 419 209	6,47	9011221	6,45	9407988	6,49
55–59	14 190 116	4,98	6865439	4,91	7324677	5,05
60–64	11 118 462	3,90	5288527	3,78	5829935	4,02
65–69	9 532 702	3,35	4409658	3,15	5123044	3,53
70–74	8 780 521	3,08	3887793	2,78	4892728	3,37
75–79	7 424 947	2,61	3057402	2,19	4367545	3,01
80–84	5 149 013	1,81	1929315	1,38	3219698	2,22
85–89	2 887 943	1,01	926654	0,66	1961289	1,35
90–94	1 175 545	0,41	303927	0,22	871618	0,60
95–99	291 844	0,10	58667	0,04	233177	0,16
100 лет и больше	48 427	0,02	9860	0,01	38567	0,03
Всего, всех возрастов	**284 796 887**	**100**	**139813108**	**100**	**144983779**	**100**

Изучив табл. 5.1 и проанализировав числа, можно узнать очень многое о разных аспектах населения в США. В случае с полом обратите внимание, что женщин немного больше, чем мужчин, потому что в 2001 году население состояло из 51% женщин (разделите общее число женщин на общую численность населения и умножьте на 100%) и 49% мужчин (разделите общее число мужчин на общую численность населения и умножьте на 100%). Теперь что касается возраста: процентное отношение всех жителей младше 5 лет составляло 6,8%. Самая многочисленная группа — люди в возрасте от 40 до 44 лет, составившие 8% населения. Затем можно рассмотреть существующие взаимосвязи между полом и возрастом, сравнив разные части таблицы. Например, сравним процент женщин и мужчин в возрастной группе старше 80 лет. Поскольку данные поделены на возрастные группы по пять лет, придется посчитать самостоятельно. Процент женщин в возрасте от 80 лет и выше составляет 2,22% + 1,35% + 0,6% + 0,16% + 0,03% = 4,36%. Процент мужчин старше 80 лет равен 1,38% + 0,66% + 0,22% + 0,04% + 0,01% = 2,31%. Отсюда видно, что в возрастной группе старше 80 лет женщин почти в два раза больше, чем мужчин. Похоже, такие данные подтверждают мнение о том, что женщины живут дольше, чем мужчины.

Если вам известно количество элементов в каждой группе, вы всегда сможете сами подсчитать проценты. Но если же вы знаете только проценты без общего числа в группе, то восстановить исходное количество элементов в группе не получится. Например, вы, возможно, слышали, что 80% опрошенных людей предпочитают сырные крекеры Cheesy сырным крекерам Crummy. Но сколько всего человек было опрошено? Их вполне могло быть всего 10, ведь 8 из 10 тоже равно 80%, так же как и 800 из 1000. Эти две дроби (8 из 10 и 800 из 1000) в статистике имеют разное значение, потому что в первом случае результат основывается на слишком малом количестве данных, а во втором случае данных в основе много. (Подробнее о точности данных и пределе погрешности см. главу 10.)

Получив перекрестные таблицы, в которых представлена информация по двум дискретным переменным, вы можете провести статистическую проверку и определить, действительно ли между этими переменными существует важная взаимосвязь. (Подробнее о таких статистических проверках читайте в главе 18.)

Подытоживание числовых данных

В случае с *числовыми данными* измеряемые характеристики, такие как высота, вес, IQ, возраст или доход, представлены в виде чисел. Поскольку данные имеют числовое значение, то для того, чтобы подытожить их, существует больше способов, чем для работы с дискретными данными. Подобные итоги часто подводятся в средствах массовой информации, поэтому знание того, что означают такие статистические показатели, а также что они скрывают относительно данных, поможет лучше понять исследование, о котором вам пытаются рассказать в обычной жизни.

Находим центр

Самый распространенный способ подытожить набор числовых данных — описать его центр. Например, можно спросить: “Каково типичное значение?” или “Где находится середина данных?”. Центр в наборе данных на самом деле можно определить несколькими способами, и выбор метода в значительной степени влияет на выводы, которые люди делают по поводу данных.

Средняя зарплата игроков NBA

Игроки NBA зарабатывают огромные деньги. По сравнению с большинством людей это действительно так. Но сколько именно они получают и действительно ли это так много, как кажется? Ответ зависит от того, как вы будете подытоживать информацию. Часто можно слышать о таких игроках, как Шакил О'Нил, который за сезон 2001–2002 годов заработал 21,4 млн. долларов. Типична ли такая сумма для игроков NBA? Нет. В указанное время Шакил О'Нил был самым высокооплачиваемым игроком NBA. Так сколько же получает обычный игрок лиги? Один из способов ответить на этот вопрос — вычислить *среднюю* зарплату. Среднее значение — это, наверное, самый популярный статистический показатель всех времен. Это один из вариантов определить, где находится "центр" данных.

Чтобы найти среднее значение в наборе данных, обозначаемое $\bar{x}$, нужно проделать следующее.

1. **Сложите все числа в наборе данных.**
2. **Разделите результат на количество чисел в наборе данных, обозначаемое *n*.**

Например, данные о заработках тринадцати игроков команды Los Angeles Lakers (за исключением тех, кто ушел в начале сезона) за 2001–2002 годы представлены в табл. 5.2.

Таблица 5.2. Зарплата игроков команды Los Angeles Lakers в сезоне 2001–2002 годы

Игрок	Зарплата ($)
Шакил О'Нил	21 428 572
Коб Брайант	11 250 000
Роберт Хорри	5 300 000
Рик Фокс	3 791 250
Линдси Хантер	3 425 760
Дерек Фишер	3 000 000
Самаки Уокер	1 400 000
Митч Ричмонд*	1 000 000
Брайан Шоу*	963 415
Девин Джордж	834 250
Марк Медсен	759 960
Джелани Маккой	565 850
Станислав Медведенко	465 850
Всего	**54 184 907**

* без учета лимита затрат на зарплату

Сложив все зарплаты, мы получим общую сумму в 54 184 907 долларов. Разделив ее на общее число игроков ($n = 13$), окажется, что средняя зарплата составляет 4 168 069,77 долларов. Довольно неплохой результат для средней зарплаты? Но обратите внимание, что Шакил О'Нил находится на первом месте в списке и его зарплата самая высокая во всей лиге. Если взять среднюю зарплату всех игроков Lakers помимо Шака, то она

составит 32 756 335/12 = 2 729 694,58 долларов. Сумма по-прежнему немаленькая, но уже намного меньше, чем средняя зарплата всех игроков вместе с Шакилом О'Нилом. (Конечно, фанаты скажут, что это всего лишь свидетельствует о том, насколько он важен для команды. И такой вопрос представляет собой только верхушку айсберга непрекращающихся споров по поводу статистики, которые так любят вести спортивные болельщики.)

Итак, за сезон 2001–2002 года средняя зарплата игроков команды Lakers составляла около 4,2 млн. долларов. Но действительно ли среднее значение отображает ситуацию? Иногда оно может слегка сбивать с толку, и это как раз тот случай. Все потому, что каждый год несколько ведущих игроков (таких как Шак) получают намного больше всех остальных (и, кстати, снова подобно Шаку, они чаще оказываются выше других спортсменов). В статистике это называется *выбросом* (т.е. числом в наборе данных, которое либо очень большое, либо очень маленькое по сравнению со всеми оставшимися данными). Способ подсчета среднего значения может повлиять на то, что высокие выбросы будут тянуть среднее вверх (как это сделала зарплата Шакила О'Нила в предыдущем примере) или низкие выбросы потянут среднее вниз.

Помните, как во время школьных экзаменов вы и большинство других учеников в классе получили плохие отметки, а парочка "ботаников" заработала по 100 баллов? Помните, как учитель не менял шкалу оценивания, чтобы отразить низкие результаты большинства учеников? Возможно, он использовал среднее значение, а в таких условиях оно в действительности не соответствует центру полученных баллов.

Что же еще можно использовать помимо среднего значения, чтобы показать, какой может быть зарплата "обычного" игрока NBA или каким был балл "обычного" ученика в вашем классе? Еще один статистический показатель, который используется для указания центра в наборе данных, называется медианой. Медиана до сих пор остается в тени статистики в том смысле, что ее используют совсем не так часто, как следовало бы, хотя в последнее время сообщения о ней встречаются все чаще.

Делим зарплаты до медианы

Медиана в наборе данных — это значение, которое находится точно посередине. Вот как нужно находить ее в наборе данных.

1. **Расположите числа по возрастающей.**
2. **Если в наборе данных насчитывается нечетное количество чисел, выберите то, которое расположено точно посередине**.

   Это и есть медиана.

3. **Если в наборе данных четное количество чисел, найдите среднее значение двух чисел, которые расположены посередине, это и будет медианой.**

Зарплаты игроков команды Los Angeles Lakers за 2001–2002 годы (см. табл. 5.2) уже расположены от самой низкой (внизу) до самой высокой (вверху). Поскольку список содержит имена и зарплаты тринадцати игроков, то середина — это седьмая срока сверху (или снизу), т.е. медиана зарплаты — это сумма, которую заработал Самаки Уокер, получивший в этом сезоне 1,4 млн. долл.

Эта медиана зарплаты игроков Lakers намного ниже средней зарплаты в 4,2 млн. долл. Но поскольку при подсчете средней зарплаты для этой команды учитывались выбросы (такие нетипичные значения, как зарплата Шакила О'Нила), то медиана зарплаты больше скажет вам о реальной средней зарплате в команде. (Заметьте, что только три

игрока получили больше, чем среднюю зарплату в 4,2 млн. долл., а зарплаты, бо́льшие медианы, имеют шесть игроков.) На медиану не влияют зарплаты тех игроков, которые получали очень много (как это было в случае со средним значением). (Кстати, самая низкая зарплата игрока Lakers в сезоне 2001–2002 годов составила 465 850 долларов — большие деньги в представлении многих людей, но сущие гроши по сравнению с тем, что, как вы думали, получают игроки лиги NBA!)

Правительство США часто использует медиану, чтобы обозначить центр данных. Например, Бюро переписей США сообщило, что в 2001 году медиана дохода семьи составила 42 228 долларов, что на 2,2% меньше, чем в 2000 году, когда медиана дохода была равна 43 162 долларов.

Определяем центр: сравниваем среднее с медианой

Теперь представим, что вы оказались в команде NBA и пытаетесь оговорить будущую зарплату. Если вы представляете владельцев, то хотите показать, как много все здесь зарабатывают и сколько денег вы тратите, значит, нужно принять в расчет всех суперзвезд и использовать среднее значение. Но если вы стоите на стороне игроков, то предпочтете взять медиану, потому что она реальнее отражает то, что получают игроки, находящиеся на средних строчках. Пятьдесят процентов игроков получают больше медианы, а 50% — меньше. Именно поэтому значение и называется медианой — это как разделительная линия между полосами магистральной автодороги, указывающая четко на середину.

Гистограмма — это вид графика, упорядочивающий и отражающий числовые данные в виде изображения, на котором представлены группы данных и количество или процентное отношение данных в каждой группе. (Подробнее о гистограммах и других видах визуального представления данных см. главу 4.) Если выбросы данных находятся вверху, то на гистограмме будет наблюдаться *скос вправо*, а среднее значение окажется больше медианы. (Пример гистограммы со скосом вправо показан на верхнем изображении рис. 5.1). Если выбросы находятся снизу, то для гистограммы характерен *скос влево*, и среднее значение будет меньше медианы. (Гистограмма в центре рис. 5.1 показывает, что такое скос влево.) Если данные *симметричны* (имеют примерно одинаковую форму по обе стороны от середины), то среднее значение и медиана тоже примерно равны. (Нижняя гистограмма на рис. 5.1 — пример симметричных данных на гистограмме.)

Среднее значение в наборе данных зависит от аномальных значений, в отличие от медианы. Если вы видите сообщение о среднем значении, поинтересуйтесь еще и медианой, чтобы можно было сравнить эти два статистических показателя и лучше понять, что на самом деле происходит с данными и что действительно типично для них.

Учитываем вариации

В наборе данных всегда существуют вариации, независимо от того, какие характеристики вы измеряете, ведь не всякий элемент совокупности постоянно сохраняет одно и то же значение каждой переменной. Именно благодаря изменчивости статистика и является столь увлекательной дисциплиной. Например, стоимость жилья варьируется от дома к дому, из года в год, от штата к штату. Доход семьи варьируется от семьи к семье, от страны к стране и от года к году.

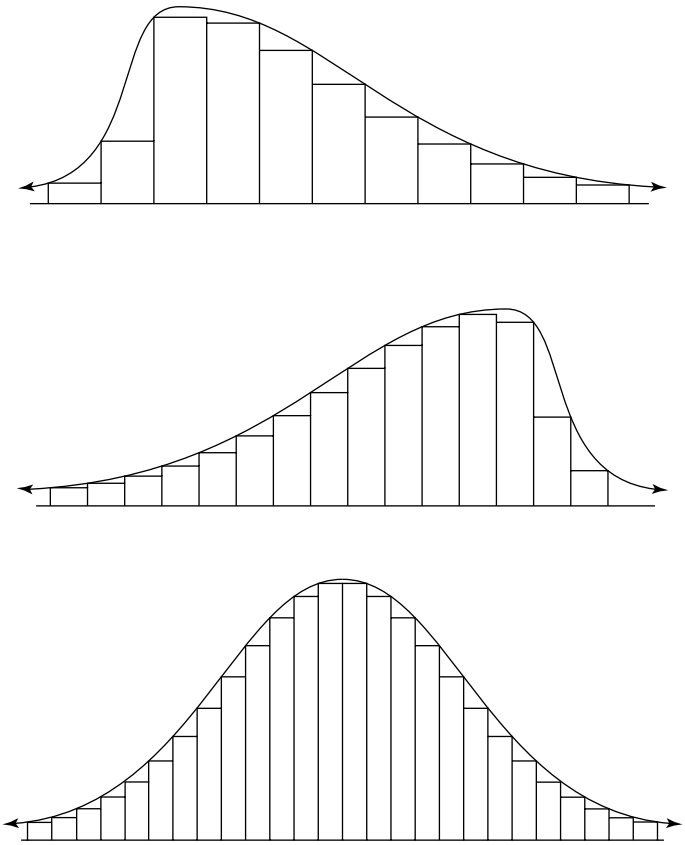


Рис. 5.1. Данные со скосом вправо, влево и симметричные данные

Количество ярдов, которые преодолевает квотербек за игру, варьируется от игрока к игроку, от игры к игре, от сезона к сезону. Количество времени, которое уходит у вас на дорогу до работы, варьируется изо дня в день. Весь фокус в работе с изменчивостью заключается в том, чтобы уметь измерять ее самым выгодным способом.

Что же все-таки означает стандартное отклонение

До сих пор самым распространенным мерилом изменчивости остается стандартное отклонение. *Стандартное отклонение* отражает типичное расстояние от любой точки в наборе данных до центра. Грубо говоря, это среднее расстояние от центра, и в данном случае центром является среднее значение. Чаще всего о стандартном отклонении не говорят в отдельности, если уж о нем упоминают (а это бывает не так уж часто), то, скорее всего, это делается мелким шрифтом, как правило, в скобках, например "($s = 2{,}68$)".

Стандартное отклонение *генеральной совокупности данных* обозначается греческой буквой σ. Стандартное отклонение выборки из генеральной совокупности обозначается буквой *s*. Поскольку чаще всего стандартное отклонение всей совокупности неизвестно, тогда выходит, что в формулы, в которых требуется это значение, подставлять нечего. Но не волнуйтесь. Выход найдется.

Поэтому, работая со статистическими показателями, поступайте так, как делают все статистики — всякий раз, когда они наталкиваются на неизвестную величину, они просто дают ее приближенное значение — и готово! Значит, в тех случаях, когда у неизвестно, вместо него используется *s*.

Когда я в этой книге использую термин *стандартное отклонение*, то имею в виду *s*, т.е. стандартное отклонение выборки. (Когда речь будет идти о стандартном отклонении генеральной совокупности, я вам скажу!)

Вычисление стандартного отклонения

Формула стандартного отклонения такова:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}},$$

Чтобы найти стандартное отклонение *s*, поступайте так.

1. **Найдите среднее значение в наборе данных.**

   Чтобы найти среднее значение, сложите все числа и разделите результат на количество чисел в наборе, *n*.

2. **От каждого числа отнимите среднее.**

3. **Возведите в квадрат все разности.**

4. **Найдите сумму всех результатов, полученных на этапе 3.**

5. **Разделите сумму квадратов (полученную на этапе 4) на количество чисел в наборе данных минус 1 ($n - 1$).**

6. **Найдите квадратный корень из получившегося числа.**

Статистики в формуле *s* делят на $(n - 1)$, а не на *n*, чтобы характеристики стандартного отклонения позволяли использовать его в теории. Например, разделив на $(n - 1)$, можно гарантировать, что стандартное отклонение в среднем не будет *смещено* (относительно цели). Если вам когда-нибудь удастся получить всю генеральную совокупность данных и вы захотите найти стандартное отклонение совокупности у, то воспользуйтесь той же формулой, что и для *s*, только в знаменателе должно быть *n*, а не $(n - 1)$!

Разберем такой небольшой пример. Предположим, у вас есть четыре числа: 1, 3, 5 и 7. Их среднее значение составляет $16/4 = 4$. Вычитая среднее из каждого числа, вы получите $(1 - 4) = -3$, $(3 - 4) = -1$, $(5 - 4) = +1$ и $(7 - 4) = +3$. Возведем каждый результат в квадрат и получим 9, 1, 1 и 9. Сложим эти числа, их сумма будет равна 20. В этом примере $n = 4$, значит, $(n - 1) = 3$, поэтому делим 20 на 3 и получаем 6,67. Наконец, извлекаем квадратный корень из 6,67, что даст нам 2,58, это и будет стандартное отклонение выбранного набора данных. Итак, для набора данных 1, 3, 5 и 7 типичное расстояние от среднего значения составляет 2,58.

Поскольку для подсчета стандартного отклонения нужно проделать много шагов, чаще всего за вас это будет делать компьютер. Но зная, как вычисляется стандартное отклонение, вы сможете лучше понять этот статистический показатель и заметить, когда он указан неверно.

Интерпретация стандартного отклонения

Истолковать стандартное отклонение в виде отдельно взятого числа довольно сложно. По сути, небольшое стандартное отклонение означает, что значения в наборе данных близки к середине, а большое стандартное отклонение говорит о том, что значения в наборе данных расположены дальше от середины.

В некоторых ситуациях (например, в производстве продукции или контроле качества), может требоваться минимальное стандартное отклонение. У конкретной детали автомобиля, которая для идеальной работы должна иметь сантиметры в диаметре, стандартное отклонение не должно быть большим. Большое стандартное отклонение в этом случае означало бы, что большинство деталей окажутся на свалке, потому что не подойдут при сборке — или же у автомобиля возникнет масса проблем на дороге.

В ситуациях же, когда вы просто наблюдаете и записываете данные, большое стандартное отклонение — это не всегда плохо. Оно всего лишь отражает большую степень изменчивости в изучаемой группе. К примеру, если изучить заработную плату всех сотрудников конкретной компании, начиная от студента-практиканта и заканчивая генеральным директором, то стандартное отклонение может быть очень большим. С другой стороны, если сузить группу и учитывать только студентов-практикантов или только руководителей высшего звена, то стандартное отклонение будет меньше, потому что зарплата у представителей каждой из этих двух групп не слишком отличается.

Определяя величину стандартного отклонения, следите за единицами. Например, стандартное отклонение, равное 2 и измеряемое в годах, равно стандартному отклонению 24 в месяцах. Кроме того, когда будете рассматривать стандартное отклонение, обращайте также внимание на величину среднего значения. Если среднее количество групп новостей в Интернете, в которых зарегистрирован пользователь, составляет 5,2, а стандартное отклонение при этом равно 3,4, то вообще-то говоря, это довольно значительная изменчивость. Но если вы говорите о возрасте пользователей групп новостей, среднее значение которого равно 25,6 лет, то стандартное отклонение в 3,4 будет сравнительно небольшим.

Еще один способ интерпретировать стандартное отклонение — это использовать его вместе со средним значением при описании места самой высокой концентрации данных. Если данные распределяются в виде колоколообразной кривой (когда большинство единиц расположены ближе к середине, а по краям их становится все меньше), то для толкования стандартного отклонения можно воспользоваться так называемым эмпирическим правилом (см. главу 4). *Эмпирическое правило* гласит, что в пределах одного стандартного отклонения в любую сторону от среднего значения должно находиться около 68% данных; в пределах двух стандартных отклонений должно быть около 95% данных, а в пределах трех стандартных отклонений от среднего значения должно помещаться около 99% данных.

В ходе исследования вопроса о том, как люди находят друзей в киберпространстве, например, было установлено, что среднее значение возраста пользователей групп новостей равно 31,65 лет, а стандартное отклонение составляет 8,61 лет. Данные были распределены в виде колоколообразной кривой. Согласно эмпирическому правилу, возраст около 68% пользователей групп новостей находится в пределах 1 стандартного отклонения (8,61 лет) от среднего значения (31,65 лет). Значит, около 68% пользователей были в возрасте от 31,65 – 8,61 лет до 31,65 + 8,61 лет, т.е. между 23,04 и 40,26 лет.

Возраст около 95% пользователей был равен от 31,65 – 2(8,61) до 31,65 + 2(8,61), т.е. от 14,43 до 48,87 лет. И наконец, возраст около 99% пользователей в Интернете составлял от 31,65 – 3(8,61) до 31,65 + 3(8,61) или от 5,28 до 57,48 лет. (Подробнее о применении эмпирического правила см. главу 8.)

Большинство людей не морочат себе голову, пытаясь учесть 99% значений в наборе данных, как правило, они довольствуются 95%. Учитывать еще одно дополнительное стандартное отклонение от среднего значения только для того, чтобы захватить каких-то 4% данных (99% – 95%) кажется многим лишней тратой сил.

Свойства стандартного отклонения

Вот некоторые характеристики, которые помогут вам интерпретировать стандартное отклонение.

- ✓ Стандартное отклонение никогда не может выражаться отрицательным числом. (Это объясняется способом вычисления и тем фактом, что оно измеряет расстояние. А расстояния не бывают отрицательными.)
- ✓ Самое маленькое из возможных значений стандартного отклонения — это 0, и такое бывает только в идеальных (выдуманных) случаях, когда все числа в наборе данных равны (т.е. отклонений от среднего значения нет).
- ✓ На стандартное отклонение влияют выбросы (очень большие или очень маленькие нетипичные числа в наборе данных). Это можно объяснить тем, что стандартное отклонение основано на *расстоянии* от среднего значения, хотя среднее значение тоже зависит от нетипичных.
- ✓ Стандартное отклонение измеряется в тех же единицах, что и исходные данные.

В поддержку стандартного отклонения

Стандартное отклонение — это одно из тех явлений, о котором нечасто говорят в средствах массовой информации, и в этом заключается большая проблема. Если вы выясните только, где находится центр данных, но не будете знать, насколько изменчивы эти данные, то это будет лишь часть правды. Более того, вы, возможно, упустите из виду самое интересное. Разнообразие придает вкус жизни, но не зная, насколько разнообразны или изменчивы данные, вы не сможете сказать, не слишком ли они пресные.

Не зная стандартного отклонения, вы также не сможете установить, все данные близки к среднему значению (как диаметр деталей автомобиля, которые сходят с конвейера, если все работает так, как надо) или же они разбросаны в широком диапазоне (как зарплата игроков NBA). Если вам говорят, что средняя начальная зарплата сотрудника компании Statistix составляет 70 тыс. долларов, вы можете подумать: “Ух-ты! Здорово!”. Но если стандартное отклонение зарплат в Statistix равно 20 тыс. долларов, то, воспользовавшись эмпирическим правилом и предположив, что данные имеют колоколообразную форму, вы установите, что будете получать от 30 до 110 тысяч (это 70 тысяч плюс или минус два стандартных отклонения, каждое из которых равно 20 тыс. долларов). В компании Statistix наблюдается заметная изменчивость в том, что касается оплаты труда, значит, по большому счету, средняя начальная зарплата в 70 тыс. долларов мало о чем говорит.

С другой стороны, если стандартное отклонение составляет всего 5 тысяч, можно уже намного лучше представить, чего ожидать от начальной зарплаты в этой компании.

Без стандартного отклонения невозможно эффективно сравнить два набора данных. Что, если у двух наборов данных примерно одинаковые средние значения и медианы? Значит ли это, что данные тоже похожи? Совсем нет. Например, у наборов данных 199, 200, 201 и 0, 200, 400 одинаковое среднее значение, равное 200, и одинаковые медианы, составляющее тоже 200. Но у них совсем разные стандартные отклонения. У первого набора данных оно очень маленькое по сравнению со вторым набором.

Журналисты часто не говорят о стандартном отклонении. Единственное объяснение, которое я могу придумать, — это то, что люди, должно быть, о нем не спрашивают. Возможно, широкая общественность еще просто не готова к стандартному отклонению. Но упоминание об этом понятии может встречаться в СМИ чаще по мере того, как все больше людей поймут, как много стандартное отклонение может рассказать о результатах. Во многих профессиях стандартное отклонение часто вычисляется и используется, потому что этот статистический показатель является принятым и широко распространенным способом оценить изменчивость.

Если не попасть в диапазон

Очень часто в средствах массовой информации говорится о диапазоне как о способе оценить изменчивость набора данных. *Диапазон* — это самое большое значение набора данных минус самое маленькое значение. Найти диапазон очень просто, нужно всего лишь расположить числа по порядку (от самого маленького до самого большого) и быстро произвести вычитание. Возможно, именно поэтому диапазон упоминается так часто. Ведь объяснить его популярность познавательной ценностью никак нельзя.

Диапазон набора данных не содержит практически никакой информации. Он зависит от двух чисел в наборе, оба из которых могут быть крайними величинами (выбросами). Я бы советовала забыть о диапазоне и попытаться вычислить стандартное отклонение, которое является более информативным критерием изменчивости набора данных.

Как и можно было ожидать, зарплатам игроков NBA свойственна значительная изменчивость. Типичный пример этому — зарплаты отдельной команды, Los Angeles Lakers, в сезоне 2001–2002 годов. Вспомним табл. 5.2, в которой указано, какие деньги получали 13 игроков команды. Средняя зарплата составляет 4 168 069,77 долларов, а медиана зарплаты равна 1 400 000. Диапазон зарплат от самой большой, равной 21 428 572 долларам (Шакил О'Нил) до самой маленькой в 465 850 долларов (Станислав Медведенко) равен 21 428 572 – 465 850 = 20 962 722. Класс — вот это разброс! Он говорит о том, какая большая разница сохраняется между самыми высокооплачиваемыми и самыми низкооплачиваемыми игроками, тут уж никаких сомнений. Но многое ли говорит этот диапазон об общей изменчивости зарплат во всей команде? Не совсем. Стандартное отклонение равно 5,98 млн долларов, это по-прежнему очень большая сумма, но поскольку стандартное отклонение подсчитывается с учетом всех зарплат в команде (а не только самой большой и самой маленькой), то для статистики оно играет намного более важную роль, чем диапазон.

Столкнувшись с итоговой статистикой, отыщите там стандартное отклонение, чтобы можно было понять, велика ли изменчивость данных. Если его нет, требуйте или обратитесь к источнику (пресс-релизу, статье из журнала, непосредственно к исследователям), где его можно найти. Не доверяйте диапазону, это слишком приблизительная оценка изменчивости, поэтому она мало что может означать.

Определяем свое местоположение: процентили

Любому интересно, какое положение он занимает по сравнению с другими. В школе оценка, полученная за контрольную, значит меньше, чем то, как этот результат соотносится с оценками других детей в классе. В таких экзаменах, как GRE и ACT, общее количество баллов часто остается неизменным из года в год, а вот оценки студентов постоянно варьируются, потому что каждый год меняется сам тест. Значит, зная свой балл, вы всегда будете знать, что он значит по сравнению с другими учащимися, сдававшими экзамен вместе с вами. Другими словами, вы выясните свое относительное положение в группе.

Понимание процентилей

Самый распространенный способ оценить относительное положение — это использовать *процентили*. Процентиль — это процентное отношение элементов в наборе данных, расположенных ниже оцениваемой точки. Если у вас 90-й процентиль, к примеру, то это значит, что 90% людей, сдававших экзамен вместе с вами, набрали меньше балов, чем вы. Также это означает, что 10% получили более высокие результаты, потому что общая сумма должна равняться 100%. (Ведь каждый экзаменующийся должен показать какой-либо результат относительно вашего, правильно?)

Процентиль — это *не* балл сам по себе. Предположим, что после сдачи экзамена GRE вам сообщили, что у вас 80-й процентиль. Это не значит, что вы правильно ответили на 80% вопросов. Суть в том, что баллы 80% других студентов были меньше ваших, а 20% студентов получили более высокие результаты, чем вы.

Подсчет процентилей

Чтобы вычислить k-й процентиль (где k — любое число от единицы до ста), сделайте следующее.

1. **Расположите все данные в наборе по возрастающей.**
2. **Умножьте процент k на общее количество чисел *n*.**
3. **Округлите результат до ближайшего целого числа.**
4. **Отсчитайте числа слева направо (от самого маленького до самого большого), пока не дойдете до значения, полученного на этапе 3.**

Например, предположим, на экзамене вы выполнили 25 тестов, и результаты, полученные по ним, расположить по возрастающей, то это будет выглядеть так: 43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99. Представим теперь, что вы хотите подсчитать 90-й процентиль для своих результатов. Поскольку данные уже расположены по порядку, вам нужно умножить 90% на общее количество чисел в ряду, получится $90\% \times 25 = 0{,}90 \times 25 = 22{,}5$. Округлим этот результат до ближайшего

целого числа, получится 23. Это значит, что, отсчитывая слева направо (от самого маленького до самого большого числа в наборе данных), вы идете до 23-го числа в наборе. Этим числом будет 98, значит, это и есть 90-й процентиль для этого набора данных.

50-й процентиль — это точка среди данных, при которой 50% данных находятся ниже, а 50% — выше нее. Вы уже знаете ее под другим именем — медиана. И действительно, медиана — это особый процентиль, 50-й.

Высокий процентиль — это не всегда очень хорошо. К примеру, если у вашего города 90-й процентиль в том, что касается уровня преступности в городах такого же размера, это значит, что в 90% городов, похожих на ваш, уровень преступности ниже, чем в вашем городе, а это не очень хорошая ситуация.

Интерпретация процентилей

Правительство США часто указывает процентили среди других статистических показателей. Например, Бюро переписей США сообщило, что медиана дохода семьи в 2001 году была равна 42 228 долларам. Кроме того, Бюро приводит различные процентили дохода семьи, например, 10-й, 20-й, 50-й, 80-й, 90-й и 95-й. В табл. 5.3 показаны значения этих процентилей.

Таблица 5.3. Доход семьи в США за 2001 год

Процентиль	Доход семьи в 2001 году ($)
10-й	10 913
20-й	17 970
50-й	42 228
80-й	83 500
90-й	116 105
95-й	150 499

Изучив эти процентили, вы поймете, что значения в нижней половине доходов ближе друг к другу, чем в верхней. Разница между 50-м и 20-м процентилем составляет около 25 000 долларов, а разрыв между 50-м и 80-м процентилем равен 41 000 долларов. Также разница между 10-м и 50-м процентилем составляет всего около 31 000 долларов, а между 50-м и 90-м процентилем этот разрыв равен целых 74 000 долларов.

Взглянув на эти процентили и на то, как они распределены среди данных, вы сможете сказать, что этот набор данных, если его представить на гистограмме, имел бы скос вправо. (По сути, *гистограмма* — это столбиковая диаграмма, в которой данные разбиваются на группы, после чего показывается число в каждой группе. Подробнее о гистограммах см. главу 4.) Это объясняется тем, что более высокие доходы имеют больший разброс, чем меньшие доходы, значения которых находятся ближе друг к другу. В данном отчете среднее значение не показано, потому что на него очень повлияли бы такие выбросы (семьи с очень высокими доходами), в результате чего среднее бы отклонилось вверх, искусственно завысив доходы семей в США.

Процентили используются в средствах массовой информации и во многих общественных документах. Они могут сообщить дополнительную интересную информацию о данных, в том числе и насколько ровно или неровно распределены данные, насколько они симметричны, а также некоторые важные моменты о данных, например, где находится медиана. Кроме того, процентили могут рассказать и о том, каково ваше положение (ваш балл, доход и т.д.) в наборе данных. Иногда значение среднего не важно, если вы знаете, насколько выше или ниже среднего находитесь. Подробнее об этой и других сферах применения процентилей см. главу 8.

Независимо от того, какие именно данные подытоживаются или какой статистический показатель используется, помните, что итоговая статистика не может предоставить всей информации о данных. Но если статистические показатели выбраны правильно и не вводят в заблуждение, то они быстро сообщают массу информации. Ошибки или упущения все же могут случаться, поэтому обязательно будьте готовы искать дополнительные, не такие популярные статистические показатели, чтобы с их помощью выяснить, какова же истина, скрытая за данными.

Часть III

Определение шансов

The 5th Wave

Рич Теннант

"Итак, давайте рассмотрим статистические вероятности данной ситуации. Нас четверо, он один. Филипп, вероятно, начнет вопить, Нора, вероятно, упадет в обморок, ты, вероятно, начнешь кричать на меня за то, что я не запер машину, а я, вполне вероятно, побегу как сумасшедший, как только он двинется на нас."

В этой части...

Приготовьтесь бросать кости! В этой части раскрываются некоторые секреты азартных игр (и первое правило — остановиться в нужный момент!). Знание основ вероятности поможет вам определиться, на что рассчитывать во время игры или в ситуации, которая требует от вас скорейшего принятия решения. Возможно, вы даже удивитесь, что вероятность и ваша интуиция — это не всегда одно и то же!

Глава 6

Каковы шансы? Правила вероятности

В этой главе...

- Использование вероятности в повседневной жизни и в работе
- Принцип действия вероятности
- Борьба вероятности и вашей интуиции
- Связь вероятности со статистикой

В этой главе вы узнаете, как вероятность используется в обычной жизни и на работе, а также познакомитесь с некоторыми правилами вероятности. Знайте, что вероятность и интуиция — это не всегда одно и то же. Учитесь избегать самых распространенных заблуждений, связанных с вероятностью, а еще выясните, какое отношение вероятность имеет к статистике.

Рискуем с вероятностью

Часто можно услышать выражение: "Каковы шансы, что это случится?". Например, вы читаете о том, что два торнадо обрушились на один и тот же городок в Канзасе с интервалом в пятьдесят лет. В самолете вы встречаетесь с другом, которого не видели многие годы. Вы прокалываете две шины за один день. Ваша команда-неудачница выигрывает чемпионат по баскетболу. Происходят странные вещи, и иногда вы просто удивляетесь: "Каковы шансы? Кто бы мог такое предвидеть? Какова вероятность того, что это повторится?". Все эти вопросы связаны с вероятностью.

Но вероятность касается не только изучения странностей в жизни (хотя это, несомненно, веселое увлечение тех, кто работает с ней). На самом деле вероятность связана с систематической работой с неизвестным, изучением вариантов развития событий, составлением самых вероятных сценариев или разработкой запасного плана на тот случай, если самые вероятные сценарии окажутся бесполезными.

Жизнь — это последовательность непредсказуемых событий, но вероятность можно использовать для того, чтобы попытаться предсказать, высоки ли шансы того, что определенные события произойдут. Укажем еще некоторые, более приземленные сферы обычной жизни, в которых вы можете встретиться с вероятностью.

- ✓ Метеорологи предсказывают на сегодня вероятность дождя — 80% (наверное, лучше пойти на работу в плаще).
- ✓ По опыту вы знаете, что небольшое превышение скорости увеличивает ваши шансы попасть в "зеленую волну" по дороге на работу (конечно, пока вас за это не оштрафуют).

- ✓ По дороге на работу вы думаете, скажется ли ваш помощник Боб сегодня больным, ведь сегодня пятница, а он в 75% случаев болеет именно по пятницам. (Вы также размышляете над тем, каковы шансы, что Боб нашел другую работу. По-видимому, вероятность такого события намного ниже.)
- ✓ В перерыв вы покупаете лотерейный билет, потому что "Кто-то же должен выиграть, и это вполне могу быть я!" (Кстати, ваши шансы сорвать джек-пот в этот раз равны 1 на 89 миллионов, так что на многое не рассчитывайте.)
- ✓ По телевизору вы слышите сообщение о новейших исследованиях в медицине, где вам говорят, что если в течение дня немного вздремнуть, то это снизит шансы бессонницы на 35%. (Окончания рассказа вы уже не слышите, потому что успеваете уснуть.)
- ✓ Вечером вы смотрите игру своей любимой бейсбольной команды, которая снова выигрывает, и начинаете мечтать о том, какие же у нее шансы победить в чемпионате мира.

Кроме того, вероятность используется практически в каждом виде профессиональной деятельности, ею интересуются маркетинговые компании и инвестиционные фирмы, правительственные организации и производственные заводы, больницы и рестораны. Далее перечислены лишь некоторые примеры того, как можно применять вероятность в работе.

- ✓ Небольшая компания проводит исследование, чтобы выяснить, нравится ли покупателям ее продукция и можно ли реализовывать ее в режиме онлайн. Если компания права, то она сможет заработать кучу денег, если же ошибается, то ей грозит банкротство.
- ✓ Компания, производящая картофельные чипсы, должна убедиться, что пачки заполняются в соответствии со всеми требованиями: если чипсов будет слишком мало, то у компании начнутся большие неприятности из-за неправильного позиционирования продукции. Если же чипсов слишком много, то компания понесет убытки. Делается выборка пачек, и на основе этой выборки компания рассчитывает вероятность того, что с оборудованием что-то не так.
- ✓ Господин А.Я. Надеющийся решил осуществить свою мечту и теперь претендует на пост губернатора, но прежде чем ввязываться в неприятности, связанные с необходимостью собрать миллионы долларов для проведения кампании, он проводит опрос, чтобы определить свои шансы победить на выборах.
- ✓ Фармацевтическая компания разработала новое лекарство, понижающее артериальное давление. Исходя из клинических испытаний на добровольцах, компания определяет вероятность того, что человек, принимающий это лекарство, поправит свое здоровье и/или будет ощущать какие-либо побочные эффекты.
- ✓ Инженер-генетик с помощью вероятности пытается предсказать генетические модели и результаты в самых разных сферах, от выведения новой культуры до выявления наследственных болезней на ранних этапах жизни.

- ✓ Менеджер ресторана думает о вероятности в связи с тем, когда и сколько клиентов придут к нему в ресторан. В соответствии с этим он дает распоряжение готовить блюда.
- ✓ Биржевой брокер использует вероятность каждый день, принимая решения. Он постоянно думает о том, как изменится цена определенных акций, что следует делать — покупать или продавать, а также что нужно сообщить клиентам.

Получаем преимущество: основы вероятности

Вероятность существует везде, и все же иногда ее трудно понять, потому что она, как кажется, неподвластна интуиции. Первый шаг на пути к тому, чтобы получить преимущество, — это понять некоторые основные правила вероятности и то, как эти правила можно применять. Когда статистики говорят о вероятности, то речь идет о вероятности *исхода*, т.е. одного конкретного результата случайного процесса. А что такое *случайный процесс*, спросите вы? Это любой процесс, у которого возможен не один-единственный исход, а несколько вариантов результата. Например, если вы один раз бросаете кубик с шестью гранями, то исход (цифра на выпавшей грани) может быть одним из шести возможных вариантов: 1, 2, 3, 4, 5 или 6.

Излагаем правила

Познакомьтесь со следующими правилами вероятности.

- ✓ Вероятность исхода — это процент того, сколько такой исход может наблюдаться. Часто его можно подсчитать, разделив количество раз, которое исход может наблюдаться, на общее количество всех возможных исходов. Например, вероятность того, что на единожды брошенном кубике выпадет цифра 1, составляет 1 из 6 или 1/6 (или 16,7%).
- ✓ Любая вероятность — это число (процент) от 0 до 100%. (Заметьте, что в статистике проценты часто выражаются в виде пропорций — чисел между 0 и 1.) Если вероятность исхода равна 0%, то он не может произойти *никогда*, ни при каких условиях. В большинстве случае вероятность не равна ни 0%, ни 100%, а находится где-то посередине.
- ✓ Сумма вероятностей всех исходов равна 1 (или 100%).
- ✓ Чтобы определить вероятность целого ряда исходов, нужно сложить вероятности каждого исхода отдельно. Например, вероятность выбросить нечетное число (1, 3 или 5), один раз бросив кости, равна сумме вероятностей выбросить 1, 3 и 5: $1/6 + 1/6 + 1/6 = 1/2$ или 50%.
- ✓ *Дополнение* события — это все возможные исходы *за исключением* тех, что составляют это событие. Вероятность дополнения события равна 1 минус вероятность события. Например, выбросить 1, 2, 3, 4 или 5 — это дополнение к тому, чтобы выбросить 6 за одно метание кости, значит, вероятность выбросить любое из чисел: 1, 2, 3, 4 или 5 равна 1 минус вероятность выбросить 6, т.е. $1 - 1/6 = 5/6$.

Если дополнение события определить затруднительно, то часто бывает проще найти вероятность самого события и отнять ее из 1. Зачем отнимать эту вероятность из 1? Потому что сумма вероятностей всех исходов равна 1, значит, вероятность дополнения события плюс вероятность самого события тоже должна быть равна 1.

Бросаем кости

Во время игры в кости бросают два кубика, и по условиям игры число 7 имеет большое значение. В этой игре любой исход состоит из двух чисел, выпавших на кубиках (например, комбинация 6 и 2 является исходом). Числа на двух кубиках складываются, чтобы получить сумму (см. табл. 6.1). Сумма в 7 очков выпадает чаще всего, поэтому у нее самая высокая вероятность выпадения. Банкомет (игрок, бросающий кости) бросает кости, и первый бросок называется *исходным* (например, комбинация 6 и 2 определяет исходный бросок, равный 8). Если сумма исходного броска равна 7, то банкомет отстраняется от игры, и все теряют свои ставки. Если сумма исходного броска не равна 7, то банкомет продолжает бросать кости до тех пор, пока не выпадет либо 7, либо сумма, выпавшая во время исходного броска (в нашем примере это 8). Любой человек за столом может сделать ставку на то, выпадет ли 7 до того, как снова будет получена сумма исходного броска. Именно поэтому все игроки в кости в таком приподнятом настроении и подбадривают банкомета. Они надеются, что тот принесет им удачу и выбросит ту комбинацию, на которую они сделали ставку.

Вы можете использовать перечисленные в предыдущем разделе правила вероятности, чтобы изучить исходы сумм двух кубиков и определить их вероятность. Знаете, какая сумма занимает второе место по частоте выпадений?

Когда бросаются два кубика, то у каждого из них есть шесть возможных результатов, вместе они дают 36 (6×6) возможных комбинаций двух чисел или 36 возможных пар. Поскольку в нашем примере исход — это сумма, полученная на двух кубиках, то у вас есть 11 разных возможных исходов в диапазоне от 2 (т.е. 1 + 1) до 12 (т.е. 6 + 6). В табл. 6.1 показаны 36 возможных результатов, а также 11 разных сумм, выпавших на двух кубиках.

Воспользовавшись первым правилом вероятности (см. раздел “Излагаем правила”), вы можете подсчитать вероятность для каждой из возможных сумм. Перечень всех исходов и их вероятностей называется *моделью вероятности*. Например, сумма, равная семи, может получиться несколькими способами: (1, 6), (2, 5), (3, 4), (4, 3), (5, 2) и (6, 1). Имея 36 возможных комбинаций для двух кубиков, вероятность того, что выпадет сумма в 7, равна 6/36 или 1/6. Точно так же вы можете определить вероятность получить суммы от 2 до 12. Модель вероятности для суммы двух кубиков показана в табл. 6.2. Таким образом, следующая после 1/6 самая высокая вероятность в 5/36 характерна для двух сумм, которые расположены по обеим сторонам от 7 (6 и 8). Обратите внимание, что сумма всех вероятностей в табл. 6.2 равна 1. Также заметьте, что вероятности постоянно увеличиваются по мере того, как сумма кубиков растет с 2 до 3, 4, 5, 6, и достигает пика, когда сумма на двух кубиках равна 7 (именно поэтому количество комбинаций, в результате которых может получиться сумма, равная 7, больше, чем для любой другой суммы). И дальше вероятности постоянно уменьшаются, когда сумма переходит от 8 к 9 и т.д. вплоть до 12.

Таблица 6.1. Исходы сумм на двух кубиках

Цифры на кубиках	Сумма	Цифры на кубиках	Сумма	Цифры на кубиках	Сумма	Цифры на кубиках	Сумма	Цифры на кубиках	Сумма	Цифры на кубиках	Сумма
1, 1	2	2,1	3	3, 1	4	4, 1	5	5, 1	6	6, 1	7
1, 2	3	2, 2	4	3, 2	5	4, 2	6	5, 2	7	6, 2	8
1, 3	4	2, 3	5	3, 3	6	4, 3	7	5, 3	8	6, 3	9
1, 4	5	2, 4	6	3, 4	7	4, 4	8	5, 4	9	6, 4	10
1, 5	6	2, 5	7	3, 5	8	4, 5	9	5, 5	10	6, 5	11
1, 6	7	2, 6	8	3, 6	9	4, 6	10	5, 6	11	6, 6	12

Таблица 6.2. Модель вероятности для суммы на двух кубиках

Сумма на двух кубиках	Вероятность
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Выигрыш в любой азартной игре основан на вероятности. Например, в крэпсе можно делать дополнительные ставки на то, какой будет сумма определенного броска. Если вы делаете ставку на то, что в конкретном броске банкомет выбросит сумму в 2 и это действительно происходит, то вы выиграете больше, чем если бы делали ставку на то, что выпадет сумма в 8. Почему? Потому что, согласно табл. 6.2, вероятность получить сумму в 2 на двух кубиках меньше, чем сумму в 8. Именно поэтому это и называется азартной игрой. (Подробнее о вероятности и азартных играх см. главу 7.)

Вычислить вероятности суммы на двух кубиках было довольно легко. Но другие случаи могут быть намного сложнее, например, вероятность разных исходов в покер, таких как фул-хаус, флеш-рояль или две пары. Однако важно помнить, что ранг карт на руках игрока в покер напрямую зависит от вероятности получить такой набор карт. Высший ранг в покере — это флеш-рояль (десятка, валет, дама, король и туз одной масти). Это объясняется тем, что вероятность собрать такие карты самая низкая.

Модели и имитация

Не все вероятности можно подсчитать с помощью математики. В тех случаях, когда математика не помогает, для оценки вероятности используются другие методы или же применяются уже известные вероятности для того, чтобы сделать предположения об окружающем мире. Например, для того, чтобы определить вероятность урагана на побережье США, предположительное место и время стихии, используются сложные компьютерные модели. Такие модели основаны на данных, полученных об ураганах в прошлом, а также на текущих погодных условиях и других переменных. Ученые преобразуют информацию в сложную математическую модель, которая пытается предсказать вероятный урон, который может нанести ураган. В этой сфере еще многое предстоит сделать, но прогресс наблюдается постоянно. Подобные модели могли бы спасать жизни, собственность и экономить миллионы долларов, если бы люди заблаговременно могли знать, чего ожидать, и готовиться к этому.

Другие модели основываются на данных наблюдений. В ходе опроса американских общин, проводимого в 2001 году. Бюро переписей США, изучались семьи в Колумбусе, Огайо, чтобы понять, что представляет собой община. Одной из изученных характеристик был состав семьи (женатые пары, другие семьи, одинокие люди и другие жители, не имеющие семьи). Данные подытожены на рис. 6.1. Эти статистические данные о *выборке* семей могут служить моделью вероятности для получения информации обо *всех* семьях в Колумбусе, штат Огайо.

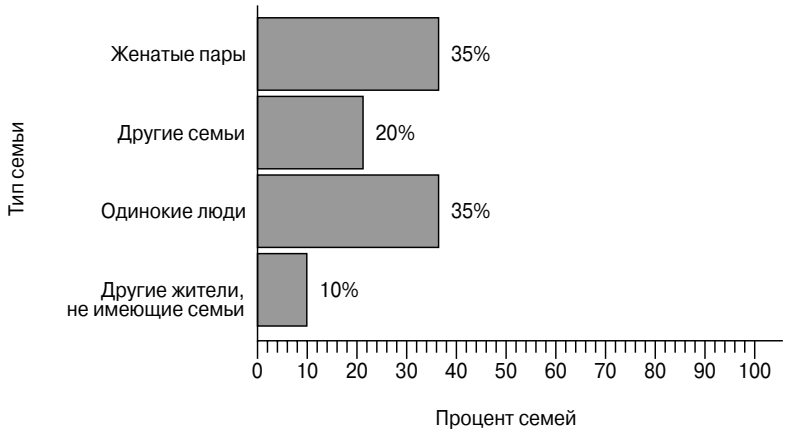


Рис. 6.1. Состав семей в г. Колумбус, штат Огайо, 2001 год

Например, поскольку 35% опрошенных представляли собой женатые пары, то можно сказать, что вероятность того, что выбранная наугад семья в Колумбусе окажется женатой парой, составляет 35%. Также можно воспользоваться правилами вероятности и сформулировать другие утверждения относительно жителей Колумбуса в 2001 году. К примеру, какова вероятность того, что в выбранном наугад доме живет семья любого типа? Это будет сумма вероятностей попасть на женатую пару (35%) или семьи, которая относится к категории "Другие семьи" (20%). Значит, вероятность того, что в выбранном наугад доме в городе Колумбус, штат Огайо, живет семья, равна 35% + 20% = 55%. (Следовательно, вероятность найти дом, в котором нет семьи, равна 100% – 55% или 45%).

Модель вероятности на рис. 6.1 нельзя использовать для других местностей помимо Колумбуса, Огайо, потому что в ходе данного опроса выборка семей была сделана только в Колумбусе. Поэтому использовать данные для описания иной совокупности было бы неправомерно. (Подробнее об опросах и о том, как они связаны с совокупностями, см. главу 16.)

Имитация — это еще один способ оценить вероятность в том случае, когда применить формулу невозможно. В ходе *имитации* (*моделирования*) процесс повторяется снова и снова в одних и тех же условиях (обычно с применением компьютера), при этом ведется запись исходов. Вероятность любого исхода оценивается в процентном отношении раз, сколько такой исход наблюдался в ходе имитации. Например, спортивный болельщик, у которого было слишком много свободного времени, сымитировал на своем компьютере тысячи соревнований Национальной атлетической ассоциации колледжей

NCAA и использовал эти имитации, чтобы предсказать, что с вероятностью 95% баскетбольный чемпионат NCAA в 2002 году выиграет Duke. Как это часто бывает в вопросах везения, предсказание оказалось ошибочным (эту команду удалили на первых этапах соревнований), что еще раз доказало: единственное, в чем можно быть уверенным, — это неуверенность.

Интерпретация вероятности

Вероятность можно толковать двумя способами: как краткосрочный шанс или как долгосрочное процентное отношение. В короткой перспективе вероятность события — это процентный шанс того, что событие произойдет в следующий раз. Например, метеоролог утверждает, что на завтра вероятность осадков равна 40%. Или вероятность для бейсболиста отбить подачу в среднем составляет 0,291 (т.е. в среднем у него 29,1% шанса отбить следующую подачу).

Кроме того, вероятность означает процент того, сколько раз событие произойдет в долгосрочной перспективе (в течение долгого периода времени в ходе повторяющихся попыток при неизменных условиях). Значит, 40% вероятности того, что завтра будет дождь, связан с данными наблюдений, сделанными в течение многих дней за прошедшие годы (дождь шел в 40% таких дней). Среднюю вероятность для бейсболиста отбить подачу 0,291 можно понимать как отношение удачного приема подач к общему числу подач (в этом случае предполагается, что он отобьет мяч 291 раз из 1000 подач).

Избегаем заблуждений

Основные правила вероятности кажутся совершенно понятными, но очень часто вероятность действует вопреки интуиции. В этом разделе вы узнаете о самых распространенных заблуждениях, которые связаны с вероятностью.

Кажущаяся большая вероятность

Если бы нужно было записать последовательность результатов, которые вы получите, шесть раз подбросив *правильную монету* (т.е. симметричную), то, скорее всего, вы вряд ли бы записали что-то вроде ОРРРРО (где “О” значит “орел”, а “Р” — “решка”), потому что это не кажется слишком “случайным”. Но такая последовательность орлов и решек имеет ровно столько же шансов выпасть, как и любая другая. Это объясняется тем, что вероятность получить орел такая же, как и вероятность выбросить решку. Значит, если бы вам нужно было сравнить вероятность получить два орла (из шести подбрасываний) с вероятностью получить шесть орлов (из шести подбрасываний), то значения были бы разными. Вероятность получить два орла (из шести подбрасываний) выше, потому что это можно сделать бо́льшим количеством способов, чем каждый раз выбрасывать именно орел.

В случае с лотереей последовательность 1, 2, 3, 4, 5, 6 имеет столько же шансов выиграть, как и любая другая комбинация шести цифр, даже если совсем непохоже, что это *когда-нибудь* случится. Благодаря этому факту вы поймете, что все другие комбинации так же *маловероятны*, как и эта. Но если вы сделаете ставку на такую комбинацию и выиграете, то, наверное, не захотите ни с кем делиться своим выигрышем.

Предсказания на короткий или долгий период

Вероятность хорошо подходит для предсказаний долгосрочных моделей, но в краткосрочной перспективе она работает не так удачно. В долгосрочной перспективе вам известно, что если только вероятность события не равна 0, то когда-нибудь оно произойдет, и в зависимости от вероятности вы можете примерно представить себе, долго ли придется ждать. Однако вы не знаете точно, *когда* случится это событие. Именно поэтому вероятность так интересна, а заядлые игроки возвращаются к игре снова и снова.

Например, если я шесть раз подброшу монету, и шесть раз подряд выпадет орел, каким, как вы думаете, будет исход следующего броска, орел или решка? Возможно, вы считаете, что у меня выпадет решка, значит, вероятность выбросить на этот раз решку выше. Но на самом деле вероятность выбросить при следующем броске решку по-прежнему та же, что и во всех предыдущих случаях. Если монету подбрасывать много раз, то можно ожидать, что в 50% случаев выпадет орел, а в 50% — решка. Но невозможно предсказать, *когда именно* этот орел или решка выпадет при броске. (Значит, хотя и кажется, что должна выпасть решка, вероятность получить орел или решку при следующем броске по-прежнему равна 50%.) В конечном итоге решки начнут выпадать, но невозможно предсказать, когда именно.

Думаем 50 на 50

Одно из распространенных заблуждений — полагать, что каждая ситуация с двумя возможными исходами — это ситуация "50 на 50" (другими словами, 50% вероятности, что вы получите любой из двух результатов, как это было при броске монеты). Многие люди думают, что только потому, что возможны два исхода, каждый из них может произойти (один шанс из двух). На самом деле очень часто это не так. Не всякая ситуация похожа на подбрасывание монеты. Во многих случаях у одного из исходов вероятность больше.

Например, вспомните об автоматическом сигнале светофора на пешеходном переходе через шумную улицу. Точно ли 50% раз повторится сигнал "Идите"? Нет. Если это улица с интенсивным движением, то светофор будет останавливать поток машин реже, а пешеходам придется ждать дольше, пока им представится возможность перейти через дорогу. Возьмем, к примеру, спорт: баскетболист стоит на линии штрафного броска. Разве его шансы попасть в корзину равны 50 на 50? (Он либо попадет, либо нет). Но шансы будут равны 50 на 50, только если общее процентное отношение удачных штрафных бросков этого игрока составляет 50% из многих попыток. Скорее всего, процент все же выше.

Интерпретация редких событий

Вероятность может быть очень спорным вопросом, особенно если дело касается редких событий. *Редкое событие* — это событие с небольшой вероятностью. Но что конкретно это значит? А это значит, что для каждой конкретной ситуации или человека это событие маловероятно, но если ситуация в течение долгого периода времени повторится достаточное количество раз или с достаточным количеством людей, то такое событие обязательно когда-нибудь где-нибудь с кем-нибудь произойдет. Это характерно для ситуаций, в которых есть группа людей с редким для одного города заболеванием, и требуется определить, произошло это по какой-то причине (из-за загрязнения воздуха, воды, почвы и т.д.) или случайно (то, о чем большинство людей не задумываются).

Поскольку кажется маловероятным, что редкое событие действительно произойдет, то естественно, всегда хочется свалить вину на кого-то или что-то. В одних ситуациях это правильно, а в других — проявление слепого случая. Разве повышение средней температуры воздуха в течение трех лет подряд говорит о глобальном потеплении? Если на ферме две коровы родили двухголовых телят, означает ли это, что эти коровы чем-то больны? Сколько шин нужно проколоть, чтобы это можно было называть тенденцией? Если вы проанализируете какое-то событие, которое уже произошло, и скажете: "Каковы были шансы, что это событие произойдет именно здесь?", значит, подобное событие было неожиданным, вы не были уверены, что оно обязательно где-нибудь и когда-нибудь случится.

Например, если вы будете достаточно долго подбрасывать правильную монету, то в конце концов у вас случайно выпадет ряд орлов. Когда-нибудь такое должно было бы произойти. Здесь некого винить, кроме случая. Однако средства массовой информации склонны видеть закономерность, если событие повторяется два или больше раз, например, похищение детей в разных частях страны, пожары в ночных клубах или случаи заболевания редкой болезнью в одном городе. Может быть, такие ситуации нужно изучать с целью установить возможные причины, но все же СМИ не должны забывать, что иногда события повторяются просто случайно, и за этим не скрывается никакая сенсация. Интересно также отметить, что люди по-разному оценивают вероятность редких событий в зависимости от того, хорошее оно, например, выигрыш в лотерею ("Это обязательно с кем-то произойдет, так почему бы не со мной!") или плохое, например, получить удар молнией во время соревнования по гольфу ("Такое может произойти один раз на миллион. Мне это не грозит!") Возможно, это просто человеческая природа. ***Памятка***: человеческая природа не придерживается законов вероятности.

Чтобы избежать самых распространенных заблуждений, связанных с вероятностью, не забывайте о следующем.

- ✓ Вероятность неэффективна для предсказания краткосрочного поведения. Она эффективна в случае с долгосрочными моделями.
- ✓ Если возможны всего два исхода, то совсем не обязательно, что у каждого из них 50% шансов произойти.
- ✓ Если где-то наблюдается целый ряд редких событий, то это могло произойти просто случайно. Редкие события обязательно произойдут с кем-то, где-то, когда-то.
- ✓ Нельзя считать себя удачливым только потому, что процесс повторяется снова и снова при заданных условиях (как в случае с азартными играми). У вероятности нет памяти.
- ✓ Последовательности исходов, которые кажутся "чем-то бо́льшим, чем простая случайность", часто имеют ту же вероятность, что и последовательности, похожие на "случайные". Например, вы, возможно, думаете, что у последовательности ОРРРРО меньше шансов выпасть, чем у ОРРРОР, потому что она не кажется "случайной". На самом деле вероятность у таких последовательностей одинакова, потому что в каждом исходе есть четыре орла и две решки (и при определении вероятности в данном случае порядок не имеет значения).

Связь вероятности со статистикой

Вероятность — это интересно, но какое отношение она имеет к статистике? Хороший вопрос. Это не совсем очевидно, но вероятность и статистика идеально подходят друг другу. Относительно выборки элементов собираются данные, после чего подсчитываются статистические показатели, чтобы подытожить эту информацию. Но на этом вы не останавливаетесь. Следующий шаг — сделать определенное предположение, обобщение, вывод или принять решение касательно совокупности, из которой была сделана эта выборка. И тут в дело вступает вероятность.

Оценка

Часто данные собирают для того, чтобы оценить пропорции или средние значения генеральной совокупности. Например, врачи оценивают шансы того, что у человека случится сердечный приступ, сначала собирая информацию о весе пациента, его индексе массы тела, поле, возрасте, наследственности, питании, спортивной подготовке и т.д. Затем они сравнивают эту информацию с данными, которые были получены у выборки людей с похожими характеристиками, после чего определяется вероятность (или уровень риска) для пациента за определенное время пережить сердечный приступ. Инженеры оценивают среднее количество автомобилей, которые проедут в час пик по определенному отрезку автомагистрали, записывая данные с помощью специальных приборов, встроенных в тротуар. Когда данные собраны, вероятность используется для определения того, какая часть информации о выборке будет варьироваться от выборки к выборке, изо дня в день, из часа в час и т.д.

Предположение

Статистика применяется для того, чтобы делать самые разные предположения — обо всем, начиная от погоды и предполагаемой численности населения и заканчивая распространением болезни или будущей стоимостью акций на бирже. Данные собираются в течение определенного времени, затем они анализируются и создается модель, которая не только хорошо подходит для этих данных, но также позволяет сделать некоторые предположения на будущее. Вероятность помогает людям использовать такие модели и оценить, насколько точными могут быть предположения, учитывая имеющиеся данные. Кроме того, вероятность помогает ученым определить, каким будет самый вероятный сценарий развития событий в конкретных условиях.

Например, Бюро переписей США обнародовало свои предположения, касающиеся численности населения страны. В настоящее время можно рассмотреть прогнозы на 2100 год. В 2000 году предполагаемая численность населения на 2003 г. была равна 282 798 000, а по данным на май 2003 года (о чем говорится на сайте Бюро переписей) численность населения в стране уже составляла 291 065 455. Значит, на середину года прогноз уже отстал от действительности на 8,3 млн. человек, но это только 2,8% населения в данной ситуации. Оценить общую численность населения Соединенных Штатов Америки в будущем очень непросто. Сложно пересчитать всех, кто живет в стране в данный момент! (Кстати, согласно данным Бюро переписей, ожидается, что в 2100 году численность населения США составит 570 954 000 человек.)

Решение

Принятие многих решений связано со статистикой и вероятностью. Метод лечения часто выбирается с учетом того, сколько процентов людей почувствовали улучшение благодаря именно такому лечению. Вероятность того, что оно поможет следующему человеку, будет оцениваться на основе процентного отношения других пациентов, которым данное лечение помогло. В большинстве документов, которые пациент обязан подписать перед операцией, описываются возможные побочные эффекты или осложнения, а также приводятся факты, как часто они наблюдаются. (Подробнее о медицинских исследованиях см. главу 17.)

Проверка качества

Еще целый ряд решений, связанных с вероятностью, принимается в производстве. Многие компании, выпускающие продукцию, проводят определенный контроль качества, т.е. выбирают продукцию, которая сходит с конвейера, и оценивают ее качество согласно существующим критериям. Вероятность применяется для того, чтобы решить, нужно ли приостановить процесс производства из-за проблем с качеством продукции. Расхождения между проверенной продукцией и критериями могут объясняться случайной изменчивостью или тем, что была сделана нерепрезентативная выборка. Точно так же эти расхождения могут означать, что что-то не так с самим процессом. Если необоснованно остановить производство, это повлечет за собой большие затраты денег и времени, но не остановить процесс вовремя — значит, понести убытки в том, что касается удовлетворения клиентов компании. Значит, вероятность используется для принятия довольно важных решений в производстве. (Подробнее о контроле качества см. главу 19.)

При переносе результатов с выборки на всю совокупность вероятность используется для оценки точности таких обобщений. Это нужно, чтобы решить, какой вывод наиболее вероятен и почему. Принимая решение о ситуации с неизвестным исходом, вероятность используют для оценки собранных доказательств, для того, чтобы сделать выбор на основе такой оценки, а также чтобы определить шансы того, что принятое решение окажется правильным. (Подробнее см. главу 14.)

Глава 7

Азарт и победа

В этой главе...

- Почему казино зарабатывают деньги
- Вероятность в основе азартных игр
- Хватаем деньги и бежим: советы для игроков

Лас-Вегас — одно из самых восхитительных мест на земле. И все же дешевые закусочные и величественные римские гладиаторы, окружающие Дворец Цезаря, — это не самые привлекательные в городе достопримечательности (хотя я настоятельно рекомендую посетить и увидеть и то, и другое!). Лас-Вегас — это, наверное, рай для азартных игроков, место, куда нужно отправиться, если вы чувствуете себя везучим и хотите сорвать большой куш. Тот факт, что огромные толпы народу, играющие в Лас-Вегасе, проигрывают все, совершенно не важен для новых полчищ потенциальных победителей, которые ежедневно садятся в самолеты и летят в Неваду с надеждой и предвкушением богатства. Ведь как бы там ни было, но кто-то же ведь должен выиграть? И это вполне можете оказаться вы. Вряд ли эта глава поможет вам выиграть большие деньги в Вегасе (или в любом другом месте, где вы решите попытать счастья с Госпожой Удачей), но все же с ее помощью вы поймете, с чем вам придется иметь дело, вы найдете здесь советы, касающиеся того, как выиграть побольше или, по крайней мере, проиграть поменьше.

Почему казино до сих пор в игорном бизнесе

Казино — прекрасное место, здесь царит восхитительная атмосфера, везде — ярко раскрашенные автоматы, мерцающие огоньки, счастливые крупье и помощники, а также никаких часов или окон (казино не хотят, чтобы посетители обращали внимание на время). Все хорошо продумано, от планировки здания (чтобы попасть в туалет, вам нужно пройти мимо несметного числа игровых автоматов) до узора на ковре (они специально сделаны яркими и утомительными для глаз, потому что владельцы не хотят, чтобы вы смотрели вниз — вы должны смотреть на место действия и переходить от стола к столу).

Игорный бизнес превратился в настоящую науку, и те, кто руководит игорными заведениями, очень хорошо овладели этой дисциплиной. Они предлагают вам отличное развлечение и шанс выиграть большие деньги, от вас же требуется только платить. Многих людей манит шанс выиграть тысячи долларов или новую машину, всего один раз дернув рычаг игрового автомата или только однажды собрав идеальный набор карт в блэк-джек. Конечно же, такой шанс существует, но если бы все выигрывали много, то казино уже давно были бы вынуждены отойти от дел. Поэтому им приходится придумывать способы, как забрать ваши деньги, что они и делают, устанавливая правила игр, которые дают им незначительную фору в каждом раунде, следя за тем, что если кто-то действительно

выигрывает, то выигрывает много (о чем и рассказывает затем всем подряд), и приглашая вас оставаться в казино как можно дольше. Они знают, что чем дольше вы играете, тем выше их шансы забрать ваши деньги.

О том, как обыграть казино практически в любой игре, написаны сотни книг. Каждый автор хочет убедить вас, что именно его стратегия поможет вам сорвать банк. Но на самом деле лучшее, что способны сделать такие книги — это помочь проиграть не слишком много, потому что, учитывая правила игр, заведение (казино) всегда имеет преимущество. В некоторых играх, таких как блэк-джек, это преимущество меньше, а в других больше, как, например, в случае с игровыми автоматами, которые, по слухам, приносят некоторым казино 80% от общей выручки. Самое главное, что вам нужно запомнить по поводу любой азартной игры — это то, что всегда нужно уметь вовремя остановиться. Если бы все так поступали, то казино уже давно бы прогорели. Но, конечно же, такого не будет, потому что остановиться в нужный момент может далеко не каждый.

С другой стороны, существует еще одна хорошая стратегия — сократить расходы и остановиться до того, как вы проиграете слишком много, вместо того, чтобы надеяться, что, в конце концов, удача повернется к вам лицом и вы отыграете все потраченные деньги. Часто в таких ситуациях вы проигрываете вдвойне. Казино утверждают, что вы останетесь и будете играть несмотря ни на что, повышая их шансы в конечном итоге получить еще больше ваших денег. Судя по тому, сколько роскошных казино сейчас строится в Лас-Вегасе, похоже, их владельцы не так уж и неправы.

Что можно сделать, чтобы свести к минимуму шансы на проигрыш или проиграть не так много? Еще до того, как вы подключитесь к игре, установите собственные границы (например, остановиться, когда вы уже потратили определенную сумму) и всегда придерживайтесь их. Играя, сумейте победить азарт до того, как зайдете слишком далеко. Знайте меру еще до того, как начнете играть.

Знание вероятности помогает в лотерее

Два самых важных инструмента, которые вы можете использовать в азартной игре, — это информация о ваших шансах выиграть и правильное понимание того, что эти шансы означают. Несомненно, озвученная вероятность может не согласовываться с вашей интуицией, но ведь когда дело касается денег, вы же верите фактам, правда? (Подробнее об использовании вероятности в статистике см. главу 6.)

Ниже приводятся некоторые распространенные *заблуждения*, касающиеся вероятности.

- ✓ Любую ситуацию, у которой есть только два возможных исхода, можно назвать ситуацией 50 на 50 (50% шансов выиграть и 50% шансов проиграть).
- ✓ Комбинация номеров 1, 2, 3, 4, 5, 6 в лотерее никогда не будет выигрышной, эти числа не достаточно случайны.
- ✓ Отличная мысль — купить не один лотерейный билет, а сто. Таким образом ваши шансы выиграть заметно возрастают.
- ✓ Если у Джо и Сью уже есть три девочки, то у них очень неплохой шанс в следующий раз родить мальчика.
- ✓ Чем дольше вы играете на игровом автомате, тем выше ваши шансы выиграть.

В этом разделе я развею все эти заблуждения, и вы сможете реально взглянуть на любую игру, лучше понять, на что можно рассчитывать, и как себя нужно вести. Возможно, некоторые факты разрушат волшебство и азарт, которые связаны в вашем понимании с игрой, но в то же время это, вероятно, объяснит, почему вы никогда не встретите статистика (ни одного из тех, кого я знаю) среди профессиональных или сиюминутных азартных игроков. Более того, Лас-Вегас не стал бы возражать, если бы статистики больше не проводили там свои конференции, ведь когда в последний раз они заходили в казино, то оставили там совсем немного денег! (Лично я играю только на игровых автоматах, где нужно бросать пятицентовые монеты. По крайней мере, 20 долларов хватит надолго.)

Шансы 50 на 50

Вы собираетесь бросать правильную монету (т.е. монету, у которой не нарушена симметрия), с одной стороны у нее орел, с другой — решка. Каковы шансы выбросить орел? Пятьдесят процентов. Каковы шансы выбросить решку? Пятьдесят процентов. Если бы вы делали ставку на исход подбрасывания этой монеты, то ваши шансы выиграть (или проиграть) были бы равны 50 на 50. Почему так? Потому что у вас есть два возможных исхода, орел или решка, и у каждого из них есть равные шансы произойти. Вы знаете семейную пару, ожидающую ребенка. Естественно, у них может родиться либо мальчик, либо девочка. Оба этих исхода равновероятны, значит, у этой пары 50 на 50 шансов, что родится девочка (или мальчик). С другой стороны, если вы купите один из тысячи лотерейных билетов в мотогонках, то у вас тоже есть два возможных исхода — победить или проиграть. Но означает ли это, что ваши шансы на победу составляют 50 на 50? Нет. Почему? Потому что вы не единственный, кто купил билет!

Немного лучше разобраться в этих тонкостях помогут четыре правила вероятности.

- ✓ Вероятность того, что состоится определенный исход, представляет собой процентное отношение количества раз, которые этот исход ожидается в долгосрочной перспективе, если точно те же условия будут повторяться снова и снова.
- ✓ Любая вероятность — это число от 0 до 1. Вероятность, равная 0, значит, что исход невозможен. Вероятность, равная 1, означает, что исход гарантирован.
- ✓ Все вероятности всех возможных исходов должны в сумме давать 1. Это значит, что вероятность того, что исход *не* произойдет, равна 1 минус вероятность того, что исход *действительно* произойдет.
- ✓ Вероятность события (комбинации исходов) равна сумме вероятностей отдельных исходов, из которых состоит событие.

В случае с монетой возможны два исхода, орел или решка. Количество способов выбросить орел равно 1, количество способов получить решку тоже равно 1. Общее число возможных исходов 2: орел или решка. Значит, в данной ситуации шанс выбросить орел равен 0,5 или 50%. То же касается и решки. Это ситуация 50 на 50.

Но если внимательно посмотреть на первое правило, то вы поймете, почему не всегда два исхода имеют шансы произойти 50 на 50. Действительно, нужно учитывать количество способов, какими можно реализовать исход. В случае с билетом на мотогонках вы можете либо выиграть, либо проиграть. Количество способов выиграть — 1, потому что организаторы вытянут только один выигрышный лотерейный билет. Количество способов проиграть — 999, потому что все оставшиеся билеты проигрышны. Общее количе-

ство исходов — 1000. Это значит, что ваши шансы выиграть составляют 1/1000 = 0,001, а шанс проиграть равен 999/1000 = 0,999. Естественно, в том, что касается вашего билета, у вас есть всего два возможных исхода (выиграть или проиграть), но у этих исходов *неравные* шансы произойти, значит, такая ситуация, несомненно, не относится к числу ситуаций 50 на 50.

В жизни очень мало ситуаций 50 на 50. Чтобы оказаться в одной из них, у вас должно быть всего два возможных исхода *и* вероятности обоих должны быть равными, т.е. составлять 50%. В большинстве случаев с двумя возможными исходами они имеют разную вероятность.

Вытягиваем выигрышные номера

Итак, вы готовы сыграть в лотерею Powerball. Вы слышали, что на сегодня джек-пот составляет 200 млн. долларов, поэтому на этот раз можете купить больше лотерейных билетов, чем обычно, чтобы увеличить свои шансы на выигрыш. И вы готовы выбрать числа. (Вам нужно выбрать пять разных чисел от 1 до 53, затем вы выбираете одно число от 1 до 42 как специальный номер (Powerball), причем он может совпадать или не совпадать с одним из тех пяти, что вы уже выбрали.) Чтобы выиграть большой джек-пот, нужно правильно выбрать (в любом порядке) все первые пять чисел плюс правильно назвать специальное число. Какую же комбинацию выбрать? Номер на футболке вашего брата, день рождения мамы, четыре цифры из номера социальной страховки, возраст вашей собаки в месяцах или числа, которые вы видели во сне прошлой ночью? Эти варианты ничем не хуже других, потому что любая выбранная комбинация может выиграть точно так же, как и все остальные.

Все кажется разумным, пока вы не задумаетесь над комбинацией 1, 2, 3, 4, 5 и числом Powerball 6. Складывается впечатление, что такая комбинация ни за что не выиграет, потому что эти числа недостаточно случайны. Ну, вот как раз тот случай, когда интуиция вас подводит. На самом деле у такой комбинации точно те же шансы на победу, как и у любой другой. Возьмем для примера случай, когда возможные числа выбираются из 1, 2, 3, 4, и вы должны выбрать 2 числа. Шесть возможных исходов таковы: 1-2, 1-3, 1-4, 2-3, 2-4, 3-4. Ваш шанс выиграть равен 1 из 6. Обратите внимание, что у комбинации 1-2 те же шансы на победу, что и у любой другой. То же касается и комбинации 1, 2, 3, 4, 5 и числа Powerball 6. Такие комбинации, как 23-16-05-24-18 с числом Powerball 12, возможно, кажутся более вероятными, но помните, что для того, чтобы сорвать джек-пот, вам нужно угадать все числа до единого.

Такая комбинация, как 1, 2, 3, 4, 5 с числом Powerball 6 кажется маловероятной, но на самом деле у нее такие же шансы на победу, как и у любой другой. Такая комбинация позволяет понять, насколько *малы* в действительности шансы выиграть джек-пот. (Реальные шансы сорвать джек-пот в лотерее Powerball с условиями, описанными выше, равны 1 на 120 526 770.)

Обычно в лотерее вероятность выиграть называется шансом на победу, и в данном случае оба эти термина обозначают одно и то же. Но то, как используются понятия шансов в спортивных ставках (лошадиные скачки, футбольные матчи, бокс и т.д.), отличается от описанного здесь. Более сложная разновидность шансов в пари не охвачена данной книгой.

Покупка лотерейных билетов — лучше меньше, да лучше

Лотерейный билет Powerball стоит всего один доллар и дает вам шанс выиграть многомиллионный джек-пот. Вы полагаете, что кто-то все же должен в конечном счете сорвать банк, поэтому решаете тоже рискнуть. Ведь если не играть, то ничего и не выиграешь. Если вы реально оцениваете свои шансы выиграть и проиграть, то покупка время от времени нескольких лотерейных билетов может быть дешевым развлечением.

Однако проблемы начинаются тогда, когда люди покупают множество билетов, думая, что тем самым увеличивают свои шансы на победу. Хотя приобретение ста билетов (а не одного) действительно в сто раз увеличивает шанс выиграть, но вы должны также понимать, что шансы выиграть много очень малы — почти равны нулю. И если число, очень близкое нулю, умножить на сто, оно все равно останется очень близким нулю. Если вы не можете себе позволить проиграть эти сто долларов (а у вас почти все шансы именно проиграть), тогда вообще не нужно участвовать в лотерее.

Чтобы лучше понять вероятность выиграть джек-пот в лотерее Powerball, посмотрите на рис. 7.1 Здесь показаны выигрыши и шансы получить их в конкретном розыгрыше лотереи. (В большинстве розыгрышей Powerball шансы получить одни и те же суммы остаются неизменными.) Серыми кружочками обозначены пять шариков, выбранных от 1 до 53, а черный кружок обозначает число Powerball. Шансы выиграть 3 доллара равны 1 из 70 или 0,0142 (около 1,5%). Заметьте, что это скорее вероятность, чем реальный шанс, но примем в расчет и ее. Поднимаясь вверх по шкале, выигрыши возрастают, а вот шансы получить их снижаются, причем резко. Например, угадать четыре числа из пяти — это примерно 1 шанс из 12 000, а правильно назвать все пять чисел — 1 шанс из 3 млн. Наконец, угадать пять чисел плюс число Powerball — для этого у вас есть 1 шанс из более чем 120 млн.

Вы можете выиграть более 10 млн долларов!		
Угадывание	**Выигрыш**	**Шансы**
+	Jackpot	1:120 526 770
	$100 000	1:2 939 ,677
+	$5 000	1:502 195
	$100	1:12 249
+	$100	1:10 685
	$7	1:261
+	$7	1:697
+	$4	1:124
	$3	1:70

Рис. 7.1. Выигрыши и шансы получить их в лотерее Powerball

Почему шансы так резко снижаются, если мы добавляем всего одно число к выигрышным? Это можно проиллюстрировать на маленьком примере. Предположим, вам нужно выбрать 2 числа от 0 до 9. Если правильно угадать нужно только одно число, то ваши шансы — 1 из 10. Если же нужно правильно назвать оба, то шансы падают до 1 из 45. Это объясняется тем, что 10 вариантов для первого числа мы умножаем на 9 вариантов второго числа (потому что повторения запрещены), что дает нам 90 вариантов. Но теперь эти 90 вариантов нужно разделить на 2, потому что числа можно назвать в произвольном порядке и все равно выиграть. (Например, комбинации 1-0 и 0-1 не считаются двумя разными вариантами.) Шансы резко меняются именно из-за этого умножения. Если помимо правильного угадывания пяти чисел вам еще нужно правильно назвать и число Powerball, шансы умножаются еще на 42 (потому что у вас есть 42 варианта для числа Powerball).

В целом шанс получить *любой* выигрыш (а не только крупную сумму), как утверждают устроители лотереи, равен 1 из 36. Поскольку шанс выиграть вообще — это то же, что шанс выиграть любой приз, сложите вероятности всех выигрышей, у вас получится 1 из 36. Здесь используется четвертое правило вероятности (см. раздел “Шансы 50 на 50” чуть выше в этой главе).

Прежде чем ввязываться в какую-либо азартную игру, всегда определите шансы или вероятность победить. И не тратьте денег больше, чем вы можете себе позволить не слишком заметно для кошелька. В случае с играми, где возможен большой приз, шансы выиграть всегда крайне малы, и купив больше билетов или сыграв множество раз, вы не увеличите эти шансы настолько, чтобы это могло оправдать затраты. Как говорится, лучший способ удвоить деньги в азартной игре — это сложить их пополам и сунуть в карман!

Предсказываем пол ребенка

У Джо и Сью уже есть три дочери, и вскоре они ожидают появления четвертого ребенка. На этот раз они очень хотят, чтобы родился мальчик. Друзья и родственники думают, что сейчас шансы на это больше, потому что у них уже были три девочки подряд. Но правы ли они? Это напоминает людей за игрой в кости, подбадривающих банкомета (человека, который бросает кубики), потому что он “на подъеме” и не может ошибиться. Действительно ли существует полоса везения или неудач, и могут ли после определенных событий вырасти или уменьшиться шансы повторения того, что уже было?

Во многих ситуациях, особенно в азартных играх, полосы везения или неудач не существует, потому что всякий раз, когда вы приступаете к игре, все начинается с нуля, и исход предыдущего раза никак не влияет на исход в этот раз или в будущем. Другими словами, когда события независимы друг от друга, вероятность того, что тенденция сохранится, не меняется всякий раз, когда происходит игра.

В случае с Джо и Сью это может оказаться неожиданностью, но вероятность родить мальчика по-прежнему осталась 50 на 50, такой же, как и была, независимо от того, что у них уже есть три дочери. Точно так же, если вы трижды подбрасываете монету, и всякий раз выпадает орел, не стоит ожидать, что в четвертый раз у решки будет больше шансов. Шансы по-прежнему 50 на 50.

Вероятность работает только в предсказании долгосрочных моделей, а не на короткий период. Если вы знаете, что у вас 50% шансов выбросить орел, это значит, что, если вы бросаете монетку много раз, то в половине случаев выпадет орел, а в половине — решка. Но невозможно предсказать, *когда* именно выпадет то или другое. Просто в долгосрочной перспективе эти исходы приближаются к своим реальным вероятностям. Такое явление называется *законом средних чисел*, о котором речь пойдет в следующем разделе. Многие люди используют это понятие, чтобы объяснить, почему заканчивается полоса везения или неудач, даже хотя таких полос в действительности не существует.

Пытаемся выиграть на игровом автомате

Однорукий бандит (т.е. игровой автомат) — это мощная сила. У таких машин есть лоток, сделанный из специального материала, благодаря чему звук сыплющихся монет разносится далеко вокруг. Автоматы мигают и пищат каждый раз, когда вы что-то выигрываете, они пищат даже тогда, когда вы бросаете в щель монету, просто чтобы придать дополнительный антураж особой атмосфере казино. Кое-кто говорит, что выигрышные машины (те, в которых самые большие призы) расположены возле входа в казино и в конце ряда. Другие же утверждают, что самые строгие "бандиты" находятся возле столов, где играют в блэк-джек, потому что люди не хотят отвлекаться на писк. Владельцы казино никогда не выдают свои секреты, поэтому невозможно сказать наверняка, но несомненно одно: игровые автоматы могут очень быстро выманить все ваши деньги. Не нужно никаких специальных умений — для того, чтобы дернуть за рычаг, хватит нескольких секунд, что будет стоить от пяти центов (в моем случае) до тысяч долларов (есть такие специальные автоматы в некоторых казино). Все траты и рывки рычага делаются в погоне за огромным джек-потом, который может ожидать вас при следующей попытке.

Одно из самых распространенных заблуждений, связанных с игровыми автоматами, состоит в том, что чем дольше играешь, тем выше шансы на победу. Именно на это делают ставку казино. Но на самом деле справедливо как раз обратное утверждение, и все из-за закона средних чисел. *Закон средних чисел* гласит, что в долгосрочной перспективе средние числа приблизятся к ожидаемым значениям. С точки зрения статистики, *ожидаемое значение* — это средневзвешенное среднее всех исходов, основанное на их вероятностях.

Что касается любой азартной игры в казино, то у заведения всегда немного больше шансов выиграть в конкретном раунде. Это значит, что в долгосрочной перспективе вы должны ожидать, что у вас очень большие шансы проиграть небольшую сумму и очень маленькие шансы выиграть большой приз. Казино делает ставку на то, что в долгосрочной перспективе все усредняется в их пользу, даже если принимать в расчет большие джек-поты.

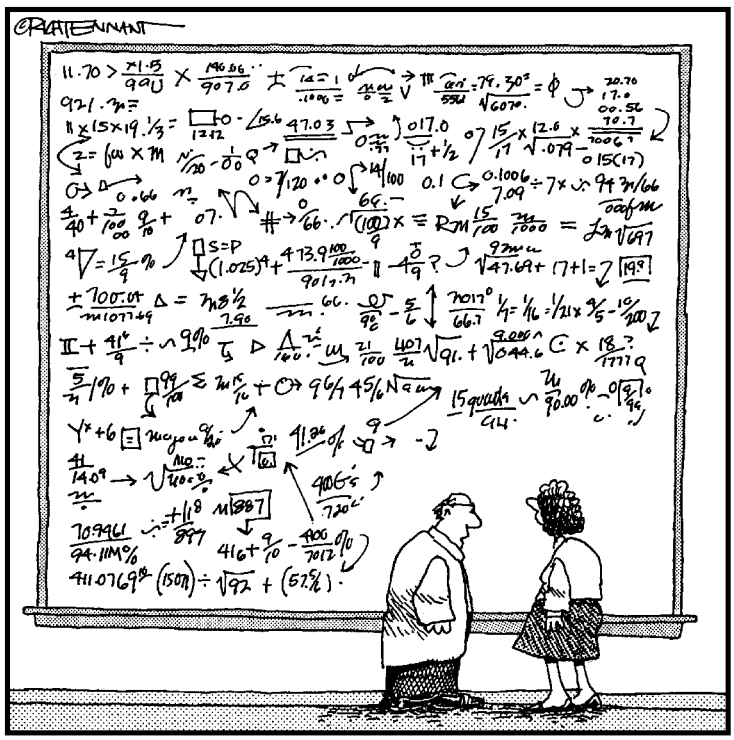
Лично для вас это значит, что каждая игра будет заканчиваться для вас с небольшим отрицательным ожидаемым значением. И чем дольше вы играете, тем больше денег потеряете, потому что общее ожидаемое значение — это сумма ожидаемых значений всех игр, а все эти ожидаемые значения — отрицательные числа. Теперь вы знаете, почему казино предлагает бесплатные алкогольные напитки во время игры и почему вы не увидите часов на стенах или окон, чтобы уследить за сменой времен года, пока сидите у игрового автомата. Казино делают ставку на то, что вы совершенно забудете о законе средних чисел во время игры.

Единственный шанс — это вовремя остановиться, до того, как вступит в силу закон средних чисел. Помните, вероятность всегда на стороне казино, потому что они удерживают ее в долгосрочной перспективе. Вам не обмануть матушку-природу, точно так же вам не обмануть вероятность. И если вы все же выиграете, то хватайте свои деньги и убегайте!

Часть IV

Разбираемся в результатах

The 5th Wave — Рич Теннант



"А что именно мы хотим здесь сказать?"

В этой части...

Эта часть поможет вам понять основы статистики, т.е. ту информацию, благодаря которой вы сможете разобраться в более сложных вопросах, присущих любой статистической операции. Вы узнаете, как оценивать изменчивость разных выборок, как вывести формулу для измерения точности статистического показателя, а также как определить положение элемента относительно остальной совокупности (что называется мерой относительного положения). Освоив все это, вы сможете уверенно вычислять и разбираться во всех существующих статистических данных.

Глава 8

Мера относительного положения

В этой главе...

- Изучение колоколообразной кривой
- Самые распространенные случаи использования эмпирического правила
- Вычисление и интерпретация нормированных параметров
- Подробнее о процентилях

Единственный способ качественно интерпретировать статистические результаты — иметь то, с чем их можно сравнить, что позволит взглянуть на полученные результаты в иной плоскости. Например, предположим, что какая-то студентка медицинского факультета по имени Роди проходит стандартизованный тест на получение сертификата физиотерапевта и набирает 235 баллов. Что в таком случае означают эти 235 баллов? Ничего, если это все, что вам известно. Нужно определить положение этого балла, выяснив, где он находится относительно других баллов, полученных по результатам теста. Лампочка, работающая больше 1200 часов — это чудо природы или обычная электрическая лампочка? Вы не ответите на этот вопрос, не зная срока службы большинства лампочек. Предположим, средний балл Боба на экзамене по математике по окончанию курса равен 78. Это тройка или четверка? Все зависит от того, как его средний балл соотносится со средними баллами других учеников этого класса (и от доброты преподавателя!).

В этой главе вы прочтете, как находить и интерпретировать относительное положение отдельных результатов. Ваша главная задача — понять, где находится элемент относительно всех других элементов совокупности. В главах 9 и 10 я объясню, как определять и интерпретировать относительное положение результатов выборки (например, среднего выборки и доли выборки в генеральной совокупности). В данном случае цель — определить положение среднего или пропорции вашей выборки по сравнению с совокупностью всех возможных значений среднего или пропорции выборки.

Распрямляем колоколообразную кривую

Первый шаг в определении положения конкретного результата — это создание перечня или изображения всех возможных значений, которые может принимать переменная в совокупности, с указанием того, как часто это происходит. Все это называется *распределением*.

Существует много разновидностей распределения. Например, баллы в одном классе (назовем его классом господина Среднего) распределены равномерно, т.е. каждой группе присуждается равное количество баллов (см. верхнее изображение на рис. 8.1), а вот в другом классе (пусть это будет класс господина Усредненного) баллы расположены по-

лярно, т.е. ученики получают либо очень высокие, либо очень низкие оценки (как показано на нижнем изображении рис. 8.1). (Большинство распределений находятся все же где-то между этими двумя крайностями.) Обратите внимание, что для любого вида распределения общее число процентов должно быть равно 100%, потому что каждое значение балла обязательно должно попадать в это распределение.

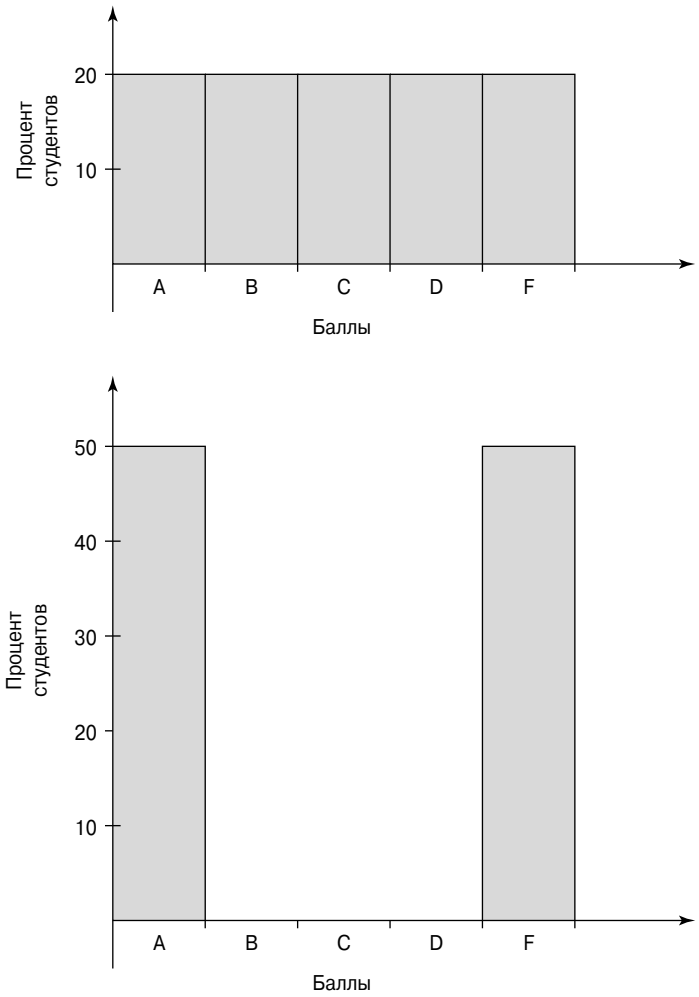


Рис. 8.1. Распределение баллов в двух классах

Колоколообразная кривая описывает данные, характеризующие переменную, которая содержит неограниченное (или очень большое) число возможных значений. Эти значения распределены среди совокупности таким образом, что при внесении их на гистограмму получившаяся в результате фигура напоминает колокол. По сути, это значит, что где-то в середине распределения у вас есть большая группа элементов, а когда вы движетесь к краю (в любом направлении), элементов становится все меньше. Многие переменные в реальном мире (например, баллы стандартизованного теста, срок годности продукции, высота, вес и т.д.) имеют распределение, которое можно представить в виде

колоколообразной кривой. Поэтому колоколообразная кривая играет довольно важную роль и выделяется среди всех других возможных распределений.

В статистике распределение, которое имеет колоколообразную форму, называется *гауссовым* или *нормальным распределением*. Оно показано на рис. 8.2. В этом примере переменная — это количество часов, которое, как ожидает некая компания (давайте назовем ее "Туши свет"), прослужат ее электрические лампочки. (Хотели бы вы быть тем самым человеком, которому нужно проверить такой забавный фактик?)

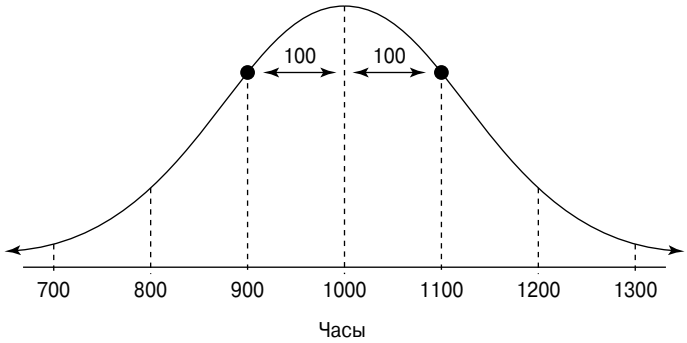


Рис. 8.2. Распределение срока службы лампочек от компании "Туши свет"

Характеристики гауссова распределения

У каждой колоколообразной кривой (гауссова распределения) есть определенные свойства. С их помощью вы можете определить относительное положение конкретного результата в распределении. Ниже перечислены свойства, присущие всем гауссовым распределениям. Подробнее о них речь пойдет в последующих разделах.

- ✓ Форма кривой симметрична.
- ✓ В центре распределения находится возвышение, склоны которого равномерно уходят вправо и влево.
- ✓ Среднее значение находится четко посередине распределения. Среднее значение совокупности обозначается греческой буквой м.
- ✓ Среднее и медиана — это одно и то же значение, что объясняется симметрией.
- ✓ Стандартное отклонение показывает обычное (почти среднее) расстояние между средним и всеми данными. Стандартное отклонение совокупности обозначается греческой буквой у.
- ✓ Около 95% значений находятся в пределах двух стандартных отклонений от среднего.

Описываем форму и центр

Гауссово распределение *симметрично*, что значит, что если сложить его пополам точно посередине, то две половинки окажутся зеркальными копиями друг друга. Поскольку кривая симметрична, то *среднее* (точка равновесия) и *медиана* (точка, по обе стороны от которой лежат половины данных) равны и находятся в центре распределения. Срок

службы электрических лампочек, показанный на рис. 8.2, имеет гауссово распределение, среднее (и медиана) которого составляет 1000 часов. (Подробнее о среднем и медиане см. главу 5, симметрия описана в главе 4.)

Определение изменчивости

Форма и среднее не единственные важные характеристики, которые нужно учитывать при работе с распределением. Изменчивость значений тоже крайне важна, даже несмотря на то, что средства массовой информации в основном игнорируют эту характеристику и приводят в своих сообщениях только среднее. На рис. 8.2 можно увидеть, что у лампочек от компании "Туши свет" диапазон службы охватывает от 700 до 1300 часов, при этом у большинства лампочек он составляет от 900 до 1100 часов. Как потребителю, нужна ли вам при покупке лампочки такая изменчивость? Наверное, нет. Конкурирующая компания (пусть она называется "Включай свет") пытается создать лампочку с меньшей изменчивостью срока службы. Среднее срока службы ее лампочек по-прежнему равно 1000 часов, но эта компания научилась изготавливать лампочки с более стабильным сроком службы, в диапазоне от 940 до 1060 часов, при этом большинство лампочек работают от 980 до 1020 часов (рис. 8.3).

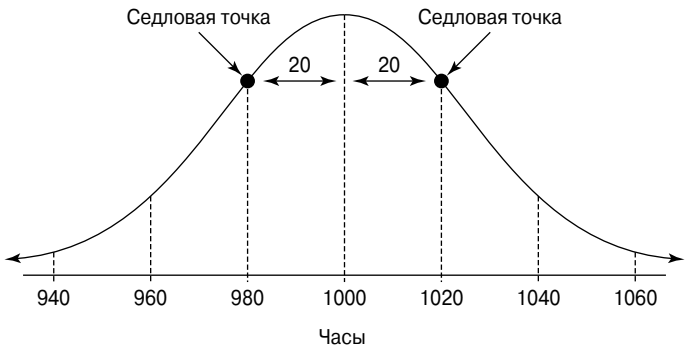


Рис. 8.3. Распределение срока службы лампочек от компании "Включай свет"

Изменчивость в распределении измеряется и обозначается количеством стандартных отклонений. (Формула стандартного отклонения описывается в главе 3.) В случае с нормальным распределением стандартное отклонение играет особенно важную роль, потому что это расстояние от среднего до того места на распределении, которое называется *седловой точкой*. В каждом нормальном распределении есть две седловых точки, на одинаковом расстоянии от среднего. Чтобы найти седловую точку, начните со среднего и двигайтесь вправо или влево до тех пор, пока кривая из перевернутой чаши (выпуклость вверх) не превратится в чашу, стоящую правильно (выпуклостью вниз). На рис. 8.2 и 8.3 седловые точки обозначены жирным. Стандартное отклонение срока службы лампочек от компании "Туши свет" (см. рис. 8.2) составляет 100 часов. Стандартное отклонение срока службы более стабильных лампочек от компании "Включай свет" (см. рис. 8.3) равно 20 часов. (Подробнее о стандартном отклонении см. главу 5.)

Перед изучением результатов проверьте шкалы как на вертикальной, как и на горизонтальной оси любого распределения и определите стандартное отклонение. В зависимости от шкалы распределение может выглядеть более

сжатым или более растянутым, чем должно. Например, рис. 8.2 и 8.3 кажутся похожими, но их шкалы совершенно разные. Лучше будет сравнивать сроки службы лампочек обеих компаний, если изобразить распределения с одинаковой шкалой, как это показано на рис. 8.4. Теперь вы видите, насколько более разбросанным является срок службы лампочек, которые выпускает компания "Туши свет", по сравнению с лампочками, изготовленными в компании "Включай свет". Срок службы лампочек от "Включай свет" больше сконцентрирован вокруг среднего.

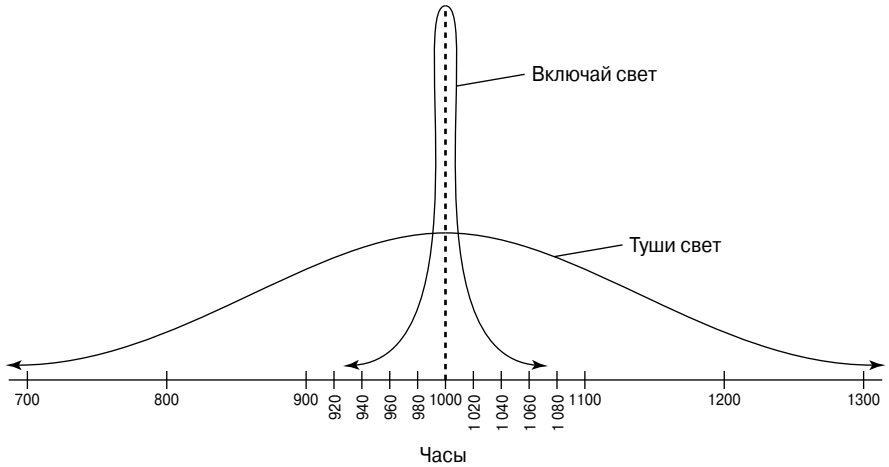


Рис. 8.4. Изменчивость срока службы лампочек от компаний "Туши свет" и "Включай свет"

Ищем большинство значений: эмпирическое правило

Если распределение в центре имеет выпуклую форму, — а гауссово распределение, естественно, подходит под эту характеристику — вы можете сделать определенные утверждения о том, где будет находиться большинство значений, при помощи 1, 2 или 3 стандартных отклонений от среднего. Правило, которое позволяет это сделать, называется *эмпирическим правилом*.

Эмпирическое правило гласит, что если у распределения выпуклая форма, тогда справедливы следующие утверждения.

- ✓ Около 68% значений находятся в пределах 1 стандартного отклонения от среднего (или между средним минус 1 стандартное отклонение и средним плюс 1 стандартное отклонение). В статистике это обозначается как $\mu \pm \sigma$.
- ✓ Около 95% значений расположены в пределах 2 стандартных отклонений от среднего (или между средним минус 2 стандартных отклонения и средним плюс 2 стандартных отклонения). В статистике это обозначается как $\mu \pm 2\sigma$.
- ✓ Около 99% (на самом деле, 99,7%) значений лежат в пределах 3 стандартных отклонений от среднего (или между средним минус 3 стандартных отклонения и средним плюс 3 стандартных отклонения). Статистики обозначают это так: $\mu \pm 3\sigma$.

Если вы не знаете среднее значение и стандартное отклонение совокупности, то в формулах эмпирического правила замените стандартное отклонение совокупности у стандартным отклонением выборки *s*. Точно так же можно заменить среднее совокупности м средним выборки $\bar{x}$. (Подробнее см. главу 3.)

На рис. 8.5 показано эмпирическое правило. Причина, по которой 68% находятся в пределах 1 стандартного отклонения от среднего, состоит в том, что большинство значений гауссова распределения скоплены в середине, ближе к среднему (как показано на рис. 8.5). Помните, что это колоколообразная форма. Если передвинуться еще на одно стандартное отклонение в любую сторону от среднего, вы захватите еще 30% значений (что в сумме и даст 95% значений), ведь теперь вы учитываете меньше из выпуклой части и больше со склонов. Наконец, передвигаясь еще на 1 стандартное отклонение в любую сторону от среднего, вы захватите оставшиеся части склонов, на которые приходится 4,7% (почти все, что осталось) значений, и в сумме получится 99,7% данных. Большинство исследователей, представляя полученные результаты, довольствуются 95%, потому что учитывать 3 стандартных отклонения от среднего, чтобы захватить оставшиеся 4,7%, не имеет смысла.

В предшествующем изложении эмпирического правила нужно особо подчеркнуть слово *приблизительно*. Эти результаты только приблизительны (но это очень хорошие приближения). Далее в этой главе (см. раздел "Обратимся к нормированному параметру") вы узнаете, как получить более точную информацию относительно того, какой процент значений в распределении находится между, ниже или выше определенных значений. Однако эмпирическое правило играет в статистике очень важную роль (поскольку захватив два стандартных отклонения в обе стороны от среднего значения, вы захватите около 95% значений).

В случае с электрическими лампочками от компании "Туши свет" (см. рис. 8.2) стандартное отклонение составляет 100 часов, а среднее значение — 1000 часов. Используя эмпирическое правило, вы можете определить относительное положение некоторых важных значений среди этих данных. К примеру, согласно представленной модели, около 68% лампочек, как ожидается, проработают от 900 до 1100 часов (1000 ± 100), около 95% лампочек должны проработать от 800 до 1200 часов ($1000 \pm 2 \times 100$), а 99,7% лампочек имеют срок службы от 700 до 1300 часов.

Чтобы ответить на другие вопросы о сроке службы лампочек, вы можете наряду с эмпирическим правилом использовать симметрию гауссова распределения. Например, какой процент лампочек от компании "Туши свет" должен проработать 1000 часов или больше? Ответ — 50%, потому что медиана находится на отметке в 1000 часов, а половина значений больше медианы. Какой процент лампочек от компании "Туши свет" должен проработать больше 1200 часов (см. рис. 8.2)? Ответ — 2,5%. Почему? Потому что 95% электрических лампочек имеет срок службы от 800 до 1200 часов, а учитывая, что общее количество лампочек, показанных на кривой, должно составлять 100%, то в оставшихся зонах двух склонов должно находиться 5% значений. Лампочки, работающие больше 1200 часов, показаны только на правом склоне, и в связи с симметрией вы можете разделить 5% пополам, что и даст в результате 2,5%. Значит, лампочки, работающие больше 1200 часов — это, скорее, чудо приро-

ды, потому что такое бывает только в 2,5% случаев (по крайней мере, у компании “Туши свет”). В случае с компанией “Включай свет” лампочка, которая работает дольше — это нечто неслыханное, потому что 1200 часов — это намного дальше, чем 3 стандартных отклонения от среднего для лампочек, изготовленных этой компанией (см. рис. 8.4).

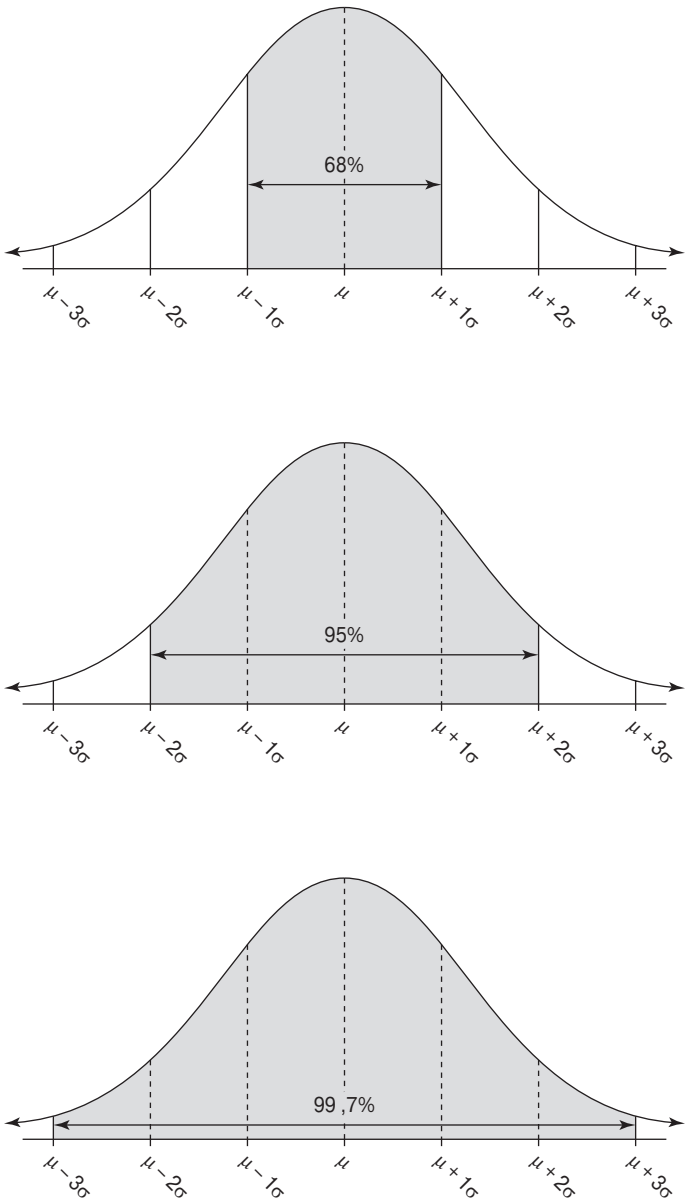


Рис. 8.5. Эмпирическое правило (68, 95 и 99,7%)

Мораль сей басни такова: если вам нравится играть в азартные игры, покупайте лампочки, выпущенные компанией "Туши свет", потому что так у вас будет больше шансов получить либо очень долговечную, либо крайне ненадежную лампочку. Другими словами, лампочки от компании "Туши свет" отличаются большой изменчивостью в том, что касается их срока службы. Если вы по натуре консерватор, тогда покупайте лампочки у компании "Включай свет" — они более стабильны и не преподнесут вам никаких сюрпризов.

Эмпирическое правило *неприменимо*, если в центре распределения нет выпуклости. Однако можно делать предположения или определять некоторые важные значения данных при помощи гистограммы и/или вычисления процентилей. (Подробные сведения о гистограммах и процентилях вы найдете в главах 4 и 5 соответственно.)

Обратимся к нормированному параметру

Представим себе, что гипотетическая студентка медицинского факультета Роди сдала стандартизованный тест на получение сертификата физиотерапевта и набирает 235 баллов. Все, что вы знаете — это то, что баллы в таком тесте имеют нормальное распределение. Каким был результат Роди — хорошим, плохим или средним? На этот вопрос ответить нельзя, не зная положения, которое занимает Роди среди других людей, проходивших тот же тест. Другими словами, вам нужно определить относительное положение баллов Роди в распределении всех баллов, полученных за тест.

Подробнее о стандартном отклонении

Определить относительное положение результата Роди можно разными способами — одни из них лучше, другие хуже. Прежде всего, вы можете оценить полученный балл в свете всех возможных результатов, зная, что для данного теста максимальный результат — 300 баллов. Так вы не сравните ее баллы ни с чьими другими, а только с возможным максимумом. Значит, вы все же не знаете, какое положение занимает результат Роди относительно других результатов. Затем можно попробовать сравнить ее результат со средним. Предположим, среднее было равно 250. Так вы получаете немного больше информации. Теперь известно, что результат Роди (235 баллов) ниже среднего, более того, он ниже среднего на 15 единиц (потому что $235 - 250 = -15$). Но какую роль в данной ситуации играет эта разница в 15 баллов?

В данном случае нужно знать, чему равно стандартное отклонение, чтобы понять относительное положение любого значения в распределении (см. раздел "Распрямляем колоколообразную кривую" выше в этой главе, где было рассказано о сроке службы электрических лампочек разных компаний, рис. 8.4). Предположим, что в случае с тестом Роди стандартное отклонение составляло 5 (как на рис. 8.6).

Распределение со стандартным отклонением, равным 5, означает, что баллы находились довольно близко друг к другу, и 15 единиц ниже среднего это действительно достаточно много. В этом случае можно сказать, что балл Роди намного ниже среднего, потому что разница в 15 единиц — это 3 стандартных отклонения ниже среднего (если 1 стандартное отклонение равно 5, то $-15/5 = -3$). И совсем небольшая часть участников теста набрали еще меньше баллов, чем она. (Известно, что 99,7% участников получили от 235 до 265 баллов, в соответствии с эмпирическим правилом, а общий процент всех возможных баллов равен 100%. Это значит, что процент тех, чей результат не попадает в диапазон от 235 до 265 баллов, равен $100 - 99{,}7 = 0{,}3\%$. Необходимо знать, какой про-

цент участников набрал еще меньше баллов, чем Роди со своими 235 баллами, а это будет половина от 0,3%. Таким образом, всего 0,15% или 0,0015 участников теста получили результат еще хуже, чем Роди.)

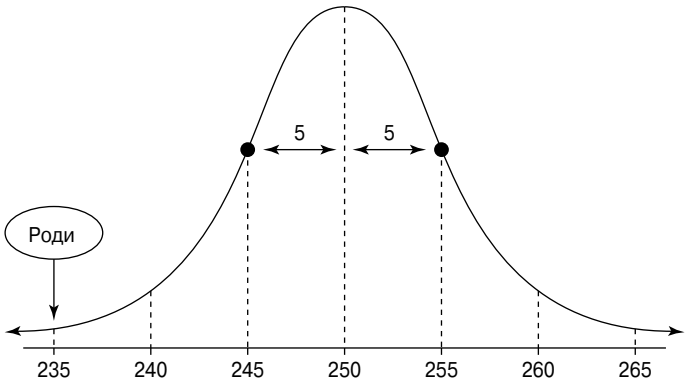


Рис. 8.6. Баллы с нормальным распределением, средним в 250 и стандартным отклонением в 5

Теперь представим, что стандартное отклонение имеет другое значение, скажем, 15, а среднее результатов теста остается прежним (250). На рис. 8.7 показано, как теперь выглядит распределение.

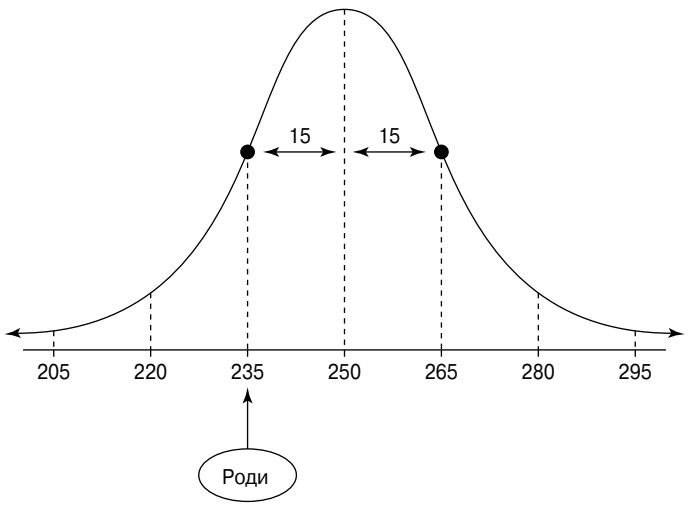


Рис. 8.7. Баллы с нормальным распределением, средним в 250 и стандартным отклонением в 15

Стандартное отклонение, равное 15, значит, что баллам присуща значительно большая изменчивость (или разброс), чем в предыдущей ситуации. В таком случае то, что результат Роди на 15 единиц ниже среднего — это совсем неплохо, потому что эти 15 единиц представляют собой всего 1 стандартное отклонение ниже среднего ($-15/15 = -1$). В этом примере 68% полученных результатов находятся между 235 и 265 баллами, о чем говорит эмпирическое правило, значит, половина из оставшихся 32% участников (кото-

рые расположены на нижних склонах) набрали меньше баллов, чем Роди. Следовательно, при таких условиях результат 16% участников хуже, чем у Роди. Ее относительное положение по-прежнему не очень хорошее, но во втором сценарии оно заметно улучшилось по сравнению с первым. Обратите внимание, что ее баллы не меняются независимо от сценария, а меняется только интерпретация этих баллов из-за разницы в стандартных отклонениях.

Относительное положение любого результата в распределении во многом зависит от стандартного отклонения. Расстояние в исходных единицах здесь не очень важно.

Очень часто в средствах массовой информации стандартное отклонение вообще не указывается. Никогда не рассматривайте какие-либо статистические результаты, сравнивая их только со средним значением, не зная, чему равно стандартное отклонение. Числа могут оказаться намного дальше от среднего, чем кажутся на первый взгляд.

Вычисление нормированного параметра

Чтобы найти, описать и интерпретировать относительное положение любого значения в гауссовом распределении (такого, как результат тестирования Роди), вам нужно превратить этот балл в то, что в статистике принято называть нормированным параметром. *Нормированный параметр* — это нормированный вариант исходного значения, показывающий число стандартных отклонений выше или ниже среднего. Формула для вычисления нормированного параметра такова:

нормированный параметр = (исходное значение – среднее)/стандартное отклонение.

Или, используя условные обозначения, нормированный параметр = $\frac{x-\mu}{\sigma}$.

Чтобы преобразовать исходное значение в нормированный параметр, выполните следующие действия.

1. **Найдите среднее и стандартное отклонение генеральной совокупности, с которой вы работаете.**

   Например, результат Роди на экзамене — 235 баллов — можно превратить в нормированный параметр согласно любому из двух возможных сценариев, описанных в предыдущем разделе. В первом случае среднее равно 250, а стандартное отклонение равно 5, значит, действия этапа 1 уже выполнены.

2. **Отнимите среднее из исходного значения.**

   В первом сценарии Роди вы найдете реальное расстояние от среднего путем вычитания: 235 – 250 = –15 (что означает, что ее результат на 15 единиц ниже среднего).

3. **Разделите результат на стандартное отклонение.**

   В ситуации с Роди расстояние равно –15. Разделив это расстояние на число стандартного отклонения, вы получите –15/5 = –3, что и будет нормированным параметром Роди. В первом сценарии (стандартное отклонение = 5) 235 баллов, которые получила Роди, находятся на расстоянии 3 стандартных отклонений ниже среднего. Во втором сценарии (стандартное отклонение = 15) нормированный параметр Роди равен (235 – 250)/15 = –15/15 = –1. Значит, в этом случае ее результат лежит в пределах 1 стандартного отклонения ниже среднего.

Чтобы при вычислении нормированных параметров не допустить ошибок, обязательно выполняйте действия этапов 2 и 3 в описанном порядке.

Свойства нормированных параметров

При интерпретации нормированных параметров полезными могут оказаться следующие свойства.

- ✓ Вследствие эмпирического правила почти все нормированные параметры (99,7% из них) находятся между –3 и +3.
- ✓ Отрицательный нормированный параметр означает, что исходное значение находилось ниже среднего.
- ✓ Положительный нормированный параметр значит, что исходное значение было выше среднего.
- ✓ Нормированный параметр, равный 0, значит, что исходным значением было само среднее.
- ✓ Значения из гауссова распределения при нормировке образуют особое гауссово распределение со средним, равным 0, и стандартным отклонением 1. Такое распределение называется *стандартным нормальным распределением* (см. рис. 8.8).

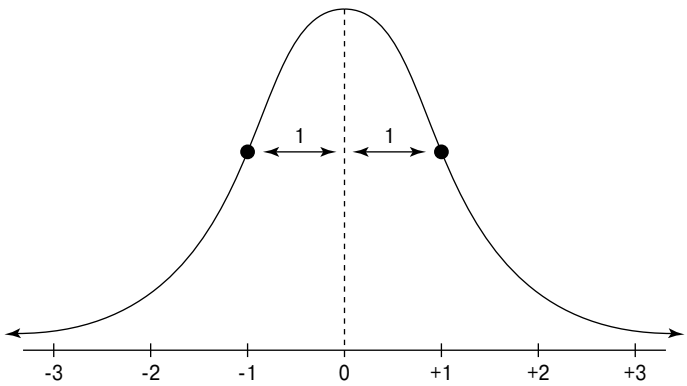


Рис. 8.8. Стандартное нормальное распределение

Нормированные параметры имеют универсальную интерпретацию, именно поэтому они настолько важны. Если вам дадут нормированный параметр, вы сможете сразу же интерпретировать его. Например, нормированный параметр +2 означает, что этот показатель находится в двух стандартных отклонениях выше среднего. Для понимания нормированного параметра не нужно знать, каким было исходное значение или среднее и стандартное отклонение. Нормированный параметр указывает на относительное положение этого значения, что в большинстве случаев и есть самое важное.

Преобразование в нормированный параметр не меняет относительное положение никакого значения в распределении, меняются всего лишь единицы. (Это аналогично изменению единиц температуры при переходе из шкалы Фаренгейта в шкалу Цельсия. Температура на улице не меняется, меняются только единицы ее измерения.) Вычитание среднего из исходного значения уравновешивает все значения относительно нуля. Если исходное значение совпадает со средним, то оно превратится в нормированный параметр, равный 0. Цена деления в любую сторону от среднего в том, что касается стандартного отклонения, остается прежней (она может быть равна 5, 15 и т.д.). Но нормированные параметры должны иметь цену деления 1. Это значит, что после вычитания среднего нужно разделить результат на стандартное отклонение. (Это похоже на то, как мы меняем дюймы на футы, разделив их на 12.)

Сравним яблоки и апельсины с помощью нормированных параметров

Распространенная сфера применения нормированных параметров — это сравнение значений из разных распределений, которых иначе просто не сравнить. К примеру, представим, что Билл подает заявление в два разных учебных заведения (назовем их Университет Данных и Огайо Стат), и для поступления ему нужно сдать экзамен по математике. Экзамены совершенно разные (даже количество вопросов в них отличается), и когда Билл узнает свои баллы, он захочет сравнить их, чтобы понять, в каком из колледжей он занимает более выгодное относительное положение по результатам экзамена по математике.

В Университете Данных Биллу говорят, что он набрал 60 баллов, распределение всех результатов нормальное, среднее равно 50, а стандартное отклонение равно 5. Из Огайо Стат сообщают, что Билл получил 90 баллов, результатам свойственно нормальное распределение со средним 80 и стандартным отклонением 10. Результаты какого экзамена лучше? Сходу сравнить 50 и 90 нельзя, потому что эти баллы подсчитаны по совсем разным шкалам. Невозможно сказать, что Билл сдал оба экзамена одинаково, потому что по результатам каждого он набрал на 10 баллов выше среднего. Здесь важную роль играет стандартное отклонение. Чтобы объективно и достоверно сравнить эти значения, нужно каждое из них превратить в нормированный параметр, чтобы оба результата находились на одной шкале (где большинство значений лежат между -3 и $+3$ с ценой деления 1).

Результат Билла в Университете Данных — 60 баллов с учетом того, что среднее равно 50, а стандартное отклонение для всей совокупности результатов экзамена равно 5. Следовательно, нормированный параметр равен $(60 - 50)/5 = 10/5 = 2$, т.е. относительное положение Боба по результатам экзамена в Университете Данных — 2 стандартных отклонения выше среднего. В университете Огайо Стат он набрал 90 баллов, при этом среднее равно 80, а стандартное нормальное распределение составляет 10. Значит, нормированный параметр Боба здесь равен $(90 - 80)/10 = 10/10 = 1$, т.е. относительное положение Боба по результатам математического экзамена в Огайо Стат — 1 стандартное отклонение выше среднего. Значит, этот результат не такой хороший, как результат, полученный в Университете Данных. Следовательно, Боб лучше сдал экзамен по математике Университете Данных.

Не сравнивайте результаты из разных распределений, прежде не превратив их в нормированные параметры. Нормированные параметры позволяют произвести объективное и достоверное сравнение по одной шкале.

Оценка результатов с помощью процентилей

Процентили используются для сравнения и определения относительного положения самыми разными способами. Вес младенцев, их рост и размеры головы, например, измеряются и интерпретируются с помощью процентилей. Кроме того, процентили используются компаниями для того, чтобы выяснить, какое положение занимает организация среди конкурентов в том, что касается продаж, прибылей, удовлетворения запросов клиентов и т.д. (Подробнее о процентилях см. главу 5.) А связь между процентилями и нормированными параметрами очень важна, что и будет видно на следующем примере.

Предположим, Роди (гипотетическая студентка медицинского факультета) сдает тест на получение сертификата физиотерапевта и набирает 235 баллов. Она пытается выяснить, что в действительности означает этот результат. Ей говорят, что у результатов теста — гауссово распределение со средним 250 и стандартным отклонением 15, что значит, что ее нормированный параметр в этом тесте равен −1,0 (одно стандартное отклонение ниже среднего). Результат ниже среднего, но, может быть, все же не слишком плох? Каждый год тест на получение сертификата отличается от предыдущего, поэтому различны и проходные баллы. Но, как правило, проходят верхние 60%, а проваливаются оставшиеся внизу 40%. Имея такую информацию, что вы скажете: прошла ли Роди этот тест? И какой проходной балл в этом году? Именно процентили помогут Роди раскрыть тайну своего нормированного параметра и узнать ответ на волнующий вопрос.

Если устроители теста всегда принимают верхние 60% сдававших тест, а оставшиеся 40% считаются провалившимися, то это значит, что у проходного балла 40-й процентиль. (Помните, что процентиль означает процент *ниже*.) Каков процентиль результата Роди? Вернемся к рис. 8.7 и воспользуемся эмпирическим правилом. Таким образом мы узнаем, что ее результат находится в 1 стандартном отклонении ниже среднего. Значит, около 68% результатов находятся в пределах от 235 до 265 (т.е. в рамках одного стандартного отклонения от среднего), а остальные (100% – 68% = 32%) не попадают в эти пределы. Это значит, что результат Роди имеет 16-й процентиль (приблизительно), поэтому она не поступает. Чтобы пройти тест, ей нужно быть на уровне как минимум 40-го процентиля.

Эмпирическое правило при определении процентилей может вам пригодиться только до этого момента. Возможно, вы заметили, что я выбрала результат Роди произвольно, так, чтобы он оказался точно на одной из отметок шкалы. Но представим, что она набрала баллы, которые на шкале находятся где-то между отметками. На помощь придет таблица стандартного нормального распределения. В табл. 8.1 показан соответствующий *процентиль* (т.е. процент значений, находящихся ниже) для любого нормированного параметра от −3,4 до +3,4. В этот диапазон попадает намного больше 99,7% ситуаций, с которыми вы вообще можете столкнуться. Обратите внимание, что с увеличением нормированного параметра процентиль тоже растет. Также заметьте, что нормированный параметр имеет 50-й процентиль в этих данных, что совпадает с медианой. (Подробнее о средних и медианах см. главу 5.)

Таблица 8.1. Нормированные параметры и соответствующие процентили из стандартного нормального распределения

Нормированный параметр	Процентиль (%)	Нормированный параметр	Процентиль (%)	Нормированный параметр	Процентиль (%)
-3,4	0,03	-1,1	13,57	+1,2	88,49
-3,3	0,05	-1,0	15,87	+1,3	90,32
-3,2	0,05	-0,9	18,41	+1,4	91,92
-3,1	0,10	-0,8	21,19	+1,5	93,32
-3,0	0,13	-0,7	24,20	+1,6	94,52
-2,9	0,19	-0,6	27,42	+1,7	95,54
-2,8	0,26	-0,5	30,85	+1,8	96,41
-2,7	0,35	-0,4	34,46	+1,9	97,13
-2,6	0,47	-0,3	38,21	+2,0	97,73
-2,5	0,62	-0,2	42,07	+2,1	98,21
-2,4	0,82	-0,1	46,02	+2,2	98,61
-2,3	1,07	0,0	50,00	+2,3	98,93
-2,2	1,39	+0,1	53,98	+2,4	99,18
-2,1	1,79	+0,2	57,93	+2,5	99,38
-2,0	2,27	+0,3	61,79	+2,6	99,53
-1,9	2,87	+0,4	65,54	+2,7	99,65
-1,8	3,58	+0,5	69,15	+2,8	99,74
-1,7	4,46	+0,6	72,58	+2,9	99,81
-1,6	5,48	+0,7	75,80	+3,0	99,87
-1,5	6,68	+0,8	78,81	+3,1	99,90
-1,4	8,08	+0,9	81,59	+3,2	99,93
-1,3	9,68	+1,0	84,13	+3,3	99,95
-1,2	11,51	+1,1	86,43	+3,4	99,97

Чтобы с помощью табл. 8.1 найти процентиль, нужно преобразовать исходное значение в нормированный параметр. Это проще, чем создавать таблицу для всех существующих гауссовых распределений со всеми возможными средними и стандартным отклонением. Вот почему стандартное нормальное распределение — отличный способ нахождения процентиля. Здесь используется нормированная шкала, и таблица подходит на все случаи жизни.

Чтобы подсчитать процентиль для данных с гауссовым распределением, необходимо выполнить следующие действия.

1. **Преобразуйте исходное значение в нормированный параметр. Для этого из исходного показателя вычтите среднее, затем разделите разность на стандартное отклонение. (Математически это записывается $\frac{x-\mu}{\sigma}$.)**
2. **Воспользуйтесь табл. 8.1 и найдите процентиль, соответствующий этому нормированному параметру.**

Итак, у результата Роди, набравшей 235 баллов, 16-й процентиль. Предположим, Клинт сдавал тот же тест и набрал 260 баллов. Проходит ли он? Чтобы выяснить это, вы можете преобразовать его результат в нормированный параметр и найти соответствующий процентиль. Его нормированный параметр равен (260 – 250)/15 = 10/15 = 0,67. С помощью табл. 8.1 можно определить, что результат Клинта находится где-то между 72,58-м и 75,80-м процентилями. (Будучи консерваторами, давайте выберем процентиль поменьше.) В любом случае Клинт успешно сдал тест, потому что его результат лучше, чем, по крайней мере, у 72% участников (в том числе и Роди). А также его процентиль выше 40, т.е. больше, чем проходной балл.

Но разве не лучше было бы, если бы вам не приходилось преобразовывать результаты каждого человека в нормированный параметр только для того, чтобы выяснить, сдал ли он тест? Почему бы просто не определить проходной балл в исходных единицах и сравнить все результаты с ним? Известно, что у проходного балла 40-й процентиль. В табл. 8.1 находим, что нормированный параметр, ближайший к 40-му процентилю, — это –0,3. Это значит, что проходной балл находится на расстоянии 0,3 стандартных отклонений ниже среднего. Но каков этот результат в исходных единицах? С помощью формулы нормированного параметра можно найти исходное значение.

Формула для преобразования исходного значения (x) в нормированный параметр (назовем его Z) такова: $Z = \frac{x-\mu}{\sigma}$.

Изменим это уравнение так, чтобы с помощью нормированного параметра (Z) найти исходное значение (x). Теперь формула выглядит так: $x = Z\sigma + \mu$.

Чтобы преобразовать нормированный параметр в значение в исходных единицах (исходное значение), необходимо выполнить следующие действия.

1. **Найдите среднее и стандартное отклонение генеральной совокупности, с которой вы работаете.**
2. **Умножьте нормированный параметр (Z) на стандартное отклонение.**
3. **Добавьте к полученному результату среднее.**

Вернемся к примеру. Проходной балл, преобразованный в нормированный параметр, равен –0,3, значит, $Z = -0{,}3$. Также известно, что среднее всех результатов $\mu = 250$, а стандартное отклонение $\sigma = 15$. Преобразуя нормированный параметр в исходное значение, получаем: $x = -0{,}3 \times 15 + 250 = -4{,}5 + 250 = 245{,}5$ (или 246). Значит, проходной балл равен 246. Участники, набравшие меньше 246 баллов, не проходят (к примеру, Роди), а все, кто получил больше 246 баллов, сдали тест (например, Клинт).

Но что, если результатам не свойственно нормальное распределение? Тогда все равно можно подсчитать процентили, но делать это придется вручную или с помощью специальной компьютерной программы (например, Microsoft Excel).

Чтобы найти k-й процентиль для данных с негауссовым распределением, необходимо выполнить следующие действия.

1. **Расположите значения по возрастающей.**
2. **Пусть n — это размер набора данных. Умножьте k на n и округлите результат до ближайшего целого числа.**
3. **Отсчитывайте данные до тех пор, пока не дойдете до того порядкового номера, который получили на этапе 2. Это и будет k-й процентиль вашего набора данных.**

   Например, предположим, у вас есть такой набор данных: 1, 6, 2, 5, 3, 9, 3, 5, 4, 5, и вы хотите найти 90-й процентиль.

   Этап 1: расположим данные по порядку и получим последовательность 1, 2, 3, 3, 4, 5, 5, 5, 6, 9.

   Этап 2: $n = 10$, $k = 90\%$, тогда k умноженное на n — это $0{,}90 \times 10 = 9$.

   Этап 3: это значит, что девятое число (от самого маленького к самому большому), т.е. число 6, и будет 90-м процентилем. Около 90% значений расположены ниже 6, а 10% значений находятся выше. (Подробнее о понятии и примерах процентилей см. главу 5.)

Глава 9

Осторожно: результаты бывают разные!

В этой главе...

- Осознание того, что результаты выборок бывают разными
- Оценка изменчивости результатов выборок с помощью центральной предельной теоремы
- Выявление факторов, влияющих на изменчивость

Статистику часто представляют себе как ряд однократных явлений, например: "Один из каждых двух браков заканчивается разводом", "Четыре из пяти опрошенных стоматологов рекомендуют жевательную резинку Trident" или "Средняя продолжительность жизни женщины, родившейся в 2000 году, составит 80 лет". Люди слышат подобные заявления и полагают, что эти результаты имеют отношение и к ним. Например, обычный человек может подумать, что вероятность того, что его брак распадется, равна 50%, его стоматолог, вероятно, советует жевательную резинку Trident, а если бы у него с женой (а вдруг они к тому времени еще не разведутся?) в 2000 году родилась девочка, то она бы прожила до 80 лет.

Но разве все эти утверждения не должны сопровождаться примечанием "плюс/минус", чтобы стало понятно, что результаты варьируются? Несомненно! Но такое бывает, к сожалению, нечасто. На самом деле, если исследователи для получения результатов не могут провести *перепись* (т.е. собрать данные обо всех представителях совокупности до единого), значит, полученные результаты у разных выборок будут разными, а изменчивость может оказаться намного большей, чем вы думаете! Возникает вопрос: в какой степени могут отличаться статистические результаты? Вы надеетесь (возможно, просто автоматически полагаете), что не сильно, а значит, полученные результаты можно применить практически к каждому. Но действительно ли это так? Совсем нет: изменчивость в статистических результатах зависит от множества факторов, и обо всех них и пойдет речь в этой главе.

Предполагаем, что результаты будут разными

На днях по телевизору я увидела рекламный ролик напитка для похудения. В нем рассказывалась поразительная история о женщине, которая за полгода сбросила более 20 кг (и еще больше года продолжала худеть). Во время ее рассказа внизу экрана на несколько секунд мелькнула надпись: "Результаты нетипичны".

Отсюда возникает вопрос: “А что же типично?”. Какой вес можно рассчитывать потерять за полгода, употребляя этот продукт, или, если вы хотите похудеть на 20 кг, то как долго нужно использовать этот напиток? Вы знаете, что вне зависимости от того, что это за результаты, у разных людей они бывают разные. Но данный рекламный ролик пытается убедить зрителя в том, что ему удастся наконец потерять 20 кг за шесть месяцев (хотя мелким шрифтом утверждалось совсем обратное). На самом деле намного лучше было бы, если бы производитель сообщил, в какой степени могут варьироваться результаты. Не помешало бы также, если бы в рекламе были представлены результаты не одного человека, а выборки нескольких человек.

Случаи из жизни (рассказы отдельных людей) очень увлекательны, но не имеют значения с точки зрения статистики!

Предположим, вы пытаетесь оценить количество людей в Соединенных Штатах Америки, поддерживающих президента. Если опросить случайную выборку из 1000 жителей США, одобряют ли они действия президента, то вы получите результаты одной выборки (например, 55% поддерживают президента). Но не стоит заявлять, что 55% населения Соединенных Штатов Америки выступают в поддержку президента, ведь ваши результаты относятся только к одной выборке из 1000 человек.

Если взять еще одну случайную выборку из 1000 человек той же совокупности и задать им тот же вопрос, то результат вполне может оказаться иным. Более того, учитывая, что население США такое большое и далеко не каждый житель является сторонником президента, то у 100 разных случайных выборок из 1000 человек каждая — если все их сделать из одной генеральной совокупности и поставить перед ними один и тот же вопрос — будет 100 разных результатов. Так как же нужно оглашать результаты выборки? Кое-кто определяет, в какой степени результаты будут варьироваться, и прилагает эту информацию.

Помните, что у разных выборок будут разные результаты. Не воспринимайте статистические данные однозначно и не пытайтесь использовать их, не выяснив, насколько отличными могут быть результаты.

Оценка изменчивости результатов выборок

Каким же образом можно измерить степень изменчивости результатов, если не делать все возможные выборки и не сравнивать полученные данные? Ведь с тем же успехом можно довести дело и до переписи! К счастью, благодаря некоторым важнейшим статистическим результатам (в основном, центральной предельной теореме), можно определить степень расхождения между средними значениями и долями разных выборок, не делая все возможные комбинации выборок (какое облегчение!). В двух словах, *центральная предельная теорема* гласит, что распределение всех средних значений (или долей) выборок можно считать нормальным, если размер выборки является достаточным. И еще: для центральной предельной теоремы неважно, каким было распределение исходной совокупности. Как такое возможно? Необходима выборка достаточно большого размера, а также нужно знать некоторые характеристики генеральной совокупности (такие как среднее и стандартное отклонение). И с этого момента в дело вступает волшебная статистическая теория.

Стандартные ошибки

Изменчивость среднего значения (или долей) выборки измеряется в стандартных ошибках. *Стандартная ошибка* — это такое же базовое понятие, как и стандартное отклонение, оба они означают типичное расстояние от среднего. Но есть и отличия. Значения исходной совокупности *отклоняются* друг от друга по естественным причинам (у людей разный рост, вес и т.д.), отсюда и название стандартного *отклонения*, которое измеряет такую изменчивость. Средние значения выборки варьируются из-за ошибок, которые происходят потому, что проводится не перепись, а только выборка, отсюда и название стандартной ошибки, которая измеряет изменчивость средних значений выборки. (Подробнее о стандартном отклонении см. главу 5. В этом разделе больше внимания будет уделено интерпретации стандартных ошибок. Подробнее о формулах стандартных ошибок см. главу 10.)

Рассмотрим пример: Бюро трудовой статистики США пытается проследить, как люди ежегодно тратят свои деньги, и для этого проводит Опрос потребительских расходов (ОПР). Делается выборка семей, которых просят рассказать о своих затратах. (Здесь может возникнуть смещение.) Типичный размер выборки при этом — 7500 человек. В табл. 9.1 показаны некоторые результаты из ОПР 2001 года. Сюда включены не только среднее количество расходов, которые делают представители выборки (среднее выборки для каждой статьи расходов), но и стандартные ошибки для каждого среднего выборки.

Таблица 9.1. Среднегодовые расходы американских семей в 2001 году

Расход	Среднее ($)	Стандартная ошибка ($)
Питание (дома)	3 085,52	42,30
Питание (вне дома)	2 235,37	38,35
Телефон	914,41	9,69
Газ и бензин (для автомобилей)	1 279,37	12,88
Материалы для чтения	141,00	2,99

Можно изучить результаты из табл. 9.1, сделав относительные сравнения. Например, обратите внимание, что около 42% всех расходов на питание идут на еду вне дома, учитывая, что 2 235,37/(3 085,52 + 2 235,37) = 2 235,37/5 320,89 = 0,42 или 42%. Стандартная ошибка для расходов на питание больше, чем ошибки по другим статьям расходов в этом перечне, потому что все семьи питаются по-разному. Но все же вас может удивить, почему значение стандартной ошибки для питания всего лишь такое, как показано в таблице. Помните, что стандартная ошибка показывает, какую степень изменчивости можно ожидать в среднем, если бы вы сделали еще одну выборку. Если размер выборки большой, то среднее не может измениться намного. Известно также, что правительство никогда не делает выборки маленького размера.

Перечень стандартных ошибок для средних значений выборки — это не та информация, которую часто предоставляют средства массовой информации. Однако вы можете (и должны, если результаты важны для вас) копать глубже и искать стандартные ошибки. Лучший способ сделать это — просмотреть статьи об исследовании и попытаться обнаружить указания на присущие им стандартные ошибки.

Выборочное распределение

Перечень всех значений, которые может принимать среднее выборки (или *выборочное среднее*), а также указание того, как часто эти значения встречаются, называется *выборочным распределением* среднего выборки. Выборочное распределение, как и многие другие виды распределений, имеет форму, центр и меру изменчивости. (Подробнее о форме, центре и изменчивости см. главу 4. Распределения описаны в главе 3.)

Согласно центральной предельной теореме, если выборки достаточно большие, то распределение всех возможных средних значений будет колоколообразным, или нормальным (гауссовым), распределением с тем же средним значением, что и генеральная совокупность. (Подробнее о гауссовом распределении см. главу 3.) Это объясняется тем, что среднее выборки расположено около общего среднего значения, т.е. среднего совокупности. Очень большие значения в выборке компенсируются очень маленькими значениями, также встречающимися в ней, вследствие *эффекта усреднения*. Изменчивость выборочного распределения измеряется стандартными ошибками. Дополнительная польза от применения среднего при оценивании (а не общего или отдельного значения) в том, что по мере увеличения размера выборки изменчивость среднего снижается. (Теми же свойствами обладает и выборочное распределение долей выборки, если речь идет о категорийных (да/нет) данных, полученных в ходе опросов.)

Использование эмпирического правила для интерпретации стандартных ошибок

Поскольку выборочное распределение средних выборки (или долей выборки в генеральной совокупности) является нормальным (колоколообразным), то, воспользовавшись эмпирическим правилом, вы сможете понять, насколько могут отличаться результаты конкретной выборки при условии, что эта выборка была достаточно большой. (Подробнее об эмпирическом правиле см. главу 8.)

В случае со средним и долями выборки эмпирическое правило гласит, что вы можете рассчитывать на следующее.

- ✓ Около 68% всех средних выборки находятся в пределах 1 стандартного отклонения от среднего совокупности.
- ✓ Около 95% выборочных средних находятся на расстоянии 2 стандартных отклонений от среднего совокупности.
- ✓ Около 99,7% средних выборки находятся в пределах 3 стандартных отклонений от среднего совокупности.
- ✓ Значения выборки для категорийных (да/нет) данных: 68, 95 или 99,7% долей выборки будут находиться на расстоянии 1, 2 или 3 стандартных отклонений от долей генеральной совокупности, соответственно.

Что говорит эмпирическое правило о том, насколько могут отличаться средние конкретных выборок? Помните, что 95% средних выборки должны находиться в пределах 2 стандартных ошибок от среднего совокупности, и ваша задача — найти среднее совокупности. Значит, если на самом деле ваша оцен-

ка представляет собой диапазон, в который входит среднее выборки плюс/минус 2 стандартных ошибки, то в 95% случаев такая оценка будет правильной. (Количество стандартных ошибок, которое добавляется или отнимается, называется *пределом погрешности*. Подробнее о пределе погрешности см. главу 10.)

Рассмотрим пример: согласно Бюро трудовой статистики США, средняя семья в 2001 году состояла из 2,5 человек (0,7 из них представляли собой несовершеннолетние дети, а 0,3 — люди старше 65 лет), у которых было 1,9 транспортных средства. (К сожалению, для этих данных стандартные ошибки не сообщаются.) Согласно табл. 9.1, средние расходы на телефон в этом году у выборки из 7500 семей составили 914,41 долларов на семью. Насколько эти результаты должны отличаться от выборки к выборке (т.е. у разных выборок из 7500 семей, если бы они также были сделаны из той же совокупности)? Стандартная ошибка для расходов на телефон в этой выборке равна 9,69 долларов. Это значит, что 95% средних расходов на телефон в данной выборке находятся в пределах $2 \times 9{,}69$ долларов (или 19,38 долларов) по обе стороны от реального среднего совокупности. Отсюда видно, какую степень различия можно ожидать от среднего телефонных расходов, если размер выборки равен 7500.

В предыдущем примере нельзя утверждать, что 95% всех семей в совокупности имеют расходы в таком диапазоне. Нет, вы высказываете предположение о том, каким может быть среднее расходов на телефон для всех семей в совокупности. Средние телефонные расходы представляют собой одно число, но поскольку вы не можете его точно определить, то оцениваете его с помощью такого диапазона значений.

Кроме того, можно использовать эмпирическое правило, чтобы дать приблизительную оценку средних телефонных расходов для всех семей в США (а не только выборки из 7500). И снова, используя второе свойство эмпирического правила, можно определить, что средние телефонные расходы всех семей в США составляют около 914,41 долларов плюс/минус $2 \times 9{,}69$ долларов = 19,38 долларов. Такая оценка будет правильной в 95% всех сделанных выборок (будем надеяться, что выборка, сделанная для Опроса потребительских расходов, была одной из них). На статистическом языке это означает, что вы почти с 95% уверенностью можете сказать, что средние годовые расходы на телефон для всех американских семей находятся где-то в пределах от 914,41 – 19,38 до 914,41 + 19,38 долларов, т.е. между 895,03 и 933,79 долларов. И чтобы добиться полной уверенности в 99,7% (а не 95%), нужно добавлять или вычитать 3 стандартных ошибки.

Подобный результат, для которого нужно добавить или отнять определенное количество стандартных ошибок, называется *доверительным интервалом* (который подробно описан в главе 11). Добавленное или отнятое количество называется *пределом погрешности*.

В средствах массовой информации, если речь идет о каких-либо количественных данных (например, доходы населения, цены на жилье или цены акций на бирже), вы почти никогда не увидите указания на предел погрешности. И тем не менее его всегда нужно было бы указывать, чтобы любой человек мог оценить точность результатов! В случае с опросами (в ходе которых собираются категорийные данные, выраженные в долях), предел погрешности, напрямую связанный со стандартной ошибкой (см. главу 10), приводится часто. Откуда взялись такие двойные стандарты? Точно никто не знает.

Подробнее о центральной предельной теореме

Заметьте, что результаты, к которым в предыдущем разделе было применено эмпирическое правило, помогают в общих чертах понять, что такое 1, 2 или 3 стандартных ошибки, а также сообщают, чего можно ожидать от среднего выборки (или долей выборки). На самом деле получить такие результаты позволяет центральная предельная теорема (ЦПТ). Статистики обожают ЦПТ, без нее они бы остались без работы. ЦПТ позволяет им хоть что-то сказать о том, какими могут быть результаты выборки, и для этого не придется изучать все возможные выборки, существующие в конкретной совокупности. Кроме того, теорема предоставляет в их распоряжение формулы подсчета стандартных ошибок, а также более конкретную информацию о том, какой процент средних значений (или долей) выборки будет находиться между любым количеством стандартных ошибок (а не только 1, 2 или 3).

Центральная предельная теорема гласит, что для любой совокупности со средним μ и стандартным отклонением σ справедливо следующее.

- ✓ Распределение всех возможных средних выборки $\bar{x}$ является *приблизительно* нормальным для выборок достаточно большого размера. Это значит, что вы можете использовать нормальное распределение, чтобы ответить на вопросы или сделать выводы относительно среднего своей выборки. (Подробнее о нормальном распределении см. главу 8.)
- ✓ Чем больше размер выборки (n), тем ближе к нормальному будет распределение среднего выборок. (Большинство статистиков согласны, что если n будет равно хотя бы 30, то чаще всего этого будет уже достаточно для дальнейшей работы.)
- ✓ Среднее распределения среднего выборок также равно μ.
- ✓ Стандартная ошибка для средних выборок равна $\frac{\sigma}{\sqrt{n}}$. По мере того как увеличивается n, ошибка уменьшается.
- ✓ Если исходным данным было свойственно нормальное распределение, то у среднего выборок тоже будет такое же нормальное распределение независимо от размера выборки.

Если стандартное отклонение генеральной совокупности у неизвестно (что в основном и бывает), то приблизительно определить его можно при помощи s, т.е. стандартного отклонения выборки, подставив его в формулу стандартной ошибки. Подробнее об этом см. главу 12.

Согласно ЦПТ, если исходные данные распределены нормально, то распределение среднего выборок тоже будет нормальным при условии, что выборки имеют достаточный размер. Это снова же объясняется эффектом усреднения.

Проверка баллов экзамена по математике

Рассмотрим результаты стандартного экзамена по математике за 2002 год для учеников средних школ. (Это та ситуация, когда среднее и стандартное отклонение совокупности известны, потому что все экзамены, которые сдавались в 2002 году, записаны и проверены.) Средний балл на экзамене по математике для юношей равен 21,2 со стандартным отклонением в 5,3. Девушки на том же экзамене набрали в среднем 20,1 балл

(со стандартным отклонением 4,8). Предыдущие исследования показали, что баллам такого экзамена свойственно приблизительно нормальное распределение. На рис. 9.1 показано распределение баллов юношей и девушек соответственно с учетом предыдущей информации. В каждом случае размер всей генеральной совокупности сдававших экзамен (юношей и девушек вместе) был около одного миллиона учеников.

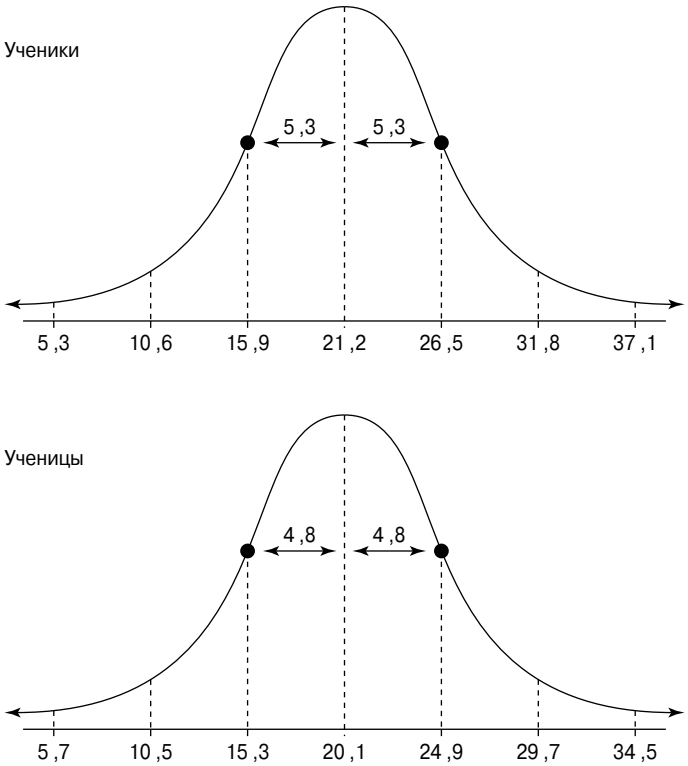


Рис. 9.1. Результаты экзамена по математике учеников и учениц средних школ в 2002 году

Согласно эмпирическому правилу (см. главу 8) около 95% юношей в ходе экзамена по математике набрали между 10,6 и 31,8 баллов, а около 95% девушек за тот же экзамен получили от 10,5 до 29,7 баллов. Результаты юношей и девушек вполне сопоставимы.

Предположим, вас интересует средний балл выборок 100 человек из общей совокупности в 500 000 юношей, сдававших в 2002 году экзамен по математике. Предположим, у вас в классе 100 учеников, и вы хотите знать, каковы их успехи по сравнению со всеми возможными классами такого размера. Каким будет распределение всех возможных средних значений выборок? Согласно ЦПТ, это будет нормальное распределение с одним и тем же средним (21,2). Стандартная ошибка будет равна 5,3, деленные на корень квадратный из 100 (потому что в этом гипотетическом примере $n = 100$).

Значит, стандартная ошибка для этой выборки будет равна $\frac{5,3}{\sqrt{100}} = 5,3/10 = 0,53$.

На рис. 9.2 показано, каким будет выборочное распределение среднего выборок из 100 юношей.

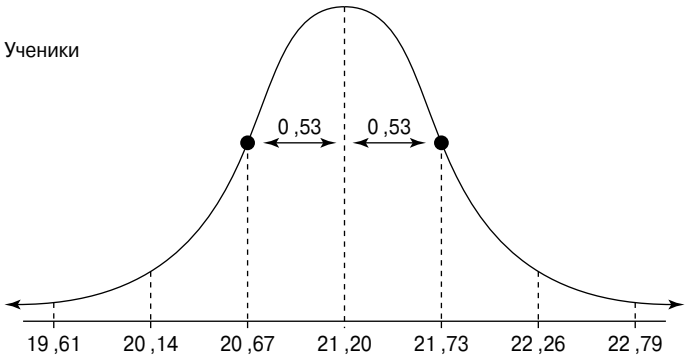


Рис. 9.2. Средний результат экзамена по математике в 2002 году для выборок из 100 учеников средних школ

Обратите внимание, насколько меньше на рис. 9.2 стандартная ошибка для среднего выборок по сравнению со стандартным отклонением исходных баллов, представленным на рис. 9.1. Это объясняется тем, что каждое среднее выборки на рис. 9.2 содержит информацию о 100 учениках, а каждый отдельный балл на рис. 9.1 содержит информацию только об одном ученике. Средние выборок отличаются не так сильно, как баллы отдельных учеников. Вот почему использовать среднее выборки для определения среднего совокупности намного лучше, чем отдельный балл (или случай из жизни).

На рис. 9.3 показано, что случится с выборочным распределением среднего выборок, если размер выборки вырастет до 1000 учеников. Стандартная ошибка уменьшается до 5,3, деленных на корень квадратный из 1000, т.е. 5,3/31,62 = 0,17. Стандартная ошибка на рис. 9.3 меньше, чем стандартная ошибка на рис. 9.2, потому что средние выборок на рис. 9.3 касаются 1000 учеников и содержат больше информации, чем средние выборок из рис. 9.2 (которые касаются 100 студентов в каждой выборке).

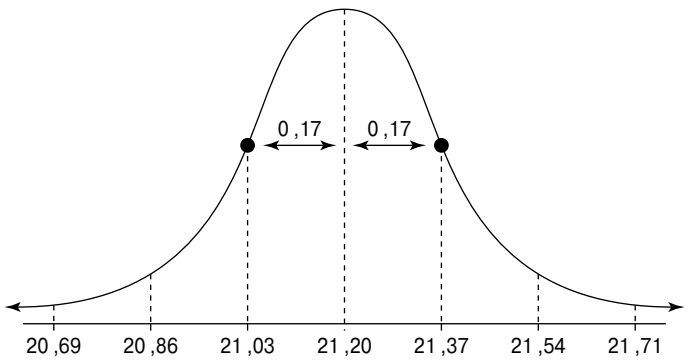


Рис. 9.3. Средние результаты экзамена по математике для выборок по 1000 учеников (2002 год)

Для сравнения на рис. 9.4 показаны одновременно все три распределения для юношей (индивидуальные баллы, средние выборок по 100 учеников и средние выборок по 1000 учеников).

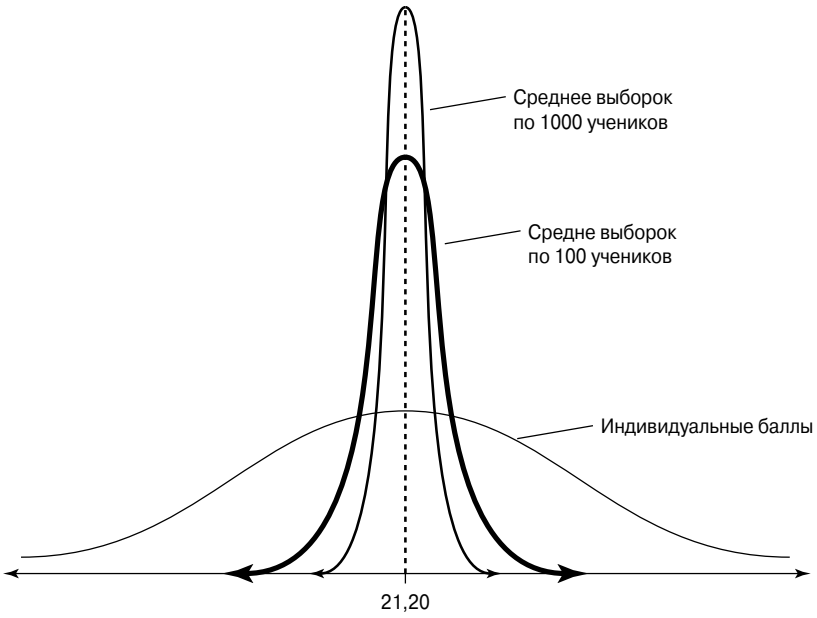


Рис. 9.4. Выборочное распределение результатов математического экзамена 2002 года для юношей с указанием индивидуальных баллов, среднего выборок по 100 учеников и среднего выборок по 1000 учеников

Чтобы ответить на вопросы о результатах выборки в таких ситуациях, как результаты экзамена по математике, можно воспользоваться центральной предельной теоремой. Например, нужно узнать, какова вероятность того, что средний балл на экзамене по математике для выборки из 100 юношей будет равен 22 или меньше. С помощью метода, описанного в главе 8, преобразуем балл 22 в нормированный параметр, вычитая среднее совокупности (21,2) и деля разность на стандартную ошибку (а не на стандартное отклонение). Формула такого преобразования следующая: $\frac{\overline{x} - \mu}{\sigma \div \sqrt{n}}$, где $\overline{x}$ — это средний балл для выборки (в данном случае 22), а $\frac{\sigma}{\sqrt{n}}$ — стандартная ошибка. Заметьте, что σ — это стандартное отклонение совокупности (5,3). Итак, в этом примере стандартная ошибка равна (см. рис. 9.2), средний балл класса из значения 22 превращается в нормированный параметр $(22 - 21{,}2)/0{,}53 = 0{,}8/0{,}53 = 1{,}51$. Требуется вычислить процентное отношение баллов, которые находятся слева от этого значения (другими словами, процентиль, который соответствует нормированному параметру 1,51). Согласно табл. 8.1 из главы 8, это процентное отношение составит около 93,32%.

Не забудьте сначала найти разность 22 – 21,2, прежде чем делить ее на 0,53, иначе у вас получится –18, а это неверно.

Группа средних всегда менее изменчива, чем группа отдельных баллов. А средние, касающиеся выборок большего размера, еще менее изменчивы, чем средние выборок меньшего размера.

Центральная предельная теорема подходит не только для среднего выборок. Вы можете воспользоваться ею, и чтобы ответить на вопросы или сделать выводы относительно долей совокупности, исходя из доли своей выборки. Те же выводы, касающиеся формы, центра и изменчивости среднего выборки подходят и для долей выборки. Конечно, формула будет немного отличаться, но в основе своей понятия остаются прежними. Прежде всего, доля совокупности обозначается $\hat{p}$, она равна количеству элементов выборки в интересующей вас категории, деленному на общий размер выборки (n).

Центральная предельная теорема гласит, что для любой совокупности данных, где p — это общее процентное отношение совокупности, справедливо следующее.

- ✓ Распределение долей всех возможных выборок ($\hat{p}$) *приблизительно* нормальное при условии, что размер выборки достаточно велик. (Подробнее о нормальном распределении см. главу 8.)
- ✓ Чем больше размер выборки (n), тем ближе к нормальному будет распределение долей выборки. (Это значит, что вы можете использовать нормальное распределение для того, чтобы ответить на вопросы или сделать выводы относительно долей своей выборки.)
- ✓ Среднее распределения долей выборки также равно p.
- ✓ Стандартная ошибка для долей выборки равна $\sqrt{\frac{p(1-p)}{n}}$. Она уменьшается, если n возрастает.

Заметьте, что стандартная ошибка для доли выборки содержит p, которое является долей совокупности. Скорее всего, это значение будет неизвестно, и приблизительно его можно определить с помощью доли выборки $\hat{p}$. (Подробнее об этом см. главу 12.)

Доли в математике

Вы можете использовать центральную предельную теорему, чтобы ответить на вопросы, связанные с долями. Например, предположим, нужно выяснить, какая доля абитуриентов не отказалась бы от занятий с репетитором по математике. Наряду со стандартизованным экзаменом ежегодно проводится и опрос учащихся, один из вопросов которого и призван выяснить, не нужна ли конкретному студенту дополнительная помощь по математике. В 2002 году 38% учащихся, сдававших экзамен, ответили на этот вопрос положительно. Это одна из тех ситуаций, когда известна доля совокупности p ($p = 0{,}38$). Исходные данные в таком случае (как и с категорийными данными) не имеют нормального распределения, потому что здесь возможны только два ответа: да или нет. Распределение совокупности ответов на вопрос о математических навыках показано на рис. 9.5 в виде столбиковой диаграммы (подробнее столбиковые диаграммы рассмотрены в главе 4).

Предположим, из этой комбинированной совокупности, насчитывающей более миллиона человек (все ученики, сдававшие экзамен в 2002 году), требуется сделать выборки по 100 учеников, и найти процент учеников, признавших необходимость улучшить свои знания по математике. Распределение всех долей выборок показано на рис. 9.6. Это нормальное распределение со средним $p = 0{,}38$ и стандартной ошибкой $= \sqrt{\frac{0{,}38(1-0{,}38)}{100}} = \sqrt{0{,}00236} = 0{,}049$ или 4,9%, т.е. около 5%.

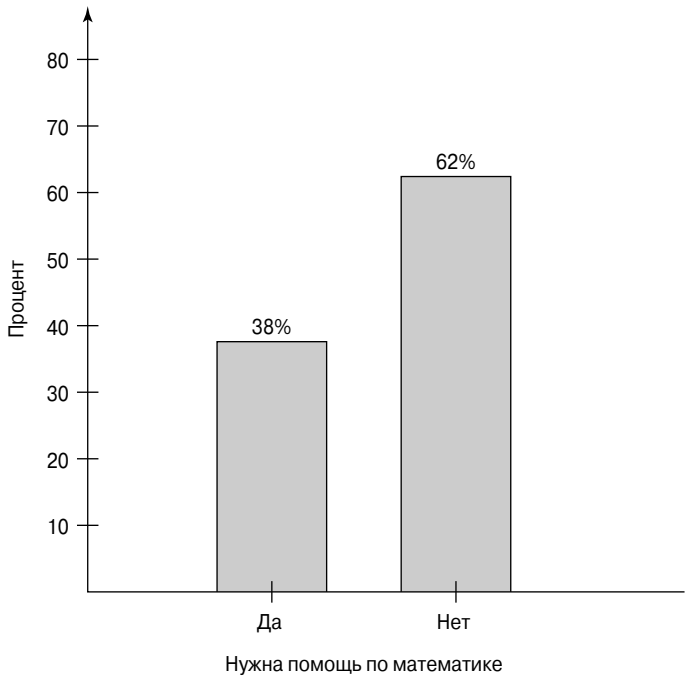


Рис. 9.5. Процентные отношения всех учеников совокупности, отвечавших в 2002 году на вопрос о дополнительной помощи по математике

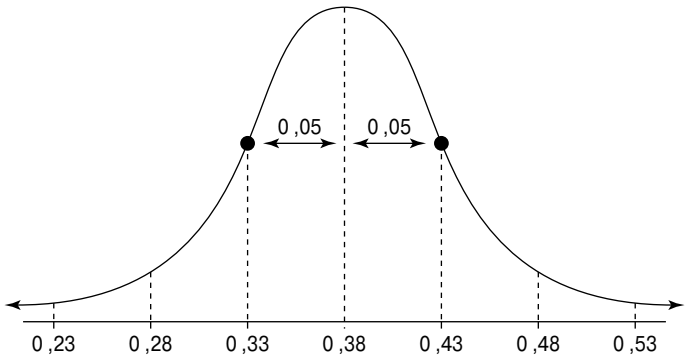


Рис. 9.6. Доля учеников, ответивших положительно на вопрос о дополнительной помощи по математике в 2002 году, для выборок по 100 человек

Воспользовавшись ЦПТ, можно определить, что некоторые доли выборок больше 0,38, некоторые меньше, но большинство из них (около 95%) находятся в пределах 0,38 плюс/минус $2 \times 0{,}05 = 0{,}10$ или $38\% \pm 10\%$. Результаты все равно в значительной степени варьируются, на 10% в обе стороны от доли совокупности.

Теперь сделаем выборки по 1000 человек из исходной совокупности и найдем долю тех, кто в каждом случае ответил, что ему нужна помощь по математике. Распределение долей выборок в данной ситуации будет во многом напоминать ситуацию, изображен-

ную на рис. 9.7. Все выглядит точно так же, как и на рис. 9.6, за исключением того, что распределение будет плотнее.

Стандартная ошибка уменьшится до $\sqrt{\frac{0,38(1-0,38)}{1000}} = \sqrt{0,000236} = 0,015$ или 1,5%.

Около 95% результатов выборки будут находиться между $0,38 - 2(0,015)$ или $0,38 + 2(0,015)$, т.е. между 0,35 и 0,41 (т.е. между 35% и 41%). Другими словами, если взять из данной совокупности несколько разных выборок размером по 1000 человек и найти долю для каждой выборки, то они будут не слишком отличаться от выборки к выборке, скорее, эти доли будут расположены близко друг к другу. Все это объясняется большим размером выборки — 1000 человек.

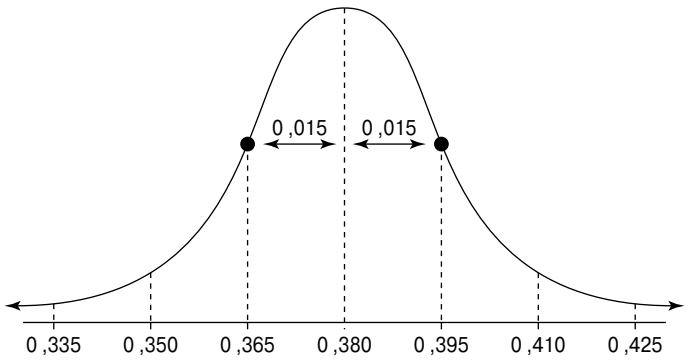


Рис. 9.7. Доля учеников, ответивших в 2002 году положительно на вопрос о дополнительной помощи по математике, для выборок по 1000 человек

Прежде чем делать выводы о процентных отношениях в выборках любого размера, выясните, насколько будут отличаться результаты. Для этого нужно найти стандартную ошибку или предел погрешности (равный порядка двух стандартных ошибок, подробнее см. главу 10). Зная ожидаемую степень изменчивости, вы сможете правильно оценить полученные результаты.

Выборка какого размера будет достаточно большой, чтобы для работы с ней можно было использовать центральную предельную теорему? Большинство статистиков согласны с тем, что оба выражения: $n \times p$ и $n \times (1-p)$ должны быть больше или равны 5. Таким образом решается проблема, когда доля очень близка либо 1, либо 0 (другими словами, те крайние случаи, когда либо все, либо никто не попадают в интересующую нас группу.) В таких крайних случаях необходима большая выборка, чтобы были представлены все группы, даже те, в которые входят не очень много людей. Для организаторов большинства опросов не составляет труда набрать достаточное количество участников, чтобы решить подобную проблему.

ЦПТ — хорошая новость для тех, кто пытается интерпретировать результаты выборки. Пока выборка имеет достаточный размер (а данные надежны и не смещены), то излагаемая информация будет близка к истине. (Но это пока данные надежны и не смещены. Подробнее о том, какие проблемы могут возникать со статистикой, см. главу 2.)

Кроме того, центральная предельная теорема позволяет ответить и на другие важные вопросы, касающиеся среднего и долей выборок. Например, если служба доставки обещает доставить товар в течение двух дней, а по результатам вашей выборки из 30 доставок получается 2,4 дня, достаточно ли таких доказательств, чтобы утверждать, что компания прибегает к ложной рекламе? Или же это только атипичная выборка задержанных доставок? (Подобные вопросы освещаются в главе 14.)

Возможно, вас беспокоит то, что для использования ЦПТ всегда нужно знать среднее совокупности (μ) или долю совокупности (p). Не волнуйтесь! Вы узнаете секрет, который уже давно известен статистикам: если определенное значение неизвестно, просто предположите его и двигайтесь дальше. (Подробнее об этом в главе 11.)

Факторы, влияющие на изменчивость результатов в выборках

На степень изменчивости среднего или долей пропорций выборки влияют два основных фактора: размер выборки и степень изменчивости в исходной совокупности.

Размер выборки

Размер выборки влияет на степень изменчивости результатов выборки. Предположим, у вас есть пруд с рыбой, и вы хотите выяснить среднюю длину рыбы в этом пруду. Если взять несколько случайных выборок по 100 рыб, а затем несколько случайных выборок по 1000 рыб, всякий раз записывая среднее выборки, то среднее каких выборок будет больше отличаться — тех, в которых 100 элементов или 1000? Среднее выборок размером 100 будет отличаться больше, потому что среднее каждой выборки основывается на меньшем количестве информации (т.е. меньшем числе рыб). На доли выборки это повлияет точно так же.

Небольшой размер выборки приводит к тому, что вычисление среднего выборки (и долей выборки) проходит с большими стандартными ошибками. Если размер выборки больше, то среднее (и доли) выборки будут определены с меньшими стандартными ошибками. Другими словами, чем больше данных вы соберете о каждой выборке, тем меньше будут отличаться результаты от выборки к выборке.

Изменчивость среднего выборки (или долей выборки) измеряется стандартными ошибками. Изменчивость среднего выборки — это $\frac{\sigma}{\sqrt{n}}$, а изменчивость долей выборки — $\sqrt{\frac{p(1-p)}{n}}$. В знаменателе каждой формулы есть n (и ничего больше).

Следовательно, с увеличением размера выборки (т.е. знаменателя) стандартная ошибка (т.е. вся дробь) уменьшается. Получение более подробной информации о выборке (благодаря большему размеру выборки) снижает изменчивость среднего выборки (и долей выборки).

Изменчивость совокупности

Если увеличивается изменчивость совокупности, то же происходит и с изменчивостью среднего или долей выборки. Представьте себе, что у вас есть два пруда с рыбой и вы хотите определить среднюю длину всей рыбы в них. Рыба в Пруде изменчивом намного сильнее отличается по длине, чем рыба из Пруда стабильного. Вы делаете выборку по 100 рыб из каждого пруда и находите среднюю длину рыбы в этой выборке. Если взять несколько выборок по 100 рыб из каждого пруда и в каждом случае определить среднее выборки, то какие средние будут отличаться больше — из Пруда изменчивого или из Пруда стабильного? Средние выборок из Пруда изменчивого будут отличаться больше, прежде всего потому, что совокупность рыбы в Пруде изменчивом больше отличается по длине.

На изменчивость долей выборки изменчивость в совокупности влияет точно так же. Например, предположим, вы хотите определить долю рыбы в Пруде стабильном, отличающейся хорошим здоровьем (назовем это p). Если рыба в Пруде стабильном здорова (т.е. p близко к 1), стандартное отклонение совокупности $p(1-p)$ будет маленьким, потому что состояние всей рыбы почти одинаковое. Если теперь вы сделаете много выборок рыбы из однородной (здоровой) совокупности и найдете процентное отношение тех, здоровье которых особенно крепкое, то процентные отношения от выборки к выборке останутся во многом одинаковыми. Значит, если p близко к 1, стандартная ошибка доли выборки мала. То же происходит, и в том случае когда почти вся рыба больна (p близко к 0). Но если около 50% рыб здоровы, а 50% больны, то изменчивость долей в разных выборках будет выше, потому что изменчивость совокупности тоже увеличилась. Более того, у совокупности, в которой $p = 0{,}5$, наблюдается наибольшая изменчивость, в результате чего стандартные ошибки долей выборок тоже максимальны.

Чем более изменчива исходная совокупность, тем бо́льшая изменчивость характерна для стандартной ошибки среднего выборки (или долей выборки). Как уже отмечалось, компенсировать такую большую изменчивость можно с помощью увеличения размера выборки.

Вспомним, что изменчивость среднего выборки находится по формуле $\frac{\sigma}{\sqrt{n}}$, а изменчивость долей выборки — это $\sqrt{\frac{p(1-p)}{n}}$.

В числителе этих формул стоит стандартное отклонение исходной совокупности для любого из двух случаев (σ — для числовых данных, а $p(1-p)$ — для категорийных). Следовательно, когда стандартное отклонение совокупности (т.е. числитель) увеличивается, то возрастает также и стандартная ошибка (т.е. вся дробь). Больше изменчивости в совокупности означает больше изменчивости в средних выборок (или долях выборок). Такую изменчивость можно компенсировать путем увеличения размера выборки, потому что если увеличивается n (знаменатель), то вся дробь, т.е. стандартная ошибка, уменьшается.

Подставить числа в формулу и найти степень точности данных (как это кажется исследователям или в чем они хотят убедить вас) может кто угодно. Но если изначально результаты были смещены, то об их точности речь уже не идет. (Хотя формулам это неизвестно, и вам придется быть начеку.) Поэтому прежде чем изучать степень изменчивости результатов, обязательно проверьте, как были сделаны выборки для определенного исследования и как собирались данные. (Подробнее об этом читайте в главе 17.)

Глава 10

Оставим место для предела погрешности

В этой главе...

- Понятие предела погрешности
- Вычисление предела погрешности
- Изучение влияния размера выборки
- Что не входит в предел погрешности

Хорошие исследователи, проводящие опрос или эксперимент, обязательно сообщают, насколько точны полученные результаты, чтобы потребители этой информации могли правильно оценить результаты. Для этого используется *предел погрешности* — т.е. мера того, насколько результаты выборки должны быть близки к изучаемому параметру всей совокупности.

Многие журналисты тоже понимают важность предела погрешности для оценки информации, поэтому в средствах массовой информации уже появляются статьи, включающие сведения и об этом критерии. Но что на самом деле означает предел погрешности?

В этой главе речь пойдет о пределе погрешности и о том, как с его помощью можно оценить точность статистической информации. Здесь рассматривается также размер выборки. Как ни странно, небольшая выборка все же может дать правильную информацию о жизни в стране — или в мире, — если все исследование было проведено правильно.

Насколько важен этот плюс/минус

Возможно, вы уже слышали о таком понятии, как предел погрешности, скорее всего, в связи с результатами опроса. Например, может быть, вам доводилось слышать, как кто-то говорит: “Предел погрешности этого исследования — плюс/минус три процента”. Что же нужно делать с такой информацией и насколько она действительно важна? Дело в том, что сами по себе результаты исследования (без предела погрешности) свидетельствуют лишь о том, что думает об изучаемом вопросе одна *выборка* участников. Ничего не сообщается обо всей *генеральной совокупности*, если бы она была изучена (вот это была бы работка!). Предел погрешности помогает понять, насколько вы близки к истине относительно всей изучаемой совокупности, и при этом вы используете информацию только об одной выборке из этой совокупности.

Как уже отмечалось в главе 3, выборка — это репрезентативная группа, отобранная из совокупности, которую вы хотите изучить. Результаты, основанные на изучении выборки, не были бы точно такими же, если бы вы изучили всю совокупность, потому что, делая выборку, вы не получаете информацию о каждом элементе совокупности. Но если исследование проведено правильно (подробнее о разработке хороших исследований

см. главу 17), то результаты выборки должны быть близки к реальным значениям для всей совокупности.

Предел погрешности не означает, что была сделана ошибка. Он говорит только о том, что вы изучили не всю совокупность целиком, значит, результаты могут немного отличаться от истины. Другими словами, вы признаете, что ваши результаты могут отличаться от результатов последующих выборок и остаются точными только в определенном диапазоне, на который и указывает предел погрешности.

Рассмотрим пример опроса, проведенного одной из ведущих исследовательских организаций, *The Gallup Organization*. Предположим, в последнем опросе участвовали 1000 человек из США, и результаты показали, что 520 опрошенных (52%) думают, что президент прав в своих поступках, а 48% не согласны с этим. Представим, что *Gallup* указывает предел погрешности в плюс/минус 3%. Итак, известно, что большинство людей в этой *выборке* одобряют действия президента, но можно ли сказать это о большинстве *всех американцев*? В данном случае нет. Почему?

Если 52% *опрошенных* одобряют президента, то можно ожидать, что процентное отношение сторонников президента в *генеральной совокупности всех американцев* будет равно 52% плюс/минус 3%. Значит, от 49 до 55% всех американцев выступают в поддержку президента. Это максимально точные результаты, которые можно получить от выборки в 1000 человек. Но обратите внимание, что 49%, т.е. нижняя граница диапазона, представляет собой меньшинство, потому что это меньше 50%. Значит, даже изучив эту выборку, нельзя наверняка сказать, что большинство американцев поддерживают президента. Вы можете лишь утверждать, что от 49 до 55% американцев выступают в поддержку президента, и это может быть большинством, а может и не быть.

На минуту задумайтесь о размере выборки. Разве не интересно, что выборка всего из 1000 американцев из совокупности более чем 288 000 000 человек может дать погрешность результатов опроса всего в плюс/минус 3%? Это невероятно! Значит, чтобы действительно понять, что происходит в больших совокупностях, нужно сделать выборку всего лишь крошечной доли от целого. Статистика — на самом деле мощный инструмент, позволяющий выяснить, что люди думают по определенным вопросам. Возможно, именно поэтому вас так часто приглашают принять участие в многочисленных опросах.

Чтобы быстро и в общих чертах понять, каков предел погрешности для выборки любого размера, разделите 1 на корень квадратный из n (размера выборки). В примере с опросом *Gallup* $n = 1000$, а квадратный корень из него равен 31,62, значит, приблизительный предел погрешности будет 1, деленное на 31,62 или около 0,03, т.е. 3%. Дальше в этой главе вы узнаете, как точнее вычислять предел погрешности.

Находим предел погрешности: общая формула

Предел погрешности — это значение "плюс/минус", которое приобщается к результатам выборки, когда вы от изучения выборки переходите к изучению всей сово-

купности, которую она представляет (генеральной совокупности). Следовательно, в общей формуле предела погрешности есть знак ±. Итак, как же получить это значение плюс/минус (иным способом в отличие от приблизительной оценки, описанной выше)? В этом разделе вы все узнаете.

Измерение изменчивости выборки

Результаты выборки варьируются, но насколько? Согласно центральной предельной теореме (см. главу 9), если выборка имеет достаточно большой размер, то распределение долей выборки (или средних значений выборок) будет представлено колоколообразной кривой (т.е. нормальное распределение — см. главу 8). Некоторые доли выборок (или средние выборок) будут переоценивать значения совокупности, а некоторые недооценят их, но большинство окажется посередине. А что такое середина? Если вы найдете среднее результатов всех возможных выборок, которые есть в совокупности, то это среднее и будет настоящей долей совокупности (в случае с категорийными данными) или настоящим средним совокупности (если данные числовые). Как правило, на изучение всех возможных выборок совокупности и усреднение полученных результатов не хватает ни времени, ни денег, но некоторые сведения о вариантах всех других выборок помогут вам определить степень изменчивости одной доли (или среднего) совокупности.

Стандартные ошибки — вот основа предела погрешности. *Стандартная ошибка* — это статистический показатель, по сути равный стандартному отклонению данных, деленному на квадратный корень из *n* (размер выборки). Это соответствует тому факту, что размер выборки в значительной мере влияет на изменчивость показателя выборки. (Подробнее о стандартных ошибках см. главу 9.)

Количество стандартных ошибок, которые вы должны добавить или отнять, чтобы получить предел погрешности, зависит от того, насколько уверенным в результатах вы хотите быть (это называется вашим *доверительным уровнем*). Обычно исследователям нужно 95% уверенности, поэтому основное правило для получения предела погрешности требует добавлять или отнимать около 2 стандартных ошибок (1,96, если быть точным). Это позволит вам подтвердить около 95% всех возможных результатов, которые могут быть получены в ходе повторных выборок. Чтобы быть уверенным на 99%, вы добавляете или отнимаете около 3 стандартных ошибок (точнее, 2,58). (Подробнее о доверительном уровне и количестве стандартных ошибок см. главу 12.)

Чтобы *точно* сказать, какое количество стандартных ошибок для подсчета предела погрешности на любом доверительном уровне нужно добавить или отнять, необходимо использовать особую колоколообразную кривую, которая называется *стандартное нормальное распределение* (подробнее см. главу 8). Для любого заданного доверительного уровня соответствующее значение в стандартном нормальном распределении (называемое *Z-значением*) показывает количество стандартных ошибок, которые нужно добавить или отнять, чтобы получить требуемый доверительный уровень. Для доверительного уровня в 95% точное *Z*-значение равно 1,96 (т.е. “около” 2), а для уровня 99% точное *Z*-значение составляет 2,58 (т.е. “около” 3). Некоторые самые распространенные доверительные уровни, а также соответствующие им *Z*-значения приведены в табл. 10.1.

Таблица 10.1. *Z*-значения для некоторых доверительных уровней (в процентах)

Уровень в процентах	*Z*-значение
80	1,28
90	1,64
95	1,96
98	2,33
99	2,58

Определение предела погрешности для доли выборки

Когда в ходе опроса людей просят выбрать правильный ответ из ряда предложенных (например, "Вы одобряете, не одобряете деятельность президента или не можете ответить?"), то статистический показатель, который используется в изложении результатов, — это доля участников выборки, которые распределяются по группам, т.е. *доля выборки* или *процент* выборки. Общая формула предела погрешности для вашей доли выборки такова: $Z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, где $\hat{p}$ — это доля выборки, n — размер выборки, а Z — подходящее Z-значение для желаемого доверительного уровня (из табл. 10.1).

Вот как нужно подсчитывать предел погрешности для процента выборки.

1. **Найдите долю выборки $\hat{p}$ и размер выборки n.**
2. **Умножьте $\hat{p} \times (1-\hat{p})$.**
3. **Разделите результат на n.**
4. **Найдите корень квадратный из полученного значения.**

   Теперь у вас есть стандартная ошибка.

5. **Умножьте результат на Z-значение, соответствующее желаемому доверительному уровню.**

   Обратитесь к табл. 10.1. Если вы хотите на 95% быть уверенным в своих результатах, Z-значение равно 1,96.

Вернемся к примеру о том, поддерживают ли американцы президента. Здесь вы можете найти предел погрешности. Сначала предположим, что вам необходим доверительный уровень в 95%, значит, $Z = 1{,}96$. Количество американцев в выборке, которые одобряют действия президента, было 520 человек. Это значит, что доля выборки $\hat{p} = 520/1000 = 0{,}52$. (Размер выборки n был равен 1000.) Предел погрешности для этого вопроса определяют следующим образом:

$$Z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1{,}96 \times \sqrt{\frac{0{,}520(1-0{,}520)}{1000}} = 1{,}96 \times \sqrt{\frac{0{,}2496}{1000}} = 1{,}96 \times \sqrt{0{,}0002} =$$

$$= 1{,}96 \times 0{,}0158 = 0{,}0310 = 3{,}10\%.$$

Доля выборки — это десятичный эквивалент процента выборки. Другими словами, если у вас есть процент выборки 5%, то в формуле нужно использовать не 5, а 0,05. Чтобы преобразовать проценты в десятичную дробь, просто разделите их на 100. После завершения всех вычислений вы сможете опять вернуться к процентам, умножив окончательный ответ на 100%.

Изложение результатов

Чтобы представить результаты этого опроса, вы могли бы сказать: "Основываясь на сделанной выборке, 52% американцев одобряют действия президента, с пределом погрешности плюс/минус 3,1%". (Звучит почти как у *Gallup*!)

А как о своих результатах сообщает организация, проводящая соцопрос? Вот как это делается:

Исходя из опроса всей выборки взрослых, мы на 95% уверены, что предел погрешности в нашем процессе выборки и его результатах не больше плюс/минус 3,1 процентных единиц.

Это похоже на длинный перечень правовых оговорок, который следует в конце рекламного объявления об аренде автомобилей. Но теперь вы понимаете, что сообщается мелким шрифтом!

Никогда не принимайте результаты опроса или исследования без предела погрешности. Предел погрешности — это единственный способ определить, насколько близка информация о выборке к реальной совокупности, которую вы изучаете. Результаты выборок варьируются, и если сделать иную выборку, то можно получить иные результаты. Поэтому необходим предел погрешности, чтобы сказать, насколько результаты выборки соответствуют действительным значениям совокупности. Когда вы в следующий раз услышите в СМИ сообщение о том, что был проведен какой-нибудь опрос, присмотритесь внимательней, чтобы понять, говорится ли там о пределе погрешности. В последнее время многие исследователи указывают предел погрешности полученных в ходе опроса результатов (но так бывает не всегда).

Определение предела погрешности для среднего выборки

Когда в ходе исследования людей просят указать числовое значение (например, "Сколько человек живет в вашем доме?"), то статистический показатель, который используется для изложения результатов, — это среднее всех ответов, данных членами выборки. Он также известен как *среднее выборки*, или *выборочное среднее*.

Общая формула предела погрешности для среднего выборки такова: $Z \times \frac{s}{\sqrt{n}}$, где s — это стандартное отклонение выборки, n — размер выборки, а Z — Z-значение, соответствующее желаемому доверительному уровню (из табл. 10.1).

Предел погрешности среднего выборки вычисляется следующим образом.

1. **Найдите стандартное отклонение выборки s и размер выборки n.**

   Подробнее о том, как вычисляется среднее и стандартное отклонение, см. главу 5.

2. **Разделите стандартное отклонение выборки на корень квадратный из размера выборки.**

   Теперь у вас есть стандартная ошибка.

3. **Умножьте результат на соответствующее Z-значение (из табл. 10.1).**

   Z–значение для доверительного уровня 95% равно 1,96.

Например, представим себе, что вы работаете менеджером в магазине мороженого и обучаете новых сотрудников заполнять большие стаканчики определенным количеством мороженого (по 300 граммов). Вы хотите оценить средний вес наполненных стаканчиков и определить предел погрешности. Вы просите каждого нового сотрудника наугад вы-

брать стаканчики и проверить их вес, записав данные в блокнот. Для выборки стаканчиков из $n = 50$ средний вес был 309 г, при этом стандартное отклонение выборки $s = 1{,}8$ г. Каков предел погрешности? (Необходим доверительный уровень 95%). Его нужно подсчитывать так:

$$Z \times \frac{s}{\sqrt{n}} = 1{,}96 \times \frac{1{,}8}{\sqrt{50}} = 0{,}5.$$

Итак, чтобы сообщить о полученных результатах, вы должны сказать, что, исходя из выборки 50 стаканчиков, вы полагаете, что средний вес всех больших стаканчиков с мороженым, которые сделали новые сотрудники, равен 309 граммов с пределом погрешности плюс/минус 0,5 г.

Заметьте, что в примере с мороженым единицы измерения — граммы, а не проценты! Работая с данными и сообщая о полученных результатах, всегда помните о единицах измерения. Также следите за тем, чтобы статистические показатели приводились с правильными единицами.

Чтобы избежать ошибки при округлении в вычислениях, на каждом этапе сохраняйте хотя бы два знака после запятой в каждой десятичной дроби. Ошибки при округлении накапливаются, и к тому времени, как вы закончите считать, результат уже может быть неверным.

Проверяем полученные результаты

Если вы хотите быть уверены в своих результатах больше, чем на 95%, тогда нужно добавлять или отнимать больше 2 стандартных ошибок. Например, чтобы быть уверенным на 99%, для получения предела погрешности вы должны добавить или отнять 3 стандартные ошибки. После этого предел погрешности будет больше, что обычно только ухудшает ситуацию. Большинство людей полагает, что добавлять или отнимать еще одну стандартную ошибку, чтобы добавить всего 4% уверенности (99% вместо 95%), неразумно. Единственный способ на 100% быть уверенным в результатах — это сделать предел погрешности таким большим (добавив или отняв множество стандартных ошибок), что он охватит все существующие возможности. Например, в конце концов, вы можете сказать что-то вроде: “Я на 100% уверен, что процентное отношение людей в совокупности, которым нравится мороженое, равно 50% плюс/минус 50%”. В таком случае вы на 100% уверены в результатах, но что они значат? Ничего.

Никогда нельзя быть абсолютно уверенным в том, что результаты выборки отражают ситуации в совокупности, даже при наличии предела погрешности (если, конечно, вы не совершаете таких сумасшедших поступков, как включение 100% всех возможностей, что было сделано в предыдущем примере с мороженым). Даже если вы уверены в своих результатах на 95%, на самом деле это означает, что если повторить процесс снова и снова, то в 5% случаев выборка не будет представлять всю совокупность, просто по случайности (а не из-за проблем с процессом выборки или чем-то еще). Значит, все результаты надо рассматривать, помня об этом. Ведь как бы там ни было, статистика никогда не призывает вас говорить, что вы абсолютно уверены!

Определяем влияние размера выборки

Два самых важных пункта относительно размера выборки таковы.

- ✓ Все эти формулы работают хорошо только до тех пор, пока выборка остается достаточно большого размера. (Тогда что такое "достаточно большой" размер? Речь об этом пойдет в данном разделе.)
- ✓ Между размером выборки и пределом погрешности существует обратная взаимосвязь.

Рассмотрим подробнее оба пункта.

Что такое "достаточно большой размер"

Почти все опросы проводятся с участием сотен или тысяч человек, и обычно этого бывает достаточно, чтобы теория, лежащая в основе статистических формул, работала исправно. Но статистики выработали несколько общих правил, чтобы наверняка удостовериться, достаточен ли размер выборки.

Для долей выборки нужно точно знать, что $n \times \hat{p}$ равно как минимум 5, и $n \times (1 - \hat{p})$ тоже равно как минимум 5. Например, в предыдущем примере опроса о поддержке президента $n = 1000$, $\hat{p} = 0{,}52$, а $1 - 0{,}52 = 0{,}48$. Значит, $n \times \hat{p} = 1000 \times 0{,}52 = 520$, а $n \times (1 - \hat{p}) = 1000 \times 0{,}48 = 480$. Оба результата намного больше 5, значит, здесь все в порядке.

Для среднего выборок вам нужно учитывать только размер выборки сам по себе. Вцелом размер выборки n должен быть больше 30, чтобы статистическая теория смогла работать. Но если он вдруг равен 29, не паникуйте, 30 — это не волшебное число, а просто общее правило.

Размер выборки и предел погрешности

Связь между размером выборки и пределом погрешности очень проста: с увеличением размера выборки предел погрешности уменьшается. Это обратная взаимосвязь, потому что оба значения движутся в разных направлениях. Если хорошенько подумать, то все логично: чем больше информации у вас есть, тем точнее должны быть результаты. (Конечно, здесь предполагается, что данные собирались и обрабатывались должным образом.)

Обратная взаимосвязь подробнее рассматривается в главе 9.

Больше — не всегда лучше!

В предыдущем примере опроса, касающегося рейтинга президента, результаты выборки всего в 1000 человек из 280 000 000 жителей Соединенных Штатов Америки имеют предел погрешности в 3% по сравнению с тем, что сказали бы все члены совокупности, если бы им предоставили такую возможность. Как это может быть?

С помощью формулы предела погрешности для доли выборки можно проследить, как резко меняется предел погрешности с изменением размера выборки.

Предположим, что в примере с поддержкой президента $n = 500$. (Помните, что в этом примере $\hat{p} = 0{,}52$.) Следовательно, предел погрешности для доверительного уровня в 95% равен

$$Z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1{,}96 \times \sqrt{\frac{(0{,}520)(0{,}480)}{500}} = 1{,}96 \times 0{,}0223 = 0{,}0438,$$

что составляет 4,38%. Если $n = 1000$ в том же примере, то предел погрешности (для доверительного уровня 95%) равен

$$1{,}96 \times \sqrt{\frac{(0{,}520)(0{,}480)}{1000}} = 1{,}96 \times 0{,}0158 = 0{,}0310,$$

что эквивалентно 3,10%. Если увеличить n до 1500, то предел погрешности (для того же доверительного уровня) будет равен

$$1{,}96 \times \sqrt{\frac{(0{,}520)(0{,}480)}{1500}} = 1{,}96 \times 0{,}0129 = 0{,}0253 \text{ или } 2{,}53\%.$$

Наконец, если $n = 2000$, предел погрешности равен

$$1{,}96 \times \sqrt{\frac{(0{,}520)(0{,}480)}{2000}} = 1{,}96 \times 0{,}0112 = 0{,}0219 \text{ или } 2{,}19\%.$$

Проанализируйте эти разные результаты и вы заметите, что после определенного момента польза от увеличения размера выборки все время уменьшается. Всякий раз, когда вы опрашиваете еще одного человека, стоимость опроса растет, а увеличение размера выборки с 1500 до 2000 снижает предел погрешности всего на 0,34% (одна треть процента!). Дополнительные расходы и усилия для такого незначительного уменьшения предела погрешности могут быть неоправданными. Больше — не всегда намного лучше!

Как ни странно, бывает, что больше — это хуже! Я объясню это в следующем разделе.

Ограничение предела погрешности

Предел погрешности — это мера того, насколько точно ваши результаты выборки отражают ситуацию в изучаемой совокупности. (Или, по крайней мере, он указывает верхний предел ошибки, который у вас должен быть.) Поскольку свои выводы о совокупности вы основываете на выборке, то нужно также сообщать, насколько эти результаты могут варьироваться хотя бы под воздействием случая.

Предел погрешности также показывает максимальное *ожидаемое* расстояние между результатами выборки и реальными результатами совокупности (как если бы они были получены в ходе переписи). Конечно, имей вы полную правду о совокупности, то вы бы не стали проводить опрос, не так ли?

Понимать, что предел погрешности *не* измеряет, не менее важно, чем то, что же он может измерить. Предел погрешности *не* измеряет ничего, кроме случайных вариаций — он не учитывает смещение или ошибки, которые происходят во время подбора участников, подготовки или проведения опроса, сбора данных или их анализе и формулировке окончательных выводов.

Большие выборки не всегда означают лучшие выборки! Хороший девиз, который нужно помнить при изучении статистических результатов, гласит: “Чем больше мусора на входе, тем больше его на выходе”. Каким бы красивым и научным ни казался предел погрешности, не забывайте, что формула, которой пользовались при его определении, ничего не говорит о качестве данных, на

которых предел погрешности основывается. Если доля выборки или среднее выборки были получены в ходе *смещенной выборки* (т.е. той, в которой одним людям отдавалось предпочтение перед другими), плохой разработки, неправильных методов сбора данных, смещенных вопросов или систематических ошибок при записи, тогда вычислять предел погрешности бессмысленно, ведь он не будет значить ровным счетом ничего. Например, опрос 50 000 человек производит впечатление, но если все это были посетители определенного Web-сайта, то предел погрешности для этих результатов не значит ничего, ведь все вычисления основаны на смещенных, вымышленных результатах! Конечно, некоторые исследователи все равно могут их обнародовать, поэтому вам нужно знать, что использовалось в формуле — хорошая информация или мусор. Если окажется, что это все же мусор, то вы знаете, что делать с пределом погрешности. Просто забудьте о нем. (Подробнее об ошибках, которые могут случиться во время опроса или эксперимента, см. главы 16 и 17 соответственно.)

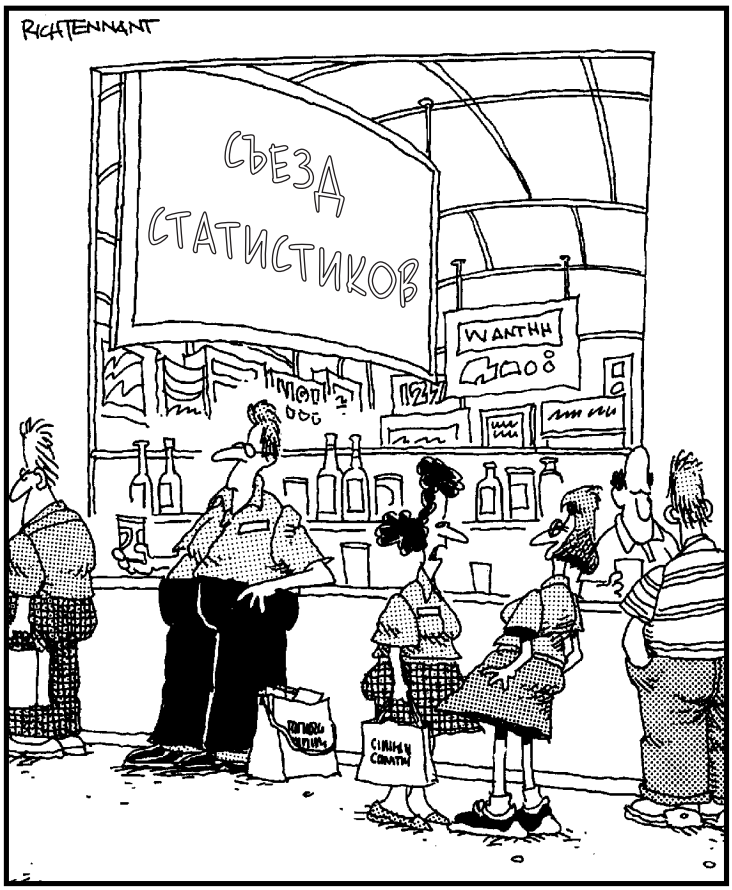
В правовых поправках, которыми сопровождается обнародование результатов опросов, *The Gallup Organization* всегда указывает, к чему относится предел погрешности. Эта организация сообщает, что помимо ошибок выборки опросы могут содержать дополнительные ошибки или смещение, что связано с формулировкой вопросов и некоторыми вопросами логистики при проведении исследований (например, потеря данных из-за обзвона номеров, которые больше не действительны). Это значит, что даже несмотря на наилучшие намерения и самое пристальное внимание к деталям и контролю над процессом, проколы все же бывают. В мире нет ничего идеального. И в данном случае важно знать, что предел погрешности не может измерить степень всех этих дополнительных ошибок.

Часть V

Уверенные предположения

The 5th Wave

Рич Теннант



"Будь осторожна. Он кажется хорошим мужем, но тут кто-то принес гистограмму его браков и разводов. Картина не очень радужная."

В этой части...

Всякий раз, когда вам сообщают какой-либо статистический показатель, вы узнаете лишь часть истории. В отдельном статистическом показателе нет самого главного — указания на то, в какой степени он будет варьироваться в разных ситуациях. Во всех хороших приблизительных оценках есть не только статистические показатели, но еще и предел погрешности. Такая комбинация показателя плюс/минус предел погрешности называется доверительным интервалом. Доверительные интервалы касаются не только отдельных статистических показателей, они содержат важную информацию о точности всей предположительной оценки.

В этой части вы найдете общее описание доверительных интервалов: их функции, формулы, вычисления, факторы влияния и толкование. Кроме того, вы познакомитесь с некоторыми примерами самых распространенных доверительных интервалов.

Глава 11

Приблизительные оценки: понятие доверительных интервалов

В этой главе...

- Описание работы процесса оценивания
- Общая формула доверительного интервала
- Интерпретация результатов доверительного интервала
- Обнаружение неправильных результатов

Большинство статистических показателей используется для приблизительной оценки какой-либо характеристики изучаемой совокупности, например, средний доход семьи, процентное отношение людей, покупающих подарки к Рождеству в режиме онлайн, среднее количество мороженого, ежегодно потребляемого в Соединенных Штатах Америки (наверное, этого лучше и не выяснять). Такие характеристики совокупности называются *параметрами*. Как правило, люди хотят оценить (т.е. высказать предположение) значение параметра, сделав выборку из совокупности и используя статистические показатели этой выборки, которые и позволят им сделать качественное предположение. Так что же такое "качественное предположение"?

Самым качественным было бы полное отсутствие предположения — т.е. если бы вы приступили к работе и сразу же точно определили параметр. Но определить точное значение параметра, не проводя перепись всей совокупности, невозможно — в большинстве случаев это была бы изматывающая и дорогостоящая работа. Но статистики не боятся трудностей, поэтому часто говорят: "Быть статистиком — значит, никогда не говорить, что ты уверен. Главное — просто подобраться поближе к истине". Конечно, статистики хотят быть уверенными в том, что полученные результаты как можно точнее отражают действительность, ведь на исследования были затрачены средства и время. Добиться как можно более точных результатов намного проще, чем вы думаете. Если процесс проводится правильно (а в средствах массовой информации это бывает нечасто!), приблизительная оценка может точно отражать параметр. В этой главе вы вкратце узнаете о доверительных интервалах (т.е. разновидности предположительных оценок, которые статистики используют и рекомендуют), о том, почему ими нужно пользоваться (в отличие от однократного предположения), как интерпретировать доверительный интервал и как замечать ошибочные предположения.

Не все предположения равны

Откройте журнал или газету, включите радио или телевизор, и вы найдете массу статистических данных, многие из которых представляют собой предположительные оценки того или иного количества. Возможно, вы задумаетесь над тем, как были получены эти

показатели. В одних случаях числа прошли тщательную проверку, в других — это просто выстрел наугад. Вот некоторые примеры предположений, на которые я наткнулась всего в одном номере ведущего журнала по бизнесу. Все они поступили из разных источников.

- ✓ 26 миллионов человек хотя бы раз в году играют в гольф.
- ✓ 6,7% домов в США были куплены без предоплаты.
- ✓ Хотя на сегодняшний день устроиться на работу непросто, в некоторых областях новые кадры очень нужны: в течение следующих восьми лет потребуется 13 000 помощников анестезиологов. Зарплата — от 80 до 95 тыс. долларов в год.
- ✓ За сезон бейсболист высшей лиги использует в среднем 90 бит.
- ✓ В 2002 году 7,4 миллиона жителей США отправились в круиз. Из них около 4% обратились за медицинской помощью на корабле.
- ✓ Автомобиль Lamborghini Murcielago разгоняется с 0 до 60 миль в час за 3,7 секунды. Максимальная скорость — около 205 миль в час.

Некоторые из этих данных получить проще, чем другие. Вот кое-какие наблюдения, которые я смогла сделать по этому поводу.

- ✓ Как узнать, что 26 миллионов человек хотя бы раз в год играют в гольф? На самом деле выяснить это не так уж и сложно, потому что все игроки в гольф перед началом игры должны заполнять анкету. Значит, после изучения заполненных анкет можно сделать качественное предположение о том, сколько человек играют хотя бы раз в год. (Единственная сложность — не считать повторно тех, кого вы уже учли раньше.)
- ✓ Установить процент путешественников, которым потребовалась медицинская помощь, или домов, купленных без предоплаты, можно в ходе опроса. Если опрос провести правильно (см. главу 16), то такие данные могут быть довольно точными.
- ✓ Как выяснить, сколько помощников анестезиологов потребуется в течение следующих восьми лет? Начать можно с выяснения, сколько специалистов за этот период уйдет на пенсию, однако при этом не учитывается развитие отрасли. Сделать предположение на один-два года можно довольно точно, но вот заглянуть в будущее на восемь лет — это задача намного сложнее.
- ✓ Найти среднее количество бит, которые использует за сезон бейсболист высшей лиги, можно, опросив самих игроков, людей, отвечающих за их снаряжение, или компании, поставляющие биты.
- ✓ Определить скорость автомобиля сложнее, но это можно сделать в ходе эксперимента с использованием секундомера. При этом нужно проверить много разных машин (а не только одну) той же модели.

Не все статистические показатели равны. Чтобы определить, надежен ли какой-либо показатель, не принимайте его на веру. Подумайте о том, имеет ли он смысл и как бы вы сформулировали предположение. Если эта информация действительно важна для вас, выясните, в ходе какого процесса было сделано предположение.

Связь показателя с параметром

В рамках годового отчета, полученного на основе Текущего опроса населения, Бюро переписей США оценивает *медиану* дохода семьи в Соединенных Штатах Америки и расписывает эту информацию по штатам. Зачем оценивать медиану, а не среднее значение дохода? (Подробнее о разнице между медианой и средним см. главу 5.) Потому что доход семьи несимметричен, ведь многие находятся у нижнего предела, а положение на максимально верхнем занимает намного меньше людей.

Чтобы определить медиану дохода семьи, Бюро переписей делает случайную выборку из примерно 28 000 семей и задает им вопросы (среди которых есть и вопрос о доходе). Исходя из данных выборки из 28 000 семей, Бюро подсчитывает медиану дохода для этой выборки. В 2000 году медиана дохода для выборки была равна 42 222 долларов.

Бюро переписей использует медиану дохода выборки (статистический показатель), чтобы предположить медиану дохода для всего населения США (параметр). Но поскольку Бюро известно, что выборка не всегда абсолютно точно отражает совокупность, то результаты сопровождаются указанием предела погрешности (см. главу 10). Эти плюс/минус (±), которые добавляются к любой оценке, позволяют с правильной точки зрения взглянуть на результаты. Если вы знаете предел погрешности, то понимаете, насколько ошибочной может быть оценка из-за того, что результаты основывались на выборке из совокупности, а не на всей генеральной совокупности.

Поскольку в выборку не была включена вся генеральная совокупность и выборка может не совсем идеально отражать ситуацию в совокупности, Бюро подсчитывает предел погрешности для медианы выборки и учитывает его при формулировке предположения. В 2000 году предел погрешности для медианы дохода выборки был равен 258. Поэтому Бюро переписей США предполагает, что в 2000 году медиана дохода семьи составила 42 228 плюс/минус 258 долларов или 42 228 ± 258 долларов. Это и есть доверительный интервал для медианы дохода семей в Соединенных Штатах Америки (подробнее о доверительных интервалах см. следующий раздел).

Заметьте, что предел погрешности в предыдущем примере сравнительно мал. Это объясняется тем, что была использована выборка большого размера (вы получаете то, за что платите!). Подробнее о взаимосвязи между размером выборки и пределом погрешности см. главу 10.

Делаем самое лучшее предположение

Лучший способ оценить параметр (т.е. характеристику всей совокупности) — это определить статистический показатель плюс/минус предел погрешности, воспользовавшись данными большой выборки. Таким образом, ваши результаты будут представлять собой предположение, основанное на выборке, а также будет указан определенный индикатор, который сообщает, насколько это предположение может отличаться от выборки к выборке.

Статистический показатель плюс/минус предел погрешности — это и есть *доверительный интервал*.

- ✓ Слово *интервал* используется потому, что полученный вами результат превращается в интервал. Например, скажем, процент детей, которым нравится бейсбол, составляет 40% плюс/минус 3,5%. Это означает, что процент детей, которым нравится бейсбол, находится где-то между 40% – 3,5% = 36,5% и 40% + 3,5% = 43,5%. Значит, нижняя граница интервала — это статистика минус предел погрешности, а верхняя граница — это показатель плюс предел погрешности.
- ✓ Слово *доверительный* используется потому, что вы в определенной степени доверяете процессу, в ходе которого получили этот интервал. Это называется вашим *доверительным уровнем*.

Формулы и примеры наиболее широко используемых доверительных интервалов вы найдете в главе 13.

Уверенная интерпретация результатов

Представим, что вы — биолог-исследователь, пытаетесь поймать рыбу ручной сетью, размер которой соответствует ширине вашего доверительного интервала. (Ширина равна пределу погрешности, умноженному на два, чтобы учесть и сложение, и вычитание.)

Предположим, ваш доверительный уровень равен 95%. Что это означает? Это означает, что если вы будете снова и снова забрасывать свою сеть в воду, то выловите 95% рыбы. Ловля рыбы в данном случае значит, что ваш доверительный интервал был правильным и содержал истинный параметр (здесь этот параметр представлен рыбой).

Но значит ли это, что у вас 95% шансов поймать рыбу, если вы забросите сеть всего один раз? Нет. Непонятно? Наверняка. Объясняю: к примеру, у вас была единственная попытка забросить сеть и вы закрыли глаза перед тем, как забросить ее в воду. В этот момент у вас 95% поймать рыбу. Но протяните сеть под водой, не открывая глаз — и у вас останется всего два варианта: вы либо поймаете рыбу, либо нет. Вероятность здесь не играет роли.

Точно так же, после того, как данные были собраны, а доверительный интервал вычислен, вы либо находите истинный параметр генеральной совокупности, либо нет. Значит, вы не говорите, что на 95% уверены в том, что параметр находится в этом интервале, потому что вы ведь или нашли его, или нет. В чем вы на 95% уверены — так это в процессе, в ходе которого собирали данные и находили доверительный интервал. Вы знаете, что в результате этого процесса будут получены интервалы, которые правильно отражают среднее значение в 95% случаев. Оставшиеся 5% случаев данные, собранные в выборке, просто случайно имеют аномально высокие или низкие значения, поэтому не представляют совокупность. В таких случаях вы не нашли параметр.

Таким образом, имея правильный размер и структуру сети, вы поймаете 95% рыбы в течение заданного периода времени. Но в ходе каждой отдельной попытки вы либо поймаете рыбу, либо нет.

Доверительный уровень, размер выборки и изменчивость совокупности — все это влияет на предел погрешности, а значит, и на ширину доверительного интервала. Но предел погрешности, а, следовательно, и ширина доверительного интервала, ничего не будут значить, если данные, используемые в исследовании, были смещенными и/или ненадежными. Самый лучший выход — узнать, как собирались данные, а уж затем считать истинным полученный предел погрешности (см. главу 10).

Замечаем ошибочные доверительные интервалы

Если данные получены в ходе хорошо разработанных опросов или экспериментов (см. главы 16 и 17) и основаны на случайных выборках большого размера (см. главу 9), можно доверять качеству информации. Если предел погрешности небольшой, этот доверительный интервал предполагает точную и надежную оценку параметра. Но, к сожалению, так бывает не всегда.

Не все оценки такие точные и надежные, как вас хотят в том убедить. Например, у опроса с участием 20 000 человек, проведенного по Интернету, возможно, предел погрешности будет небольшим, но он не значит ничего, если в этом опросе принимали участие только те люди, кто случайно оказался на конкретном Web-сайте. Другими словами, такая выборка не имеет ничего общего со случайной выборкой (когда у всех элементов совокупности есть равные шансы попасть в выборку). Тем не менее, о подобных результатах сообщается очень часто, при этом приводится и предел погрешности, в результате чего исследование кажется таким надежным. Но будьте начеку и помните о таких вымышленных результатах! (Подробнее о границах предела погрешности см. главу 10.)

Прежде чем принимать решения, исходя из чьей-либо оценки, проделайте следующее.

- ✓ Выясните, как был получен статистический показатель. Он должен быть результатом научного процесса, в ходе которого собираются надежные, объективные и точные данные. (См. главы 2 и 3.)
- ✓ Ищите предел погрешности. Если он не указывается, найдите оригинальный источник.
- ✓ Помните, что если статистика ненадежна или имеет смещение, то предел погрешности не имеет смысла. (Подробнее о том, как избежать смещения при опросе, см. главу 16, а критерии качественных данных в экспериментах рассматриваются в главе 17.)

Глава 12

Вычисление точных доверительных интервалов

В этой главе...

- Ожидание определенного доверительного уровня в предполагаемых результатах
- Общие методы вычисления доверительного интервала
- Факторы, влияющие на ширину доверительного интервала

Доверительный интервал — красивое название статистического показателя, вместе с которым сообщается и предел погрешности (общие сведения о доверительном интервале можно найти в главе 11; подробнее о пределе погрешности см. главу 10). Большинство статистических показателей служат для того, чтобы высказать предположение об определенных характеристиках совокупности (называемых *параметрами*), беря за основу выборку. Поэтому каждый статистический показатель должен содержать предел погрешности. Ведь (как сказано в главе 9) у разных выборок будут разные результаты!

В этой главе вы узнаете, как вычислить собственный доверительный интервал. Вы познакомитесь с некоторыми деталями доверительных интервалов: что делает их *уже* или шире, почему можно быть более или менее уверенным в полученных результатах, а также что они измеряют, а что — нет. Имея такую информацию, вы будете знать, что искать, когда вам встретятся статистические результаты, и сможете установить, насколько они точны.

Вычисление доверительного интервала

Доверительный интервал состоит из статистического показателя плюс/минус предел погрешности (см. главу 10). Например, предположим, вы хотите узнать процент пикапов среди всех транспортных средств в Соединенных Штатах Америки (в данном случае это и будет параметром). Невозможно изучить все машины в США, поэтому вы делаете случайную выборку из 1000 транспортных средств на разных автомагистралях в разное время дня. В результате выясняется, что 7% выбранных машин были пикапами. Но ведь вы не можете сказать, что ровно 7% всех автомобилей на американских дорогах будут пикапами, поскольку известно, что этот результат основан всего на 1000 выбранных машин. И хотя 7% — это довольно близко к истинному показателю, наверняка это знать невозможно, потому что вы основываете результаты на выборке, а не на всех транспортных средствах в США.

Отношение подростков к бездымному табаку

Продолжительное исследование, которое проводит Университет Мичигана, изучает мнение подростков по разным вопросам, в том числе и о том, насколько рискованным они считают употребление бездымного табака (который обычно называется жевательным табаком). Исследование показывает, что подростков, которые считают употребление бездымного табака очень рискованным, стало больше, чем 15 лет назад. Результаты исследования приведены ниже.

- ✓ В выборке 2001 года, состоящей из 2 100 учеников двенадцатого класса, 45,4% считали, что бездымный табак наносит серьезный вред. Предел погрешности был равен плюс/минус 2%. Следовательно, 95% доверительный интервал для процентного отношения *всех* учеников двенадцатых классов, которые считают бездымный табак очень рискованным, будет 45,4% ± 2%.
- ✓ Исходя из данных выборки 1986 года, состоявшей из 3000 учеников двенадцатого класса, доверительный уровень для всех двенадцатиклассников, считавших бездымный табак вредным, был 25,8% ± 1,6%.

Так что же делать? Вы добавляете и отнимаете предел погрешности, чтобы показать, какую, на ваш взгляд, неточность могут иметь результаты. (Подробнее о пределе погрешности см. главу 10.) Эта неточность объясняется не тем, что вы что-то сделали не так, а всего лишь тем фактом, что проводилась только *выборка* (изучение части совокупности), а не *перепись* (изучение всей совокупности).

Ширина доверительного интервала — это предел погрешности, умноженный на два. Например, представим, что предел погрешности равен 5%. Значит, доверительный интервал показателя в 7% плюс/минус 5% идет от 7% – 5% = 2% вплоть до 7% + 5% = 12%. Это значит, что ширина доверительного интервала равна 12% – 2% = 10%. Более простой способ определить этот интервал — сказать, что ширина доверительного интервала составляет предел погрешности, умноженный на два. В данном случае ширина доверительного интервала равна $2 \times 5\% = 10\%$.

Ширина доверительного интервала — это расстояние от нижней границы интервала (статистика – предел погрешности) до верхней границы интервала (статистика + предел погрешности). А чтобы быстрее определить ширину доверительного интервала, можно умножить предел погрешности на два.

Ниже описаны шаги для оценки параметра вместе с доверительным интервалом, а также подсказки, где можно найти более подробную информацию о каждом этапе.

1. **Выберите доверительный уровень и размер выборки (см. главу 9).**
2. **Сделайте случайную выборку элементов из совокупности (см. главу 3).**
3. **Соберите надежные и объективные данные об элементах выборки.** Подробнее данные опросов описаны в главе 16, а данные экспериментов — в главе 17.
4. **На основе данных определите статистику, обычно это среднее или доля (см. главу 5).**
5. **Подсчитайте предел погрешности (см. главу 10).**
6. **Проанализируйте статистику плюс/минус предел погрешности и дайте окончательную оценку параметра.**

   Это называется *доверительным интервалом* для данного параметра.

Выбор доверительного уровня

Обратите внимание, что в примере об отношении подростков к бездымному табаку (см. соответствующий раздел выше) есть фраза "95% доверительный интервал". У каждого доверительного интервала (и, если уж на то пошло, у каждого предела погрешности) есть связанный с ним доверительный уровень. В данном примере доверительный уровень был равен 95%. Доверительный уровень помогает учесть другие возможные результаты, которые вы могли бы получить, если делаете предположение, исходя из однократной выборки. Если вы хотите на 95% быть уверенным в других возможных результатах, тогда ваш доверительный уровень будет 95%.

Изменчивость результатов выборки измеряется в количестве стандартных ошибок. *Стандартная ошибка* похожа на стандартное отклонение в наборе данных, единственное отличие — стандартная ошибка применяется по отношению к среднему или процентному отношению выборки, которые вы могли бы получить. (Подробнее о стандартных ошибках см. главу 10.) Каждому доверительному уровню соответствует определенное количество стандартных ошибок, которые нужно добавить или отнять. Такое количество стандартных ошибок называется *Z*-значением (потому что оно соответствует стандартному нормальному распределению). См. табл. 10.1 в главе 10.

Какой доверительный уровень обычно используют исследователи? Бывают разные уровни, от 80 до 99%. Самый распространенный доверительный уровень — 95%. Статистики любят шутить: "Почему статистикам нравится их работа? Потому что им нужно давать правильные ответы только в 95% случаев". (Броско, но довольно привлекательно?)

Быть на 95% уверенным — это значит, что если вы сделаете много-много выборок и каждый раз, исходя из результатов, определите доверительный интервал, то в 95% случаев полученные доверительные интервалы попадут точно в цель, т.е. будут действительно отражать истинный параметр. Чтобы получить 95% доверительный уровень, согласно эмпирическому правилу необходимо добавить или отнять "около" 2 стандартных ошибок. Центральная предельная теорема позволяет точнее назвать это количество, и поэтому "около 2" на самом деле означает 1,96. В табл. 10.1 из главы 10 представлены некоторые доверительные уровни и соответствующие им *Z*-значения.

Если вы хотите быть уверенными в своих результатах больше, чем на 95%, тогда нужно добавлять и отнимать больше стандартных ошибок. Например, чтобы быть уверенным на 99%, нужно получить предел погрешности, добавив и отняв три стандартные ошибки. Чем выше доверительный уровень, тем больше *Z*-значение, больше предел погрешности и шире доверительный интервал (при условии, что все остальные данные остаются прежними). За дополнительную уверенность приходится платить.

Обратите внимание на фразу: "При условии, что все остальные данные остаются прежними". Можно компенсировать увеличение предела погрешности увеличением роста выборки. Подробнее об этом см. раздел "Факторы в размере выборки".

Подробнее о ширине

При высказывании предположения с использованием доверительного интервала главная цель состоит в том, чтобы доверительный интервал был узким. Тогда можно точнее определить параметр. Если добавить и отнять большее число, результат получится менее точным. Например, предположим, вы пытаетесь определить процент машин

с полуприцепами на федеральной автомагистрали в период между 12 и 18 часами, и в результате у вас получается 95% доверительный интервал, согласно которому процентное отношение таких грузовиков равно 50% плюс/минус 50%. Интервал действительно снижается! (Разумеется, это шутка!) Однако вы забыли о главном, пытаясь дать качественное предположение.

В данном случае доверительный интервал слишком широкий. Лучше было бы сказать примерно так: 95% доверительный интервал для процентного отношения машин с полуприцепами на федеральной автомагистрали между 12 и 18 часами дня равен 50% плюс/минус 3%. Для этого потребовалась бы выборка большего размера, но дело того стоило бы.

Итак, если предел погрешности небольшой — это хорошо, значит, меньше — еще лучше? Не всегда. Чтобы максимально сузить доверительный интервал, вам придется провести намного более сложное — и дорогое — исследование, и на определенном этапе увеличение стоимости не оправдывает незначительное повышение точности. Большинство исследователей при определении процента (например, процента женщин, республиканцев или курильщиков) спокойно довольствуются пределом погрешности от 2% до 3%.

Узкий доверительный интервал — это хорошо.

Но как добиться того, чтобы доверительный интервал был достаточно узким? Обдумать этот вопрос придется до того, как собирать данные, ведь после окончания сбора данных ширина доверительного интервала уже установлена.

На ширину доверительного интервала влияют три фактора.

- ✓ Доверительный уровень (о чем уже говорилось в предыдущем разделе).
- ✓ Размер выборки.
- ✓ Степень изменчивости в генеральной совокупности.

Формула предела погрешности в том, что касается среднего выборки, такова: $Z \times \frac{s}{\sqrt{n}}$, где

- ✓ Z — значение из стандартного нормального распределения, соответствующее доверительному уровню (см. табл. 10.1 в главе 10).
- ✓ n — размер выборки (см. главу 9).
- ✓ $\frac{s}{\sqrt{n}}$ — стандартная ошибка среднего выборки (подробнее о стандартной ошибке см. главу 10).

Доверительный интервал для среднего значения будет равен $\bar{x}$ плюс/минус предел погрешности. В главе 13 приводятся формулы самых распространенных доверительных интервалов, с которыми вы можете столкнуться.

Каждый из этих трех факторов (доверительный уровень, размер выборки и изменчивость совокупности) в значительной степени воздействует на ширину доверительного интервала. Вы уже знаете, в чем заключается влияние доверительного уровня. В следующем разделе вы узнаете, как на ширину доверительного интервала влияют размер выборки и изменчивость совокупности.

Заметьте, что сама по себе статистика (например, 7% машин в примере с пикапами) не имеет отношения к доверительному интервалу, за ширину которого целиком отвечают предел погрешности и указанные выше три фактора.

Факторы в размере выборки

Связь между пределом погрешности и размером выборки очень проста: по мере увеличения размера выборки предел погрешности уменьшается. Это соответствует тому, что, как вы надеетесь, является правдой: чем больше у вас информации, тем точнее будут результаты. (Конечно, здесь предполагается, что это будет хорошая, надежная информация. Подробнее о проблемах со статистикой см. главу 2.)

Посмотрите на формулу предела погрешности для среднего выборки и обратите внимание, что в знаменателе этой дроби стоит n (это касается почти всех формул предела погрешности): $Z \times \frac{s}{\sqrt{n}}$. Когда n увеличивается, то увеличивается знаменатель дроби, из-за чего вся дробь становится меньше. Поэтому предел погрешности уменьшается, а доверительный интервал становится *уже*.

Если вас интересует высокий доверительный уровень, нужно увеличить Z-значение, а значит, и предел погрешности, в результате чего доверительный интервал расширяется, а это плохо. Компенсировать расширение доверительного интервала можно, увеличив размер выборки и тем самым снизив предел погрешности, что и приведет к сужению доверительного интервала. Увеличение размера выборки позволит сохранить необходимый доверительный уровень, вместе с тем не расширив доверительный интервал (а именно это, по большому счету, вам и нужно). Это можно выяснить даже до начала исследования: если вам известен предел погрешности, который вы хотите получить, то вы сможете заранее определить необходимый размер выборки (см. главу 9).

Если в качестве статистического показателя вы будете использовать процентное отношение (например, процент людей, предпочитающих летом ходить в сандалиях), то приблизительно прикинуть предел погрешности можно, разделив 1 на корень квадратный из n (размер выборки). Можно подобрать разные значения n и посмотреть, как при этом меняется предел выборки.

Какой же приблизительный размер выборки нужен при проведении опросов, чтобы доверительный интервал оставался узким? Используя формулу из предыдущего параграфа, вы можете кое-что быстро сравнить. У опроса из 100 человек предел погрешности будет равен около $\frac{1}{\sqrt{100}} = 0{,}10$ или плюс/минус 10% (т.е. ширина доверительного интервала составляет 20%, а это довольно много). Но если опросить 1000 человек, то предел погрешности заметно снижается и будет равен плюс/минус 3%, ширина доверительного интервала теперь составляет 6%. После опроса 2 500 человек предел погрешности будет равен плюс/минус 2% (значит, ширина уменьшится до 4%). Если представить, насколько большая вся совокупность (например, население США — свыше 280 млн человек!), то это будет довольно маленькая выборка, позволяющая, тем не менее, получить точные результаты.

Однако не переусердствуйте с размером выборки, потому что в определенный момент выгода от этого исчезает. Например, если от выборки в 2500 человек перейти к 5000, то доверительный интервал сузится примерно до $2 \times 1{,}4 = 2{,}8\%$. Всякий раз, когда вы опрашиваете еще одного человека, то стоимость опроса возрастает, а значит, добавлять еще 2 500 человек к опросу только для того, чтобы сузить интервал немногим больше чем на 1% , может оказаться неразумным.

Точность вычисления результатов зависит не только от размера выборки, но и от качества данных. Выборка большого размера, для которой характерно значительное смещение (см. главу 2), может дать узкий доверительный интервал, но при этом она не будет значить ровным счетом ничего. Это как будто вы соревнуетесь в стрельбе и равномерно выпускаете стрелы из лука, но всякий раз оказывается, что вы попадаете в цель соседа — вот насколько вы сбились с курса. Но в свете статистики оценить смещение нельзя, можно только попытаться свести его к минимуму.

Чем больше размер выборки, тем меньшим будет предел погрешности, и тем *уже* станет доверительный интервал, при условии, что все остальное останется без изменений, а качество данных будет высоким.

Учитываем изменчивость совокупности

Один из факторов, влияющих на изменчивость результатов выборки, — это тот факт, что изменчивость присуща и самой совокупности. Если бы все значения совокупности оставались неизменными, то представьте, каким бы скучным стал этот мир! (Более того, если бы не изменчивость, не существовало бы и статистики.) Например, в совокупности домов такого большого города, как Коламбус, штат Огайо, вы заметите значительную разницу не только в разновидности строений, а и в их размере и стоимости. А изменчивость в стоимости жилья в Коламбусе должна быть больше, чем изменчивость в стоимости жилья в отдельно взятом жилом массиве Коламбуса.

Это значит, что если вы делаете выборку со всего Коламбуса и находите среднюю стоимость, то предел погрешности должен быть больше, чем когда вы сделаете выборку в одном крупном жилом массиве Коламбуса, даже если доверительный уровень и размер выборки каждый раз будут одинаковыми. Почему? Потому что стоимость жилья в городе разная, и среднее выборки тоже будет меняться от выборки к выборке больше, чем если бы вы сделали выборку всего в одном жилом массиве, где цены примерно одинаковы. Это означает, что для того, чтобы получить ту же степень точности, которая будет при выборке из одного жилого массива, вам нужно будет включить в выборку со всего города больше домов.

Изменчивость измеряется стандартным отклонением. Стандартное отклонение совокупности (σ), как правило, неизвестно, поэтому вы заменяете его *s*, стандартным отклонением выборки (см. главу 4). Обратите внимание, что в формуле предела погрешности для среднего выборки *s* появляется в числителе стандартной ошибки: $Z \times \frac{s}{\sqrt{n}}$. Следовательно, с увеличением стандартного отклонения (числитель) увеличивается и стандартная ошибка (вся дробь). В результате предел погрешности возрастает, а доверительный интервал становится шире.

Усиление изменчивости исходной совокупности увеличивает предел погрешности, в результате чего расширяется доверительный интервал. Это увеличение можно компенсировать увеличением размера выборки.

Глава 13

Самые распространенные доверительные интервалы: формулы и примеры

В этой главе...

- Знакомство с формулами доверительных интервалов
- Уверенное вычисление

Если требуется найти среднее совокупности, но вы не можете сделать этого из-за ограниченности во времени или в средствах (обычно так и происходит), лучший выход в такой ситуации — сделать выборку из совокупности, найти *ее* среднее и использовать его для приблизительной оценки среднего всей совокупности. В таком случае (подробнее см. главы 11 и 12) необходимо учесть еще один критерий, показывающий, насколько точными будут ваши результаты. Ведь сделай вы другую выборку, эти результаты хоть немного, но все же изменятся. Значит, вместе со средним выборки вам нужно определить предел погрешности (т.е. то, насколько могут варьироваться результаты от выборки к выборке), а среднее выборки плюс/минус предел погрешности составляет доверительный интервал для среднего всей совокупности.

Найти доверительный интервал может быть не так уж просто, поэтому в этой главе я познакомлю вас с формулами самых распространенных доверительных интервалов (ДИ), объясню особенности вычислений и приведу некоторые примеры.

Вычисление доверительного интервала для среднего совокупности

Если измеряемая характеристика (например, доход, IQ, цена, высота, количество или вес) относится к *числовым* данным, то большинство стремится найти среднее значение для генеральной совокупности, потому что среднее — это единый числовой итог для совокупности, показывающий, где находится ее центр. Вы находите среднее совокупности, используя среднее выборки плюс/минус предел погрешности. Результат такого действия называется *доверительным интервалом для среднего совокупности*.

Формула ДИ для среднего совокупности такова: $\bar{x} \pm Z \times \frac{s}{\sqrt{n}}$, где $\bar{x}$ — среднее выборки, s — стандартное отклонение выборки, n — размер выборки, а Z — подходящее значение из стандартного нормального распределения, соответствующее желаемому доверительному уровню. Подробнее о формулах для $\bar{x}$ и s см. главу 3, значения Z для конкретных доверительных уровней описаны в главе 10 (табл. 10.1).

Чтобы найти ДИ для среднего совокупности, поступайте следующим образом.

1. **Определите доверительный уровень и найдите соответствующее *Z*-значение.**

   См. главу 10 (табл. 10.1).

2. **Найдите среднее выборки ($\bar{x}$), стандартное отклонение выборки (*s*) и размер выборки (*n*).**

   См. главу 3.

3. **Умножьте *Z* на *s* и разделите произведение на корень квадратный из *n*.**

   Это предел погрешности.

4. **Чтобы получить ДИ, возьмите $\bar{x}$ плюс/минус предел погрешности**. Нижняя граница ДИ — это $\bar{x}$ минус предел погрешности, а верхняя граница — $\bar{x}$ плюс предел погрешности.

Например, представьте себе, что вы работаете в Министерстве природных ресурсов и хотите с 95% уверенностью определить среднюю длину мальков судака в рыбопитомнике.

Поскольку вас интересует 95% доверительный интервал, то необходимое значение *Z* равно 1,96.

Предположим, вы делаете случайную выборку из 100 мальков и определяете, что их средняя длина составляет 7,5 дюйма. Стандартное отклонение (*s*) равно 2,3 дюйма. (Подробнее о вычислении среднего и стандартного отклонения см. главу 4.) Это значит, что $\bar{x} = 7{,}5$, $s = 2{,}3$, а $n = 100$.

Умножаем 1,96 на 2,3 и делим результат на корень квадратный из 100 = (10). Следовательно, предел погрешности равен плюс/минус $1{,}96 \times (2{,}3/10) = 1{,}96 \times 0{,}23 = 0{,}45$ дюйма.

Ваш 95% доверительный интервал для средней длины мальков судака в этом рыбопитомнике равен 7,5 дюймов плюс/минус 0,45 дюйма. (Нижняя граница интервала равна 7,5 – 0,45 = 4,05 дюймов, верхняя граница — это 7,5 + 0,45 = 7,95 дюймов.). Значит, вы можете с 95% уверенностью сказать, что, исходя из данных выборки, средняя длина мальков судака во всем рыбопитомнике составляет от 7,05 до 7,95 дюймов (17,9–20,2 см).

Если размер выборки небольшой (меньше 30), то вычисления потребуют небольшой модификации. Подробнее об этом речь пойдет в главе 15.

Определение доверительного интервала для доли совокупности

Если измеряемая характеристика относится к категорийным, или *дискретным*, данным (к примеру, мнение по определенному вопросу (поддержка, отрицание или равнодушие), пол, политическая партия, модель поведения (пристегиваете или нет ремень безопасности во время вождения)), то чаще всего исследователей интересует доля (или процент) людей, относящихся к определенной категории. Например, процентное отношение людей, выступающих за четырехдневную рабочую неделю, процентное отношение республиканцев, участвовавших в прошлых выборах, или процентное отношение водителей, которые не пристегивают ремень безопасности. В каждом из этих случаев задача заключается в том, чтобы дать приблизительную оценку доли совокупности, ис-

пользуя долю выборки плюс/минус предел погрешности. Результат называется *доверительным интервалом для доли совокупности*.

Формула ДИ для доли совокупности такова: $\hat{p} \pm Z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, где $\hat{p}$ — доля выборки, n — размер выборки, а Z — подходящее значение из стандартного нормального распределения, соответствующее желаемому доверительному уровню. (Подробнее о формулах и вычислении $\hat{p}$ см. главу 3. Значения Z для конкретных доверительных уровней представлены в главе 10, табл. 10.1.)

Чтобы подсчитать ДИ для доли совокупности, выполните такие действия:

1. **Определите доверительный уровень и найдите соответствующее *Z*-значение.**

   См. главу 10 (табл. 10.1).

2. **Найдите долю выборки ($\hat{p}$), разделив количество людей в выборке, обладающих необходимыми характеристиками, на размер выборки (*n*).**

   Примечание: $\hat{p}$ должно быть десятичной дробью от 0 до 1.

3. **Умножьте $\hat{p}$ на (1 – $\hat{p}$), затем разделите произведение на *n*.**

4. **Найдите корень квадратный из результата, полученного на этапе 3.**

5. **Умножьте ответ на *Z*.**

   Это будет предел погрешности.

6. **Чтобы получить ДИ, возьмите $\hat{p}$ плюс или минус предел погрешности. Нижняя граница ДИ — $\hat{p}$ это минус предел погрешности, а верхняя граница — $\hat{p}$ плюс предел погрешности.**

Например, предположим, что вы хотите определить процентное отношение количества раз, когда на определенном перекрестке загорается красный свет.

Поскольку вас интересует 95% доверительный интервал, то *Z*-значение равно 1,96.

Вы делаете случайную выборку из 100 разных поездок через этот перекресток и устанавливаете, что 53 раза попали на красный свет, значит, $\hat{p} = 53/1000 = 0{,}53$.

Берем $0{,}53 \times (1 - 0{,}53)$ и делим это произведение на 100. Получается $0{,}2491/100 = 0{,}002491$.

Корень квадратный из этого числа равен 0,0499.

Следовательно, предел погрешности равен плюс или минус $1{,}96 \times 0{,}0499 = 0{,}0978$.

Ваш 95-процентный доверительный интервал для того количества раз, когда вы на конкретном перекрестке попадете на красный свет, составляет 0,53 (или 53%) плюс/минус 0,0978 (округляем это значение до 0,10 или 10%). (Нижняя граница интервала равна $0{,}53 - 0{,}10 = 0{,}43$ или 43%, а верхняя граница будет равна $0{,}53 + 0{,}10 = 0{,}63$ или 63%.) Другими словами, вы можете с 95% уверенностью сказать, что, исходя из данных выборки, на этом перекрестке попасть на красный свет можно от 43% до 53% случаев.

Делая вычисления, связанные с процентами, используйте десятичные дроби. Закончив вычисления, преобразуйте дробь в процентное отношение, умножив ее на 100. Чтобы избежать ошибки при округлении, всегда оставляйте минимум две цифры после запятой.

Нахождение доверительного интервала для разницы двух средних значений

Цель многих опросов и исследований состоит в том, чтобы сравнить две совокупности, например, мужчин и женщин, семьи с высоким и низким доходом, республиканцев и демократов. Если сравниваемые характеристики являются числовыми (например, рост, вес или доход), то вас будет интересовать разница средних значений между двумя совокупностями. К примеру, вы хотите сравнить разность в возрасте между республиканцами и демократами или разность среднего дохода у мужчин и женщин. Найти разность средних значений двух совокупностей можно, сделав выборку в каждой совокупности и использовав разницу между средними значениями этих выборок плюс/минус предел погрешности. Результатом будет *доверительный интервал для разницы между средними значениями двух совокупностей*.

Формула ДИ для разницы между средними двух совокупностей следующая:

$$(\bar{x} - \bar{y}) \pm Z \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

где $\bar{x}$, s_1 и n_1— это среднее, стандартное отклонение и размер первой выборки, а $\bar{y}$, s_2 и n_2 — среднее, стандартное отклонение и размер второй выборки. Z — это подходящее значение из стандартного нормального распределения, соответствующее желаемому доверительному уровню. (Подробнее о формулах и вычислении среднего значения и стандартного отклонения см. главу 3. Значения Z для конкретных доверительных уровней представлены в главе 10, табл. 10.1.)

Чтобы найти ДИ для разницы между средними двух совокупностей, проделайте следующее.

1. **Определите доверительный уровень и найдите соответствующее Z-значение.**

   См. главу 10 (табл. 10.1).

2. **Найдите среднее ($\bar{x}$), стандартное отклонение (s_1) и размер первой выборки (n_1), а также среднее ($\bar{y}$), стандартное отклонение (s_2) и размер второй выборки (n_2).**

   См. главу 3.

3. **Найдите разность средних значений выборок ($\bar{x} - \bar{y}$).**

4. **Возведите в квадрат s_1 и разделите на n_1. Возведите в квадрат s_2 и разделите на n_2. Сложите результаты и извлеките из суммы корень квадратный.**

5. **Умножьте результат, полученный на этапе 4, на Z.**

   Это предел погрешности.

6. **Чтобы получить ДИ, возьмите ($\bar{x} - \bar{y}$) плюс/минус предел погрешности.**

   Нижняя граница ДИ равна ($\bar{x} - \bar{y}$) *минус* предел погрешности, а верхняя граница ДИ — это ($\bar{x} - \bar{y}$) *плюс* предел погрешности.

Предположим, вы хотите с 95% уверенностью определить разницу между средней длиной початков двух разновидностей кукурузы (при условии, что все они растут в одинаковых условиях и в течение одинакового периода времени). Назовем эти разновидности кукурузы Статзерно и Зернослад ость.

Поскольку вас интересует доверительный интервал в 95%, то Z равно 1,96.

Предположим, вы сделали случайную выборку из 100 початков Статзерна и нашли, что их средняя длина равна 8,5 дюймов со стандартным отклонением 2,3 дюйма. В случайной выборке из 110 початков Зернosладости средняя длина початков равна 7,5 дюймов со стандартным отклонением 2,8 дюймов. Это значит, что $\bar{x}$=8,5, s_1 = 2,3 и n_1 = 100, $\bar{y}$ = 7,5, s_2 = 2,8 и n_2 = 110.

Разность между средними выборок ($\bar{x} - \bar{y}$) составляет 8,5 – 7,5 = +1 дюйм. Это значит, что среднее значение для Статзерна минус среднее значение для Зернosладости является положительным значением, следовательно, из этих двух разновидностей кукурузы початки Статзерна в сделанной выборке больше. Но достаточно ли большая эта разница, чтобы ее можно было отнести ко всей совокупности? Именно это вам и поможет решить доверительный интервал.

Возводим в квадрат s_1 (2,3) и получаем 5,29. Делим это значение на 100, получается 0,0529. Возводим в квадрат s_2 (2,8) и делим на 110: 7,84/110 = 0,0713. В сумме результаты дают 0,0529 + 0,0713 = 0,1242. Корень квадратный из этой суммы равен 0,3524.

Умножим 1,96 на 0,3524 и получим 0,69 дюйма. Это будет предел погрешности.

Ваш 95% доверительный интервал для разницы между средней длиной початков этих двух разновидностей кукурузы равен 1 дюйм плюс/минус 0,69 дюйма. (Нижняя граница интервала — это 1 – 0,69 = 0,31 дюйма, а верхняя граница интервала равна 1 + 0,69 = 1,69 дюйма.) Это означает, что вы можете с 95% уверенностью сказать, что початки вида Статзерно в среднем длиннее, чем початки вида Зернosладость примерно на 0,31–1,69 дюйма (0,79–4,29 см). (Обратите внимание, что все значения в этом интервале положительны. Это значит, что, исходя из полученных вами данных, в среднем початки Статзерна всегда будут длиннее, чем початки Зернosладости.)

Заметьте, что разность ($\bar{x} - \bar{y}$) может быть отрицательной. Например, если переставить две разновидности кукурузы, то разность была бы –1. Ничего страшного, просто не перепутайте группы. Положительная разность означает, что в первой группе значения больше, чем во второй, а отрицательная разность говорит о том, что значения в первой группе меньше значений второй группы. Если вы хотите избежать отрицательных значений, всегда выбирайте в качестве первой группы ту, значения которой больше. В таком случае разность всегда будет положительной.

Если ваша выборка небольшого размера (меньше 30 элементов), в главе 15 вы найдете информацию о том, как нужно изменить принцип вычислений.

Определение доверительного интервала для разницы двух долей

Если сравниваются характеристики двух групп, относящиеся к *категорийным* (*дискретным*) данным, например, мнение по определенному вопросу (за/против), тогда нас будет интересовать разница между долями двух совокупностей — к примеру, разница между долей женщин, поддерживающих идею насчет четырехдневной рабочей неде-

ли, и долей мужчин, выступающих за то же предложение. Чтобы найти разницу между долями двух разных совокупностей, вы делаете выборку в каждой совокупности и используете разницу между долями этих совокупностей плюс/минус предел погрешности. Полученный результат называется *доверительным интервалом для разницы между долями двух совокупностей*.

Формула доверительного интервала для разницы между долями двух совокупностей выглядит так:

$$(\hat{p}_1 - \hat{p}_2) \pm Z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

где $\hat{p}_1$ и n_1 — доля и размер первой выборки, а $\hat{p}_2$ и n_2 — доля и размер второй выборки. Z — подходящее значение из стандартного нормального распределения, соответствующее желаемому доверительному уровню. (Подробнее о долях выборки см. главу 3, а Z-значения можно найти в главе 10, табл. 10.1.)

Чтобы вычислить ДИ для разницы между долями двух совокупностей, поступайте так.

1. **Определите доверительный уровень и найдите соответствующее Z-значение.**

   См. главу 10 (табл. 10.1).

2. **Найдите долю первой выборки $\hat{p}_1$, взяв общее количество в интересующей вас категории из первой выборки и разделив его на размер выборки n_1. Точно так же определите $\hat{p}_2$ для второй выборки.**

3. **Найдите разность между долями выборки ($\hat{p}_1 - \hat{p}_2$).**

4. **Умножьте $\hat{p}_1$ на $(1 - \hat{p}_1)$ и разделите на n_1. Умножьте $\hat{p}_2$ на $(1 - \hat{p}_2)$ и разделите на n_2. Сложите эти результаты и извлеките из суммы квадратный корень.**

5. **Умножьте Z на результат, полученный на этапе 4.**

   Это и будет предел погрешности.

6. **Чтобы получить ДИ, возьмите ($\hat{p}_1 - \hat{p}_2$) плюс/минус предел погрешности из этапа 5.**

   Нижняя граница ДИ — это ($\hat{p}_1 - \hat{p}_2$) минус предел погрешности, а верхняя граница ДИ равна ($\hat{p}_1 - \hat{p}_2$) плюс предел погрешности.

Выполняя любые вычисления, связанные с процентами выборок, вы должны использовать десятичные дроби. После того как вычисления закончены, преобразуйте результат в проценты, умножив его на 100. Чтобы избежать ошибок при округлении, всегда оставляйте хотя бы две цифры после запятой.

Предположим, вы работаете в торговой палате Лас-Вегаса и хотите с 95% уверенностью определить разницу между долей женщин, которые когда-либо посещали выступление двойника Элвиса Пресли, и долей мужчин, тоже ходивших на такой концерт. Это поможет вам выяснить, как лучше продвигать на рынок новые предложения в сфере развлечений.

Поскольку вам нужен 95% доверительный интервал, то Z-значение равно 1,96.

Предположим, вы сделали случайную выборку из 100 женщин, среди них оказались 53 женщины, бывавших на выступлении двойника Элвиса, значит, $\hat{p}_1 = 53/100 = 0{,}53$. Из случайной выборки в 110 мужчин 37 человек тоже ходили на подобные концерты, значит, $\hat{p}_2 = 37/100 = 0{,}34$.

Разница между этими долями выборок (женщины – мужчины) равна $0{,}53 - 0{,}34 = 0{,}19$.

Умножаем 0,53 на (1 – 0,53) и делим результат на 100, получается 0,2491/100 = 0,0025. Затем умножаем 0,34 на (1 – 0,34) и делим результат на 110, получается 0,2244/110 = 0,0020. В сумме это даст 0,0025 + 0,0020 = 0,0045. Корень квадратный из этой суммы равен 0,0671.

Умножив 1,96 на 0,0671, вы получите 0,13 или 13%, что является пределом погрешности.

Ваш 95% доверительный интервал для разности между долей женщин, бывавших на концерте двойника Элвиса, и долей мужчин, тоже посещавших такое выступление, составляет 0,19 или 19% (результат, который вы получили на этапе 3) плюс/минус 13%. Нижняя граница доверительного интервала равна 0,19 – 0,13 = 0,06 или 6%, верхняя граница доверительного интервала составляет 0,19 + 0,13 = 0,32 или 32%. Значит, вы с 95% уверенностью можете сказать, что на концерте двойника Элвиса бывал больший процент женщин, чем мужчин, и разница между этими процентами составляет от 6 до 32%, о чем говорят данные ваших выборок. Так как вы думаете, разве мужчины действительно призн*а*ются, что ходили на концерт двойника Элвиса? В связи с этим результаты могут оказаться немного смещенными. (Когда я в последний раз была в Вегасе, то мне показалось, что я действительно видела Элвиса — он водил такси в районе аэропорта...)

Обратите внимание, что результат ($\hat{p}_1 - \hat{p}_2$) может оказаться отрицательным. Например, если переставить местами мужчин и женщин, то разница между этими значениями была бы равна –0,19. Положительная разность означает, что значения в первой группе больше значений во второй, отрицательная разность подразумевает, что значения в первой группе меньше значений второй группы. Избежать отрицательных значений можно, если в качестве первой группы всегда выбирать ту, у которой значения больше.

Часть VI

Переходим к проверке гипотез

The 5th Wave — Рии Теннант

В этой части...

Многие статистики формулируют свои утверждения так: “Четыре из пяти опрошенных стоматологов советуют эту жевательную резинку”, “Наши подгузники впитывают на 25 процентов больше влаги, чем подгузники ведущей марки”. Но как проверить, действительно ли это так? Исследователи (те, которые знают, чем занимаются) прибегают к так называемой проверке гипотез.

В этой части вы узнаете основные критерии для проверки гипотезы, научитесь подбирать их, использовать и интерпретировать результаты (постоянно помня о том, что вы пытаетесь сделать утверждение обо всей совокупности, имея в своем распоряжении данные только об одной выборке). Кроме того, вы получите некоторые советы и познакомитесь с примерами самых распространенных критериев для проверки гипотезы.

Глава 14

Утверждения, проверки и выводы

В этой главе...

- Проверка утверждений других людей
- Использование статистики в качестве доказательства
- Изучение доказательств и принятие решений
- Осознание того, что вы могли ошибиться

Вы постоянно слышите утверждения, в которых используется статистика. В средствах массовой информации они встречаются в избытке.

- ✓ Двадцать пять процентов всех женщин в США страдают варикозным расширением вен. (Некоторые утверждения лучше не озвучивать, правда?)
- ✓ Количество экстази, которое потребляют подростки, за последние годы снизилось. В течение года это снижение составило от одной девятой до одной трети, в зависимости от того, в каком классе они учатся.
- ✓ Шестимесячный малыш спит в среднем от 14 до 15 часов в сутки. (Ага, это точно!)
- ✓ Приготовить пирог из полуфабриката такой-то марки можно всего за 5 минут.

Многие утверждения содержат числа, которые, на первый взгляд, были взяты "с потолка". В некоторых утверждениях сравнивается один продукт или группа с другими. Возможно, иногда вы задумываетесь, справедливы ли такие фразы, и делаете правильно. Не все утверждения кардинально влияют на жизнь (ну, например, какой вред в том, чтобы пользоваться мылом, очищающим не на 99,99%?), но некоторые все же имеют огромное значение — к примеру, какое лекарство от рака помогает лучше всего, какой минифургон самый безопасный или стоит ли налаживать выпуск определенных лекарств. Хотя многие высказывания подтверждаются серьезными научными (и статистически проверенными) исследованиями, но это относится не ко всем. В этой главе вы узнаете, как с помощью статистики определить, действительно ли высказанное утверждение, и докопаться до правды о том, какой именно процесс *должны* были использовать исследователи, чтобы подтвердить все высказанные утверждения.

Отвечаем на утверждения: некоторые советы

В наш век информации (и больших денег) очень многое зависит от того, можете ли вы доказать свои утверждения. Компании, заявляющие, что их продукция лучше ведущего бренда, должны быть в состоянии подтвердить это, иначе на них подадут в суд.

Нужно убедительно доказать, что лекарства, одобренные Управлением по контролю над продуктами питания и лекарственными средствами, не имеют побочных эффектов, представляющих угрозу для жизни. Чтобы избежать рекламаций и не потерять бизнес, производители должны гарантировать, что их продукция изготавливается в соответствии с требованиями.

В результате исследования тоже может быть сделано утверждение, имеющее жизненно важное значение, например, какой самый лучший метод лечения рака, каковы самые распространенные побочные эффекты определенного вида хирургического вмешательства, каков коэффициент выживаемости после определенного лечения и действительно ли новое экспериментальное лекарство увеличивает продолжительность жизни. Исследование, в ходе которого ученые попытаются ответить на эти вопросы, должно быть продуманным, чтобы можно было принять правильное решение (или, по крайней мере, самое информированное решение с точки зрения статистики). Если это правило не соблюдается, исследователи могут потерять репутацию, доверие и финансирование. (А иногда на них оказывают давление, чтобы были опубликованы фиктивные результаты, что приводит к появлению новых проблем.)

Каковы варианты?

Будучи потребителем такой информации и столкнувшись с каким-либо утверждением (например, "Наше мороженое выбрали 80% опрошенных"), вы можете поступить следующим образом.

- ✓ Автоматически поверить ему (или, наоборот, сразу же отвергнуть).
- ✓ Провести собственное исследование и подтвердить или опровергнуть утверждение.
- ✓ Копнуть глубже и попытаться найти больше информации, чтобы принять затем самостоятельное решение.

Слепо поверить результатам (или тут же отвергнуть их) — это неразумно. Единственная ситуация, когда это можно сделать — если источник уже зарекомендовал себя с положительной точки зрения или если результат не очень важен (ведь вы же не будете проверять *все* услышанные утверждении до единого). Подробнее об оставшихся двух вариантах мы поговорим в следующих разделах.

Сторонимся случаев из жизни

Второй вариант реакции на утверждение — самостоятельная проверка — это как раз то, что предпочитают многие организации, например, *The Gallup Organization*, которая проводит собственные опросы, Институт страхования безопасности на дорогах, который проверяет транспортные средства в условиях аварий и сообщает о своих результатах, *Consumer Reports*, проверяющий качество и стоимость продукции, а также Институт правильного домашнего хозяйства, проверяющий продукцию, прежде чем одобрить ее.

Самостоятельная проверка может быть эффективной, если ее провести правильно, когда точные, объективные данные собираются в ходе хорошо разработанного исследования (подробнее о разработке исследований см. главы 16 и 17).

Часто к такому подходу прибегают в профессиональной деятельности. Например, конкурент может высказать утверждение о своей продукции, которое, как вам кажется, не соответствует действительности и должно быть проверено. Или же вы уверены, что

ваша продукция лучше той, что производит конкурент, поэтому вы захотите сравнить их. Многие производители, кроме того, осуществляют собственный контроль качества (см. главу 19), поэтому проверяют свою продукцию, чтобы удостовериться, что она изготавливается в соответствии со всеми требованиями.

И все же хотя группы, имеющие ресурсы и знания для правильного проведения исследования, могут прибегнуть к подобному подходу для проверки утверждения, но если процесс провести неправильно, то его результаты все равно могут вводить в заблуждение.

Один из способов проверки утверждений, которым пользуются средства массовой информации, — это отправка людей "в поле" для самостоятельной проверки продукции. Это избитый и ненаучный (хотя и веселый) способ проверки гипотезы. Например, представим себе, что какая-то телепередача решила, что общественность должна знать, действительно ли на приготовление пирога из полуфабриката известной марки потребуется пять минут. Возможно, на самом деле приготовление займет больше или меньше времени. С точки зрения статистики, интересующая нас в данном случае переменная относится к числовым данным — время на приготовление, — а совокупность включает в себя все пироги, приготовленные с использованием материала известной марки. Изучаемый параметр — это *среднее* время приготовления всех пирогов этой марки. (Параметр — это число, которое подытоживает данные о совокупности и, как правило, как раз и является темой утверждения.) В данном случае утверждение состоит в том, что среднее время приготовления — 5 минут. Задача — проверить это утверждение. Сколько пирогов будет использовано в передаче? Догадались? (Всего один!)

Камеры снимают студию с разных углов, ведущий бодро говорит о том, как весело готовить этот пирог, какой он красивый, при этом не забывая следить за временем (ведь скоро будет перерыв на рекламу). В конце передачи говорится, что на выпекание пирога ушло, скажем, 5,5 минут, что очень близко к тому, что говорилось в утверждении, но не совсем. Напоследок прозвучит комментарий, что если украсить этот самый пирог такой-то марки плиткой Snickers, то вы сделаете правильный выбор (и это действительно так, кстати говоря).

Если бы для таких телепередач требовался статистик, который бы отвечал за статистические доказательства, я бы ухватилась за такую возможность. Главное, что я хочу донести до вас — результаты выборок отличаются (от человека к человеку, от пирога к пирогу), подробнее об этом см. главу 9. Измерение и понимание такой изменчивости — вот где начинается настоящая статистическая работа. Следовательно, чтобы результаты проверки какого-либо утверждения можно было считать надежными, они должны основываться на реальных данных (т.е. больше чем на одном наблюдении). Многие не понимают, что для того, чтобы правильно проверить утверждение, нужно делать выборку намного большего размера, чем 1 элемент (или даже 2 или 3), ведь результаты выборок варьируются.

Вы не можете (или, по крайней мере, не должны) делать серьезные выводы, исходя только из одного случая из жизни, которым на самом деле и является выборка размером в 1 элемент. В статистике выборка размером в 1 элемент просто не имеет смысла. Имея всего одно значение, невозможно измерить изменчивость (найдите в главе 5 формулу стандартного отклонения, и поймете, о чем идет речь). В этом и заключается проблема большинства телепередач, в которых проверяется какое-либо утверждение в ходе всего одной или двух попыток. Такая проверка ненаучна, и у зрителей складывается неверное представление о том, как нужно проверять гипотезу. И если сделать вывод о времени приготовления одного пирога из полуфабриката, не имея научного обоснования, земля от этого не перевернется, но подумайте, как часто в вашей жизни бывало, что мнение всего одного человека влияло на то, какое решение вы принимали.

Не доверяйте результатам исследований, основанным на крайне маленьких выборках, особенно на тех, выборка в которых состояла всего из 1 элемента. Например, если в ходе исследования проверяется одна упаковка мяса, одна детская игрушка или правильное приготовление одним фармацевтом одного рецепта в отдельно взятый день, держитесь от таких результатов подальше. Они могут стать интересными историями и указать на проблемы, застуживающие дополнительного изучения, но результаты подобных исследований совершенно ненаучны, поэтому не нужно делать на их основе никаких выводов.

Копаем глубже

Копать глубже, чтобы получить больше информации — вот как нужно реагировать на утверждения, которые представляют для вас какую-либо значимость. При этом вы сможете получить нужные сведения, которые пригодятся для проверки сомнений и принятия информированного решения.

Самая большая разница между статистически правильной проверкой утверждения и неподготовленной проверкой утверждения состоит в том, что при хорошей проверке используются данные, собранные научным, объективным способом, основанные на изучении случайных выборок достаточно большого размера для того, чтобы дать точную информацию. (Подробнее об этом см. главу 2.) Большинство научных исследований, в том числе в сфере медицины, фармацевтики, инженерного дела и управления, проводятся с использованием статистических опытов для проверки, подтверждения или опровержения различных утверждений. Поскольку вы — потребитель массы подобной информации, то вам нужно знать, что искать для проверки исследования, понимания его результатов и принятия собственных решений по поводу сделанных утверждений.

Возможно, вас интересует, насколько вы как потребитель защищены относительно утверждений, высказываемых исследователями. Правительство США контролирует многие исследования и производство (например, Управление по контролю над продуктами питания и лекарственными средствам контролирует исследование и распространение лекарств, министерство сельского хозяйства США контролирует производство продуктов питания и т.д.). Но в некоторых областях, таких как диетические добавки (витамины, травяные и минеральные добавки и т.д.), контроль недостаточен.

Как потребителю всех результатов, то и дело встречающихся в современном обществе, вам нужно иметь в своем распоряжении информацию, чтобы принимать качественные решения. Лучше всего сначала связаться с исследователем (или журналистом) и выяснить, подтверждены ли их утверждения научными данными. Если ответ утвердительный, тогда изучите описание и результаты исследований, а затем критически оцените полученную информацию (подробнее об этом см. главы 16 и 17).

Проверяем гипотезу

Проверка гипотезы — это статистическая процедура, предназначенная для проверки утверждения. Обычно утверждение касается параметра совокупности (т.е. одного числа, характеризующего всю совокупность). Поскольку параметры чаще всего неизвестны, то все хотят высказаться о том, какими могут быть эти значения. К примеру, утверждение

о том, что 25% (или 0,25) всех женщин страдают варикозным расширением вен — это утверждение о доле (это и есть *параметр*) всех женщин (т.е. *генеральной совокупности*), у которых наблюдается варикозное расширение вен (это *переменная*, потому что варикозное расширение вен может быть, а может и нет).

Неужели вы думаете, что кто-то точно знает, что процентное отношение женщин с варикозным расширением вен действительно составляет 25%? Нет, это всего лишь утверждение, а не констатация факта. Остерегайтесь таких высказываний.

Определяем, что нужно проверить

Говоря точнее, в утверждении о варикозном расширении вен параметр, т.е. доля совокупности (p) равен 0,25. (Это утверждение называется *нулевой гипотезой*.) Например, исходя из собственных наблюдений, вы можете предположить, что реальная доля женщин, у которых наблюдается варикозное расширение вен, ниже 0,25. Или же вы можете подумать, что из-за популярности высоких каблуков эта доля может быть больше, чем 0,25. Или, если вы просто задаете вопрос, действительно ли эта доля равна 0,25, то ваша альтернативная гипотеза звучит так: “Нет, она не равна 0,25” .

Помимо проверки гипотезы относительно категорийных переменных (наличие или отсутствие варикозного расширения вен — это категорийная переменная), вы можете проверить и гипотезы, касающиеся числовых переменных, например, средняя продолжительность времени, которое тратят на дорогу люди, работающие в Лос-Анджелесе, или их средний доход. В таких случаях изучаемый параметр — это среднее совокупности (которое обозначается μ). И снова утверждение заключается в том, что этот параметр равен определенному значению.

Можно также проверять гипотезы, которые касаются сразу нескольких параметров. Например, требуется сравнить средний доход или время в дороге для людей из двух или более крупных городов. Или вы решите проверить, есть ли какая-то взаимосвязь между временем в дороге и доходом. На все эти вопросы можно ответить с помощью проверки гипотез. Хотя в каждой ситуации детали различны, но общая идея остается неизменной. В этой главе приводится пример единичной выборки с определением среднего и долей (большие выборки). В главе 15 вы найдете дополнительные сведения о самых распространенных критериях для проверки гипотез.

Выдвижение гипотезы

Для проверки любого утверждения нужны две гипотезы. Первая называется *нулевой гипотезой* и обозначается H_0. Согласно нулевой гипотезе, параметр совокупности *равен* заявленному значению. Например, если утверждается, что среднее время для приготовления пирога из полуфабриката известной марки равно пяти минутам, то в этом случае нулевую гипотезу статистики записали бы так: $H_0 : \mu = 5$.

Какова альтернатива?

Прежде чем переходить непосредственно к проверке, необходимо сформулировать две гипотезы — одна из них нулевая. Но если окажется, что нулевая гипотеза ошибочна, каким будет альтернативный вариант? На самом деле существует три разновидности второй (или альтернативной) гипотезы, которая обозначается как H_a.

Вот эти три варианта и их обозначения.

- ✓ Параметр совокупности не равен заявленному значению ($H_a : \mu \neq 5$).
- ✓ Параметр совокупности больше заявленного значения ($H_a : \mu > 5$).
- ✓ Параметр совокупности меньше заявленного значения ($H_a : \mu < 5$).

Выбор альтернативной гипотезы для проверки зависит от того, к какому выводу вы хотите прийти при условии, что у вас достаточно доказательств для опровержения нулевой гипотезы (утверждения).

Например, если вы хотите проверить, права ли компания, когда утверждает, что для приготовления ее пирога нужно 5 минут, и стремитесь выяснить, больше или меньше этого значения реальное время, которое нужно для выпечки, то вам нужно выбрать вариант "не равно". Тогда гипотеза для проверки будет такой: $H_0 : \mu = 5$ или $H_a : \mu \neq 5$.

Если вы хотите проверить, не будет ли реальное время больше того, о котором сказано в утверждении (т.е. не прибегает ли компания к ложной рекламе), тогда вы выбираете вариант "больше чем" и работаете с такими двумя гипотезами: $H_0 : \mu = 5$ или $H_a : \mu > 5$.

Наконец, предположим, вы работаете в компании, продвигающей этот пирог на рынок, и считаете, что его можно приготовить меньше, чем за 5 минут (компания сможет использовать это в своей маркетинговой программе). Тогда вам нужен вариант "меньше чем", а две рабочих гипотезы выглядят так: $H_0 : \mu = 5$ или $H_a : \mu < 5$.

Различение гипотез

Как узнать, какую гипотезу назвать H_0, а какую — H_a? Как правило, нулевая гипотеза свидетельствует о том, будто не происходит ничего нового, предыдущий результат остается таким же, каким он был прежде, или у групп остается прежнее среднее значение (т.е. разница между ними равна нулю). В целом вы полагаете, что высказанные утверждения истинны, пока не было доказано обратное.

В каком-то смысле проверка гипотезы напоминает судебное разбирательство. В суде H_0 — это вердикт о невиновности обвиняемого, а H_a — вердикт о виновности. В случае с судебным разбирательством вы полагаете, что обвиняемый невиновен до тех пор, пока обвинение не сможет убедительно доказать, что он *виновен*. Присяжные говорят, что представленные доказательства не вызывают сомнений, значит, они отвергают H_0 "не виновен" в пользу H_a "виновен".

В общем, проверяя гипотезу, вы выдвигаете H_0 и H_a и полагаете, что истинно утверждение H_0, пока собранные свидетельства (т.е. ваши данные и статистические показатели) не докажут обратное. И если вы соберете достаточно доказательств против H_0, значит, вы отвергаете H_0 в пользу H_a. Исследователь обязан собрать доказательства против H_0 до того, как отклонить ее. (Вот почему H_0 часто называют исследовательской гипотезой, потому что H_a — это гипотеза, которую исследователь стремится подтвердить.) Если предпочтение отдается гипотезе H_a, то исследователь может сказать, что он получил *статистически значимый* результат, т.е. его результаты опровергают предыдущее утверждение, а значит, происходит что-то новое и непривычное.

Очень часто люди проводят проверку гипотез, потому что намерены доказать, что H_0 ошибочна, и поддерживают альтернативную гипотезу. (Зачем проводить исследование, только чтобы доказать то, что остается неизменным?) Результаты, о которых вам говорят в средствах массовой информации, как

правило, относятся к тем, что смогли доказать ошибочность гипотезы H_0. Именно это может стать сенсацией. Во многих случаях это хорошо, потому что исследователи и производитель должны прилагать все усилия для того, чтобы избежать отрицательной рекламы, к которой приведет отзыв продукции, судебный иск или правительственное расследование. Ведь если какое-либо из их утверждений (H_0) будет опровергнуто человеком, проводившим независимую проверку гипотезы, то исследователей и производителей признают виновными в ложной рекламе или ложных утверждениях, а это не есть хорошо.

Сбор доказательств: выборка

После того, как гипотезы выдвинуты, мы переходим к следующему этапу, на котором необходимо собрать доказательства и определить, имеют ли они отношение к утверждению, которое содержится в гипотезе H_0. Помните, что утверждение делается относительно генеральной совокупности, но проверить всю совокупность невозможно. Лучшее, что обычно можно сделать — это провести выборку. Как и в любой другой ситуации при сборе статистических данных, качество данных имеет чрезвычайное значение. (Множество примеров неправильной статистики приводится в главе 2.)

Хорошие данные начинаются с хорошей выборки. Приступая к выборке, нужно помнить о двух вещах: избегать смещений и быть точным. Чтобы избежать смещений, сделайте случайную выборку (т.е. у любого элемента совокупности должен быть шанс быть выбранным) достаточно большого размера, чтобы результаты были точными (см. главу 3).

Сбор доказательств: статистика

После того, как выборка сделана, наступает этап правильного решения математических задач. В вашей нулевой гипотезе содержится утверждение о том, что собой представляет какой-либо параметр совокупности (к примеру, доля всех женщин с варикозным расширением вен или средний расход бензина у американского грузовика). На языке статистики это значит, что собранные данные оценивают изучаемую переменную, а те статистические показатели, которые вы высчитываете, будут содержать и статистику выборки, наиболее точно оценивающую параметр совокупности. Другими словами, если вы проверяете утверждение о доле женщин с варикозным расширением вен, то вам нужно вычислить долю женщин с этим заболеванием в своей выборке. Если вы проверяете утверждение о среднем расходе бензина у американского грузовика, то интересующий вас статистический показатель — это средний расход бензина у грузовиков в вашей выборке. (Подробнее о вычислении статистики см. главу 5.)

Стандартизация доказательств: статистика критерия

После того как уже получены все статистические показатели выборки, может показаться, что аналитическая часть закончена и можно переходить к формулированию выводов — но это не так. Проблема в том, что невозможно представить результаты в перспективе, если рассматривать их в привычных единицах. Это объясняется тем, что ваши результаты основаны только на одной выборке, а результаты разных выборок будут отличаться, как вам уже известно. Такую изменчивость нужно учитывать, иначе сделанные выводы могут оказаться совершенно неверными. (Насколько отличаются результаты выборок? Изменчивость выборки измеряется стандартными ошибками. Подробнее этот вопрос рассмотрен в главе 9.)

Предположим, было сделано утверждение, что 25% всех женщин страдают варикозным расширением вен, а в вашей выборке из 100 женщин это заболевание обнаружено только у 20%. Стандартная ошибка для процентного отношения вашей выборки равна 4% (согласно формулам в главе 9), что означает, что ваши результаты будут варьироваться вдвое сильнее, т.е. на 8%, о чем гласит эмпирическое правило (см. главу 10). Значит, разница в 5% между утверждением и результатом вашей выборки (25% – 20% = 5%) не такая уж и большая. Она соответствует расстоянию меньше чем в 2 стандартных ошибки от утверждения. Следовательно, вы принимаете утверждение H_0, потому что собранные вами данные не опровергли его.

Однако представим, что процентное отношение основывается на выборке не из 100, а из 1000 женщин. Это снижает степень вариации результатов, потому что у вас теперь больше информации. Стандартная ошибка теперь равна 0,012 или 1,2%, а предел погрешности вдвое больше, т.е. 2,4% в любую сторону. Значит, разница в 5% между результатами вашей выборки (20%) и утверждением (25%) становится серьезнее. Это *намного* больше, чем 2 стандартных ошибки от утверждения. Ваши результаты, основанные на выборке в 1000 человек, не должны так сильно отличаться от утверждения. Значит, какой вывод можно сделать? Утверждение (H_0) считается ложным, потому что ваши данные его не подтверждают.

Количество стандартных ошибок, на которые статистический показатель в ту или другу сторону отличается от среднего, называется *нормированным параметром* (см. главу 8). Определяя нормированный параметр, вы от вычисленного статистического показателя отнимаете среднее значение и делите результат на стандартную ошибку. В случае с проверкой гипотезы в качестве среднего значения используется значение из H_0. (Это потому, что вы полагаете, будто H_0 истинно, пока не будет доказано обратное.) Такой нормированный вариант статистического показателя называется *статистикой критерия* и является главным компонентом проверки гипотезы. (В главе 15 представлены формулы самых распространенных критериев для проверки гипотезы.)

Общий принцип преобразования статистического показателя в статистику критерия (нормированный параметр) для средних значений/долей таков:

1. **Вычтите из своего статистического показателя заявленное значение (которое представлено в виде H_0).**
2. **Разделите результат на стандартную ошибку статистического показателя (см. главы 9 и 10).**

Статистика критерия представляет собой расстояние между реальными результатами выборки и заявленным значением совокупности и измеряется в количестве стандартных ошибок. В случае с отдельным средним или долей эти нормированные расстояния должны иметь стандартное нормальное распределение, если была сделана выборка достаточного размера (см. главы 8 и 9). Значит, для интерпретации статистики критерия в таких случаях вам нужно проверить, попадает ли она в стандартное нормальное распределение (*Z*-распределение).

Хотя статистический показатель выборки никогда не будет точно равен значению совокупности, но вы предполагаете, что если H_0 истинно, то оба этих значения будут близки. Это значит, что если расстояние между утверждением и статистическим показателем выборки небольшое (измеряемое стандартными ошибками), то ваша выборка близка к утверждению, и данные подтверждают H_0. Но если расстояние все увеличивается, то данные все меньше поддерживают H_0. В определенной точке вы, исходя из своих данных, будете вынуждены отказаться от H_0 и выбрать H_a. Когда именно это произойдет? Об этом читайте в следующем разделе.

Изучение доказательств и принятие решений: p-значения

Чтобы проверить истинность утверждения, вы анализируете статистику критерия, полученную на основе своей выборки, и определяете, подтверждает ли она высказанное утверждение. А как это определить? Чаще всего — проверив, где находится ваша статистика критерия на стандартном нормальном распределении (*Z*-распределении), см. главу 9. У *Z*-распределения среднее равно 0, а стандартное отклонение равно 1. Если ваша статистика критерия ближе к 0 или хотя бы находится в том диапазоне, где должны находиться большинство результатов, тогда вы говорите, что утверждение (H_0), возможно, истинно, и результаты выборки это подтверждают. Если ваша статистика критерия находится на склонах стандартного нормального распределения, тогда вы заявляете, что результаты вашей выборки вряд ли могли оказаться здесь в таком распределении, значит, они не подтверждают утверждение (H_0). Но какое расстояние от 0 можно считать как "слишком много"? Пока вы рассматриваете выборку достаточно большого размера, то знаете, что, согласно центральной предельной теореме (см. главу 10), ваша статистика критерия находится где-то на стандартном нормальном распределении. Если нулевая гипотеза справедлива, то большинство (95%) выборок будут иметь статистику критерия, которая находится в пределах примерно 2 стандартных ошибок от утверждения. Если H_a — это вариант "не равно", а статистика критерия не попадает в этот диапазон, то любое утверждение H_0 будет опровергнуто (см. рис. 14.1).

Если альтернативная гипотеза — это вариант "не равно", то вы опровергаете H_0, только если статистика критерия оказывается на левом склоне распределения. Точно так же, если H_a — это вариант "больше чем", то вы опровергаете H_0, только если статистика критерия попадает на правый склон.

Рис. 14.1. Статистики критерия и ваше решение

Основы *p*-значений

Вывод можно сформулировать точнее, если отметить, на каком расстоянии от стандартного нормального распространения находится статистика критерия, чтобы все могли знать, какие были получены результаты, т.е. насколько убедительны доказательства против утверждения. Сделать это можно, если поместить статистику критерия на стандартное нормальное распределение (*Z*-распределение) и определить вероятность того, что эта статистика могла иметь такое значение или быть больше (в том же направлении).

Для этого воспользуемся табл. 8.1 (см. главу 8). P-значение указывает, какова вероятность того, что вы бы получили имеющиеся результаты выборки, если бы нулевая гипотеза была истинной. Чем дальше расположена статистика критерия на склонах стандартного нормального распределения, тем меньше будет *p*-значение, а значит, тем больше у вас доказательств против нулевой теории.

Все *p*-значения — это вероятность от 0 до 1.

Чтобы определить *p*-значение для вашей статистики критерия (среднее/доли, большие выборки), выполните следующие действия:

1. **Определите местоположение статистики критерия на стандартном нормальном распределении (см. табл. 8.1 в главе 8).**
2. **Найдите процентную вероятность того, что она может иметь такое же или большее значение в том же направлении.**

   а) Если H_a — это вариант "меньше чем", найдите процентиль из табл. 8.1 в главе 8, который соответствует вашей статистике критерия.

   б) Если H_a — вариант "больше чем", найдите процентиль из табл. 8.1 в главе 8, соответствующий вашей статистике критерия, а затем отнимите этот результат из 100%. (В этом случае вас интересует процентное отношение вправо от статистики критерия, а процентили показывают процентное отношение влево. См. главу 5.)
3. **Если (и только если) H_a — это вариант "не равно", умножьте этот процент на 2.**

   Таким образом будут учтены варианты как "больше чем", так и "меньше чем".
4. **Преобразуйте процентное отношение в вероятность, разделив его на 100 или переместив запятую в десятичной дроби на два знака влево.**

Для интерпретации *p*-значения выполните следующие действия.

- ✓ Если *p*-значение небольшое (как правило, меньше 0,05), опровергайте H_0. Ваши данные не поддерживают H_0 и не вызывают сомнений.
- ✓ Если *p*-значение большое (как правило, больше 0,05), вы не можете опровергнуть H_0. У вас нет для этого достаточных доказательств.
- ✓ Если *p*-значение находится на границе между отклонением и принятием или рядом с ней, такие результаты маргинальны. (Они могут сдвинуться в любую сторону.)

Обычно статистики принимают H_0, если только доказательства против этого утверждения не вызывают сомнений, как и в суде. Какая вероятность соответствует моменту перемены решения? Она может быть довольно спорной (понятие "маленькое *p*-значение" у каждого человека свое). По мнению большинства статистиков, если *p*-значение меньше 0,05, учитывая собранные данные, нужно опровергнуть H_0 и выбрать H_a. Могут быть и более жесткие критерии, например, 0,01, тогда для опровержения H_0 потребуется больше доказательств. Каждый читатель принимает собственное решение. Вот почему исследователи должны указывать *p*-значения, а не просто свои решения, чтобы люди могли самостоятельно сделать выводы исходя из собственных представлений. Например, если ваше *p*-значение равно 0,026, и при этом вы проверяете $H_0 : p = 0{,}25$ или $H_a : p < 0{,}25$, как это было в примере с варикозным расширением вен, то читатель, для

которого пороговая точка равна 0,05, мог бы сделать вывод, что H_0 ошибочно, потому что *p*-значение (0,026) меньше чем 0,05. А вот для читателя, установившего для себя пороговую точку в 0,01, было бы недостаточно доказательств (исходя из вашей выборки) для отклонения H_0, потому что *p*-значение, равное 0,026, больше чем 0,01.

Осторожно: интерпретации могут отличаться!

Многие действительно предпочитают заранее определять для себя такую пороговую точку, чтобы проверить гипотезу. Это называется *уровнем значимости* (α). Обычные значения α — это 0,05 или 0,01. Вот как в таком случае трактуются результаты.

- ✓ Если *p*-значение больше или равно α, то H_0 принимается.
- ✓ Если *p*-значение меньше α, то H_0 опровергается.
- ✓ *p*-значения на границе (очень близко к α) считаются маргинальными.

Некоторые исследователи не устанавливают пороговой точки, а просто приводят полученные *p*-значения и интерпретируют результаты, исходя из размера *p*-значений. Другими словами, получаем следующее.

- ✓ Если *p*-значение меньше 0,01 (очень мало), результаты считаются крайне значимыми со статистической точки зрения — H_0 отклоняется.
- ✓ Если *p*-значение находится между 0,05 и 0,01 (но не близко к 0,05), результаты считаются статистически значимыми — H_0 отклоняется.
- ✓ Если *p*-значение близко к 0,05, результаты считаются маргинально значимыми — решение принять невозможно.
- ✓ Если *p*-значение больше (но не очень близко к 0,05), результаты считаются незначимыми — H_0 принимается.

Если вам сообщили о результатах, которые считаются статистически значимыми, выясните *p*-значение и принимайте решение сами. Пороговые точки и окончательные решения у всех исследователей разные.

Все могут быть неправы: ошибки при проверке

Итак, вы твердо решили принимать или опровергать H_0. Теперь вам предстоит испытать последствия своего решения, т.е. выяснить, как отреагируют на него люди.

- ✓ Если вы делаете вывод, что утверждение неверно, но на самом деле оно окажется *истинным*, последуют ли за этим судебные иски, штраф, ненужные изменения в продукции или бойкот со стороны потребителей, чего никак не должно произойти?
- ✓ Если вы решаете, что утверждение истинно, но на самом деле это не так, что произойдет в таком случае? Будет ли продукция и дальше производиться по-прежнему? Будет ли принят новый закон, предприняты новые меры, поскольку вы заявили, что все в порядке?

Каждая проверка гипотезы оказывает влияние (иначе зачем ее проводить?).

Значит, последствия будет иметь любое решение. А вы ведь можете ошибаться! В такой ситуации уместным будет девиз сериала "Секретные материалы": "Истина где-то рядом". Но дело в том, что наверняка неизвестно, в чем заключается истина, вот почему, в первую очередь, и проводится проверка гипотез.

Ложная тревога: ошибки первого рода

Предположим, компания утверждает, что среднее время доставки ее продукции составляет два дня, а группа потребителей проверяет это утверждение и делает вывод, что оно ложно. Оказывается, что среднее время доставки — больше двух дней. Это серьезное дело. Если группа может доказать свои выводы, тогда поступит правильно, если проинформирует общественность о ложной рекламе. Но что, если группа ошибается? Даже если исследование хорошо разработано, собраны качественные данные и проведен правильный анализ, группа все равно может ошибаться.

Почему? Потому что ее выводы основаны на выборке доставок, а не на всей совокупности, ведь результаты разных выборок отличаются друг от друга (см. главу 9). Допустим, ваша статистика критерия находится на склоне стандартного нормального распределения, тогда, если утверждение истинно, такие результаты необычны, потому что вы ожидали, что они будут намного ближе к середине стандартного нормального распределения (*Z*-распределения). Но только потому, что результаты выборки необычны, это не означает, что они невозможны. *P*-значение, равное 0,04, говорит о том, что, если утверждение истинно, то вероятность получить конкретную статистику критерия (на склоне стандартного нормального распределения) равна 4% (меньше 5%). Поэтому вы отклоняете H_0, ведь шансы так малы. Но ведь шансы есть шансы!

Возможно, ваша выборка, хотя и сделана случайно, просто оказалась одной из тех атипичных выборок, результаты которых не попадают в стандартное нормальное распределение. Значит, H_0 может быть истинным, но полученные результаты заставили вас сделать другой вывод. Часто ли такое бывает? В 5% случаев (ваша пороговая точка для опровержения H_0).

Опровержение H_0 в тех случаях, когда этого делать не следует, называется *ошибкой первого рода*, хотя такое название мало о чем говорит. Лучше назовем ошибки первого типа *ложной тревогой*. В случае с доставкой, если потребительская группа совершает ошибку первого рода, опровергая утверждение компании, то она поднимает ложную тревогу. А что в результате? Недовольная и обиженная компания по доставке.

Упущения в расследовании: ошибки второго рода

С другой стороны, представим себе, что компания действительно опаздывает с доставкой. Кто может сказать, что потребительская группа способна это обнаружить? Если вместо 2 дней реальное время доставки составляет 2,1 дня, то заметить такую разницу будет довольно сложно. Если же время доставки равно 3 дням, сравнительно небольшая выборка может показать, что что-то не в порядке. Все зависит от значений посередине, например, 2,5 дня. Если H_0 действительно ложно, нужно это выяснить и опровергнуть утверждение. Если вы не опровергаете H_0, когда должны это сделать, это называется *ошибкой второго рода*, или назовем ее еще *упущением в расследовании*.

В том, чтобы суметь определить ситуации, когда H_0 ложно, и избежать ошибок второго рода, важнейшую роль играет размер выборки. Чем больше у вас информации, тем менее изменчивыми будут результаты (см. главу 8), а значит, тем больше у вас шансов обнаружить проблемы, которые присущи утверждению.

Такая способность заметить, когда H_0 действительно неверно, называется *мощностью критерия*. Мощность — это довольно сложный вопрос, но вам важно понять, что чем больше размер выборки, тем мощнее критерий. Допустить ошибки второго рода, имея мощный критерий, почти невозможно.

Возьмем любой статистически значимый результат величиной с крупинку соли, независимо от качества проведенного исследования. Каким бы ни было принятое решение, оно может оказаться ошибочным. Но если исследование разработано правильно (подробнее опросы рассматриваются в главе 16, об экспериментах см. главу 17), то вероятность ошибки значительно снижается.

Статистики советуют принять меры для того, чтобы свести к минимуму вероятность допустить ошибки первого или второго рода.

- ✓ Установить низкую пороговую вероятность для отклонения H_0 (уровень значимости α) (например, 5 или 1%), тем самым уменьшив шансы поднять ложную тревогу (т.е. свести к минимуму ошибки первого рода).
- ✓ Сделать выборку большого размера, чтобы не упустить расхождения и ответвления, которые действительно существуют (т.е. свести к минимуму ошибки второго рода).

Делаем выводы об их выводах

Даже если вы никогда сами не проводили проверку гипотезы, знание того, как их нужно проводить, отточит ваше умение критически смотреть на вещи. После завершения проверки исследователи публикуют результаты и представляют средствам массовой информации пресс-релизы, в которых говорится об их находках. Здесь тоже нужно быть внимательным. Одни исследователи добросовестно относятся к изложению своих результатов и указывают на недостатки, тогда как другие обращаются со своими выводами намного вольнее (намеренно или нет — это уже другой вопрос).

Этапы проверки гипотезы: общая картина

Любой критерий для проверки гипотезы содержит несколько этапов и включает в себя ряд процедур. В этом разделе вы в общих чертах узнаете о них. Подробнее о самых распространенных проверках гипотез, в том числе о проверке утверждений относительно одного параметра совокупности, а также связанных со сравнением двух совокупностей, см. главу 15.

Вспоминаем основные этапы проверки гипотезы (средние/доли, большие выборки)

Ниже кратко изложены вычисления, которые нужно сделать при проверке гипотезы. (Конкретные формулы для определения статистики критерия, необходимые для любой распространенной проверки гипотезы, вы найдете в главе 15.)

1. **Сформулируйте нулевую и альтернативную гипотезу.**

   а) Нулевая гипотеза H_0 гласит, что параметр совокупности равен какому-то определенному значению.

   б) Существует три возможных варианта альтернативной гипотезы. Выберите тот, который подходит для данного случая, если данные *не* подтверждают H_0.

      i. H_a: Параметр совокупности *не равен* ($\neq$) заявленному значению.

      ii. H_a: Параметр совокупности *меньше* ($<$) заявленного значения.

      iii. H_a: Параметр совокупности *больше* ($>$) заявленного значения.

2. **Сделайте случайную выборку элементов совокупности и определите статистические параметры выборки.**

   Так вы сможете лучше оценить параметр совокупности (см. главу 4).

3. **Преобразуйте статистический параметр выборки в статистику критерия, превратив его в нормированный параметр (все формулы статистик критерия представлены в главе 15):**

   а) Вычтите число, о котором идет речь в нулевой гипотезе, из своего статистического параметра выборки. Это расстояние между утверждением и вашим результатом.

   б) Разделите это расстояние на стандартную ошибку статистического показателя (подробнее о стандартных ошибках см. главу 10). Таким образом расстояние будет преобразовано в нормированные единицы.

4. **Найдите *p*-значение для своей статистики критерия.**

   а) Определите процентные шансы того, что это значение может находиться в том же направлении:

      i. Если H_a представляет собой вариант "меньше чем", найдите процентиль из табл. 8.1 в главе 8, который соответствует вашей статистике критерия.

      ii. Если H_a — это вариант "больше чем", найдите процентиль из табл. 8.1 (см. главу 8), соответствующий вашей статистике критерия, затем вычтите его из 100%. (Так вы найдете процент вправо от статистики критерия.)

   б) Тогда (и только тогда), когда H_a представляет собой вариант "не равно", умножьте этот процент на 2.

   в) Преобразуйте процент в вероятность, разделив его на 100 или передвинув запятую в десятичной дроби на два знака влево. Это будет ваше *p*-значение.

5. **Изучите полученное *p*-значение и сделайте вывод.**

   а) Меньшие *p*-значения свидетельствуют против H_0. Делаем вывод, что H_0 ложно (другими словами, опровергаем утверждение).

   б) Большие *p*-значения свидетельствуют в пользу H_0. Таким образом, опровергнуть H_0 нельзя. Ваша выборка подтверждает утверждение. Какова пороговая точка для определения величины *p*-значения? Большинство людей считает, что для этого вполне подходит 0,05. Таким образом, *p*-значения меньше 0,05 говорят, что истинность H_0 вызывает серьезные сомнения. Ваша пороговая точка называется уровнем значимости (α).

Если сравниваются две совокупности, большинство исследователей стремятся сравнить группы согласно определенному параметру, например, средний вес мужчин по сравнению со средним весом женщин или доля женщин, выступающих против определенного вопроса, по сравнению с такой же долей мужчин. В таком случае гипотезы формулируются таким образом, чтобы можно было изучить разницу между средними значениями или долями, а нулевая гипотеза гласит, что такая разница равна нулю (т.е. в группах одинаковое среднее значение или доля). В главе 15 изложены формулы и примеры таких проверок гипотез как для больших, так и для малых выборок.

Другие критерии для проверки гипотез

В мире исследований проводятся различные проверки гипотез. Самые распространенные описаны в главе 15 (наряду с простыми формулами, подробным объяснением и примерами). Но существует огромное количество разновидностей проверок, результаты их сообщаются ежедневно — многие озвучиваются, приводятся в пресс-релизах, вечерних выпусках новостей и в Интернете. Хотя исследователи проверяют гипотезу разными способами, основные идеи (такие как *p*-значения и интерпретация их результатов) остаются неизменными.

Самый важный элемент, общий для всех разновидностей проверок гипотез, — это *p*-значение. У всех *p*-значений одна интерпретация, независимо от особенностей проверки. Поэтому всякий раз, когда вам встречается *p*-значение, будьте уверены, что маленькое *p*-значение означает, что исследователь получил "статистически значимый" результат, т.е. нулевая гипотеза была опровергнута.

Независимо от типа использованной проверки гипотезы, любые сделанные выводы зависят от того, правильно ли происходил сбор данных и как тщательно проводился анализ. Даже при самых удачных условиях данные могут оказаться нерепрезентативными по чистой случайности или же истину может оказаться слишком сложно установить, следовательно, можно принять неверное решение. Но именно поэтому статистика так увлекательна — вы не знаете, правильно ли то, что вы делаете, но вы наверняка знаете, что вы все делаете правильно. Вы находите в этом хоть какой-то смысл?

Работа с меньшими выборками: *t*-распределение

Если вы работаете со средними значениями/долями, а выборка имеет небольшой размер (меньше 30 элементов), то в вашем распоряжении будет меньше информации, на основе которой можно делать выводы. Еще один недостаток такой ситуации заключается в том, что вы не можете полагаться на стандартное нормальное распределение (*Z*-распределение) при сравнении статистик критерия, потому что здесь еще не действует центральная предельная теорема. (Центральная предельная теорема может работать, если размер выборки достаточно большой и средние результаты образуют колоколообразную кривую. Подробнее об этом см. главу 8.) Вы уже знаете, что нельзя принимать в расчет результаты, основанные на выборках очень маленького размера (особенно тех, которые состоят всего из 1 элемента). Так что же делать в таких промежуточных ситуациях, когда выборка не слишком маленькая, чтобы от нее отмахнуться, но и не слишком

большая, чтобы основывать свои доказательства на стандартном нормальном распределении? В таком случае нужно использовать другой тип распределения, который называется *t-распределением*, или распределением Стьюдента. (Возможно, вы уже слышали о *t-критерии*, или *критерии Стьюдента,* проверки гипотез. Именно отсюда и происходит это понятие.)

По сути, *t*-распределение — это более краткая, утолщенная версия стандартного нормального распределения (*Z*-распределения). Идея заключается в том, что вы должны поплатиться за то, что обладаете меньшим количеством информации, а ваше наказание — это распределение с утолщенными склонами. Чтобы попасть в цель (т.е. в те магические 5%, при которых H_0 опровергается), имея меньшую выборку, вам придется собрать еще больше убедительных доказательств, чем обычно. На рис. 14.2 сравниваются стандартное нормальное распределение (*Z*-распределение) и *t*-распределение.

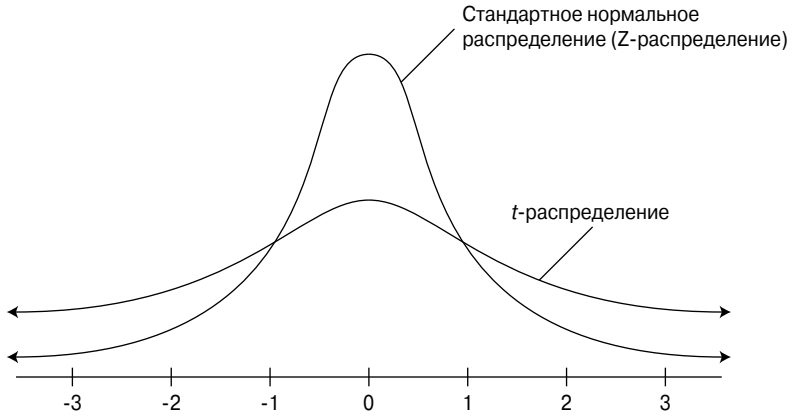


Рис. 14.2. Сравнение стандартного нормального распределения (Z) и распределения Стьюдента (t)

У выборки любого размера есть свое распределение Стьюдента. Поэтому наказание за меньший размер выборки, например, 5, будет строже, чем наказание за больший размер выборки, скажем, 10 или 20. У выборок меньшего размера более короткое, утолщенное *t*-распределение, чем у больших выборок. И, как можно было ожидать, чем больше размер выборки, тем больше *t*-распределение похоже на стандартное нормальное распределение (*Z*-распределение). Тот момент, когда они становятся максимально похожи, и соответствует выборке в 30 элементов. На рис. 14.3 показано, как может выглядеть *t*-распределение для выборок разного размера и как они соотносятся со стандартным нормальным распределением (*Z*-распределением).

Для любого *t*-распределения характерно то, что статистики называют *степенью свободы*. Если вы проверяете среднее одной совокупности с размером выборки *n*, то степень свободы для соответствующего распределения Стьюдента равна $n - 1$. Например, если размер вашей выборки равен 10, то для определения статистики критерия вы используете *t*-распределение со степенью свободы в $10 - 1$ или 9, что обозначается как t_9, а не *Z*-распределение. (Для любой проверки с использованием *t*-распределения число степеней свободы будет указано в формуле, касающейся размера выборки. Подробнее см. главу 15.)

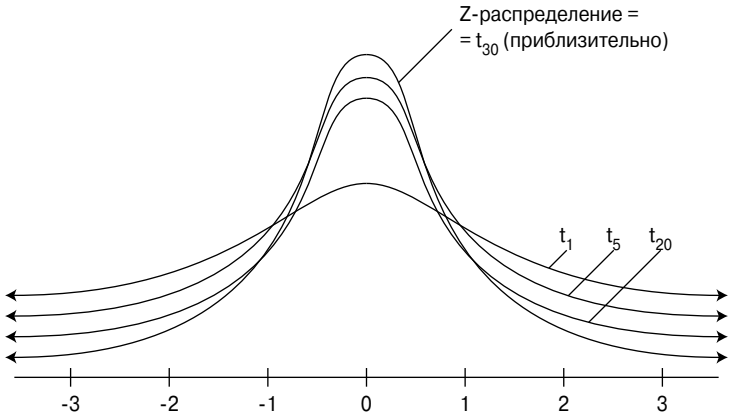


Рис. 14.3. Распределение Стьюдента для выборок разного размера

Вы используете выборку меньшего размера, и *t*-распределение наказывает вас за это. Какое наказание? Большее *p*-значение, чем то, что соответствовало бы такой же статистике критерия в случае со стандартным нормальным распределением. Это объясняется утолщенными склонами распределения Стьюдента. Под статисткой критерия на более плоском склоне *Z*-распределения остается меньше места. А вот в случае с *t*-распределением под той же статистикой критерия остается больше места, что и отражает *p*-значение. Большее *p*-значение говорит о том, что у вас меньше шансов опровергнуть H_0. Если вы собрали меньше данных, значит, ваши доказательства должны быть намного убедительнее, поэтому *p*-значения в полной мере выполняют возложенную на них задачу!

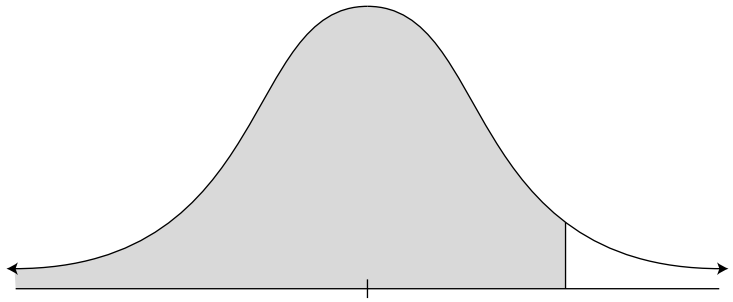
Поскольку у выборки любого размера есть свое распределение Стьюдента с собственной *t*-таблицей, в которой собраны *p*-значения, статистики разработали одну сокращенную таблицу, с помощью которой можно в общих чертах охарактеризовать полученные результаты (см. табл. 14.2). Компьютеры также могут подсказать вам точное *p*-значение для выборки любого размера.

Предположим, имеется выборка из 10 элементов, статистика критерия (которая называется *t-значением*) равна 2,5, а в качестве альтернативной гипотезы H_a выбран вариант "больше чем". Поскольку размер выборки равен 10, то для определения *p*-значения вы используете *t*-распределение с числом степеней свободы $10 - 1 = 9$. Это значит, что вы должны найти ту строку в *t*-таблице (табл. 14.2), которая соответствует столбику с числом степеней свободы 9. Ваша статистика критерия (2,5) находится между двумя значениями: 2,262 (97,5-й процентиль) и 2,821 (98-й процентиль).

Каково *p*-значение? Что-то между $100\% - 97{,}5\% = 2{,}5\% = 0{,}025$ и $100\% - 98\% = 2\% = 0{,}02$. (Не забывайте, что для варианта "больше чем" нужно процентиль вычитать из 100%.) Точно неизвестно, чему равно *p*-значение, но поскольку и 2%, и 2,5% — это меньше, чем обычная пороговая точка в 5%, то вы опровергаете H_0.

Обратите внимание, что для варианта "меньше чем" в качестве альтернативной гипотезы ваша статистика критерия была бы отрицательным числом (влево от 0 на *t*-распределении). В таком случае для определения *p*-значения вам нужно найти процент ниже, т.е. влево от статистики критерия. Но в табл. 14.2 отрицательные статистики критерия отсутствуют. Не волнуйтесь! Процентное отношение ниже (влево) от отрицательного *t*-значения — это то же самое, что и процентное отношение вправо (выше) от положительного *t*-значения, что объясняется симметрией.

Таблица 14.2. Распределение Стьюдента



Число степеней свободы	90-й процентиль	95-й процентиль	97-й процентиль	98-й процентиль	99-й процентиль
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
60	1,296	1,671	2,000	2,390	2,660
Z-значения	1,282	1,645	1,960	2,326	2,576

Значит, чтобы найти *p*-значение для отрицательной статистики критерия, найдите положительный эквивалент своей статистики критерия в табл. 14.2, определите соответствующий процентиль и вычтите его из 100%.

Например, если ваша статистика критерия равна –2,5, а число степеней свободы — 9, найдите в табл. 14.2 значение +2,5 (это значение находится между 97,5-м процентилем и 98-м процентилем). Вычтя эти числа из 100%, вы устан*о*вите, что *p*-значение представляет собой что-то среднее между 2 и 2,5%. (Обратите внимание, что такой подход к отрицательным числам отличается от подхода, представленного в табл. 8.1, которая составлялась иначе.)

Если в качестве альтернативной гипотезы (H_a) вы выбрали вариант “не равно”, то умножьте полученный процент на два.

При работе с любыми типами гипотез (больше чем, меньше чем или не равно) преобразуйте процент в вероятность, разделив его на 100 или передвинув запятую в десятичной дроби на два знака влево.

Представленая в табл. 14.2 таблица распределения Стьюдента не включает все возможные статистики критерия, поэтому просто выбирайте то значение, которое ближе всего к вашему, найдите столбик, в котором оно стоит, и определите соответствующий процентиль. После этого можете прикинуть свое *p*-значение.

В последней строке табл. 14.2 показаны соответствующие значения из стандартного нормального распределения (*Z*-распределения) для данных процентилей. Обратите внимание, что с увеличением числа степеней свободы на *t*-распределении (т.е. когда вы движетесь вниз по конкретному столбцу), числа все больше и больше приближаются к этому значению в последней строке. Это подтверждает то, что вам уже известно: если размер выборки увеличивается, *t*-распределение и *Z*-распределение становятся все больше похожи друг на друга.

Глава 15

Самые распространенные критерии проверки гипотез: формулы и примеры

В этой главе...

- Подробнее о самых распространенных критериях проверки гипотез
- Вычисление статистик критерия
- Использование результатов для принятия информированных решений

В рекламном ролике новой продукции или в газетной статье о последних открытиях в медицине то и дело можно наткнуться на утверждения о той или иной совокупности. Например, "Мы обещаем доставлять нашу продукцию менее чем за два дня" или "По результатам двух последних исследований было установлено, что диета, обогащенная клетчаткой, позволит вам снизить риск раковых заболеваний на 20%". Когда делается какое-либо утверждение (которое также называется *нулевой гипотезой*), касающееся генеральной совокупности (к примеру, среднее время, которое затрачивается на дорогу на работу и домой — 6 часов в неделю, или процент людей в США, которым нравится новое реалити-шоу — 30%), вы можете проверить его с помощью так называемого *критерия проверки гипотезы*. Кроме того, использовать проверку гипотезы можно и для сравнения двух генеральных совокупностей (например, среднее время на дорогу, затрачиваемое людьми, работающими в первую смену, по сравнению с теми, кто работает во вторую смену, или доля женщин с сотовыми телефонами по сравнению с долей мужчин). Общая информация о критериях проверки гипотезы изложена в главе 14.

Для проверки гипотезы нужно сформулировать гипотезы (утверждение и альтернативный ему вариант), сделать выборку (или выборки), собрать данные, вычислить необходимые статистические показатели и воспользоваться ими для того, чтобы проверить, истинным ли было утверждение. На самом деле вы сравниваете результаты своей выборки с заявленным параметром совокупности и проверяете, насколько они совпадают. Например, если в среднем для выборки из 1000 человек на дорогу уходит 5,2 часа, то 5,2 — это статистика выборки. Если утверждение гласит, что среднее время в дороге для всех работников в генеральной совокупности составляет 6 часов в неделю, заявленный параметр совокупности в данном случае — это среднее совокупности. Чем ближе статистика выборки к заявленному значению параметра генеральной совокупности, тем больше можно доверять утверждению. Но все же остается важный вопрос: "Сколько это — достаточно близко?" В этой главе представлены формулы, которые используются в самых распространенных критериях проверки гипотез, поясняются необходимые вычисления и приводится несколько примеров.

Проверка среднего одной совокупности

Такая проверка используется в случае с числовой переменной (например, возраст, доход, время и т.д.), если изучается только одна совокупность или группа (скажем, американские семьи или все студенты колледжей). Например, доктор Рут говорит, что работающие матери проводят со своими детьми в среднем 11 минут в день. (А отцы и того меньше — 8 минут.) Переменная здесь — время — числовая, а совокупность — это все работающие матери.

Нулевая гипотеза состоит в том, что среднее совокупности, μ, равно определенному заявленному значению μ_0. Нулевую гипотезу можно записать так: $H_0 : \mu = \mu_0$. Значит, в примере с доктором Рут нулевая гипотеза — это $H_0 : \mu = 11$ минут, и μ_0 здесь — 11 минут. Обратите внимание, что μ обозначает среднее количество минут в день, которые работающие матери уделяют своим детям. Альтернативная гипотеза H_a будет $\mu > \mu_0$, $\mu < \mu_0$ либо $\mu \neq \mu_0$. В нашем примере три варианта H_a будут такими: $\mu > 11$, $\mu < 11$, $\mu \neq 11$. (Подробнее об альтернативных гипотезах см. главу 14.) Если вам кажется, что работающие матери проводят с детьми больше времени, чем в среднем 11 минут, то в качестве альтернативной гипотезы вы выбираете вариант $H_a : \mu > 11$.

Формула статистики критерия для среднего одной совокупности выглядит так: $\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$. Для вычисления поступайте следующим образом.

1. **Найдите среднее выборки $\bar{x}$ и стандартное отклонение выборки s. Размер выборки обозначим n.**

   Подробнее о вычислении среднего и стандартного отклонения см. главу 4.

2. **Найдите разность $\bar{x}$ и μ_0.**

3. **Вычислите стандартную ошибку $s/\sqrt{n}$. Запомните результат.**

4. **Разделите результат, полученный на этапе 2, на стандартную ошибку, которую вы нашли на этапе 3.**

   В примере с доктором Рут предположим, что вы сделали случайную выборку из 100 работающих матерей и выяснили, что они проводят с детьми в среднем 11,5 минут в день. Стандартное отклонение равно 2,3 минут. Это значит, что $\bar{x} = 11{,}5$, $n = 100$, а $s = 2{,}3$.

   Берем $11{,}5 - 11 = 0{,}5$.

   Делим 2,3 на корень квадратный из 100 (который равен 10) и получаем стандартную ошибку 0,23.

   Делим 0,5 на 0,23 и получаем 2,17 (округляем результат до 2,2). Это ваша статистика критерия.

   Это значит, что среднее вашей выборки находится в пределах 2,2 стандартных ошибок от заявленного среднего совокупности. Если бы утверждение ($H_0 : \mu = 11$ минут) было истинным, был бы такой результат выборки необычным? Чтобы понять, поддерживает ли ваша статистика критерия H_0, определим *p*-значение. Для вычисления *p*-значения нужно найти вашу статистику критерия (в данном случае это 2,2) на стандартном нормальном распределении (*Z*-распределении) — см. табл. 8.1 в главе 8 — и отнять обнаруженный процентиль из 100%, поскольку в качестве H_0 вы выбрали вариант “больше чем”. В данном случае процентное отношение будет равно 100% – 98,61% = 1,39%. Значит, *p*-значение (деленное на 100) равно 0,0139. (Подробнее о вычислении *p*-значений см. главу 14.) Такое *p*-значение в 0,0139

(1,39%) намного меньше, чем 0,05 (5%). Это значит, что, если утверждение (об 11 минутах) истинно, то ваши результаты будут необычными. Следовательно, отклоняем утверждение ($\mu = 11$ минут) и принимает H_a ($\mu > 11$ минут).

Вывод таков: согласно данной (гипотетической) выборке утверждение доктора Рут об 11 минутах несколько занижено. Реальное среднее значение — больше 11 минут в день. Подробнее о вычислениях при проверке гипотез и соответствующих выводах читайте в главе 14.

Если бы размер выборки n был меньше 30, то вы бы искали свою статистику критерия не на стандартном нормальном распределении (Z-распределении), а на t-распределении. Подробнее об этом см. табл. 14.2 в главе 14. Дополнительную информацию о том, как вычисляется p-значение для вариантов "меньше чем" и "не равно", вы также найдете в главе 14.

Проверка доли одной совокупности

К такой проверке прибегают, если рассматриваемая переменная относится к дискретным (скажем, пол, политическая партия, за/против и т.п.) и изучается всего одна группа или совокупность (например, граждане США или зарегистрированные избиратели). При проверке изучается доля (p) элементов генеральной совокупности, которым присуща определенная характеристика, например, доля людей, имеющих при себе сотовые телефоны. Нулевая гипотеза выглядит как H_0: $p = p_0$, где p_0 — это определенная заявленная величина. Например, если утверждение гласит, что 20% людей всегда берут с собой сотовый телефон, то $p_0 = 0{,}20$. Альтернативной гипотезой может быть один из следующих вариантов: $p > p_0$, $p < p_0$ или $p \neq p_0$. (Подробнее об альтернативных гипотезах см. главу 14.)

Формула для вычисления статистики критерия одной совокупности выглядит так:

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Выполните следующие действия.

1. **Найдите долю выборки $\hat{p}$. Для этого количество людей в выборке, обладающих интересующей вас характеристикой (например, количество людей в выборке, имеющих сотовый телефон) разделите на размер выборки n.**
2. **Найдите разность $\hat{p}$ минус p_0.**
3. **Подсчитайте стандартную ошибку: $\sqrt{\frac{p_0(1-p_0)}{n}}$. Запомните ответ.**
4. **Разделите результат, полученный на этапе 2, на результат из этапа 3.**

Чтобы истолковать статистику критерия, определите ее положение на стандартном нормальном распределении (см. табл. 8.1 в главе 8) и подсчитайте p-значение (подробнее о вычислении p-значений см. главу 14)

Например, предположим, что производители зубной пасты Бескариесная утверждают, что четыре из пяти стоматологов советуют своим пациентам использовать пасту именно этой марки. В данном случае генеральная совокупность — это все стоматологи, а p — доля тех из них, кто рекомендует указанную пасту. У нас имеется утверждение о том, что p равно "четырем из пяти", т.е. $p_0 = 4/5 = 0{,}80$. Вы подозреваете, что на самом

деле эта доля меньше, чем 0,80, поэтому есть две гипотезы: $H_0 : p = 0{,}80$ и $H_a : p < 0{,}80$. Предположим, что 150 из 200 пациентов, попавших в выборку, получили от своих стоматологов совет использовать зубную пасту Бескариесная.

Для определения статистики критерия начнем с того, что $\hat{p} = 150/200 = 0{,}75$. Кроме того, $p_0 = 0{,}80$ и $n = 200$.

Найдем разность $0{,}75 - 0{,}80 = -0{,}05$.

Затем определяем стандартную ошибку, которая равна корню квадратному из $[(0{,}80 \times [1 - 0{,}80])/200]$ = корень квадратный из $(0{,}16/200)$ = корень квадратный из $0{,}0008 = 0{,}028$.

Статистика критерия равна $-0{,}05$, деленные на 0,028, что дает $-0{,}05/0{,}028 = -1{,}79$, округляем этот результат и получаем $-1{,}8$.

Это значит, что результаты вашей выборки находятся на расстоянии 1,8 стандартных ошибок ниже заявленного значения для совокупности.

Часто ли будут повторяться такие результаты, если H_0 справедливо? Процентная вероятность того, что значение будет равно или смещено (в данном случае влево) $-1{,}8$, равна 3,59%. (См. табл. 8.1 в главе 8 и используйте соответствующий процентиль, поскольку H_a — это вариант "меньше чем". Подробнее этот вопрос рассмотрен в главе 14.) Теперь разделим результат на 100, чтобы получить *p*-значение, которое будет равно 0,0359. Поскольку *p*-значение намного меньше чем 0,05, у вас есть все основания опровергнуть H_0.

Согласно сделанной выборке утверждение о том, что четыре из пяти (80%) стоматологов рекомендуют пользоваться пастой Бескариесная, ложно, реальный процент тех, кто дает такой совет, меньше заявленного.

При проведении большинства проверок гипотез, в которых дело касается долей, используются довольно большие выборки. Поскольку чаще всего это делается с помощью опросов, то очень редко когда можно встретить выборку небольшого размера. Подробнее о том, как находить *p*-значения для вариантов "больше чем" и "не равно", см. главу 14.

Сравнение средних двух отдельных совокупностей

Эту проверку проводят, если переменная относится к числовым (например, доход, уровень холестерина или расход горючего на километр) и сравниваются две совокупности или группы (к примеру, мужчины и женщины, спортсмены и не спортсмены, легковые автомобили и внедорожники). Необходимо сделать две отдельных случайных выборки, по одной из каждой совокупности, чтобы собрать данные, необходимые для этой проверки. Нулевая гипотеза гласит, что средние двух совокупностей равны, другими словами, их разность равна 0. Обозначить нулевую гипотезу можно так: $H_0 : \mu_x - \mu_y = 0$, где μ_x представляет среднее первой совокупности, а μ_y — это среднее второй совокупности.

Формула статистики критерия для сравнения двух средних такова:

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}}.$$

Для вычисления проделайте следующее.

1. **Найдите средние двух выборок ($\bar{x}$ и $\bar{y}$) и стандартные отклонения (s_x и s_y). Пусть n_1 и n_2 обозначают размер выборок (они не должны быть равны).**

   Подробнее о вычислениях см. главу 4.

2. **Найдите разность средних двух выборок, $\bar{x} - \bar{y}$.**

3. **Определите стандартную ошибку, $\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}$. Запомните результат.**

4. **Разделите результат, полученный на этапе 2, на результат, который вы нашли на этапе 3.**

Для интерпретации статистики критерия найдите ее на стандартном нормальном распределении (см. табл. 8.1 в главе 8) и подсчитайте *p*-значение (подробнее о вычислении *p*-значений см. главу 14).

Например, предположим, вы хотите сравнить поглощающую способность двух марок бумажных салфеток (назовем их Статабсорбент и Губкоматик). Сделать это можно, определив среднее количество унций жидкости, которые могут впитать салфетки обеих марок и не промокнуть. H_0 гласит, что разница между средней поглощаемостью равна 0 (не существует), а H_a утверждает, что разница не равна 0. Другими словами, $H_0 : \mu_x - \mu_y = 0$, а $H_a : \mu_x - \mu_y \neq 0$. Здесь не указано, салфетки какой марки впитывают больше влаги, поэтому нужно использовать вариант "не равно" (см. главу 14).

Представим себе, что вы делаете случайную выборку по 50 салфеток каждой марки и измеряете их поглощающую способность. Предположим, средняя поглощающая способность салфеток Статабсорбент (x) равна 3 унции со стандартным отклонением в 0,9 унции, а средняя поглощающая способность салфеток Губкоматик (y) составляет 3,5 унции со стандартным отклонением 1,2 унции.

Учитывая все данные, можно сказать, что $\bar{x} = 3$, $s_x = 0{,}9$, $\bar{y} = 3{,}5$, $s_y = 1{,}2$, $n_1 = 50$ и $n_2 = 50$.

Разность между средними выборок (Статабсорбент – Губкоматик) равна $(3 - 3{,}5) = -0{,}5$ унции. (Отрицательное значение говорит лишь о том, что среднее второй выборки больше среднего первой.)

Стандартная ошибка равна: $\sqrt{\frac{0{,}9^2}{50} + \frac{1{,}2^2}{50}} = \sqrt{\frac{0{,}81}{50} + \frac{1{,}44}{50}} = \sqrt{0{,}045} = 0{,}2121.$

Разделим разность, $-0{,}5$, на стандартную ошибку 0,2121, в результате получим $-2{,}36$, что можно округлить до 2,4. Это и есть ваша статистика критерия.

Чтобы найти *p*-значение, найдем $-2{,}4$ на стандартном нормальном распределении (*Z*-распределение) — см. табл. 8.1 в главе 8. В данном случае шансы оказаться слева от $-2{,}4$ равны процентилю, т.е. 0,82%. Поскольку в качестве H_a выбран вариант "не равно", то вы умножаете этот процентиль на 2 и получаете $2 \times 0{,}82\% = 1{,}64\%$. Наконец, преобразуем этот результат в вероятность, для чего разделим на 100, и получится *p*-значение 0,0164. Это *p*-значение меньше, чем 0,05. Т.е. у вас достаточно доказательств, чтобы опровергнуть H_0.

Делаем вывод: между поглощающими способностями этих двух марок бумажных салфеток имеется статистически значимая разница, о чем говорят результаты ваших выборок. Похоже, лидирует Губкоматик, потому что у этих салфеток среднее значение поглощаемости больше.

Будучи проницательным статистиком, не верьте тем рекламным роликам, в которых показана всего одна салфетка из единственной пачки (т.е. выборка размером 1) и утверждается, что она впитывает влагу лучше салфеток другой марки. А также не доверяйте утренним телепередачам, в которых производители опрашивают двух-трех человек на улице и делают сравнения на основе такой информации. При правильном проведении опроса проверка гипотезы дает результаты, которые не только интересны, *но и* могут быть обобщены. (Подробнее о недоверии к случаям из жизни см. главу 14.)

Обычно проверки гипотез, связанные со сравнением средних значений двух отдельных совокупностей, делаются с использованием выборок довольно большого размера, потому что чаще всего в их основу положены опросы. Но если обе выборки окажутся размером меньше 30 элементов, то для определения *p*-значения вам нужно будет использовать *t*-распределение (с равным числом степеней свободы или в зависимости от того, какая выборка меньше). (Подробнее о *t*-распределении см. табл. 14.2 в главе 14.)

Проверка разности средних (парные данные)

Такой тип проверки используется, когда переменная числовая (например, доход, уровень холестерина или расход топлива в литрах на километр), а элементы в выборке либо каким-то образом объединены в пары (к примеру, часто используются идентичные близнецы), либо некоторые люди были опрошены дважды (скажем, до и после проверки). Парные тесты обычно используются в случае с исследованиями, в которых выясняется вопрос, какой метод лечения, лекарство или технология работают лучше. При этом не учитываются другие факторы, которые могут повлиять на результаты. (Подробнее см. главу 17.)

Например, представим себе, что исследователь хочет выяснить, не является ли более результативным метод обучения чтению с помощью компьютерных игр по сравнению с проверенным фонетическим методом. Делается случайная выборка из 20 учеников, они разбиваются на 10 пар в соответствии со своим уровнем навыков чтения, возрастом, IQ и т.д. Исследователь выбирает наугад одного ученика из каждой пары, который затем будет учиться читать с помощью компьютерной игры, а второй ученик использует традиционный метод. В конце исследования все ученики сдают одинаковый тест на развитие навыков чтения. Данные показаны в табл. 15.1.

Таблица 15.1. Баллы за чтение после компьютерной игры и фонетического метода

№ ученической пары	Баллы ученика после компьютерной игры	Баллы ученика после фонетического метода	Парные разности (компьютерные баллы минус фонетические баллы)
1	85	80	+5
2	80	80	+0
3	95	88	+7
4	87	90	–3
5	78	72	+6
6	82	79	+3
7	57	50	+7
8	69	73	–4
9	73	78	–5
10	99	95	+4

Данные расположены попарно, но вас интересует только разность баллов, набранных за чтение (компьютерные баллы – фонетические баллы) в каждой паре, а не сами баллы. Разности баллов в каждой паре, или *парные разности*, представляют собой новый набор данных, с которыми вы будете работать. Если обе методики обучения чтению одинаковы, то среднее таких парных разностей должно быть равно 0. Если компьютерный метод эффективнее, то среднее парных разностей должно быть положительным (потому что компьютерные баллы должны быть больше фонетических). Значит, необходимо проверить гипотезу для среднего одной совокупности, где нулевая гипотеза гласит, что среднее (парных разностей) равно 0, а альтернативная гипотеза утверждает, что среднее (парных разностей) > 0.

Нулевая гипотеза обозначается $H_0 : \mu_d = 0$, где μ_d — это среднее парных разностей.

Формула статистики критерия для парных разностей такова: $\frac{d-0}{s/\sqrt{n}}$. Вычисления выполняются следующим образом.

1. **Для каждой пары данных найдите парную разницу, для чего из первого значения вычтите второе.**

   Эти разности и будут вашим новым набором данных.

2. **Найдите среднее $\bar{d}$ и стандартное отклонение s всех разностей.**

   Пусть n представляет количество парных разностей, имеющихся в вашем наборе.

3. **Найдите стандартную ошибку: $s/\sqrt{n}$. Запомните результат.**

4. **Разделите $\bar{d}$ на стандартную ошибку из этапа 3.**

Помните, что если H_0 истинно, то $\mu_d = 0$, поэтому оно не включено в эту формулу.

В случае с баллами за чтение вы можете проделать все описанные шаги и выяснить, какой метод обучения чтению лучше.

Найдите разности в каждой паре. Они показаны в табл. 15.1 в столбце 4. Обратите внимание, что здесь важен знак в каждом случае, он показывает, какой из использованных методов оказался эффективнее для конкретной пары.

Необходимо вычислить среднее и стандартное отклонение разностей (столбец 4 в табл. 15.1). (Подробнее о вычислении среднего и стандартного отклонения см. главу 4.) Определяем, что среднее разностей равно +2, а стандартное отклонение составляет 4,64. Обратите внимание, что в данном случае $n = 10$.

Стандартная ошибка — это 4,64, деленные на корень квадратный из 10 (3,16). Итак, 4,64/3,16 = 1,47. (Помните, что здесь n — это количество пар, т.е. 10.)

На последнем этапе вы делите среднее разностей, т.е. +2, на стандартную ошибку, т.е. 1,47, и получаете +1,36, Это и есть статистика критерия. Другими словами, средняя разница для этой выборки — это 1,36 стандартных ошибки выше нуля. Достаточно ли этого, чтобы сказать, что разница в баллах применима для всей совокупности?

Поскольку n меньше 30, то вы ищете 1,36 на t-распределении с числом степеней свободы 10 – 1 = 9 (см. табл. 14.2 в главе 14) и определяете p-значение. В данном случае p-значение больше 0,05, потому что в таблице 1,36 ближе к значению 1,38, значит, соответствующее p-значение будет больше 0,10. Это объясняется тем, что 1,38 находится в

столбце на 90-м процентиле, а поскольку в качестве H_a выбран вариант "больше чем", то вы берете 100% – 90% = 10% = 0,10. Теперь можно сделать вывод, что для опровержения H_0 недостаточно доказательств, поэтому компьютерный метод нельзя назвать более эффективным при обучении чтению. (Возможно, это объясняется отсутствием дополнительных доказательств из-за небольшого размера выборки.)

Во многих парных экспериментах наборы данных будут маленькими из-за затрат и времени, необходимых для проведения подобных исследований. Это означает, что при определении *p*-значения вместо стандартного нормального распределения (см. табл. 8.1 в главе 8) часто используется *t*–распределение (см. табл. 14.2 в главе 14).

Сравнение долей двух совокупностей

Такая проверка используется, когда переменная дискретная (скажем, курящий/некурящий, политическая партия, мнение за/против по определенному вопросу и т.д.), а вас интересует доля элементов с определенной характеристикой — например, доля курящих. В данном случае сравниваются две совокупности или группы (например, доля курящих женщин и доля курящих мужчин). Чтобы провести эту проверку, необходимо сделать две отдельных выборки, по одной из каждой совокупности. Нулевая гипотеза гласит, что доли в двух совокупностях одинаковы, другими словами, их разность равна 0. Обозначить нулевую гипотезу можно как $H_0 : p_1 - p_2 = 0$, где p_1 — это доля из первой совокупности, а p_2 — доля из второй совокупности.

Формула статистики критерия для сравнения двух долей такова:

$$\frac{\left(\hat{p}_1 - \hat{p}_2\right) - 0}{\sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

Выполните следующие действия:

1. **Найдите доли $\hat{p}_1$ и $\hat{p}_2$ в выборках. Пусть n_1 и n_2 — это размеры двух выборок (они не должны быть равны).**
2. **Найдите разность двух долей ($\hat{p}_1 - \hat{p}_2$).**
3. **Определите общую долю выборки $\hat{p}$, представляющую собой общее число элементов из обеих выборок, которые обладают интересующей вас характеристикой (например, общее количество курящих, как мужчин, так и женщин, в выборке), деленное на общее число элементов в обеих выборках ($n_1 + n_2$).**
4. **Найдите стандартную ошибку: $\sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$. Запомните результат.**
5. **Разделите результат, полученный на этапе 2, на результат, который вы нашли на этапе 4.**

Для интерпретации этой статистики критерия найдите ее на стандартном нормальном распределении (табл. 8.1 в главе 8) и определите *p*-значение (подробнее о *p*-значениях см. главу 14).

Например, рассмотрим те рекламные объявления о лекарствах, которые фармацевтические компании публикуют в журналах. На первой странице показана идиллическая картина, где ярко светит солнце, цветут цветы, а люди улыбаются — их жизнь изменилась

благодаря новому лекарству. Компания утверждает, что ее лекарство может победить симптомы аллергии, поможет людям улучшить сон, снизить артериальное давление или вылечить любое другое недомогание. Утверждения могут показаться слишком радужными, чтобы быть правдивыми, но когда вы переворачиваете страницу, то видите там примечание мелким шрифтом, где фармацевтическая компания доказывает свои утверждения. (Именно здесь обычно зарыта статистика!) Там же вы можете найти таблицу, в которой показано побочное воздействие лекарства по сравнению с контрольной группой (т.е. испытуемыми, которые принимали ненастоящий препарат, чтобы можно было честно сравнить их результаты с теми, кто использовал настоящее лекарство. Подробнее об этом см. главу 17.) Например, препарат Adderall борется с синдромом дефицита внимания и гиперактивности. Сообщается, что у 26 из 374 испытуемых (7%), принимавших это лекарство, в качестве побочного эффекта наблюдалась рвота, по сравнению с 8 из 210 испытуемых (4%), принимавших *плацебо* (ненастоящее лекарство). Заметьте, что пациенты не знали, какой препарат получают. В данной выборке рвота наблюдалась у большего количества человек, но достаточно ли большое это процентное отношение, чтобы утверждать, что во всей совокупности будет тот же эффект? Это можно проверить.

В этом примере у нас есть H_0: $p_1 - p_2 = 0$ и H_a: $p_1 - p_2 > 0$, где p_1 — это доля испытуемых, которых вырвало после приема Adderall, а p_2 — доля испытуемых, почувствовавших тошноту после приема плацебо.

Почему в H_a используется знак “>”, а не “<”? Дело в том, что H_a — это сценарий, по которому больше людей, принимавших Adderall, испытывали тошноту, чем те, кто использовал плацебо (если бы было наоборот, это очень заинтересовало бы Управление по контролю над продуктами питания и лекарственными средствами). Но важна также и последовательность групп. Вы должны расположить их так, чтобы первой шла группа испытывавших Adderall; если H_0 истинно, то после того, как вы вычтете из доли Adderall долю плацебо, число получится положительным. Если поменять группы местами, ответ будет отрицательным.

Теперь переходим к вычислению статистики критерия.

Прежде всего, $\hat{p}_1 = 26/374 = 0{,}07$, а $\hat{p}_2 = 8/210 = 0{,}04$. Размеры выборок — $n_1 = 374$ и $n_2 = 210$ соответственно.

Теперь найдем разность этих долей выборок, получается $0{,}07 - 0{,}04 = 0{,}03$.

Общая доля выборки $\hat{p}$ равна $(26 + 8)/(374 + 210) = 34/584 = 0{,}058$.

Стандартная ошибка равна

$$\sqrt{0{,}058 \times (1 - 0{,}058) \times \left(\frac{1}{374} + \frac{1}{210}\right)} = \sqrt{0{,}058 \times 0{,}942 \times 0{,}0074} = \sqrt{0{,}0004} = 0{,}02.$$ Уф!

Наконец, разделим разность, полученную на этапе 2, на 0,02, получится $0{,}03/0{,}02 = 1{,}5$. Это и есть статистика критерия.

P-значение — это процентная вероятность оказаться на значении 1,5 или дальше от него (в данном случае вправо), т.е. $100\% - 93{,}32\% = 6{,}68\%$, что в виде вероятности записывается как 0,0668. (См. табл. 8.1 в главе 8.) Это *p*-значение немного больше 0,05, значит, формально у вас нет достаточных доказательств для того, чтобы опровергнуть H_0. Это значит, что рвота у тех, кто принимал новое лекарство, наблюдается не чаще, чем у

тех, кто употреблял плацебо (это пример результата, который статистики называют *маргинальным*).

Р-значение, которое очень близко к той самой волшебной, но несколько спорной пороговой точке — это *маргинальный результат* для статистиков. В предыдущем примере *p*-значение, равное 0,0668, считается маргинальным результатом. Это значит, что результат находится как раз на границе между принятием и опровержением H_0. В этом же и состоит прелесть упоминания *p*-значения: проанализировав его, вы можете решить, какой вывод стоит сделать. Чем меньше *p*-значение, тем сильнее доказательства, опровергающие H_0, но сколько доказательств будет достаточно? Мнения у людей разные. Если вам встретится отчет об исследовании, в котором был получен статистически значимый результат, который очень важен для вас, выясните также *p*-значение, чтобы принять решение самостоятельно. Подробнее см. главу 14.

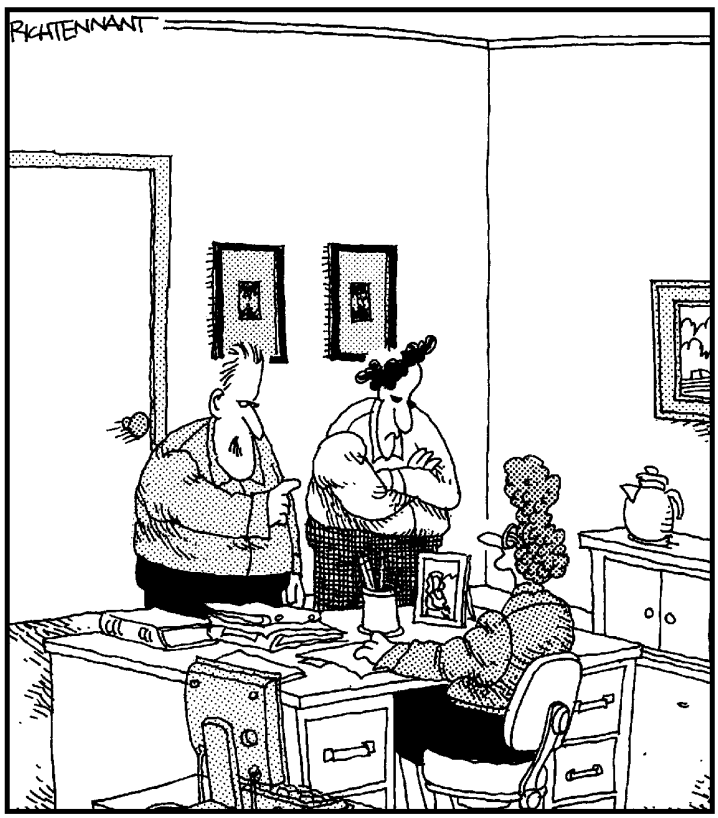
Чаще всего проверка гипотез, касающаяся долей двух отдельных совокупностей, проводится с использованием довольно больших выборок, потому что обычно они делаются на основе опроса, следовательно, вам вряд ли попадется случай, когда изучаемые выборки были небольшими.

Часть VII

Статистические исследования: взгляд изнутри

The 5th Wave

Рич Теннант



"Мы с Тедом провели вместе более 120 человеко-часов, анализируя данные опроса, и вот что обнаружили: Тед одалживает ручки и никогда их не возвращает, он специально скрипит стулом, чтобы позлить меня, а я, очевидно, разговариваю во сне."

В этой части...

Многие статистические показатели, с которыми вы сталкиваетесь каждый день, основаны на результатах опросов, экспериментов и исследований. К сожалению, нельзя верить всему, что слышишь и видишь.

В этой части вы узнаете, что на самом деле происходит за кулисами таких исследований: как они разрабатываются и проводятся, как собираются (или должны собираться) данные. Вы также научитесь замечать ошибочные результаты.

Глава 16

Опросы, опросы и еще раз опросы

В этой главе...

- Влияние опросов
- За кулисами опросов
- Интерпретация результатов опросов
- Признаки необъективных и неточных результатов опросов

Кажется, в наш век информации опросы достигли пика популярности. Все хотят знать, что думают люди по тем или иным вопросам, начиная от цен на лекарства и заканчивая методами воспитания детей или рейтингами президента и телепередач. Опросы играют в жизни любой страны заметную роль, это средство помогает быстро получить информацию о том, что вы думаете, что чувствуете и как проживаете свою жизнь. Кроме того, это метод быстро распространить полученную информацию. Опросы используются для того, чтобы изучить спорные вопросы, привлечь внимание, упрочить политическое положение, подчеркнуть важность темы или просветить, убедить в чем-либо общество.

Результаты опросов могут быть эффективными, потому что когда людям сообщают, что "такой-то процент населения делает то-то и то-то", они считают такие результаты истинными, поэтому принимают решения и основываются в своем мнении исходя из подобной информации. Но в действительности многие опросы не содержат правильной, полной или хотя бы сбалансированной информации. В этой главе речь пойдет о влиянии опросов и о том, как они используются. Вы узнаете, как готовятся и проводятся опросы, чтобы в будущем вы могли знать, на что обращать внимание, изучая результаты каких-либо опросов. В главе также рассказывается о том, как нужно толковать результаты опросов и как замечать смещенную, неточную информацию, чтобы вы самостоятельно могли определить, каким результатам верить, а какие лучше оставить без внимания.

Влияние опросов

Опрос — это инструмент для сбора данных посредством вопросов и ответов, он используется для сбора информации о мнениях, моделях поведения, демографических показателях, стиле жизни и других характеристиках изучаемой совокупности.

Вы ежедневно сталкиваетесь с опросами и их результатами. В США для них даже создана специальная телепрограмма — игровое шоу *Family Feud* основано исключительно на опросах и способности игроков назвать самые распространенные ответы, которые дали участники опроса. Игроки в этом шоу должны правильно определить ответы, полученные от респондентов на различные вопросы, например, "Назовите животного, которого можно увидеть в зоопарке" или "Назовите известного человека по имени Джон".

По сравнению с другими видами исследований, такими как медицинские эксперименты, опросы сравнительно просты в проведении и далеко не такие дорогостоящие. Они быстро дают результаты, на основе которых часто можно придумать сенсационные заголовки для газет или увлекательные статьи для журналов. Люди любят опросы, потому что им кажется, что результаты опросов отражают ответы таких же, как они сами (даже хотя лично их, возможно, никогда не приглашали стать участником опроса). И многие испытывают любопытство по поводу того, что чувствуют другие, что они делают, куда ходят и о чем волнуются. Изучение результатов опроса позволяет людям каким-то образом почувствовать себя частью большой группы. Вот на что делают ставку *интервьюеры* (люди, проводящие опрос) и вот почему они тратят столько времени на проведение опросов и сообщение результатов своих исследований.

Добираемся до источника

Кто сегодня проводит опросы? Очень многие, практически любой, у кого есть подходящий вопрос. Ниже перечислены некоторые группы, проводящие опросы и сообщающие о полученных результатах.

- ✓ Агентства новостей (например, ABC News, CNN, Reuters).
- ✓ Политические партии (правящие и те, кто стремится попасть во власть).
- ✓ Профессиональные опросные организации (такие как The Gallup Organization, The Harris Poll, Zogby International и т.д.).
- ✓ Представители журналов, теле- и радиопрограмм.
- ✓ Отраслевые организации (например, Американская медицинская ассоциация, которая часто проводит опросы своих членов).
- ✓ Группы особых интересов (скажем, Национальная оружейная ассоциация).
- ✓ Ученые-исследователи (проводящие исследования по самым разным темам).
- ✓ Американское правительство (которое проводит Опрос американского общества, Опрос жертв преступности и многие другие виды опросов, над которыми работает Бюро переписей).
- ✓ Любой желающий (который может легко провести собственный опрос через Интернет).

Не все, кто проводит опросы, надежны и имеют на это право, поэтому если вас просят принять участие или сообщают результаты, обязательно проверьте источник любого опроса. Группы, особо заинтересованные в результатах, должны либо привлекать независимую организацию к проведению опроса (или хотя бы наблюдению за ним), либо сообщить задаваемые вопросы широкой общественности. Кроме того, группы должны подробно объяснить, как разрабатывался и проводился опрос, чтобы вы могли принять информированное решение относительно надежности результатов.

Изучение актуальных тем

Темы многих опросов возникают благодаря текущим событиям, вопросам и сферам интересов, ведь своевременность и важность для публики — вот два самых привлекательных свойства любого опроса. Вот всего несколько примеров тех тем, которые изучаются современными опросами, наряду с полученными результатами.

- ✓ Влияет ли деятельность известных людей на политические пристрастия американских граждан? (По результатам CBS News, свыше 90% американцев ответили отрицательно.)
- ✓ Каков процент американцев, назначавших свидания в Интернете? (По данным CBS News, всего 6% одиноких пользователей Интернета.)
- ✓ Многие ли американцы страдают от боли? (Согласно CBS News, три четверти людей младше 50 часто или хотя бы иногда испытывают боль.)
- ✓ Сколько человек ищет в Интернете информацию о здоровье? (Около 98 миллионов, сообщает The Harris Poll.)
- ✓ Насколько оптимистичны инвесторы в настоящее время? (Согласно опросу, проведенному The Gallup Organization, ситуацию лучше назвать инвесторским пессимизмом.)
- ✓ Каким был худший автомобиль тысячелетия? (Как сказали слушатели радиопередачи Car Talk на станции NPR, это была модель Yugo.)

Читая предыдущие результаты опроса, не думали ли вы, прежде всего, о том, что они значат для вас, а уж затем задумывались, правдивы ли они? Некоторые из вышеперечисленных данных надежнее и достовернее других, и вы должны сначала подумать, стоит ли доверять результатам, но не принимать их на веру.

Выбор худшего автомобиля тысячелетия

Субботним утром на Национальном общественном радио США транслируется радиопередача Car Talk. Ее ведущие — братья Клик и Клак — дают мудрые и ценные советы дозвонившимся, тем, кто столкнулся с разными автомобильными проблемами. На Web-сайте этой передачи регулярно проводятся опросы по самым разным темам, связанным с автомобилями, например, "У кого на бампере есть наклейки, и что на них написано?". Тысячи человек отвечают на задаваемые вопросы, но, конечно же, эти слушатели не представляют всех владельцев автомобилей. Они представляют только тех, кто слушает эту радиопередачу, заходит на конкретный Web-сайт и отвечает на вопросы. Наверное, вам не терпится узнать результаты опроса, которые предложены в следующей таблице. Но прежде чем посмотреть, что ответили другие, может быть, вы захотите сами высказать свое мнение? (Однако помните, что эти результаты отражают только мнение поклонников передачи Car Talk, которые не пожалели времени, чтобы зайти на Web-сайт и поучаствовать в опросе.) Обратите внимание, что проценты в сумме не дадут 100%, потому что в таблице представлены только первые десять названий, получившие максимальное количество голосов.

Место	Тип машины	Процент проголосовавших
1	Yugo	33,7
2	Chevy Vega	15,8
3	Ford Pinto	12,6
4	AMC Gremlin	8,5
5	Chevy Chevette	7,0
6	Renault LeCar	4,3
7	Dodge Aspen/Plymouth Volare	4,1
8	Cadillac Cimarron	4,0
9	Renault Dauphine	3,6
10	Volkswagen (VW) Bus	2,7

Влияние на жизнь

Если в одних опросах весело поучаствовать или поразмышлять об их результатах, то другие могут напрямую повлиять на вашу жизнь или работу. Такие жизненно важные опросы нужно тщательно изучить, прежде чем приступать к действию или принимать важное решение. Опросы такого уровня могут заставить политиков изменить или принять новые законы, подтолкнуть исследователей начать работу над новейшими вопросами, стимулировать производителей изобретать новую продукцию или менять привычные методы ведения дел, а также они могут оказать влияние на поведение и образ мыслей людей. Вот некоторые примеры недавних опросов, которые могут коснуться и вас.

- ✓ **Подростки водят машину в нетрезвом виде**: недавний опрос агентства Reuters, в котором участвовало 1 119 подростков из Онтарио, Канада, учеников 7–13 классов, показал, что в течение прошлого года 15% из них садились за руль, выпив не меньше двух бокалов спиртного.
- ✓ **Страдает детское здравоохранение**: по данным опроса 400 педиатров, проведенного Национальным детским медицинским центром в Вашингтоне, округ Колумбия, детские врачи уделяют каждому пациенту всего 8–12 минут.
- ✓ **Преступления не учитываются**: как сообщает Министерство юстиции США после проведении опроса 2001 года, всего о 49,4% преступлений было заявлено в полицию. Жертвы объясняли нежелание обращаться в полицию причинами, которые указаны в табл. 16.1.

Таблица 16.1. Причины, по которым жертвы преступлений не обращались в полицию

Причина	Процент
Считают это личным делом	19,2
Преступнику не повезло/он не довел дело до конца	15,9
О преступлении сообщили в другие органы	14,7
Не считали преступление достаточно серьезным	5,5
Решили, что полиция не захочет разбираться	5,3
Отсутствие доказательств	5,0
Страх мести	4,6
Слишком неудобно/долго сообщать об этом	3,9
Решили, что полиция будет необъективной/неэффективной	2,7
У украденной собственности не было идентификационного номера	0,5
Не знали о преступлении до самого последнего времени	0,4
Другие причины	22,3

Чаще всего о преступлении не сообщали в полицию, потому что жертва считала это личным делом (19,2%). Обратите внимание, что почти 12% причин связаны с самим процессом сообщения (например, что это займет слишком много времени или что полиция не захочет этим заниматься, будет предвзятой и неэффективной).

- ✓ **Изменения в лечении рака груди**: какой вариант лечения рака груди лучше выбрать женщине — лампектомию (когда удаляется только опухоль, а большая часть груди остается) или мастектомию (когда удаляется вся грудь)? Самое распространенное мнение — для большинства женщин лучшим вариантом будет лампектомия. Но согласно последнему опросу хирургов, на вопрос, что бы они сами выбрали, если бы страдали ранней стадией рака груди, 50% предпочли бы мастектомию.
- ✓ **Использование сотовых телефонов за рулем повышает опасность**: недавний опрос среди покупателей и продавцов автомобилей касался тем, которые волнуют их больше всего. Данные показывают, что самая распространенная проблема, волнующая автолюбителей (53%), — это водители, которые говорят по телефону, ведя машину. Затем идут высокие цены на бензин (это беспокоит 50% респондентов) и насилие на дорогах (что вызывает тревогу у 38% опрошенных).
- ✓ **Виртуальные преступления захлестнули сферу бизнеса**: Институт компьютерной безопасности недавно провел опрос среди американских компаний с целью выяснить, как преступления, совершенные в киберпространстве, повлияли на их бизнес. Девяносто процентов респондентов сталкивались с нарушением компьютерной безопасности в прошедшем году, и 80% из них признали, что понесли из-за этого финансовые потери. 75% опрошенных заявили о том, что сотрудники используют Интернет в личных целях (например, посещают порносайты, скачивают пиратские программы и используют электронную почту не для работы).
- ✓ **Сексуальные домогательства создают большие проблемы на работе**: последний опрос, проведенный Комиссией по равным трудовым возможностям, показал, что от 40 до 70% женщин и от 10 до 20% мужчин, по их словам, стали жертвами сексуальных домогательств на работе. В последние годы количество жалоб со стороны мужчин утроилось.

В предыдущих примерах затрагиваются некоторые важные вопросы, но вы сами должны решить, верить или нет заявленным результатам. Вы должны уметь отличать надежную и достоверную информацию. Правило номер один: не нужно автоматически верить всему, что вы читаете!

За кулисами: подробнее об опросах

Опросы и их результаты — это часть вашей повседневной жизни, и вы используете эти результаты для принятия решений, которые влияют на вашу жизнь. (Некоторые решения могут даже в корне изменить ее.) Поэтому важно критически подходить к изучению результатов. Прежде чем предпринимать какие-то шаги или принимать решение на основе результатов опроса, вы должны решить, надежны ли они, достоверны ли и проверены. Хороший способ приступить к развитию в себе таких навыков детектива — это заглянуть за кулисы и выяснить, как разрабатываются, проводятся и анализируются опросы.

Процесс опроса можно разделить на десять этапов.

1. **Установите цель опроса.**
2. **Определите целевую совокупность.**
3. **Выберите тип опроса.**

4. **Сформулируйте вопросы.**
5. **Выберите время проведения опроса.**
6. **Сделайте выборку.**
7. **Соберите данные.**
8. **Проверяйте, проверяйте и проверяйте.**
9. **Систематизируйте и проанализируйте данные.**
10. **Сделайте выводы.**

На каждом этапе есть свои трудности и проблемы, но любой из них жизненно важен для получения точных и объективных результатов. Такая последовательность этапов позволит вам разработать, спланировать и провести опрос, но ею также можно воспользоваться и для критического рассмотрения чьего-либо опроса, если его результаты важны для вас.

Планирование и разработка опроса

Цель любого опроса — ответить на вопросы, касающиеся целевой совокупности (в статистике называемой *генеральной*). *Целевая совокупность* — это вся группа людей, относительно которых вы хотите сделать выводы. Чаще всего провести опрос во всей целевой совокупности невозможно, потому что в таком случае исследователям пришлось бы потратить слишком много времени и денег. (Если вы проводите опрос всей целевой совокупности, это называется *переписью*.) Обычно лучшее, что можно предпринять — это сделать выборку из целевой совокупности, опросить этих людей, а затем сделать выводы относительно целевой совокупности, исходя из данных, полученных о выборке.

Кажется просто, да? Но это не так. Как только вы осознаете, что нельзя опросить всех представителей целевой совокупности, возникает множество потенциальных проблем. К сожалению, многие опросы проводятся без предварительного обдумывания этих проблем, что приводит к ошибкам, вводящим в заблуждение результатам и неверным выводам.

Формулировка цели опроса

Руководствоваться в данном случае нужно всего лишь здравым смыслом, но в действительности многие опросы разрабатываются и проводятся так, что исследователи ни на шаг не приближаются к своей цели или же решают лишь некоторые из поставленных задач. Очень легко потеряться в вопросах и забыть, что же вы на самом деле собирались выяснить. Формулируя цель опроса, будьте предельно точны. Подумайте, какие выводы вы бы хотели сделать, если бы вам пришлось писать отчет, и пусть это поможет вам определить цели опроса.

Чем точнее вы сформулируете цель опроса, тем проще вам будет определить вопросы, с помощью которых вы достигнете этой цели, и тем легче вам будет писать отчет о проделанной работе.

Определение целевой совокупности

Представим, например, что вы хотите провести опрос, чтобы определить, насколько люди поглощены электронной перепиской личного характера на рабочем месте. Может показаться, что в данном случае целевая совокупность — это пользователи электрон-

ной почты на работе. Но вы же хотите выяснить, *насколько широко* электронная почта используется на работе, поэтому нельзя опрашивать только пользователей этой услуги, иначе результаты получатся смещенными против тех, кто не пользуется электронной почтой в работе. Но нужно ли включать сюда еще и тех, у кого в течение рабочего дня вообще нет доступа к компьютеру? (Видите, как легко можно ошибиться при подготовке опроса?)

Целевая совокупность, которая, наверное, была бы самой удачной в данном случае — это все люди, которые на рабочем месте имеют компьютеры с доступом в Интернет. У всех в этой группе, по крайней мере, есть возможность воспользоваться электронной почтой, хотя лишь некоторые из тех, у кого на работе есть электронная почта, действительно работают с ней, причем из них лишь некоторые используют электронную почту в личных целях. (А ведь именно это вы хотите выяснить — насколько интенсивно люди ведут личную электронную переписку во время работы.)

Нужно четко определять свою целевую совокупность. Именно определение позволяет сделать правильную выборку, а также подталкивает к нужным выводам, чтобы результаты не оказались слишком обобщенными. Если исследователь четко не установит целевую совокупность, это может послужить признаком и других проблем в опросе.

Выбор типа опроса

Следующий шаг в разработке опроса — выбор того типа опроса, который наиболее подходит для конкретной ситуации. Опросы можно проводить по телефону, по почте, при личном общении или через Интернет. Но не все виды применимы в определенных ситуациях. Например, предположим, вы хотите установить некоторые факторы, связанные с уровнем неграмотности в Соединенных Штатах Америки. Рассылать анкету по почте не имеет смысла, потому что те, кто не умеет читать, не смогут принять участие в опросе. В таком случае больше подойдет опрос по телефону.

Выбирайте тот тип опроса, который больше всего подходит для данной целевой совокупности, чтобы можно было получить самые достоверные и информативные данные. Изучая результаты опроса, обязательно проверьте, был ли проведенный опрос самым подходящим в конкретном случае.

Формулировка вопросов

После того, как была четко установлена цель опроса и вы выбрали подходящий тип, следующее, что нужно сделать — это сформулировать вопросы. То, как именно задаются вопросы, может в значительной степени повлиять на качество собранных данных. Один из самых распространенных источников смещения — это формулировка вопросов. *Наводящие вопросы* (т.е. вопросы, которые сформулированы так, что наталкивают на определенный ответ) могут во многом сказаться на ответах участников, а эти ответы не всегда будут точно отражать действительные ощущения респондентов по исследуемой теме. Например, вот два способа задать один и тот же вопрос об облигациях займа, выпущенных школьным округом (оба являются наводящими вопросами).

> *Не думаете ли вы, что незначительное увеличение налога с продаж будет выгодным капиталовложением в улучшение качеств образования для наших детей?*

Не думаете ли вы, что нужно прекратить увеличивать бремя, лежащее на наших налогоплательщиках, и не требовать роста налога с продаж, чтобы финансировать расточительную школьную систему?

Из формулировки каждого из этих наводящих вопросов вы легко можете понять, какой ответ от вас хотят услышать интервьюеры. Исследования показывают, что формулировка вопросов действительно влияет на результаты опросов. Лучше всего задавать вопрос нейтрально. Например, в данном случае спросить можно было бы следующим образом.

Школьный округ предлагает повысить налог с продаж на 0,01%, чтобы собрать средства для строительства новой школы. Что вы думаете об увеличении налога? (Возможные ответы: активно поддерживаю, поддерживаю, все равно, против, активно против.)

В хорошем опросе вопросы всегда сформулированы нейтрально, чтобы избежать смещения. Лучше всего оценить нейтральность вопросов — это спросить самого себя, можете ли вы определить, какой ответ от вас ожидается. Если да, то это наводящий вопрос, который может дать результаты, вводящие в заблуждение.

Если результаты опроса важны для вас, попросите у исследователей копию использованных вопросов, чтобы вы сами могли оценить их качество.

Время проведения опроса

В опросах, как и в жизни, время решает все. Текущие события постоянно определяют мнения людей, и если одни интервьюеры пытаются выяснить, что люди думают по поводу этих событий, то другие используют события в своих целях, чаще всего негативных, и строят на них свои политические платформы или делают источником сенсационных заголовков и споров. Время проведения опроса тоже может вызвать смещение результатов вне зависимости от изучаемой темы. Например, представьте себе, что ваша целевая совокупность для опроса — это люди, работающие целый день. Если проводить телефонный опрос, чтобы выяснить мнение офисных работников об интенсивности использования электронной почты в их работе, и при этом звонить им домой днем, между 8 и 17 часами, то результаты будут смещенными, потому что как раз в это время большинство офисных работников находятся на работе.

Проверьте, когда проводился опрос (дата и время), и выясните, не было ли в этот период каких-либо событий, которые могли повлиять на результаты. Также проверьте, проводился ли опрос в то время, когда принять в нем участие могли большинство представителей целевой совокупности.

Делаем выборку

После разработки опроса наступает время выбирать тех людей, которые будут в нем участвовать. Поскольку обычно времени и денег на проведение переписи (т.е. опроса всей целевой совокупности) не хватает, вам нужно выбрать определенную группу из совокупности, называемую *выборкой*. Сам процесс выборки может во многом повлиять на точность и качество результатов.

Для того чтобы сделать хорошую выборку, нужно учитывать три критерия.

- ✓ **Хорошая выборка представляет целевую совокупность**. Чтобы представлять целевую совокупность, выборка должна быть сделана в этой целевой совокупности, только в целевой совокупности и ни в чем другом. Предположим, вы хотите узнать, сколько часов в день в среднем американцы проводят перед телевизором. Если попросить студентов из общежития местного университета описать свои зрительские привычки, это не поможет. Студенты представляют только часть целевой совокупности. Если попросить об этом слушателей радио, то такая выборка тоже не будет представлять целевую совокупность, результаты будут говорить только о тех людях, кто в тот момент слушал радио, смог дозвониться по указанному номеру и вообще был достаточно заинтересован для того, чтобы стремиться поучаствовать в опросе. Точно так же опрос в Сети будет представлять только тех, у кого есть доступ к Интернету и кто зашел на сайт, где была помещена анкета для опроса.

  К сожалению, многие люди, проводящие опросы, не тратят время и деньги на то, чтобы сделать репрезентативную выборку участников для исследования. Это приводит к смещению результатов. Если вы изучаете результаты какого-либо опроса, прежде всего выясните, как делалась выборка.

- ✓ **Хорошая выборка делается случайно**. *Случайная* выборка — это выборка, при которой у всех членов целевой совокупности есть равные шансы быть выбранными. Проще всего это себе представить в виде шапки (или корзины), в которую сложены отдельные бумажки с именами всех людей. Если эти бумажки тщательно перемешать, прежде чем вытягивать, то результаты и будут случайной выборкой из целевой совокупности (в данном случае, совокупности людей, чьи имена написаны на бумажках в шапке). Случайная выборка искореняет возможность смещения.

  Уважаемые опросные организации, таки как The Gallup Organization, используют метод случайного набора номеров при обзвоне членов своей выборки. Конечно, в нее не попадают те, у кого нет телефона, но поскольку в большинстве американских семей сегодня есть хотя бы один телефон, то смещение, вызванное отсутствием в выборке людей без телефона, очень мало.

  Остерегайтесь опросов, проведенных с участием выборки большого размера, если эта выборка не была случайной. Опросы в Интернете — самый наглядный пример. Кое-кто может сказать, что для участия в опросе на данный Web-сайт зашло 50 000 человек, а значит, администратор собрал массу информации. Но вся эта информация смещена, потому что представляет точки зрения только тех, кто знал об опросе, решил принять в нем участие и имеет доступ к Интернету. В подобных случаях меньше было бы лучше: организатор опроса должен был опросить меньшее количество людей, но сделать эту выборку случайной.

- ✓ **Хорошая выборка достаточно большая, чтобы результаты были точными.** Если у вас есть выборка большого размера, если она представляет целевую совокупность и была сделана случайно, тогда собранную информацию можно считать довольно точной. Но насколько точной — это зависит от размера выборки, хотя чем больше размер, тем точнее информация (пока эта информация остается хорошей). Точность большинства опросов измеряется в процентах. Этот процент называется *пределом погрешности*

и говорит о том, насколько, с точки зрения исследователя, будут различаться результаты, если много раз повторить опрос в разных выборках того же размера. Подробнее об этом см. главу 10.

Формула для получения быстрой и приблизительной оценки точности опроса — это единица, деленная на корень квадратный из размера выборки. Например, выборка из 1 000 человек (сделанная случайно) точна вплоть до $\frac{1}{\sqrt{1000}} = 0{,}032$ или 3,2 процента. (Заметьте, что если ответы давали не все участники, то размер выборки нужно заменить на количество отвечавших. Подробнее см. раздел "Проверяем, проверяем и еще раз проверяем" далее в этой главе.)

Проведение опроса

Опрос разработан, участники выбраны. Теперь мы переходим непосредственно к процессу опроса, т.е. еще одному важному этапу, на котором может произойти множество ошибок и возникнуть смещение.

Сбор данных

Во время опроса у участников могут возникнуть проблемы с пониманием вопросов, они могут давать ответы, которых нет среди предлагаемых вариантов (если варианты предусмотрены), или же ответы могут быть неточными, а то и явно ошибочными. (В качестве примера таких ошибок третьего рода, когда участники дают неверную информацию, можно привести опрос людей, касающийся того, мошенничали ли они когда-нибудь со своими налоговыми декларациями.) Ошибки третьего рода называются *искажением ответа* — участники дают смещенный ответ.

Некоторые возможные проблемы при сборе данных можно свести к минимуму или упразднить, если тщательно обучать персонал, проводящий опрос. Благодаря правильному обучению любые проблемы, возникающие в ходе опроса, решаются четко и ясно, значит, при сборе данных не происходит никаких ошибок. Проблемы с непонятными вопросами или неполным перечнем возможных ответов можно устранить, если предварительно провести пилотное исследование нескольких участников, а затем, исходя из их реакции, определить, были ли какие-то трудности с вопросами. Персонал также можно обучать создавать обстановку, в которой каждый респондент чувствует себя достаточно защищенным и может говорить правду. Если гарантировать конфиденциальность, это тоже поможет получить ответы большего числа людей.

Проверяем, проверяем и еще раз проверяем

Любой, кто когда-либо отмахивался от опроса или отказывался "ответить на несколько вопросов" по телефону, знает, что заставить людей участвовать в опросе не так уж и просто. Если исследователь хочет свести к минимуму возможное смещение, лучше всего этого можно добиться, если привлечь к участию как можно больше людей. Для этого нужно проверять результаты один, два или даже три раза. Предлагайте денежные призы, купоны, конверты с уже оплаченной маркой и заполненным адресом, возможность выиграть призы и т.д. Важна любая мелочь.

Что может убедить человека принять участие в опросе? Если стимул со стороны исследователя вас не трогает (или если вас не охватывает чувство вины, когда вы взяли две блестящих монеты, но позже выбросили анкету), возможно, вас просто не интересует изучаемая тема. Вот где возникает смещение. Если в опросе участвуют только те, у кого

есть твердое мнение по конкретной проблеме, это значит, что будут учтены только их точки зрения, потому что другие, которым на самом деле наплевать на изучаемый вопрос, не отвечали, и их мнение "Мне все равно" нигде не учитывается. Или же им было не все равно, но у них просто не хватило времени сказать об этом. Как бы там ни было, их мнение остается неучтенным.

Например, предположим, в опросе о том, стоит ли изменить правила выгула собак в парке, участвует 1000 человек. Кто будет отвечать? Скорее всего, респондентами будут те, кто либо активно поддерживает, либо активно противится предлагаемым правилам. Представим себе, что по 100 человек из каждой из этих двух групп были единственными респондентами. Это означает, что 800 мнений не были учтены. Предположим, что ни одному из этих 800 не было никакого дела до изучаемого вопроса. Если бы вы учли их мнение, то результаты выглядели бы так: 800/1000 = 80% "все равно", 100/1000 = 10% в поддержку новых правил и 100/1000 = 10% против новых правил. Но если не учитывать голоса этих 800 респондентов, то исследователи могли бы сказать: "Из числа ответивших 50% выступили за новые правила, а 50% были против них". Впечатление от результатов складывается совсем другое (и крайне смещенное), чем то, что вы получили бы, если бы были учтены все 1000 участников опроса.

Ложные результаты опроса

По результатам исследования, опубликованным в журнале по психологии *Journal of Applied Social Psychology*, был сделан вывод, что если вы лжете кому-то ради самого человека, выслушивающего эту ложь, то ложь становится более приемлемой в социальном отношении, а если ложь говорится в интересах самого лжеца, то такая ложь менее социально приемлема. Звучит интересно и кажется разумным, но может ли это подходить ко всем? Из того, как представлены результаты, похоже, что так. Но если посмотреть, кто же в действительности принимал участие в опросе, давшем такие результаты, у вас сложится впечатление, что выводы, возможно, слегка амбициозны, если не сказать больше.

Авторы начали с опроса 1105 женщин. Из них 659 отказались участвовать, большинство из них сказали, что у них нет времени. Еще 233 оказались, с точки зрения исследователей, либо "слишком молодыми", либо "слишком старыми", а 33 не подошли для опроса, потому что исследователи столкнулись с языковым барьером. В конечном итоге было опрошено 180 женщин. Средний возраст в финальной группе участников был равен 34,8 лет.

Ну что ж, с чего начать? Исходный размер выборки, 1105 человек, был достаточно большим, но случайно ли они выбирались? Заметьте, что вся выборка состояла из женщин, что интересно, ведь из результатов исследования нельзя сделать выводы о том, что в подобных ситуациях ложь более приемлема для женщин. Затем, 659 из отобранных для участия отказались сотрудничать (что вызывает смещение). Это большой процент (60%), но, учитывая тему исследования, вас это удивлять не должно. Исследователи должны были свести проблему к минимуму, например, гарантировав, что все ответы будут анонимными. Неизвестно, делалась ли последующая проверка (и это, скорее всего, говорит о том, что ее не было). Отбрасывать 233 человека потому, что они "слишком молоды" или "слишком стары", в корне неверно, если, конечно, ваша целевая совокупность не ограничивается определенной возрастной группой. Если бы это было так, то выводы нужно было делать только об этой возрастной группе. Наконец, последняя капля: вычеркнуть 33 человека, которые оказались "неподходящими" (по словам исследователей) из-за языкового барьера. Можно было бы пригласить переводчика, ведь выводы были сделаны не только о тех, кто говорит по-английски. Нельзя, чтобы участники опроса представляли крошечный микрокосм общества (молодые женщины, владеющие английским), а вы, исходя из данных об этом крошечном микрокосме, делали выводы обо всем обществе. Начинать с выборки размером в 1105 человек, а закончить всего 180 женщинами — это грубейшая ошибка.

Доля ответивших в опросе — это отношение, которое находится делением количества респондентов на количество людей, которые были изначально выбраны для участия в исследовании. Статистики считают, что хорошая доля ответивших — это где-то около 70%. Но во многих случаях доля ответивших и не приближается к этому значению, если только опрос не проводит уважаемая организация, например, The Gallup Organization. Изучая результаты опроса, выясните долю ответивших. Если она слишком низкая (намного меньше 70%), то результаты могут оказаться смещенными, и на них можно не обращать внимания. Не обманитесь, если вам говорят, что в опросе участвовало много респондентов, но на самом деле доля ответивших была низкой. В таком случае, возможно, ответили многие, но еще больше, наверное, отказались отвечать.

Обратите внимание, что во многих статистических формулах (в том числе и в тех, что описаны в этой книге) предполагается, что размер вашей выборки равен количеству респондентов, потому что статистики хотят донести до вас, как важно не довольствоваться смещенными данными из-за отказов отвечать. Однако в действительности статистики знают, что не всегда удается убедить каждого участвовать в опросе, как бы вы ни старались. Значит, какое число вы примете в качестве *n* в формулах: исходный размер выборки (количество людей, с которыми вступили в контакт) или реальный размер выборки (количество ответивших)? Нужно использовать количество ответивших. Но заметьте, что в случае с опросом с низкой долей ответивших результаты нельзя обнародовать, потому что они, вполне вероятно, будут смещенными. Вот как важно провести проверку! (Говорят ли вам об этом, когда сообщают результаты? Очень редко.)

В том, что касается качества результатов, намного лучше изначально сделать выборку меньшего размера и тщательнее проверить то, что получится, чем выбирать большую группу потенциальных респондентов, а получить низкую долю ответивших.

Интерпретация результатов

Цель любого опроса — получить информацию о целевой совокупности. Эта информация может включать в себя мнения, демографические данные, стиль жизни или модели поведения. Если опрос был разработан и проведен объективно и аккуратно, при этом исследователи не забывали о поставленных целях, то данные должны содержать полную информацию о том, что происходит с целевой совокупностью (в рамках указанного предела погрешности). Следующий этап — организация данных для того, чтобы получить четкое представление о том, что происходит, анализ данных для установления взаимосвязей, расхождений или других интересующих отношений, и, наконец, выводы, основанные на полученных результатах.

Организация и анализ

После завершения опроса необходимо организовать и проанализировать данные (другими словами, кое-что посчитать и построить некоторые графики). На основе данных, полученных в ходе опроса, можно создать разные виды изображений и вычислить многие статистические показатели, в зависимости от собранной информации. (Числовые данные, например, доход, отличаются по своим характеристикам от дискретных данных, таких как пол, поэтому, как правило, представляются иначе.) Подробнее о том, как можно организовать и подытожить данные, см. главы 4 и 5 (соответственно). В зависимости от изучаемого вопроса, данные можно анализировать по-разному, в том числе делать предположения о показателях совокупности, проводить проверку, касаю-

щуюся совокупности, или искать взаимосвязи. Это всего некоторые из видов анализа. Подробнее о каждом из них см. главы 13, 15 и 18 соответственно.

Избегайте графиков и статистических показателей, сбивающих с толку. Не все данные опросов организованы и проанализированы правильно и объективно. Подробнее о проблемах со статистикой см. главу 2.

Делаем выводы

Выводы — это самое интересное в любом опросе, именно ради них исследователи и начинают всю затею. Если опрос был тщательно разработан и проводился правильно, выборка делалась продуманно, а данные были организованы и подытожены без ошибок, то результаты будут объективно и точно отражать действительную ситуацию в целевой совокупности. Но даже если опрос был проведен правильно, то исследователи могут неверно истолковать или преувеличить значимость результатов, т.е. скажут больше, чем должны. Знаете, как говорят: “Лучше один раз увидеть, чем сто раз услышать”? Некоторые исследователи виновны в противоположном, т.е. “Лучше один раз услышать, чем сто раз увидеть”. Другими словами, они утверждают, что видели в результатах то, во что сами хотели бы поверить. Поэтому нужно знать, где проходит грань между разумными выводами и ошибочными результатами, и научиться замечать, когда кто-то переходит эту черту.

Вот некоторые самые распространенные ошибки, которые можно допустить при подведении итогов опроса.

- ✓ Сделать предположение по поводу большей совокупности, чем была в действительности представлена в ходе исследования.
- ✓ Утверждать, что существует разница между двумя группами, хотя на самом деле такой разницы нет.
- ✓ Сказать, что “эти результаты ненаучны, но...”, а затем говорить о них, как будто они имеют научную ценность.

Анонимность или конфиденциальность

Если бы вам нужно было провести опрос, чтобы выяснить, насколько активно люди пользуются электронной почтой на работе, то доля ответивших могла бы представлять собой проблемы, потому что многие не хотят обсуждать вопрос личной переписки на рабочем месте или, по крайней мере, говорят не очень откровенно. Можно было бы подтолкнуть их к искренности, если сообщить, что во время и после проведения опроса их частная жизнь останется секретом.

Сообщая результаты опроса, исследователи, как правило, не связывают собранную информацию с именами респондентов, потому что тем самым они вмешались бы в их частную жизнь. Возможно, вам уже встречались такие понятия, как “анонимный” и “конфиденциальный”. Эти два слова имеют совершенно разный смысл.

Если результаты *конфиденциальны*, то это значит, что я могу связать информацию о вас с вашим именем в своем отчете, но обещаю не делать этого. Если результаты *анонимны*, это значит, что у меня нет никакой возможности связать информацию с именем, даже если бы я и захотела.

Если вас просят принять участие в опросе, выясните, что исследователь собирается делать с вашим ответом и будет ли ваше имя каким-то образом связано с опросом. (В хороших опросах это всегда объясняется участникам с самого начала.) Только после этого решайте, стоит ли соглашаться на участие в исследовании.

Чрезмерный восторг?

В 1998 году поисковая система Excite выпустила пресс-релиз, в котором утверждалось, что их сайт был назван самым популярным по результатам исследования, проведенного исследовательской организацией Intelliquest с помощью издания *USA Today*. Опрос проводился с участием 300 пользователей Сети, выбранных из группы в 30 000 специалистов по проверке технологии, работавших на Intelliquest. (Обратите внимание, что это *не* случайная выборка пользователей Сети!) В выводах говорилось, что потребители назвали Excite самым опытным сайтом, тем самым сделав его самым популярным в Интернете, благодаря чему эта поисковая система обошла даже Yahoo! и других конкурентов.

Excite утверждала, что, исходя из результатов этого опроса, она лучше Yahoo!. Но реальные результаты говорят о другом. По шкале от 0 до 100% средний балл за качество обслуживания клиентов у Excite был равен 89%, а у Yahoo! — 87%. Несомненно, результат Excite хорош, к тому же, он немного выше, чем у Yahoo!, но разница между баллами, набранными двумя компаниями, на самом деле находится в рамках предела погрешности для данного исследования, который равен плюс/минус 3,5%. Другими словами, с точки зрения статистики, и Excite, и Yahoo! заняли первое место. Поэтому в данном случае нельзя сказать, какая компания победила. (Подробнее об изменчивости выборки и пределе погрешности см. главы 9 и 10.)

Чтобы при подведении итогов не допустить самых распространенных ошибок, поступайте следующим образом.

1. **Проверьте, правильно ли была сделана выборка и не выходят ли выводы за пределы совокупности, представленной данной выборкой.**
2. **По возможности, прежде чем читать результаты, ознакомьтесь с правовыми оговорками по данному опросу.**

   Так на вас меньше повлияют результаты, о которых вы читаете, если они не основаны на научном методе. Теперь, когда вы знаете, что такое научный опрос (так в средствах массовой информации называется точный и объективный опрос), вы можете использовать эти критерии, чтобы самостоятельно решить, стоит ли доверять результатам.
3. **Не забывайте о статистически ошибочных выводах.**

   Если вам сообщают о разнице между двумя группами по результатам опроса, обязательно убедитесь, что заявленная разница больше, чем предел погрешности. Если разница находится в рамках предела погрешности, то можно ожидать, что результаты опроса были получены случайно, а так называемую "разницу" нельзя отнести ко всей совокупности. (Подробнее об этом см. главу 14.)
4. **Не слушайте тех, кто говорит: "Результаты не научны, но..."**

Вот что самое главное в опросах: помните о недостатках и не учитывайте информацию, которая поступает в ходе опросов, где не такие недостатки не обращали внимания. Провести плохой опрос легко и дешево, но вы получаете то, за что платите. Прежде чем изучать результаты какого-либо опроса, выясните, как он был разработан и проведен, чтобы можно было оценить качество полученных результатов.

Глава 17

Эксперименты: открытия в медицине или результаты, вводящие в заблуждение?

В этой главе...

- Какие недостатки бывают у экспериментов
- Как проводятся эксперименты
- Результаты, сбивающие с толку

Похоже, в наш век информации открытия в медицине появляются и исчезают очень быстро. Сегодня вы слышите о многообещающем новом методе лечения какого-то заболевания, а чуть позже узнаете, что на последнем этапе тестирования лекарство не оправдало возложенных на него надежд. Фармацевтические компании забрасывают телезрителей рекламными роликами таблеток, после чего миллионы людей отправляются на прием к врачу и требуют новейшие и самые действенные лекарства от своих болезней, иногда даже не зная, для чего предназначены эти препараты. В Интернете любой желающий найдет подробности какого угодно метода лечения, описания болезней или их симптомов, а также обнаружит массу информации и советов. Но насколько всему этому можно верить? Как решить, какие варианты лучше всего подходят именно для вас, если вы заболели, нуждаетесь в операции или вам срочно нужна медицинская помощь?

В данной главе вы многое узнаете об экспериментах, этой движущей силе медицинских исследований и других расследований, при которых нужно делать сравнения — например, сравнивать, какой строительный материал самый лучший, какой безалкогольный напиток предпочитают подростки, какой джип будет самым безопасным в случае аварии и т.д. Вы узнаете, в чем разница между экспериментами и исследованием по данным наблюдений, а также выясните, какие эксперименты подойдут для вас, как они должны проводиться, какие в них совершаются ошибки и как можно заметить результаты, вводящие в заблуждение. Мы ежедневно сталкиваемся с огромным количеством новостей, сообщений и "профессиональных советов", поэтому нужно научиться использовать все имеющиеся навыки критического мышления, чтобы разобраться с подчас противоречивой информацией.

Отличительные черты экспериментов

Хотя существует множество разнообразных видов исследований, но все они сводятся, по сути, к двум разновидностям: экспериментам и исследованиям по данным наблю-

дений. В этом разделе вы узнаете о том, чем же именно эксперименты отличаются от других видов исследований.

Исследование по данным наблюдений — это исследование, в котором ученый наблюдает за испытуемыми и записывает информацию. Не происходит никакого вмешательства, никаких изменений, не навязываются никакие ограничения или контроль. *Эксперимент* — это исследование, при котором испытуемые не просто наблюдаются в естественном состоянии, но и намеренно подвергаются определенному воздействию под постоянным контролем, а все полученные результаты записываются.

Рассмотрим эксперименты

Основная задача эксперимента — выяснить, вызовет ли определенное воздействие изменения в реакции. (Здесь часто используется слово "причина".) В ходе эксперимента это делается путем создания контролируемой ситуации — контролируемой настолько, что исследователь может определить, вызывает ли какой-либо фактор или комбинация факторов изменения в реакции, а если да, то в какой степени этот фактор (или комбинация факторов) влияет на реакцию.

Например, чтобы получить разрешение на новое лекарство, представители фармацевтической компании проводят эксперименты и выясняют, помогает ли это лекарство снизить артериальное давление, какая доза лучше всего подходит для определенных совокупностей пациентов, какие побочные эффекты могут наблюдаться (если они есть) и в какой степени эти побочные эффекты характерны для каждой совокупности.

Рассмотрим исследования по данным наблюдений

В определенных ситуациях оптимальный вариант работы — это исследования по данным наблюдений. Наиболее распространенным их видом являются опросы (см. главу 16). Если задача заключается лишь в том, чтобы определить, что думают люди, и собрать некоторую демографическую информацию (такую как пол, возраст, доход и т.д.), то нет ничего лучше опросов, если они разработаны и проведены правильно.

В других ситуациях, особенно если необходимо установить причинно-следственную взаимосвязь (речь о которой подробнее пойдет в главе 18), исследования по данным наблюдений не подходят. Например, предположим, на прошлой неделе вы приняли пару таблеток витамина С — именно благодаря этому вам удалось избежать простуды, которая свирепствует в офисе. Возможно, несколько дополнительных часов сна или лишний раз вымытые руки уберегли вас от болезни? Или же вам просто повезло? Когда смешано столько переменных, как можно выяснить, какая именно из них повлияла на исход ситуации, и вы не простудились?

Изучая результаты исследования, сначала определите, какова была его цель и подходит ли для ее достижения выбранный тип исследования. Например, если для установления причинно-следственных взаимосвязей вместо эксперимента было проведено исследование по данным наблюдений (см. главу 18), любые сделанные выводы нужно принимать с опаской.

Не забываем об этике

Проблема с экспериментами в том, что некоторые из них разработаны не совсем этично. Вот почему нужно было собрать так много доказательств того, что курение вызывает рак легких, и почему компании, производящие сигареты, только в последнее вре-

мя начали выплачивать огромные компенсации жертвам. Нельзя заставлять испытуемых курить, чтобы проверить, что с ними произойдет. Вы можете изучать только тех, у кого обнаружен рак легких, и попытаться докопаться до того, какие *факторы* (изучаемые переменные) могли вызвать это заболевание. Но поскольку невозможно контролировать все интересующие вас факторы — да и любую переменную, если уж на то пошло, — в случае с исследованиями по данным наблюдений очень сложно выделить одну конкретную причину.

Хотя причины рака и других болезней нельзя установить этично, проводя эксперименты на людях, лечение рака можно проверить с помощью экспериментов. Медицинские исследования, для которых нужны эксперименты, называются *клиническими испытаниями*. Подробнее см. сайт `http://www.clinicaltrials.gov`.

Опросы и другие исследования по результатам наблюдений подходят, если вы хотите узнать мнение людей, изучить их стиль жизни, не вмешиваясь в него, или выяснить некоторые демографические переменные. Но если вы стремитесь выяснить причину определенного исхода или поведения (т.е. почему что-то происходит), намного лучше для этого подойдет эксперимент. Если провести эксперимент невозможно (потому что это неэтично, слишком дорого или недостижимо), тогда его можно заменить большим количеством исследований по результатам наблюдений, в ходе которых изучались разные факторы и были сделаны такие же выводы. (Подробнее о причинно-следственных связях см. главу 18.)

Разработка хорошего эксперимента

От того, как эксперимент был разработан, зависит качество полученных результатов. Большинство исследователей обязательно напишут о своих экспериментах самые хвалебные пресс-релизы, поэтому нужно уметь разбираться во всех деталях, чтобы понять, верить ли тому, что вам говорят.

Чтобы определить, надежен ли эксперимент, проверьте, соблюдались ли при его выполнении следующие шаги.

1. **Была сделана выборка достаточно большого размера, чтобы результаты оказались точными.**
2. **В качестве испытуемых были выбраны люди, которые наиболее точно представляют целевую совокупность.**
3. **Испытуемые были в случайном порядке распределены по исследуемой и контрольной группам.**
4. **Осуществлялся контроль над возможными внешними переменными.**
5. **Для того чтобы избежать смещения, проводился двойной слепой эксперимент.**
6. **Были собраны хорошие данные.**
7. **Был проведен правильный анализ данных.**
8. **Сделанные выводы не выходят за пределы проведенного исследования.**

В последующих разделах каждый из этих критериев будет рассмотрен подробнее.

Выбираем размер выборки

Размер выборки в значительной степени влияет на точность результатов. Чем больше выборка, тем точнее результаты и тем мощнее статистические критерии (т.е. они смогут обнаружить реальные результаты, если те существуют). Подробнее см. главы 10 и 14.

Маленькие выборки не подходят для серьезных выводов

Вас может удивить то количество статей, которые были написаны об исследованиях, проведенных на очень маленьких выборках. Это очень насущный вопрос для статистиков, которые знают, что для того, чтобы выявить самые заметные расхождения между группами, потребуются выборки большого размера (по крайней мере, больше 30; см. главу 10). Если выборки были маленькие, а сделанные на их основе выводы очень серьезны, то либо исследователи использовали неправильную проверку гипотезы при анализе данных (т.е. вместо *Z*-распределения им нужно было работать с *t*-распределением; см. главу 14), либо разница оказалась настолько заметной, что выборки небольшого размера вполне хватило для ее обнаружения. Но второй вариант все же бывает нечасто.

Остерегайтесь выводов, при которых значительные результаты основываются на маленьких выборках (особенно выборках меньше 30 элементов). Если результаты важны для вас, попросите копию исследовательского отчета и выясните, как именно анализировались данные. Кроме того, проанализируйте выборку испытуемых, чтобы понять, действительно ли она представляет совокупность, о которой исследователи делают свои выводы.

Что такое "размер выборки"

Задавая вопросы о размере выборки, конкретно объясните, что вы понимаете под этим термином. Например, вы можете спросить, сколько именно испытуемых было выбрано для участия в исследовании, а также можете поинтересоваться, сколько человек завершили эксперимент. Эти два числа могут оказаться совершенно разными. Убедитесь, что исследователи могут объяснить, почему каждый конкретный участник решил бросить или не смог (по какой-либо причине) закончить эксперимент.

Например, в статье под названием "Марихуана названа эффективным средством при лечении рака" из газеты *New York Times* в первом же абзаце говорится, что марихуана "намного эффективнее" любого другого лекарства в борьбе с побочным действием химиотерапии. Но, поинтересовавшись подробностями, вы узнаете, что результаты основаны на исследовании всего 29 пациентов (15 принимали лекарство, а 14 — плацебо). Что еще больше сбивает с толку, оказывается, что только 12 из 15 пациентов изучаемой группы дошли до конца эксперимента. Так что же случилось с оставшимися тремя?

Иногда исследователи делают выводы, основываясь только на тех испытуемых, которые закончили эксперимент. Это может вводить в заблуждение, потому что данные не содержат информации о тех, кто отказался от участия (неизвестно почему) в исследовании. Таким образом данные могут оказаться смещенными. Подробнее о размере выборки, который нужен для достижения определенного уровня точности, см. главу 12.

Точность — это не единственный критерий "хороших" данных. Нужно позаботиться о том, чтобы избежать смещения, для чего необходима случайная выборка (подробнее о том, как делаются случайные выборки, см. главу 3.)

Выбор участников

Первый этап при разработке эксперимента — это выборка участников, которых исследователи называют *испытуемыми*. Хотя исследователи хотели бы, чтобы испытуемые выбирались случайно из совокупностей, которые они представляют, но в большинстве случаев это недостижимо. Например, предположим, группа офтальмологов хочет проверить новый метод лазерной терапии на близоруких пациентах. Им нужна случайная выборка испытуемых, поэтому они наугад выбирают пациентов с близорукостью по записям врачей. Они обзванивают всех и говорят: "Мы экспериментируем с новым методом лазерной хирургии для лечения близорукости, и вас наугад выбрали для участия в исследовании. Когда вы можете прийти на операцию?".

Очевидно, что такой подход будет не слишком успешным при работе с очень многими пациентами (хотя некоторые, возможно, ухватятся за этот шанс, особенно если им не придется платить за операцию). Суть в том, что получить действительно случайную выборку участников эксперимента обычно намного сложнее, чем сделать случайную выборку участников опроса.

Случайное распределение испытуемых по группам

После того, как выборка была сделана, исследователи распределяют испытуемых по разным группам. Это будет одна или несколько *исследуемых групп*, в которых изучаются разные дозировки лекарства или особенности нового метода лечения, и *контрольная группа*, участники которой не подвергаются никакому лечению или принимают ненастоящее лекарство.

Важность случайного распределения по группам

Предположим, исследователь хочет определить, какое воздействие оказывают физические упражнения на сердечный ритм. Испытуемые в исследуемой группе пробегают пять миль, при этом сердечный ритм у них измеряется до и после пробежки. Испытуемые в контрольной группе все время сидят на диване и смотрят повтор сериала "Симпсоны". В какую группу вы лично хотели бы попасть? Если вы не помешаны на идее пробежать пять миль, то, наверное, предпочли бы более простой вариант и добровольно изъявили бы желание поваляться на диване. (А может, вы так ненавидите "Симпсонов", что лучше пробежали бы пять миль, только бы не смотреть очередную серию?) Как добровольное распределение участников по группам повлияло бы на результаты исследования?

Набор добровольцев производит побочный эффект

Чтобы найти участников для эксперимента, исследователи часто приглашают добровольцев и предлагают им такие стимулы, как деньги, бесплатное лечение или последующее медицинское обслуживание. Проводить медицинские исследования на людях очень сложно, но необходимо, чтобы действительно узнать, эффективно ли лечение, какой должна быть дозировка лекарства и каковы побочные эффекты. Чтобы в правильном количестве выписывать новое лекарство в реальной жизни, врачам и пациентам нужно, чтобы такие исследования можно было обобщить для всей совокупности. Чтобы привлечь таких представительных участников, исследователям приходится проводить широкую рекламную кампанию и отбирать достаточное количество испытуемых, которые бы представляли самые разные слои совокупности. В будущем именно для самых разных слоев совокупности будет выписываться новое лекарство.

Если попасть в исследуемую группу захотят только люди, помешанные на здоровье (у которых, возможно, и так идеальный сердечный ритм), то исследователь будет изучать влияние физических упражнений (пробежка в пять миль) только на здоровых и активных людях. Он не узнает, как это упражнение повлияет на сердечный ритм лентяев. Такое неслучайное распределение испытуемых по группам оказало бы огромное воздействие на выводы, которые исследователь сделает после завершения эксперимента.

Чтобы избежать серьезного смещения в результатах эксперимента, испытуемые должны быть распределены по группам в случайном порядке, а не сами выбирать, в какую группу хотят попасть. Помните об этом, оценивая результаты эксперимента.

Контролируем эффект плацебо

Используя ненастоящее лекарство, исследователи берут в расчет так называемый эффект плацебо. *Эффект плацебо* — это реакция, которая появляется у людей (или им кажется, что она появляется) просто потому, что они убеждены, что получают определенное "лекарство" (даже если оно ненастоящее, скажем, сахарные пилюли). Подробнее об эффекте плацебо см. главу 3.

Увидев рекламу лекарства в журнале, ищите подробности. Мелким шрифтом вы можете увидеть напечатанную таблицу, в которой перечислены побочные эффекты, наблюдавшиеся у участников исследуемой группы, по сравнению с побочными эффектами, отмеченными у контрольной группы. Если контрольная группа принимает плацебо, то вам может показаться, что они не должны испытывать никаких побочных эффектов, но это не так. Группы, принимающие плацебо, часто заявляют о наличии побочных эффектов, причем процентные отношения таких случаев довольно высоки. Это объясняется тем, что мозг человека играет с ним шутку, и все эти участники подвергаются эффекту плацебо. Если вы хотите объективно оценить отмеченные побочные эффекты лекарства, тогда необходимо учитывать и те побочные эффекты, о которых заявляет контрольная группа — побочное действие, которое вызвано исключительно эффектом плацебо.

В некоторых ситуациях, например, когда испытуемые страдают от очень серьезного заболевания, предлагать им ненастоящее лекарство было бы неэтично. В 1997 году американское правительство подверглось жесткой критике за финансирование исследования ВИЧ, в ходе которого изучалась дозировка азидотимидина. В то время считалось, что этот препарат на две третьих снижает риск передачи ВИЧ от беременной женщины ребенку. Данное исследование, в котором участвовали 12 000 беременных женщин в Африке, Таиланде и Доминиканской республике, было разработано совершенно неправильно. Исследователи давали половине женщин разные дозировки азидотимидина, а вторая половина участниц получала сахарные пилюли. Конечно, если бы американское правительство знало, что половина испытуемых получает плацебо, оно бы не стало финансировать это исследование ВИЧ.

Вот как нужно поступать в подобных ситуациях. Когда использование ненастоящего лекарства вызывает этические проблемы, новое лекарство сравнивают с уже привычным, которое считается эффективным в лечении конкретного заболевания. После того, как исследователи соберут достаточно данных, чтобы можно было установить, какое ле-

карство действует лучше, они прекращают эксперимент и переводят всех участников на лучший вариант лечения, снова же по этическим причинам.

Изучая результаты эксперимента, убедитесь, что ученые сравнивали исследуемую группу с контрольной, чтобы можно было наверняка сказать, что участники этих двух групп испытали на себе разное воздействие препарата. Контрольная группа может получать либо ненастоящее лекарство, либо стандартный вариант лечения, в зависимости от ситуации.

Не забываем о внешних переменных

Представим себе, что вы участвуете в исследовании, в ходе которого нужно установить, какие факторы повлияли на то, простудились вы или нет. Если исследователи фиксируют только тот факт, была ли у вас простуда за определенный период времени, и интересуются вашим поведением при этом (сколько раз в день вы мыли руки, сколько часов спали и т.д.), то это исследование по результатам наблюдений. Проблема такого типа исследований состоит в том, что, не контролируя другие факторы, которые могли оказать воздействие, исследователи не смогут выяснить, какие именно ваши действия (если вообще такие были) на самом деле позволили получить конкретный исход.

Самый большой недостаток исследований по результатам наблюдений заключается в том, что они не выявляют реальных причинно-следственных взаимосвязей из-за наличия того, что в статистике называется внешними переменными. *Внешняя переменная* — это переменная или фактор, которые не контролировались в ходе исследования, но могли повлиять на результат.

Например, статья имеет сенсационный заголовок: “Исследование гласит: чем старше мать, тем дольше ее продолжительность жизни”. В первом абзаце рассказывается о том, что женщины, родившие первого ребенка после 40, имеют намного больше шансов прожить до 100 лет, чем те, кто впервые родили в более молодом возрасте. Но если выяснить подробности исследования (проведенного в 1996 году), то вы узнаете, что, прежде всего, в нем участвовали всего 78 женщин из пригорода Бостона, которые дожили до 100 лет, и их данные сравнивались с 54 женщинами, родившимися в то же время (1896 год), но умершими в 1969 году (это был самый давний период, за который исследователи смогли найти данные в компьютере). Эта так называемая “контрольная группа” прожила ровно 73 года, не больше и не меньше. Из тех женщин, что дожили до 100 лет, 19% родили первого ребенка после 40, а среди тех, кто умерли в 73 года, в таком возрасте рожали всего 5,5%.

Сделанные в данном случае выводы просто поражают. А как же тот факт, что в “контрольную группу” входили только матери, которые умерли в 1969 году в возрасте 73 лет? Как же все остальные матери, умершие до 73 или в возрасте от 73 до 100? Возможно, контрольная группа (будучи настолько ограниченной) включала женщин, между которыми существовала некая связь. Может быть, именно эта связь стала причиной их смерти в одном и том же году, и, возможно, именно эта связь объясняла, почему большинство из них рожали первого ребенка в юном возрасте. Кто знает? А другие переменные, которые могли повлиять как на возраст рожениц, так и на продолжительность их жизни, — такие переменные, как финансовое положение, стабильность в семье или другие социально-экономические факторы? В период Великой депрессии женщинам, участвовавшим в этом исследовании, было по 33 года, это тоже могло повлиять на продолжительность их жизни или на то, когда они родили детей.

Как исследователи решают проблему внешних переменных? Главное слово здесь — "контроль". Они контролируют все внешние переменные, которые могут предугадать. В экспериментах с людьми исследователям приходится иметь дело с множеством внешних переменных. Например, в исследовании вопроса о том, как влияют разные виды и громкость музыки на количество времени, которое покупатели проводят в бакалейных магазинах (да, они думают и об этом), исследователям приходится заранее предугадать многие внешние переменные, а затем проконтролировать их. Какие еще факторы, помимо громкости и вида музыки, могли бы повлиять на продолжительность пребывания в магазине? Мне на ум приходят несколько: пол, возраст, время суток, есть ли со мной дети, сколько у меня денег, день недели, насколько чисто и уютно в магазине, насколько любезны продавцы, и (самое важное) какой у меня мотив — закупить продукты на всю неделю или приобрести всего одну пачку печенья?

Как можно проконтролировать так много внешних переменных? Некоторые — еще при разработке исследования, например, время суток, день недели и причину для похода в магазин. Но другие факторы (скажем, восприятие обстановки в магазине) зависят исключительно от участников исследования. Идеальный метод контроля над всеми этими индивидуальными внешними переменными — использовать пары людей, которые были составлены согласно важным переменным, или же просто использовать одного человека дважды: один раз с музыкой, второй — без нее. Такой вид эксперимента называется *метод согласованных пар*. (Подробнее об этом см. главу 14.)

Прежде чем поверить какому-либо сенсационному заявлению о прорыве в медицине (да и вообще любому заявлению), выясните, как проводилось исследование. Исследования по результатам наблюдений не контролируют внешние переменные, поэтому их результаты не настолько статистически значимые (что бы там ни говорили статистические показатели), как результаты хорошо разработанного эксперимента. В тех случаях, когда эксперимент провести нельзя (ведь никто не может заставить вас рожать ребенка до или после 40), убедитесь, что исследование по результатам наблюдений основано на данных достаточно большой выборки, которая представляет широкие слои совокупности.

Двойной слепой эксперимент

Хорошо разработанные эксперименты — это двойные слепые эксперименты. *Двойной слепой* означает, что ни испытуемые, ни исследователи не знают, кто находится в исследуемой группе, а кто — в контрольной. Испытуемые не должны знать, какое лекарство получают, чтобы исследователи могли выявить эффект плацебо. Но почему же и исследователям нельзя знать, кто какое лекарство принимает испытуемый? Чтобы они не относились к испытуемым по-разному, ожидая (или не ожидая) от конкретных групп определенной реакции. Например, если исследователь знает, что вы относитесь к исследуемой группе, в которой изучаются побочные эффекты лекарства, он может ожидать, что вы почувствуете тошноту, и будет уделять вам больше внимания, чем если бы вы были участником контрольной группы. В результате данные могут получиться смещенными, а результаты — вводить в заблуждение.

Если исследователь знает, кто какое лекарство получает, а испытуемым это не известно, то исследование называется *слепым* (а не двойным слепым). Слепые исследования лучше, чем ничего, но самый хороший вариант — это двойной слепой эксперимент. Интересно: если эксперимент двойной слепой, то хоть кто-нибудь знает, какие испы-

туемые какое лекарство принимают? Конечно! Обычно за это отвечает третья сторона — еще один лаборант.

Анализируя эксперимент, выясните, был ли он двойным слепым. Если нет, результаты могут оказаться смещенными.

Сбор хороших данных

Какие данные можно назвать "хорошими"? Для оценки качества данных статистики используют три критерия, каждый из которых действительно является мощным измерительным прибором, который применяется в процессе сбора данных.

Чтобы решить, хорошие ли данные были получены в ходе эксперимента, поинтересуйтесь такими характеристиками.

- ✓ **Надежность (если провести несколько замеров, результаты останутся прежними).** Очень часто весы в ванной показывают неправильную информацию. Вы становитесь на них и видите одно число. Вы ему не верите, поэтому сходите с весов, потом возвращаетесь на них и получаете новое число. (Если во второй раз оно меньше, то вы, скорее всего, на этом и остановитесь. Если же нет, то вы будете продолжать до тех пор, пока не увидите число, которое вас удовлетворит.)

  Другими словами, ненадежные данные получаются из-за использования ненадежных измерительных приборов. Такими измерительными приборами могут быть не только весы, а и менее конкретные вещи, например, вопросы опроса, которые дают ненадежные результаты, если сформулированы двусмысленно (подробнее об этом см. главу 16).

  Изучая результаты эксперимента, выясните, как собирались данные. Если измерения не были надежными, то данные могут оказаться неточными.

- ✓ **Объективность (в данных нет систематических ошибок, которые каким-либо образом влияют на полученные величины).** Смещенные данные — это данные, которые систематически переоценивают или недооценивают реальный результат. Смещение может произойти практически на любом этапе разработки или проведения исследования. Его причиной может быть плохой измерительный прибор (как весы в ванной, которые "всегда" показывают на пять фунтов больше), вопросы, которые наталкивают участников на нужный ответ, или исследователи, которые знают, какое лечение получает испытуемый, и заведомо ожидают от него определенной реакции.

  Смещение — самая главная проблема с данными. И что еще хуже, его никак нельзя измерить. (Например, предел погрешности не измеряет смещение. Подробнее о пределе погрешности см. главу 10.) Но все же можно предпринять определенные шаги, чтобы свести к минимуму смещение, о чем уже говорилось в главе 16 и в разделе "Случайное распределение испытуемых по группам" чуть выше в этой главе.

  Помните, что на любом этапе разработки и проведения любого исследования может случиться смещение, поэтому оценивайте результаты, не упуская никаких признаков смещения. Если исследование было в значительной степени смещено, то вы можете проигнорировать результаты.

- ✓ **Достоверность (данные измеряют то, что и должны измерять).** Чтобы проверить достоверность данных, вам нужно отойти немного в сторону и взглянуть на всю картину целиком. Спросите себя: измеряют ли эти данные то, что должны измерять? Или же исследователям нужно было собирать совершенно другие данные? Важно также, подходит ли для определенной ситуации выбранный измерительный прибор. Например, если спросить учеников, какой балл по математике у них был в школе, это не будет достоверным критерием оценки реальных баллов. Более достоверным мерилом была бы проверка аттестата каждого ученика. Недостоверным будет и определение ситуации с преступностью только на основе количества преступлений, в данном случае нужно использовать *уровень преступности* (количество преступлений на душу населения).

Прежде чем принимать результаты эксперимента, выясните, какие данные как именно измерялись. Убедитесь, что исследователи собирали именно те данные, которые нужны для достижения целей исследования.

Правильный анализ данных

После того как данные были собраны, их помещают в волшебный ящик, который называется *статистический анализ*. Выбор аналитического метода не менее важен (в том, что касается качества результатов), чем любой другой аспект исследования. Правильный анализ нужно планировать заранее, на этапе разработки эксперимента. Тогда после сбора данных у вас не возникнет серьезных проблем с анализом.

Как же нужно выбирать правильный аналитический метод? Спросите себя: "После того, как данные будут проанализированы, смогу ли я ответить на вопрос, который меня интересовал с самого начала?" Если ответ отрицательный, тогда тип анализа выбран неверно.

Основные виды статистического анализа — это доверительные интервалы (используются, когда вы пытаетесь оценить значение совокупности или разницу между двумя значениями совокупности), проверка гипотезы (используется, если вы хотите проверить утверждение об одной или двух совокупностях, например, о том, что одно лекарство эффективнее другого), анализ корреляции и регрессии (используется, когда вы хотите показать, может ли одна переменная вызвать изменения в другой переменной). Подробнее о каждом из этих видов анализа см. главы 13, 15 и 18 соответственно.

Выбирая, каким образом вы будете анализировать данные, проследите за тем, чтобы данные и выбранный вид анализа были совместимы. Например, если вы хотите сравнить исследуемую и контрольную группы в том, насколько новая диета эффективна для снижения веса (по сравнению с уже существующей), то вам нужно собрать данные о том, какой вес потерял каждый человек (а не только выяснить, сколько он весит в конце эксперимента).

Делаем правильные выводы

Самые грубые ошибки, которые допускают исследователи при формулировке выводов, это:

- ✓ преувеличение результатов;
- ✓ создание взаимосвязей или озвучивание объяснений, которые не подтверждаются статистикой;
- ✓ отнесение полученных результатов к тем совокупностям, которые не были изучены в ходе исследования.

Подробнее эти проблемы рассматриваются далее.

Преувеличение результатов

Очень часто в средствах массовой информации результаты подаются намного сенсационнее, чем они есть на самом деле. Прочитав заголовок или услышав сообщение об исследовании, обязательно выясните, как оно проводилось и какие именно выводы были сделаны.

В пресс-релизах результаты тоже часто преувеличиваются. Например, в одном пресс-релизе, выпущенном Национальным институтом изучения наркомании, исследователи утверждали, что уровень употребления экстази в 2002 году был ниже, чем в 2001 году. Однако, изучив действительные статистические результаты, приведенные в отчете, вы узнаете, что в 2002 году в исследуемой выборке было меньшее процентное отношение подростков, признавшихся в употреблении экстази, чем в 2001 году, но эту разницу не сочли статистически значимой (в отличие от многих других результатов). Это значит, что хотя в 2002 году в сделанной выборке было меньше подростков, употреблявших экстази, эта разница не может объяснить изменение переменной в разных выборках иначе, чем простой случайностью. (Подробнее о статистической значимости см. главу 14.)

Заголовки и вводные абзацы в пресс-релизах и газетных статьях часто преувеличивают реальные результаты исследований. Сенсационные результаты, яркие находки и важные открытия — вот что попадает в выпуски новостей в наше время, поэтому репортеры и другие представители СМИ постоянно находятся в поиске того, из чего можно сделать сенсацию. Как можно отличить правду от преувеличения? Лучше всего выяснить все детали исследования.

Установление несуществующих взаимосвязей

Исследование, в котором связывался возраст рожениц с продолжительностью их жизни — вот еще один пример того, какие ошибки могут допускаться при толковании результатов. Действительно ли результаты этого исследования по данным наблюдений означают, что если родить в старшем возрасте, то проживешь дольше? “Нет”, — говорят исследователи. Они объясняли результаты так: если женщина рожает ребенка в старшем возрасте, это может “замедлить” ее биологические часы, что, возможно, приведет к тому, что замедлится и процесс старения.

Хочется спросить этих исследователей: “Тогда почему вы *так* и не скажете, а вместо этого обращаете внимание на возраст?”. В этом исследовании нет никаких данных, которые бы позволили мне сделать вывод, что женщины, родившие ребенка после 40, стареют медленнее других, поэтому, как мне кажется, исследователи поспешили с подобным утверждением. Или же они должны были четко объяснить, что это пока только теория, которая требует дополнительного изучения. Исходя из данных этого исследования, предположение ученых кажется скорее случайностью, чем закономерностью. (Хотя, родив своего ребенка в 41 год, я все же надеюсь на лучшее!)

Часто в пресс-релизе или в газетной статье исследователь объясняет, *почему*, на его взгляд, исследование дало именно такие результаты и какие последствия для общества в целом они будут иметь. (Это объяснение может последовать в ответ на вопросы репортера, которые позже были вырезаны из статьи, а слова исследователя просто свободно цитируются.) Многие из таких последующих объяснений — не более чем теория, которую еще нужно проверить. В подобных случаях вы должны остерегаться выводов, объяснений или связей, которые не были подтверждены в ходе исследования.

Применение результатов к людям, не изученным при исследовании

Выводы можно делать только о той совокупности, которая была представлена в вашей выборке. Если вы делаете выборку только из мужчин, то нельзя делать выводы относительно женщин. Если в вашей выборке здоровые молодые люди, то нельзя делать выводы обо всех. Но многие исследователи грешат именно этим. Это распространенная практика, которая дает результаты, вводящие в заблуждение. Остерегайтесь подобного!

Вот как можно определить, правильны ли выводы исследователей.

- ✓ Выясните, какова целевая совокупность (т.е. группа, о которой исследователь хочет сделать вывод).
- ✓ Установите, как была сделана выборка и является ли она репрезентативной для данной целевой совокупности (а не какой-то более узкой группы).
- ✓ Проверьте сделанные исследователями выводы. Убедитесь, что они не касаются более широкой совокупности, чем та, что была в действительности изучена.

Принятие информированных решений об экспериментах

Только потому, что кто-то заявляет, будто он провел "научное исследование" или "научный эксперимент", это не значит, что все было сделано правильно, а результатам можно доверять. Работая консультантом по статистике, я очень часто сталкивалась с некачественными экспериментами. Хуже всего то, что если эксперимент проведен плохо, то с этим ничего нельзя поделать, кроме как проигнорировать результаты — именно так вам следует поступить.

Вот некоторые советы, которые помогут вам принять информированное решение о том, стоит ли доверять результатам эксперимента, особенно если эти результаты очень важны для вас.

- ✓ Когда вы в первый раз слышите о результатах, возьмите карандаш и запишите как можно больше из того, что увидели или услышали, где именно вы встретили эту информацию, кто проводил исследование и каковы основные результаты.
- ✓ Изучите источник, пока не найдете человека, проводившего первоначальное исследование, затем попросите у него копию отчета.
- ✓ Изучите отчет и оцените эксперимент с точки зрения восьми этапов хорошего эксперимента, описанных в разделе "Разработка хорошего эксперимента" в этой главе. (Для этого вам совсем не обязательно досконально понять все, что написано в отчете.)
- ✓ Внимательно обдумайте выводы, которые исследователь делает о своих находках. Многие стремятся преувеличить результаты, сделать выводы, не подкрепленные статистическими доказательствами, или попытаться отнести результаты к более широкой совокупности, чем та, что в действительности была изучена.
- ✓ Никогда не бойтесь сомневаться в словах средств массовой информации, исследователей и даже собственных экспертов. Например, если у вас есть вопрос относительно медицинского эксперимента, задайте его своему врачу. Он будет только рад, что у него такой образованный и думающий пациент!

Глава 18

Поиск связей: корреляции и ассоциации

В этой главе...

- Понятие статистики связей
- Разница между ассоциацией, корреляцией и причинностью
- Предположения на основе известных связей
- Обнаружение неверных результатов

Похоже, все просто сговорились сообщать информацию о новейших связях, корреляциях, ассоциациях и обнаруженных взаимоотношениях. Многие из таких связей касаются медицинских исследований. Работа исследователя в области медицины заключается в том, чтобы сказать, что вы должны, а чего не должны делать, чтобы прожить долгую и здоровую жизнь.

Вот некоторые последние находки, которые обнародовал Национальный институт здоровья США.

- ✓ Сидячие виды деятельности (например, просмотр телепередач) связаны с увеличением веса и возрастанием риска диабета у женщин.
- ✓ Выражение гнева может иметь обратную взаимосвязь с риском сердечного приступа или удара. (Те, кто не сдерживают свой гнев, рискуют меньше.)
- ✓ У мужчин потребление спиртного в умеренных дозах снижает риск сердечных заболеваний.
- ✓ Незамедлительное лечение позволяет приостановить развитие глаукомы.

Репортеры обожают рассказывать об установлении новейших взаимосвязей, потому что такие истории могут сделать настоящую сенсацию. Хотя время от времени некоторые рекомендации, похоже, меняются, например, сегодня по радио говорят, что цинк помогает побороть простуду, а завтра сообщают, что предыдущее мнение ни на чем не было основано. Многие взаимосвязи, которые встречаются в средствах массовой информации, описываются как причинно-следственные, но разве можно верить подобным утверждениям? (К примеру, действительно ли если женщина родила первого ребенка в зрелом возрасте, то это позволит ей прожить дольше?) Не слишком ли вы скептичны и больше ничему не верите?

В этой главе вы найдете информацию о связях и корреляциях. Вы узнаете, когда между двумя факторами существует корреляция, ассоциация или причинно-следственная взаимосвязь, а также поймете, когда и как можно делать предположения, исходя из таких взаимосвязей. Помимо этого вы научитесь правильно оценивать утверждения иссле-

дователей и самостоятельно принимать решения о тех сообщениях в прессе и на телевидении, в которых говорится о новейших корреляциях.

Представим общую картину: диаграммы и схемы

Мое внимание привлекла одна статья в журнале *Garden Gate*. Заголовок гласил: "Чтобы измерить температуру, сосчитайте трели сверчка". Как утверждают авторы статьи, вам всего лишь нужно найти сверчка, посчитать количество трелей, которые он исполнит за 15 секунд, добавить к этому числу 40 — и все! Вы только что определили температуру по шкале Фаренгейта.

Национальное бюро метеорологической службы даже разработало собственную программу для преобразования данных. Вы вводите количество трелей, которые издает сверчок за 15 секунд, а программа показывает температуру в четырех разных единицах измерения, в том числе в градусах Цельсия и Фаренгейта.

Утверждение о том, что частота трелей сверчка связана с температурой, подтверждается многочисленными исследованиями. Для наглядности я взяла некоторые данные и представила их в табл. 18.1. Обратите внимание, что каждое наблюдение состоит из двух связанных между собой переменных, в данном случае это количество трелей сверчка за 15 секунд и температура в этот момент (по шкале Фаренгейта). Статистики называют такой тип двухмерных данных *бивариантными*. Каждое наблюдение состоит из пары данных, полученных одновременно.

Таблица 18.1. Трели сверчка и данные о температуре (выдержка)[1]

Количество трелей за 15 секунд	Температура по шкале Фаренгейта
18	57 (13,9)
20	60 (15,6)
21	64 (17,8)
23	65 (18,3)
27	68 (20,0)
30	71 (21,7)
34	74 (23,3)
39	77 (25,0)

Меня также заинтересовал недавний пресс-релиз, выпущенный Медицинским центром Государственного университета Огайо. В нем говорится, что аспирин может предотвратить образование полипов у больных раком толстой кишки. Поскольку один из моих близких родственников умер от этой болезни, меня порадовала мысль о том, что исследователи делают успехи в данной области, и решила узнать об этом подробнее. Данные об исследовании взаимосвязи между аспирином и полипами подытожены в табл. 18.2.

[1] В скобках указана температура по шкале Цельсия. — *Прим. пер.*

Таблица 18.2. Результаты исследования связи между аспирином и образованием полипов

Группа	Процент отметивших появление полипов*
Аспирин	17
Плацебо	27

*общий размер выборки = 635 (они были в случайном порядке примерно пополам распределены по двум группам)

Всего в необработанных данных этого исследования 635 строк, в каждой из которых приводится порядковый номер участника, группа, к которой он был отнесен (использовавшая аспирин или плацебо), а также указывается, было ли за изучаемый период у него отмечено образование полипов (да или нет). Например, первая строка этого набора данных может выглядеть так:

№22292 ГРУППА=АСПИРИН ПОЯВЛЕНИЕ ПОЛИПОВ = НЕТ

Если посмотреть на необработанные бивариантные данные в этом большом наборе данных, то вы, вероятно, едва ли сможете заметить какую-либо взаимосвязь между переменными: 635 строк данных никто (за исключением компьютера) не смог бы толком изучить. В случае с трелями сверчка и температурой, даже если вы и заметите общую модель в необработанных данных (например, определите, что с увеличением количества трелей температура, похоже, тоже растет), установить точную взаимосвязь все же очень непросто.

Чтобы разобраться в каких-либо данных, нужно прежде всего организовать их с помощью таблицы, диаграммы или графика (см. главу 4). Если данные бивариантны, а вас интересуют связи между двумя переменными, то диаграммы и графики тоже должны иметь два измерения, как и сами данные. Только так вы сможете заметить возможные взаимосвязи между переменными.

Изображение бивариантных числовых данных

Если обе переменные количественные (т.е. числовые, например, показатели роста или веса), то бивариантные данные, как правило, представляются в виде графика, который в статистике называется *диаграммой рассеивания*. У такой диаграммы два измерения: горизонтальное (называемое *осью x*) и вертикальное (которое называется *осью y*). Обе оси числовые, т.е. имеют числовую шкалу.

Построение диаграммы рассеивания

Занесение наблюдений (или точек) на диаграмму рассеивания напоминает поиск города на карте, когда соответствующий квадрат обозначается буквой и цифрой. У каждой точки есть две координаты. Первая соответствует первому элементу данных в паре (это координата *x*, т.е. расстояние, на которое вы передвигаетесь влево или вправо). Вторая координата соответствует второму элементу в паре данных (это координата *y*, т.е. расстояние, на которое вы передвигаетесь вверх или вниз). Соедините две координаты, и вы получите точку, которая представляет сделанное наблюдение. На рис. 18.1 показана диаграмма рассеивания, составленная на основе данных из табл. 18.1 о связи трелей сверчка с температурой. Поскольку при составлении табл. 18.1 я располагала данные в соответствии с их координатами *x* (количество трелей), то точки на диаграмме рассеивания (слева направо) идут в том же порядке, что и в табл. 18.1.

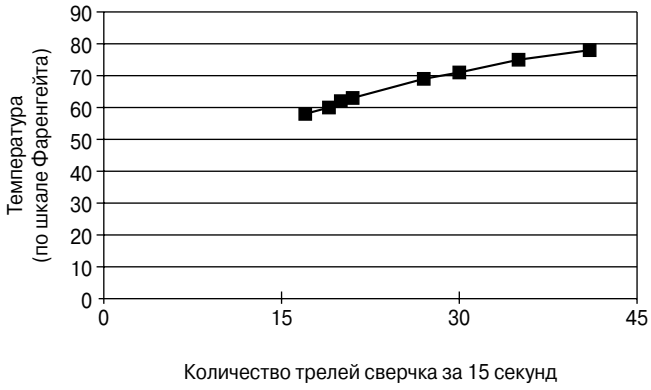


Рис. 18.1. Диаграмма рассеивания для данных о связи между трелями сверчка и температурой

Толкование диаграммы рассеивания

Интерпретировать диаграмму рассеивания нужно, двигаясь слева направо.

- ✓ Если при движении слева направо данные напоминают уходящую вверх прямую, это указывает на *положительную линейную* (или *пропорциональную*) *зависимость*. По мере увеличения *x* (с перемещением на одну единицу вправо) увеличивается и *y* (поднимается вверх).
- ✓ Если при движении слева направо данные напоминают уходящую вниз прямую, это указывает на *отрицательную линейную* (или *обратную*) *зависимость*. Другими словами, с увеличением *x* (при перемещении на одну единицу вправо) *y* уменьшается (опускается).
- ✓ Если данные вообще не образуют прямую (даже отдаленно), это означает, что линейной взаимосвязи между ними не существует.

Судя по рис. 18.1, можно сказать, что между количеством трелей сверчка и температурой существует положительная линейная зависимость, т.е. с ростом количества трелей растет и температура. Будет ли это причинно-следственной взаимосвязью (другими словами, действительно ли повышение температуры заставляет сверчков стрекотать чаще)? Вопрос неясен, потому что данные были получены только в ходе исследования по данным наблюдений, а не в результате эксперимента (см. главу 17).

Визуальное отображение бивариантных данных показывает возможные ассоциации или взаимосвязи между двумя переменными. Но только потому, что из вашего графика или диаграммы кажется, что что-то происходит, это не означает, что в действительности существует причинно-следственная взаимосвязь. Например, если посмотреть на диаграмму рассеивания данных потребления мороженого и уровня убийств, то между этими переменными тоже заметна положительная линейная зависимость. Но никто ведь не будет утверждать, что употребление мороженого подталкивает к убийству или что убийство влияет на уровень потребления мороженого. Если с помощью диаграммы или графика вас пытаются убедить в причинно-следственной взаимосвязи, копните глубже и выясните, как разрабатывалось исследование, как собирались данные, и лишь после этого оцените полученный результат, воспользовавшись критериями, описанными в главе 17.

Помимо положительной или отрицательной зависимости могут существовать и другие тенденции. Между переменными может наблюдаться кривая взаимосвязь или разные виды экспоненциальной зависимости, но в книге речь о них не идет. Однако есть и хорошая новость: очень часто зависимость между переменными бывает именно положительной или отрицательной линейной.

Изображение бивариантных дискретных данных

Если обе переменных относятся к дискретным (например, пол респондента или то, поддерживает ли он президента), данные, как правило, подытоживаются в виде таблицы, которая в статистике называется *таблицей с двумя входами* (другими словами, строки в такой таблице обозначают значения первой переменной, а столбцы — значения второй переменной).

Например, в исследовании взаимосвязи между аспирином и образованием полипов в толстой кишке обе переменных были дискретными: принимал ли больной аспирин (да или нет) и появились ли у него новые полипы (да или нет). Заметьте, что табл. 18.2 в разделе "Представим общую картину: Диаграммы и схемы" является двусторонней.

Более приятный с визуальной точки зрения способ организации двухмерных данных — это использование столбиковой диаграммы или ряда секторных диаграмм. На рис. 18.2 представлена столбиковая диаграмма, в которой показано процентное отношение пациентов, отметивших появление новых полипов в группе, принимавшей аспирин, по сравнению с теми, кто использовал плацебо. На рис. 18.3 показаны две секторные диаграммы, одна для группы, принимавшей аспирин, а вторая — для группы, использовавшей плацебо. Каждая секторная диаграмма показывает процент больных в соответствующей группе, у которых отмечено появление новых полипов.

Поскольку столбцы на диаграмме значительно отличаются по размеру, да и обе секторные диаграммы во многом различны между собой, действительно складывается впечатление, что между употреблением аспирина и появлением новых полипов у испытуемых в данном исследовании существует какая-то взаимосвязь. (Здесь главное словосочетание — "складывается впечатление". Чтобы наверняка сказать, что обнаруженная разница в выборках действительно применима к соответствующим совокупностям, необходимо провести проверку гипотезы для двух долей. Подробнее об этом см. главу 15.)

Скептически относитесь к человеку, который делает выводы о взаимосвязи между двумя переменными, используя только диаграмму или график. Внешность обманчива (см. главу 4). Чтобы доказать, что подобная взаимосвязь статистически значима, нужно провести статистические измерения и проверки (см. главу 14).

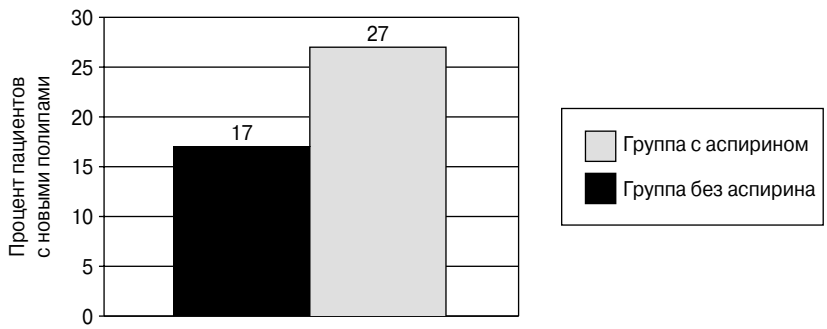


Рис. 18.2. Столбиковая диаграмма, показывающая результаты исследования связи между аспирином и появлением полипов

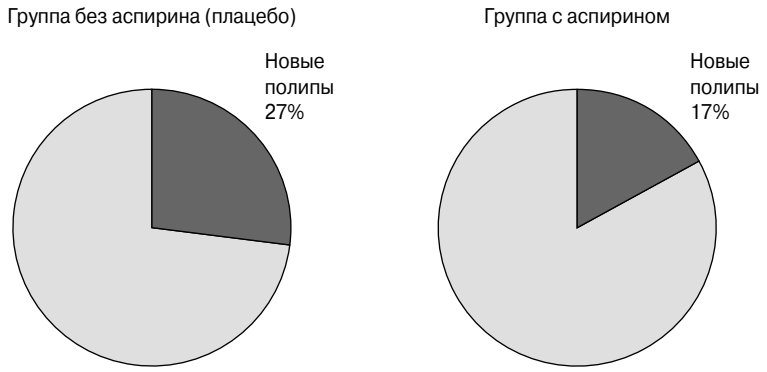


Рис. 18.3. Секторные диаграммы, показывающие результаты исследования связи между аспирином и появлением полипов

В случае с исследованием влияния аспирина на образование полипов вы можете подумать, что вторая переменная (полипы) относится к числовым, поскольку это значение показано в процентах. Но на самом деле это не так. Процент — это всего лишь удобный способ суммировать данные дискретных переменных. В данном случае вторая переменная — это данные о том, появились ли новые полипы у каждого отдельного пациента (да/нет). (Это дискретная переменная.) Процент просто подытоживает данные обо всех пациентах в категориях "да" и "нет".

Количественное выражение взаимосвязи: корреляции и другие критерии

После того как бивариантные данные были организованы, необходимо провести некоторые статистические вычисления, с помощью которых можно количественно определить или измерить степень и природу данной взаимосвязи.

Количественное определение взаимосвязи между двумя числовыми переменными

Если обе переменные являются количественными или числовыми, то статистики могут измерить направление и прочность линейной зависимости между переменными *x* и *y*. Данные, "напоминающие уходящую вверх прямую", имеют положительную линейную зависимость, но эта зависимость не обязательно будет строгой. Строгие зависимости определяется тем, насколько сильно данные напоминают прямую линию. Конечно, существуют разные степени "приближенности к прямой". Кроме того, необходимо помнить о различии между положительной и отрицательной зависимостью. Можно ли все это измерить с помощью одного статистического показателя? Конечно!

Для определения прочности и направления линейной зависимости между *x* и *y* статистики используют так называемый *коэффициент корреляции*.

Вычисление коэффициента корреляции (r)

Формула для определения коэффициента корреляции (обозначаемого *r*) такова:

$$r = \frac{1}{n-1}\sum\frac{(x-\bar{x})(y-\bar{y})}{s_x s_y}.$$

Для вычисления коэффициента корреляции выполните следующие действия:

1. **Найдите среднее всех значений x (назовем его $\bar{x}$) и среднее всех значений y (назовем его $\bar{y}$).**

   Подробнее о вычислениях см. главу 5.

2. **Найдите стандартное отклонение всех значений x (пусть это будет s_x) и стандартное отклонение всех значений y (обозначим его s_y).**

   См. главу 5.

3. **Для каждой пары в наборе данных (x, y) найдите разницу между x и $\bar{x}$, а также y и $\bar{y}$, затем перемножьте полученные результаты.**

4. **Найдите сумму всех произведений.**

5. **Разделите сумму на $s_x \times s_y$.**

6. **Разделите результат на $n - 1$, где n — это количество пар (x, y).**

Например, предположим, у вас есть такой набор данных: (3,2), (3,3) и (6,4). Выполняя описанные действия, вы можете определить коэффициент корреляции. Обратите внимание, что значения x — это 3, 3 и 6, а значения y — 2, 3 и 4.

1. **$\bar{x}$ равен 12 / 3 = 4, а $\bar{y}$ равен 9 / 3 = 3.**

2. **Стандартные отклонения таковы: s_x = 1,73, s_y = 1,00.**

3. **Произведения разностей, найденных на этапе 4, равны (3 − 4)(2 − 3) = (−1)(−1)=1; (3 − 4)(3 − 3) = (−1)(0) = 0; (6 − 4)(4 − 3)= (+2)(+1) = +2.**

4. **Сложим все результаты из этапа 3, и получится 1 + 0 + 2 = 3.**

5. **Разделим результат этапа 4 на $s_x \times s_y$, и мы получим 3 / (1,73 × 1,00) = 3/1,73 = 1,73.**

6. **Разделив результат этапа 5 на (3 − 1) (что составляет 2), получим 0,87.**

   Это и есть корреляция.

Понимание корреляции

Корреляция *r* всегда находится в пределах от −1 до +1.

- ✓ Корреляция, равная −1, указывает на идеальную отрицательную линейную зависимость.
- ✓ Корреляция, близкая к −1, указывает на сильную отрицательную линейную зависимость.
- ✓ Корреляция, близкая к 0, означает, что не существует линейной зависимости.
- ✓ Корреляция, близкая к +1, говорит о сильной положительной линейной зависимости.
- ✓ Корреляция, равная +1, указывает на идеальную положительную линейную зависимость.

Многие ошибочно полагают, что корреляция, равная –1, — это всегда плохо, потому что говорит об отсутствии взаимосвязи. На самом же деле все как раз наоборот! Корреляция, равная –1, говорит о том, что данные расположены идеально по прямой, это самая сильная линейная зависимость из всех возможных. Просто такая линия направлена вниз — именно на это и указывает знак минуса!

Насколько "близко" нужно подойти к –1 или +1, чтобы можно было говорить о сильной линейной зависимости? Большинство статистиков полагают, что для этого корреляция должна быть больше +0,6 (или меньше –0,6). Но не рассчитывайте, что она всегда будет равна +0,99 или –0,99, ведь это реальные данные, а действительность не бывает идеальной.

На рис. 18.4 показаны примеры того, как могут выглядеть разные корреляции в том, что касается прочности и направления взаимосвязи.

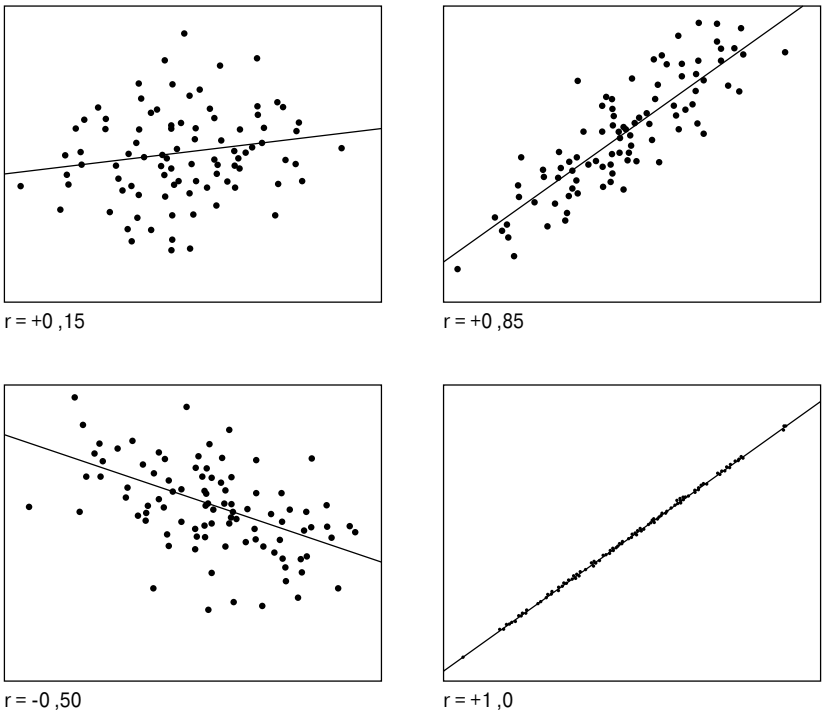


Рис. 18.4. Диаграммы рассеивания с разными корреляциями

Понимание свойств коэффициента корреляции

Вот некоторые полезные свойства корреляций:

- ✓ У корреляции нет единиц измерения. Это значит, что если изменить единицы *x* и *y*, то корреляция не изменится. (Например, изменение шкалы с Фаренгейта на Цельсия в примере с частотой трелей сверчка и температурой не повлияет на корреляцию.)
- ✓ Значения *x* и *y* в наборе данных можно поменять местами, корреляция при этом не изменится.

Количественное выражение взаимосвязи между двумя дискретными переменными

Если обе переменных относятся к дискретным (скажем, принимал ли пациент аспирин или появились ли у него новые полипы), то для описания взаимосвязи нельзя использовать слово "корреляция", потому что корреляция измеряет строгость линейной зависимости между числовыми переменными. (Такая ошибка то и дело встречается в средствах массовой информации и просто сводит статистиков с ума!)

Понятие, которое применяется для описания взаимосвязи между двумя дискретными данными, — это *ассоциация*. Между двумя дискретными переменными (например, группой испытуемых и исходом) существует ассоциация, если процент испытуемых, у которых наблюдался определенный исход, заметно отличается от процента представителей другой группы с тем же исходом. В примере с влиянием аспирина на появление полипов, о котором речь шла в первом разделе этой главы, исследователи выяснили, что в группе, принимавшей аспирин, у 17% больных раком толстой кишки появились полипы, а процентное отношение таких пациентов во второй группе было равно 27%. Поскольку эти числа довольно различны, то между двумя переменными есть ассоциация.

Но насколько большой должна быть разница между процентами, чтобы можно было говорить о значительной ассоциации между двумя переменными? Разница в выборке должна быть *статистически значимой*. Таким образом подобный вывод можно будет сделать о взаимосвязи для целой совокупности, а не только для конкретного набора данных. Для этого подойдет проверка гипотезы для двух долей (подробнее о такой проверке см. главу 15). Я анализировала данные об исследовании влияния аспирина на появление полипов с помощью такого теста и получила p-значение, меньшее 0,0001. Другими словами, эти результаты крайне значимы. (Подробнее о p-значениях см. главу 14.) Теперь понятно, почему исследователи прервали эксперимент и решили всем пациентам давать аспирин!

Объяснение взаимосвязи: ассоциация и корреляция против причинности

Если оказывается, что между двумя переменными существует либо ассоциация, либо корреляция, то это не обязательно означает, что между ними есть еще и причинно-следственная взаимосвязь. Будет ли установлена причинная ассоциация между двумя переменными, зависит от того, как проводилось исследование. Только хорошо спланированный эксперимент (см. главу 17) или несколько отдельных исследований на основе наблюдений могут подтвердить, что ассоциация или корреляция на самом деле перерастает в причинно-следственную взаимосвязь.

Похоже, аспирин действительно помогает

Выводы, сделанные исследователями в ходе изучения влияния аспирина на появление полипов, о чем рассказывалось в первом разделе этой главы, кажутся убедительными. Согласно критериям, описанным в главе 17, это исследование представляло собой хорошо спланированный эксперимент. Распределение пациентов по группам проводилось в случайном порядке, выборки были достаточно большого размера, чтобы это давало точную информацию, кроме того, осуществлялся контроль над внешними переменными. Все это значит, что исследователи в полной мере отвечают за заголовок своего

пресс-релиза: “Аспирин предотвращает появление полипов у больных раком толстой кишки”. Благодаря тщательной разработке исследования можно сказать, что между регулярным принятием аспирина больными раком толстой кишки и появлением у них полипов действительно существует причинно-следственная взаимосвязь.

Жара и пение сверчков

Влияет ли температура окружающей среды на темп пения сверчков? (Очевидно, обратное утверждение ложно, но что можно сказать о возможной причинности в таком направлении?) Кое-кто утверждает, что изменения в температуре окружающей среды влияют на частоту трелей сверчка. Однако трудно представить, какие данные, основанные на экспериментах (в отличие от исследований по результатам наблюдений) могли бы подтвердить или опровергнуть такую причинно-следственную взаимосвязь. Возможно, вы сами сможете провести эксперимент — подогреть нескольких сверчков и посмотреть, что из этого выйдет! (Но прежде чем приступить к этому, обязательно хорошо продумайте эксперимент, воспользовавшись критериями из главы 17.)

Делаем предположения: регрессия и другие методы

Итак, между двумя переменными была установлена зависимость и вы каким-то образом смогли найти ее количественные характеристики. Теперь можно создать модель, которая позволит вам воспользоваться одной переменной и предсказать вторую.

Предположения о скоррелированных данных

Если в случае с двумя числовыми переменными была установлена прочная корреляция, исследователи часто используют взаимосвязь между x и y для того, чтобы сделать предположения. Поскольку между x и y наблюдается корреляция, то между ними существует *линейная зависимость*. Это значит, что такую зависимость можно представить в виде прямой линии. Если вам известен наклон и отрезок, отсекаемый на оси y, тогда вы можете подставить значение x и предсказать среднее значение y. Другими словами, вы можете предсказать y с помощью x.

Поскольку коэффициент корреляции между количеством трелей сверчка и температурой такой высокий ($r = 0{,}98$), можно найти линию, соответствующую этим данными. Это значит найти одну прямую, которая лучше всего подходит для имеющихся данных (в том, что касается среднего расстояния от всех точек в наборе данных до полученной линии). Статистики называют такой поиск самой подходящей прямой *регрессионным анализом*.

Никогда не проводите регрессионный анализ, если вы не установили строгую корреляцию (будь то положительную или отрицательную) между двумя переменными. Бывали случаи, кода исследователи уверенно делали предположения, хотя коэффициент корреляции был равен всего 0,2! Это не имеет смысла. Прежде всего, если диаграмма рассеивания данных не напоминает прямую, то не стоит даже пытаться искать прямую, подходящую для этих данных, или делать предположения относительно генеральной совокупности.

Прежде чем изучать любую модель, в которой одна переменная используется для предположения о другой, найдите корреляцию. Если она слишком слаба, дальше можете ничего не делать.

Может показаться, что для определения самой подходящей прямой придется перепробовать множество разных линий, но это не так (визуальное наблюдение за линией на диаграмме рассеивания все же помогает представить возможный ответ). Оптимальная прямая отличается явным наклоном и имеет отрезок, отсекаемый на оси *y*, которые можно определить по формулам (использовать эти формулы совсем не трудно).

Формула для оптимальной прямой

Формула для *оптимальной* (или *регрессионной*) *прямой* выглядит так: $y = mx + b$, где m — наклон прямой, а b — отрезок, отсекаемый на оси y. Наклон прямой — это изменение y в связи с изменением x. Например, наклон 10/3 означает, что при переходе от одной точки прямой к следующей, когда x смещается вправо на 3 деления, то y поднимается вверх на 10 единиц. Отрезок, отсекаемый на оси y, — это то место на вертикальной оси, где ее пересекает прямая. К примеру, в уравнении $y = \frac{10}{3}x - 6$ прямая пересекает ось y в точке -6. Координаты этой точки $(0, -6)$, потому что пересекается ось y, а значение x отрезка, отсекаемого на оси y, всегда равно 0. Чтобы найти оптимальную прямую, нужно определить значения m и b, чтобы получить реальное уравнение прямой (например, $y = 2x + 3$ или $y = -10x - 45$).

Для выполнения всех необходимых вычислений при нахождении оптимальной прямой используют пять широко известных статистических показателей. Статистики называют их *Большой пяткой статистических показателей*:

- ✓ Среднее значений x (обозначается $\bar{x}$).
- ✓ Среднее значений y (обозначается $\bar{y}$).
- ✓ Стандартное отклонение значений x (обозначается s_x).
- ✓ Стандартное отклонение значений y (обозначается s_y).
- ✓ Корреляция между x и y (обозначается r).

(В этой главе и главе 5 содержатся формулы и подробные инструкции для вычисления этих статистических показателей.)

Определение наклона оптимальной прямой

Формула наклона m оптимальной прямой такова: $m = r\left(\frac{s_y}{s_x}\right)$, где r — корреляция между x и y, а s_y и s_x — стандартные отклонения значений y и x соответственно (подробнее о стандартном отклонении см. главу 5).

Чтобы найти наклон m оптимальной прямой, выполните следующие действия.

1. **Разделите s_y на s_x.**
2. **Умножьте полученное частное на r.**

Наклон оптимальной линии может быть отрицательным числом, потому что отрицательной может быть и корреляция. Отрицательный наклон говорит о том, что прямая направлена вниз.

В формуле наклона корреляции (безразмерному показателю) приписываются единицы измерения. Просто представляйте себе частное s_y/s_x как изменение y вследствие изменения x. Стандартные отклонения тоже измеряются в своих исходных единицах (например, температура — в градусах по Фаренгейту, а количество трелей сверчка за 15 секунд — в штуках).

Определение отрезка, отсекаемого оптимальной прямой на оси y

Формула отрезка b, отсекаемого на оси y оптимальной прямой, выглядит так: $b = \overline{y} - m\overline{x}$, где $\overline{y}$ и $\overline{x}$ — это средние значений y и x соответственно, а m — наклон (формула которого дана в предыдущем разделе).

Чтобы найти отрезок b, отсекаемый на оси y оптимальной прямой, выполните следующее:

1. **Найдите наклон m оптимальной прямой, выполнив действия, описанные в предыдущем разделе.**
2. **Умножьте m и $\overline{x}$.**
3. **Вычтите результат, полученный на этапе 2, из $\overline{y}$.**

Прежде чем находить отрезок, отсекаемый на оси y, всегда определяйте наклон. Формула отрезка, отсекаемого на оси y, содержит наклон, поэтому для вычисления b нужно знать m.

Нахождение оптимальной прямой для трелей сверчка и температуры

Хотя формула оптимальной прямой для взаимосвязи между трелями сверчка и температурой все еще вызывает споры (см. Приложение), похоже, был найден консенсус в том, что самой подходящей моделью для такой взаимосвязи является $y = x + 40$ или температура = 1 × (количество трелей за 15 секунд) + 40, где температура измеряется в градусах по Фаренгейту. Обратите внимание, что наклон этой линии равен 1, x — это количество трелей за 15 секунд, а y — температура в градусах по Фаренгейту.

В формулах наклона и отрезка, отсекаемого на оси y, присутствуют x и y, значит, вам нужно решить, какую из двух переменных обозначить x, а какую — y. При нахождении корреляций такой выбор не имеет значения, если вы постоянно придерживаетесь его, однако при поиске оптимальной прямой и при высказывании предположений очень важно, какую именно переменную вы обозначите через x, а какую — через y. Посмотрите на предыдущие формулы: если поменять x и y местами, то изменится и вся формула.

Так как же определить, где какая переменная? Как правило, x — это переменная, с помощью которой делается предположение. Статистики называют ее *объясняющей переменной*, потому что если изменить x, это объяснит, как и почему изменится y. В данном случае x — это количество трелей сверчка за 15 секунд. Переменная y называется *объясняемой переменной*, потому что она реагирует на изменение x. Другими словами, y предсказывается с помощью x. В нашем случае y — это температура.

Сравнение рабочей модели с подмножеством данных

Большая пятерка статистических показателей из подмножества данных о сверчках показана в табл. 18.3.

Таблица 18.3. Большая пятерка статистических показателей на основе данных о сверчках

Переменная	Среднее	Стандартное отклонение	Корреляция
Количество трелей (x)	$\bar{x} = 26{,}5$	$s_x = 7{,}4$	$r = +0{,}98$
Температура (y)	$\bar{y} = 67$	$s_y = 6{,}8$	

Наклон m оптимальной прямой для подмножества данных о количестве трелей сверчка по отношению к температуре равен $r \times (s_y/s_x) = 0{,}98 \times (6{,}8/7{,}4) = 0{,}98 \times 0{,}919 = 0{,}90$. Теперь, чтобы найти отрезок b, отсекаемый на оси y, вы берете $\bar{y} - m \times \bar{x}$, т.е. $67 - (0{,}90)(26{,}5) = 67 - 23{,}85 = 43{,}15$. Значит, оптимальная прямая для предсказания температуры с помощью данных о трелях сверчка такова: $y = 0{,}9x + 43{,}2$ или температура (в градусах по Фаренгейту) = 0,9 × (количество трелей за 15 секунд) + 43,2.

Заметьте, что предыдущее уравнение близко к рабочей модели $y = x + 40$, но не совпадает с ней. Почему же так происходит? Возможны несколько причин. Во-первых, "рабочая модель" — это выдуманный термин, которым обозначается нечто "не обязательно точное, но очень полезное". В течение многих лет наклон округляли до ближайшего целого числа (1), а отрезок, отсекаемый на оси y, — до ближайшего десятка (40), просто чтобы было легче запомнить и забавнее писать об этом явлении. (Для статистики в этом мало хорошего, но такая ситуация может послужить примером того, как с течением времени может исказиться статистическая информация.) Во-вторых, данные, которые я здесь использовала — это всего лишь случайное подмножество исходного набора данных (для наглядности), поэтому оно по простой случайности будет искажено (подробнее об изменчивости выборок см. главу 9). Однако поскольку данным присуща такая прочная корреляция, то разница между отдельными выборками данных будет незначительной.

Предсказание температуры по пению сверчка

Оптимальная прямая для предсказания температуры с помощью трелей сверчка, как было установлено на основе моего подмножества данных, равна $y = 0{,}9x + 43{,}2$. Любое уравнение или функция, которые используются для приблизительного определения или предсказания взаимосвязи между двумя переменными, называются *статистической моделью*. Прибегая к этой модели, вы сможете предсказать температуру при помощи трелей сверчка. Как это сделать? Выберите нужное значение x, подставьте его в модель, и найдете ожидаемое значение y.

Например, если вы хотите предсказать температуру и вам известно, что сверчок на заднем дворе издал 35 трелей за 15 секунд, подставьте 35 вместо x и получите y. Задача выглядит так: $y = 0{,}9(35) + 43{,}2 = 31{,}5 + 43{,}2 = 74{,}7$. Итак, зная, что количество трелей сверчка за 15 секунд равно 35, вы можете предположить, что ожидаемая температура составит около 75 градусов по Фаренгейту.

Только потому, что у вас есть модель, вы не можете подставлять в нее *любое* значение *x* и рассчитывать точно предположить *y*. Например, в модель нельзя подставить число больше 38 или меньше 18. Почему? Потому что в данном случае у вас нет значений *x* для этого диапазона (см. табл. 18.1). Кто вам сказал, что прямая будет работать и за пределами той области, по которой были собраны данные? Неужели вы действительно думаете, что по мере того, как температура повышается до бесконечности, сверчки будут стрекотать все быстрее и быстрее, без ограничений? В определенный момент бедный сверчок умрет от перегрева. Точно так же сверчки не выживут в слишком холодную погоду, значит, нельзя подставлять в уравнение слишком маленькие значения *x* и рассчитывать, что модель сработает.

Ни в коем случае нельзя строить предположения, используя значения *x*, которых нет в пределах собранных вами данных. Статистики называют это *экстраполяцией*. Остерегайтесь исследователей, которые пытаются высказывать утверждения, не подтвержденные никакими результатами.

Поскольку оптимальная прямая — это модель, описывающая общую взаимосвязь между *x* и *y*, то на самом деле вы предсказываете не *y*, а ожидаемое (или среднее) значение *y* для любого конкретного значения *x*.

Предположения с двумя ассоциированными дискретными переменными

После того как между двумя дискретными переменными была установлена ассоциация, можно сделать предположение (оценку), касающееся процентного отношения в каждой группе по одной из переменных, или же можно оценить разницу между процентами в двух группах. И в первом, и во втором случае для этого нужно использовать доверительный интервал (см. главы 12 и 13).

В примере с влиянием аспирина на появление полипов у больных раком толстой кишки, описанном в первом разделе этой главы, в группе, принимавшей аспирин, процент пациентов, у которых отмечено появление новых полипов, составило 17%, а во второй группе, принимавших не аспирин, а плацебо, этот процент было равен 27% (см. табл. 18.2). Здесь можно сделать следующее предположение: если вы больны раком толстой кишки, то, принимая ежедневно 325 мг аспирина, вы снижаете риск появления новых полипов.

Это предположение можно сделать более точным, например, оценить шансы появления новых полипов у больных раком толстой кишки при ежедневном употреблении аспирина, определив доверительный интервал. Поскольку 17% участников выборки, принимавших аспирин, отметили появление новых полипов, это означает, что шансы того, что у любого человека в этой совокупности (т.е. среди больных раком толстой кишки) появятся новые полипы при ежедневном употреблении аспирина, равны 0,17 плюс/минус предел погрешности, т.е. в данном случае плюс/минус 0,04. Другими словами, у больного раком толстой кишки, ежедневно принимающего 325 мг аспирина, шансы получить новые полипы находятся в пределах от 17% – 4% до 17% + 4%, т.е. от 13% до 21%. (Подробнее о формуле доверительного интервала для одной доли совокупности *p* см. главу 13, подробнее о пределе погрешности см. главу 10.)

Еще один способ сделать предположение в данной ситуации — оценить снижение риска появления полипов при условии, что пациент каждый день принимает аспирин. Это можно сделать, вычислив доверительный интервал для разности двух долей, где p_1 — доля пациентов в контрольной группе, у которых появились новые полипы, а p_2 — доля пациентов в группе, принимающей аспирин, с теми же симптомами. Здесь вас интересует разность $p_1 - p_2$. Доверительный интервал равен $0{,}27 - 0{,}17 = 0{,}10$, плюс/минус предел погрешности, или в данном случае, плюс/минус 0,03. Значит, если больной раком толстой кишки ежедневно принимает аспирин, то риск появления новых полипов снижается в пределах от 10% – 3% до 10% + 3%, т.е. от 7 до 13%. (Подробнее о формуле доверительного интервала для разности между долями двух совокупностей см. главу 13.)

Глава 19

Статистика и зубная паста: контроль качества

В этой главе...

- Как статистика помогает улучшить продукцию
- Придерживаемся требований: основы контроля качества
- Мониторинг процесса

Самые успешные производственные компании стали такими благодаря контролю качества. Они хотят, чтобы вы как потребитель были довольны их продукцией и покупали ее снова и снова. Они хотят, чтобы вы были довольны настолько, что рассказали бы всем друзьям, соседям, коллегам и даже просто прохожим на улице о том, какая замечательная у них продукция. Но как компании добиваются того, что потребитель будет доволен их предложением? Один из критериев удовлетворения клиентов — это качество продукции, причем при определении качества продукции и в улучшении качества именно статистика играет ключевую роль.

Соответствие ожиданиям

Потребители рассчитывают, что продукция будет соответствовать их ожиданиям, одно из которых состоит в том, что в упаковке находится именно то количество товара, которое указано сверху. Еще одно ожидание — это определенный уровень постоянства при покупке товара одной марки. Чего вы ожидаете от пакета картофельных чипсов? Не странно ли, что упаковка весом в 225 г может казаться такой большой, а на самом деле в ней оказывается всего несколько чипсов? (Производители утверждают, что они впускают в упаковку воздух, чтобы защитить продукцию от повреждений.) Если на этикетке сказано, что вес упаковки составляет 225 г, и она действительно весит 225 г, то вам не на что жаловаться. Но разве не будет обидно, когда вдруг выяснится, что в пакетик не доложили чипсов?

Предположим, на упаковке написано, что она весит 225 г, а на самом деле вес равен 220 г: вас это расстроит? Скорее всего, вы даже не заметите разницы. Но что, если в пакетике окажется всего 150 г? А если 125 г? В какой-то момент это определенно бросится в глаза. И как вы отреагируете? Возможно, вы поступите следующим образом.

- ✓ Махнете рукой (если, конечно, проблема не повторяется регулярно).
- ✓ Вернете товар в магазин и потребуете компенсации.
- ✓ Напишете письмо с жалобой в компанию.
- ✓ Решите больше не покупать этот товар.

- ✓ Пожалуетесь в Комитет по защите прав потребителей или другую организацию.
- ✓ Устроите бойкот товара.
- ✓ Попытаетесь устроиться на работу в компанию, чтобы быть "частью решения, а не частью проблемы".

Что ж, некоторые из этих вариантов могут показаться чрезмерными, особенно если вас обманули всего на пару картофельных чипсов. Но предположим, что вы купили машину, которая оказалась бракованной, ваш ребенок чуть не подавился деталью, отвалившейся из коляски, или же вас начало тошнить после гамбургера, который вы вчера купили в магазине. Качество может стать серьезным, жизненно важным вопросом. Хотя американское правительство разработало стандарты для многих потребительских товаров и следит за их соблюдением (например, с помощью Управления по надзору за качеством медицинских препаратов и пищевых продуктов), время от времени проблемы с производственным процессом все же возникают. Вот всего некоторые из факторов, которые могут повлиять на качество продукции в процессе производства.

- ✓ Сотрудники работают непоследовательно (из-за разницы в уровне мастерства, в образовании или условиях труда, под влиянием трудовой смены, из-за низкого боевого духа, человеческого фактора и т.д.).
- ✓ Руководство и/или контролеры непоследовательны в своей реакции на возникающие проблемы.
- ✓ Производственное оборудование работает с перебоями (из-за недостаточного ухода, износа деталей, поломок или неполадок или просто потому, что отдельные станки или производственные линии отличаются друг от друга).
- ✓ Станки и оборудование были сконструированы недостаточно точными или чувствительными.
- ✓ Сырье, используемое в производственном процессе, низкого качества.
- ✓ Недостаточен контроль над средой (температура, влажность, чистота воздуха и т.п.).
- ✓ Мониторинг проводится неэффективно.

Подталкиваемые необходимостью удовлетворять клиентов и выполнять требования правительства, успешные производители всегда стремятся повысить качество своей продукции. Одно из популярных понятий, которые используются в производственной отрасли, — *глобальное управление качеством* (Total Quality Management — TQM), которое нацелено на разработку способов непрерывного мониторинга, оценки и совершенствования производственного процесса от начала и до конца. Популярность к глобальному управлению качеством в США пришла благодаря известному и всенародно любимому статистику, доктору У. Эдвардсу Демингу, который составил известный и часто используемый перечень под названием "14 пунктов менеджмента". Согласно учению Деминга, если вы прежде всего нацелены на качество продукции, то тем самым снизите расходы, повысите производительность и станете более конкурентоспособным.

Каким же образом статистика связана с качеством продукции? Статистические показатели используются для выработки стандартов и мониторинга всех аспектов производственного процесса с тем, чтобы гарантировать выполнение всех требований. Статистика необходима для того, чтобы понять, когда нужно остановить процесс, а также выявить

проблемы еще до их появления. По сути, статистические данные говорят производителю (зачастую непрерывно) о качестве продукции как об одном из элементов философии тотального управления качеством. Статистика в мониторинге и улучшении качества продукции на протяжении всего производственного процесса называется *системой статистического контроля производственных процессов*. Предмету этой системы можно посвятить отдельную книгу, но вы легко представите, как статистический фактор связан с контролем качества, если разберетесь в некоторых основных принципах, изложенных в следующих разделах.

Качество в тюбике зубной пасты

Складывается впечатление, что потребитель уже смирился с тем, что пакетики с картофельными чипсами обычно оказываются легче, чем написано на этикетке, однако тюбики с зубной пастой все еще котируются более высоко, ведь мы ожидаем, что они будут наполнены доверху. (Производители зубной пасты должны знать, как усердно вы выдавливаете содержимое тюбика — все до последней капли.)

К счастью, производство зубной пасты (в которой есть даже собственный Совет по тюбикам) не подводит и со всей серьезностью относится к наполнению тюбиков. На Web-сайте одной из компаний, работающих в этой сфере, можно найти даже "Самые распространенные вопросы о тюбиках зубной пасты". Один из таких часто задаваемых вопросов касается того, как проверяется качество наполнения тюбика.

Специалисты в отрасли считают, что задача оборудования по наполнению тюбиков заключается в точности и постоянстве. Дозатор — вот главный инструмент для достижения этих целей. (*Дозатором* на профессиональном сленге называется аппарат, который как раз и наполняет тюбики.) Ниже приводятся некоторые из основных характеристик оборудования для наполнения тюбиков.

- ✓ Механизм точного прерывания потока, позволяющий избежать разбрызгивания.
- ✓ Система отвода воздуха при наполнении.
- ✓ Механизм, который останавливает аппарат, если тот начинает наполнять тюбики, которые по какой-либо причине отсутствуют.
- ✓ Система для быстрой очистки и замены.

Если для контроля качества при наполнении тюбиков зубной пастой необходима столь сложная система и пристальное внимание к деталям, то только представьте себе, как проверяется качество пассажирских самолетов!

Оказывается, на качество наполнения тюбика влияет несколько факторов, в том числе и те, что перечислены выше. Среди проблем, которых пытаются избежать производители зубной пасты, есть и такие: недостаточное наполнение (в основном из-за воздушных карманов) и чрезмерное наполнение (в связи с чем тюбики "расползаются", как говорят профессионалы. Еще одно следствие чрезмерного наполнения — часть пасты уходит к потребителю бесплатно, что сказывается на прибыли). Внутренние размеры тюбика тоже могут повлиять на качество наполнения. К примеру, тюбики меньшего размера (хотя в них и содержится правильное количество зубной пасты) будут деформироваться при запечатывании, а слишком большие тюбики произведут впечатление полупустых.

Качество — это точность + постоянство

Статистика, несомненно, нужна для сбора данных, которые применяются при оценке оборудования для наполнения тюбиков по каждому из критериев, описанных в предыдущем разделе. И все же роль статистики в производственном процессе нагляднее всего подчеркивается критерием качества, который принят среди производителей зубной пасты: постоянство и точность. Слова *постоянный* и *точный* связаны со статистикой прочнее, чем все другие слова в производственном процессе, по сути, они объясняют главное.

Статистика контролирует точность и постоянство с помощью контрольных диаграмм. *Контрольная диаграмма* — это специальная временная диаграмма, которая показывает значения данных в том порядке, в котором они собирались с течением времени (подробнее о временных диаграммах см. главу 4). На контрольной диаграмме линией показано расположение определенных значений производителя — или *контрольных величин* (это связано с вопросом о точности), а также обозначения того, насколько выше или ниже могут быть эти контрольные величины (это связано с вопросом о постоянстве). Значения, вносимые на диаграмму, — это вес, объем или плотность отдельного продукта, но чаще это будет средний вес, средний объем и средняя плотность выборки продукции. Верхняя и нижняя границы контрольной диаграммы называются *верхним* и *нижним контрольным пределом* (ВКП и НКП соответственно).

Например, представим, что производитель конфет выпускает их в упаковках, при этом контрольная величина — 50 штук в коробке, НКП = 45 конфет, ВКП = 55 штук. Предположим, делается выборка из 8 коробок, в них пересчитываются конфеты, и результаты оказываются следующими: 51 штука, 53 штуки, 49 штук, 51 штука, 54 штуки, 47 штук, 52 штуки и 45 штук. На рис. 19.1 показана получившаяся контрольная диаграмма для этого процесса. На первый взгляд (хотя бы временно) кажется, что ситуация под контролем.

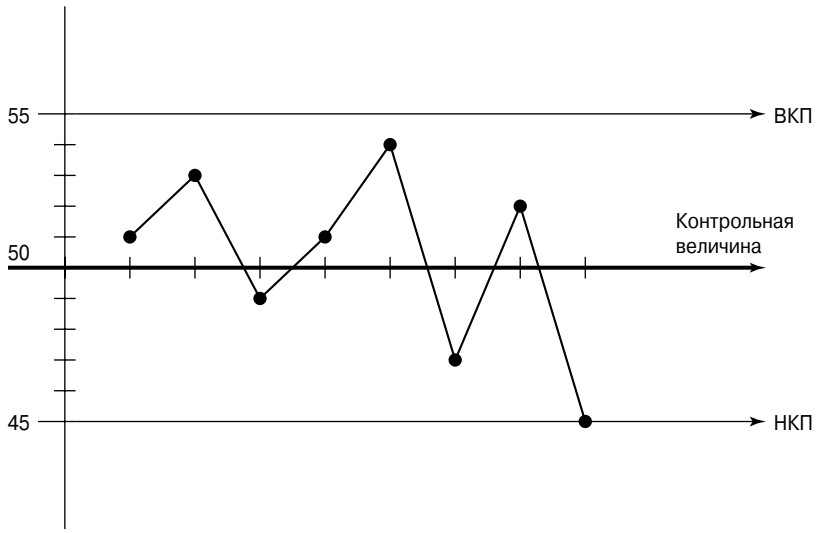


Рис. 19.1. Контрольная диаграмма для процесса наполнения коробок конфетами

Использование контрольных диаграмм для мониторинга качества

Чтобы с помощью статистики наблюдать за качеством, прежде всего нужно понять, как определить и измерить точность и постоянство. После этого вы должны установить контрольную величину, найти верхний и нижний контрольные пределы и собрать данные о процессе. Затем эти данные нужно представить в виде контрольной диаграммы, что позволит оценить процесс и понять, все ли с ним в порядке. Как раз на этом последнем этапе есть некоторый подвох: с одной стороны, вы не хотите останавливать процесс из-за ложной тревоги (а как раз это и произойдет, если вы остановите процесс, как только какое-либо из значений превысит контрольные пределы); с другой стороны, нежелательно, чтобы процесс вышел из-под контроля, что ухудшит качество продукции.

Определение точности

Если аппарат для наполнения тюбиков зубной пастой работает точно, это означает, что в конечном итоге тюбики весят столько, сколько *в среднем* и должны весить. Согласитесь, что даже самый продуманный производственный процесс не будет идеальным, в любом процессе некоторые вариации — это нормальное явление, что объясняется случайными колебаниями. Другими словами, нельзя ожидать, что все тюбики с зубной пастой будут весить точно 6,4 унции. Но если вес тюбиков все время уменьшается по сравнению с заданным или же если в них вдруг появляется много воздуха и они не весят столько, сколько должны, то потребитель это заметит, и качество продукции окажется под сомнением. Точно так же, если средний вес внезапно увеличивается, то производитель теряет деньги, потому что расходует лишнюю продукцию.

С точки зрения статистики вес продукции точен, если для него не характерно какое-либо смещение или систематическая погрешность. (*Систематическая погрешность* приводит к появлению постоянно завышенных или постоянно заниженных величин по сравнению с ожидаемым значением.) Это в полной мере соответствует понятию точности, принятому в производстве зубной пасты. В данном случае ожидаемая величина (контрольная величина) — это требования, установленные производителем, например, 6,4 унции.

Определение постоянства

Что значит, если аппарат для наполнения тюбиков зубной пастой работает с постоянством? Это означает, что *почти всегда* вес тюбика остается в контрольных пределах, поскольку опять же, любому процессу присущи вариации из-за случайных колебаний. Но если вес тюбика начинает все время увеличиваться, то контроль качества оказывается под сомнением.

С точки зрения статистики вес остается *постоянным*, если его стандартное отклонение является небольшим (см. главу 4). Насколько малым должно быть стандартное отклонение? Все зависит от требований производителя и ограничений процесса. Операторы устройств для наполнения тюбиков зубной пастой утверждают, что их аппаратура точна вплоть до 0,5%. Это значит, что большинство тюбиков, на которых написан вес в 6,4 унции, в действительности весят в пределах 0,032 унции от контрольной величины (потому что 0,5% от 6,4 — это $0{,}005 \times 6{,}4 = 0{,}032$ унции). Предположим, производители хотят, чтобы 95% тюбиков попадали в этот диапазон. Сколько стандартных отклонений нужно для того, чтобы охватить эти 95% значений вокруг контрольной величины?

Согласно эмпирическому правилу (которое можно использовать, потому что для веса тюбиков свойственно колоколообразное распределение — см. главу 8), вес 95% тюбиков находится в пределах 2 стандартных отклонений от контрольной величины. Другими словами, каждое стандартное отклонение равно 0,032/2 = 0,016 унции, в соответствии с требованиями производителя оборудования. Это довольно сдержанная оценка — требования производителя могут быть шире, чем реальные контрольные пределы, которые он устанавливает для проверки качества.

Рассчитываем на нормальное распределение

Даже хотя предполагалось, что после наполнения вес каждого тюбика с зубной пастой составит 6,4 унции, конечно же, не все они будут весить именно столько. Некоторые окажутся тяжелее, некоторые легче, но вес большинства тюбиков будет все же примерно равен 6,4 унции, при этом процентные отношения тюбиков с бо́льшим и меньшим весом равны (в контрольных пределах), если процесс находится под контролем. Помимо того, что распределение веса тюбиков напоминает холм в центре, это распределение должно быть колоколообразным или нормальным. (Подробнее о нормальном распределении см. главу 8.) Если процесс точен, среднее этого распределения будет равно контрольной величине, обозначаемой μ. Известно, что для сохранения постоянства производительности стандартное отклонение должно быть не больше 0,016 унции. Это значит, что $\mu = 6{,}4$ унции, а $\sigma = 0{,}016$ унций (см. рис. 19.2).

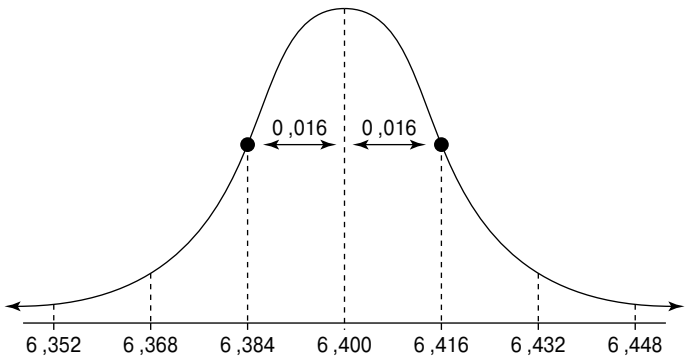


Рис. 19.2. Распределение веса отдельных тюбиков зубной пасты (если процесс под контролем)

Стандартное отклонение совокупности обозначается σ, а стандартное отклонение выборки — это *s*. (Подробнее о стандартном отклонении см. главу 5.) В большинстве случаев стандартного отклонение совокупности неизвестно, поэтому для его приблизительного определения используют *s* из данных выборки. Но в ситуации с контролем качества стандартное отклонение совокупности выпускаемой продукции устанавливается производителем. Ожидается, что если процесс находится под контролем, то стандартное отклонение будет равно этому значению.

Определяем контрольные пределы

После установления контрольной величины и ожидаемого стандартного отклонения для системы статистического контроля производственных процессов необходимо установить контрольные пределы процесса.

Если производитель будет взвешивать отдельные тюбики зубной пасты, контрольные пределы будут равны контрольной величине (среднему) плюс/минус 2 стандартных отклонения (для 95% уверенности) или же контрольной величине (среднему) плюс/минус 3 стандартных отклонения (для 99% уверенности). Формулы контрольных пределов для веса отдельных тюбиков таковы: $\mu \pm 2\sigma$ и $\mu \pm 3\sigma$ соответственно.

Производители зубной пасты установили среднее в 6,4, а стандартное отклонение в 0,016. Предположим, для них важна 95% уверенность. Это значит, что контрольные пределы равны $6,4 \pm 2(0,016) = 6,4 \pm 0,032$. НКП равен $6,4 - 0,032 = 6,368$ унции, а ВКП равен $6,4 + 0,032 = 6,432$ унции. Если для производителей важна 99% уверенность, тогда контрольные пределы для веса отдельных тюбиков составят $6,4 \pm 3(0,016) = 6,4 \pm 0,048$. НКП равен $6,4 - 0,048 = 6,352$ унции, а ВКП составляет $6,4 + 0,048 = 6,448$ унции.

Однако в большинстве случаев при мониторинге процесса вместо проверки отдельной продукции делают выборки и находят средний вес для каждой из них. Это значит, что контрольные пределы включают контрольную величину (среднее) плюс/минус 2 стандартных *ошибки* (для 95% уверенности) или контрольную величину (среднее) плюс/минус 3 *стандартных ошибки* (для 99% уверенности). *Стандартная ошибка* — это стандартное отклонение среднего выборки, она вычисляется по формуле: стандартное отклонение веса, деленное на корень квадратный из n, где n — размер выборки. (Подробнее о стандартных ошибках см. главы 9 и 10.) Стандартная ошибка обозначается $\frac{\sigma}{\sqrt{n}}$. Формулы контрольных пределов для среднего выборок таковы: $\mu \pm 2\frac{\sigma}{\sqrt{n}}$ или $\mu \pm 3\frac{\sigma}{\sqrt{n}}$ соответственно.

Стандартная ошибка всегда будет меньше стандартного отклонения. Это объясняется тем, что среднее более постоянно, чем отдельные величины, ведь оно основывается на большем количестве данных, а значит, и не будет слишком варьироваться от одной выборки к другой. Чем больше размер выборки, тем меньшей будет стандартная ошибка (потому что n стоит в знаменателе, а с увеличением n частное от деления стандартного отклонения на корень квадратный из n уменьшается). Подробнее см. главу 10.

Если процесс контролируется путем взвешивания каждого тюбика, то это то же самое, что контролировать выборку с размером $n = 1$. Значит, если $n = 1$, то формула стандартного отклонения и стандартное ошибки будет одинаковой (как и должно быть).

Предположим, размер каждой выборки тюбиков с зубной пастой равен 10. Учитывая, что стандартное отклонение установлено в 0,016, то стандартная ошибка для среднего выборки (каждая из которых равна 100) составляет $\frac{0,016}{\sqrt{10}} = 0,005$ унции. Если процесс находится под контролем, то распределение среднего веса будет нормальным, при этом среднее равно 6,4, а стандартная ошибка составляет 0,005 (см. рис. 19.3).

Если предположить, что в качестве приемлемого уровня постоянства производители зубной пасты используют 3 стандартных отклонения (и остаются консерваторами), то формула контрольных пределов выглядит как $\mu \pm 3\frac{\sigma}{\sqrt{n}} = 6,4 \pm 3 \times \frac{0,016}{\sqrt{10}} = 6,4 \pm 3(0,005)$. Нижний контрольный предел (НКП) равен $6,4 - 3(0,005) = 6,4 - 0,015 - 6,385$ унции, а верхний контрольный передел составляет $6,4 + 3(0,005) = 6,4 + 0,015 = 6,415$ унции.

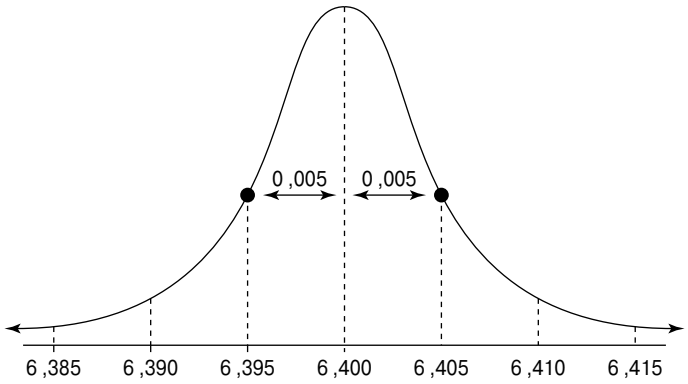


Рис. 19.3. Распределение среднего выборок по 10 тюбиков зубной пасты в каждой (если процесс под контролем)

Контрольная величина и контрольные пределы для этого конкретного процесса показаны на контрольной диаграмме на рис. 19.4.

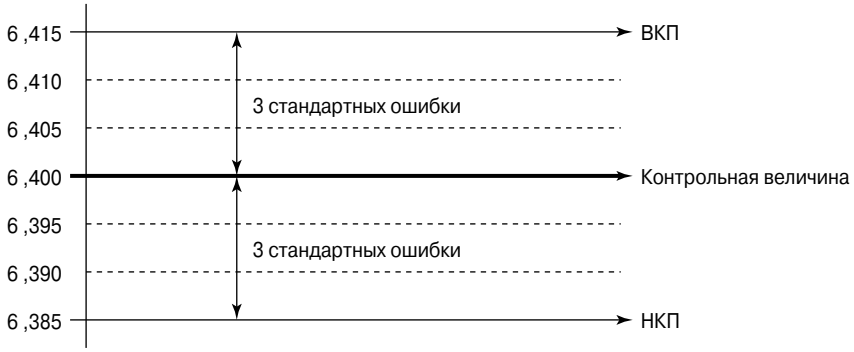


Рис. 19.4. Контрольные пределы для среднего выборок по 10 тюбиков зубной пасты (99% уверенность)

На контрольной диаграмме используются границы нормального распределения (2 или 3 стандартных ошибки выше или ниже среднего), но здесь же показано и изменение величин со временем.

Мониторинг процесса

После того как были установлены контрольные пределы, нужно проконтролировать процесс. Чаще всего для этого приходится делать выборки продукции на разных этапах, находить их средний вес и вносить его в контрольную диаграмму. Если с процессом возникли проблемы в том, что касается точности или постоянства, производственный процесс останавливается, начинается поиск проблемы, после чего вносятся необходимые коррективы.

Если процесс находится под контролем, то 68% средних выборки должны находиться в пределах 1 стандартной ошибки, 95% средних выборки — в пределах 2 стандартных ошибок, а 99% средних выборки — в пределах 3 стандартных ошибок от контрольной

величины, о чем гласит эмпирическое правило (см. главу 8). Общее среднее выборки должно равняться контрольной величине, а по обе стороны от него должно находиться равное количество значений, для которых не характерна особая модель.

В действительности если аппарат для наполнения тюбиков приходится останавливать, то замирает и вся производственная линия. Пока ремонтники ищут и устраняют проблему, теряется драгоценное время и производственные мощности. В связи с этим возрастают требования к системе статистического контроля производственных процессов, потому что она должна всегда предоставлять точную и надежную информацию. Прежде чем принять решение остановить процесс, сотрудники, отвечающие за качество продукции, должны окончательно убедиться в том, что проблема действительно существует. С другой стороны, если с производственной линией возникла проблема, они не захотят мириться с ней слишком долго. Нужно найти точку шаткого равновесия.

В связи с этим возникает следующий вопрос: как понять, что процесс "вышел из-под контроля"? Причем как это сделать без ложной тревоги, которая будет стоить компании времени и денег? Подобно многим другим вопросам, связанным со статистикой, на этот не существует единственно верного и окончательного ответа (как, например, в математике). Одни говорят, что именно это им и нравится в статистике, а другие утверждают, что именно это в статистике их раздражает.

Если контрольные пределы установлены на уровне плюс/минус 2 стандартных ошибки, тогда 95% средних значений должны находиться в этих рамках, что и будет свидетельствовать о том, что процесс находится под контролем. Но это означает, то 5% результатов могут не попасть в указанные пределы просто по случайности, и с этим придется смириться! Вот где подвох. Вы не хотите останавливать процесс, как только какое-то из значений вышло за контрольные пределы, ведь этого можно ожидать в 5% случаев (подробнее об этом см. главу 10). Значит, чтобы не допустить слишком частого повторения ложной тревоги, нужно, чтобы не одно, а намного больше значений вышли за установленные пределы, только после этого можно принять решение остановить процесс.

Вы хотите остановить процесс, только если полностью уверены в том, что где-то возникла проблема, а среднее одной выборки, оказавшееся вне контрольных пределов, не такое уж необычное явление. Но что вы скажете, если 2, 3, 4, 5 или больше результатов подряд также не попадают в установленные пределы? Где же пороговая точка? Добро пожаловать в чудесный мир туманной статистики! Вот четыре правила, которые часто применяются для того, чтобы определить, не вышел ли процесс из-под контроля и не нужно ли его остановить:

- ✓ Средние пяти выборок подряд находятся либо выше, либо ниже контрольной величины (как показано на рис. 19.5, *а*). Возможная причина: систематическое неправильное наполнение из-за проблем с процессом.
- ✓ Средние шести выборок подряд либо постоянно увеличиваются, либо постоянно уменьшаются (как на рис. 19.5, *б*). Возможная причина: продукция, сходящая с конвейера, постепенно все дальше отходит от установленного среднего значения, вероятно, из-за проблем со станками.

- ✓ Средние четырнадцати выборок подряд оказываются то выше, то ниже контрольной величины (как на рис. 19.5, *в*). Возможная причина: два разных оператора, станка или поставщика участвуют в системе, но действуют несогласованно.
- ✓ Средние пятнадцати выборок подряд находятся в пределах всего 1 стандартной ошибки от контрольной величины (как на рис. 19.5, *г*). Возможная причина: процесс более постоянен, чем это предусмотрено требованиями. (Если это стоит денег и времени, то требования нужно ослабить. Если это не приводит к потере времени и денег, то выяснить, почему процесс изменился, — и повторить это изменение в будущем — может оказаться полезно.)

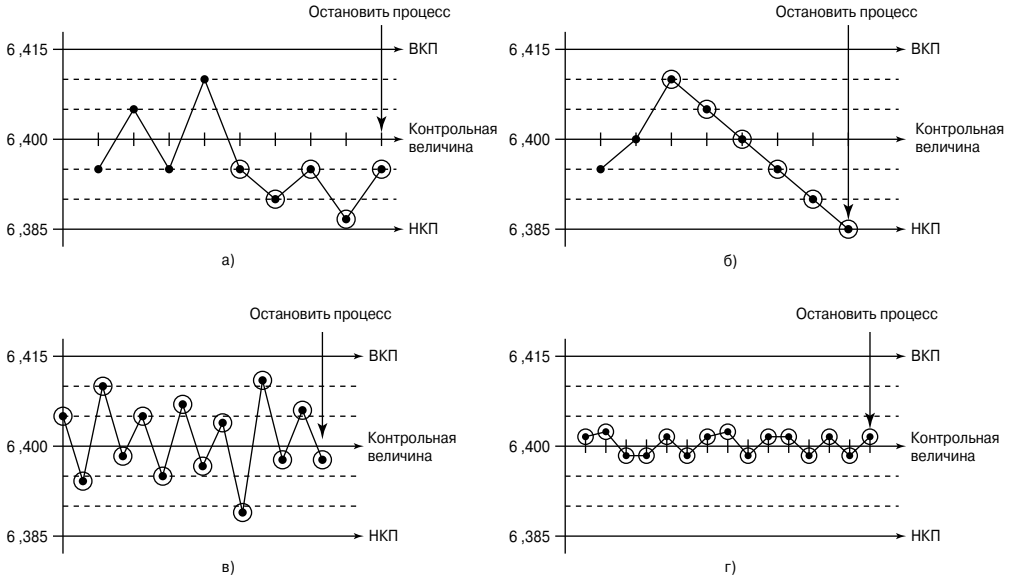


Рис. 19.5. Процессы наполнения тюбиков зубной пастой, вышедшие из-под контроля

Эти правила основаны на вероятности. Вы останавливаете процесс, когда шансы того, что он остается под контролем, слишком малы, о чем свидетельствуют полученные данные. Обратите внимание, что шансы того, что среднее одной выборки окажется выше или ниже контрольной величины, равны 50% или 0,5. Значит, в случае с первым правилом из четырех вышеперечисленных вероятность того, что средние пяти выборок подряд окажутся по одну сторону от контрольной величины, составляет $(0,5) \times (0,5) \times (0,5) \times (0,5) \times (0,5) = 0,03 = 3\%$. Такой результат меньше обычной вероятности, принятой в качестве пороговой точки (см. главу 14). Вы делаете вывод, что процесс вышел из-под контроля. Шанс того, что он все еще остается под контролем, слишком мал, учитывая все данные.

Когда вы в следующий раз откроете новый тюбик с зубной пастой, подумайте обо всех статистических показателях, которые нужно было учесть, чтобы удостовериться, что он наполнен качественно.

Часть VIII

Великолепные десятки

The 5th Wave

Рич Теннант

В этой части...

Разве может книга по статистике обойтись без собственных статистических данных? В этой части вы найдете десять критериев для проведения хорошего опроса и десять распространенных статистических ошибок.

Эта часть послужит для вас продуманным, сжатым справочником, с помощью которого вы сможете критически оценить или разработать опрос, а также обнаружить часто встречающиеся погрешности в статистике.

Глава 20

Десять критериев хорошего опроса

В этой главе...

- Критический подход к опросу
- Разработка качественного опроса

О результатах опросов можно услышать со всех сторон, и весьма вероятно, что в определенный момент жизни вас попросят принять участие в одном из них. Однако прежде чем потреблять информацию, нужно выяснить, правильно ли опрос был разработан и проведен — другими словами, не считайте результаты опроса бесспорными, пока не проверите все сами (подробнее о проблемах с опросами см. главу 16). Два важных требования к любому опросу — он должен быть *точным* (т.е. основываться на достаточном количестве данных, чтобы при другой выборке результаты не слишком изменились) и иметь минимальную степень *смещения* (систематического недооценивания или переоценивания истинных результатов, как это бывает с весами в ванной, которые всегда показывают вес на два килограмма больше!). В этой главе вы найдете десять критериев, с помощью которых сможете оценить или разработать опрос.

Целевая совокупность установлена правильно

Целевая совокупность — это вся группа элементов, которые вы хотите изучить. Например, вы стремитесь выяснить, что думают жители Великобритании о реалити-шоу. В таком случае целевой совокупностью будут все граждане этой страны.

Иногда определить целевую совокупность бывает не совсем просто. К примеру, какие возрастные группы вы хотите в нее включить? В примере с реалити-шоу вас, скорее всего, не заинтересуют дети младше определенного возраста, скажем, до 12 лет. Значит, на самом деле вашей целевой совокупностью будут все жители Великобритании в возрасте от 12 лет и старше.

Многие исследователи наплевательски относятся к точному определению целевой совокупности. Например, если Американский совет по маркетингу яиц хочет сказать: “Яйца полезны для вас!”, то здесь нужно уточнить, кто это “вы”. Например, готов ли Совет утверждать, что яйца полезны для людей с высоким уровнем холестерина? (Одно из исследований, результаты которого использовались для такого утверждения, проводилось с участием здоровых молодых людей, придерживающихся диеты с низким содержанием жиров, —вот кто имелся в виду под словом “вы”!)

Если целевая совокупность определена неточно, результаты опроса, скорее всего, необъективны. Это объясняется тем, что в изучаемую выборку могли входить люди, не относящиеся к интересующей вас совокупности, или же среди участников не было тех, кого стоило бы включить в исследование.

Выборка соответствует целевой совокупности

Проводя опрос, в поисках интересующей информации вы, как правило, все же не будете задавать вопросы всем членам целевой совокупности до единого. Для этого обычно не хватает ни времени, ни денег. Лучшее, что можно сделать в такой ситуации — провести *выборку* (отобрать группу элементов из совокупности) и собрать информацию о ней. Поскольку такая выборка элементов является единственным связующим звеном между вами и целевой совокупностью, то сделана она должна быть качественно.

Хорошая выборка представляет целевую совокупность. В ней не отдается предпочтение определенным группам из целевой совокупности, но в то же время никто и не исключается из такой выборки. Кажется, все довольно просто. Все, что нужно сделать — это составить перечень всех элементов целевой совокупности (он называется *основой выборки*) и отобрать из них определенную группу. Разве это сложно?

Вообще-то да. Предположим, целевой совокупностью являются все зарегистрированные избиратели в Соединенных Штатах Америки, которые, возможно, будут голосовать на следующих президентских выборах. Составить такой список очень сложно. Можно изучить списки избирателей, но ведь неизвестно, сколько людей пойдут голосовать на следующих выборах. Можно проверить тех, кто голосовал в прошлый раз, но многие из них могли переехать или уже умереть, в эту группу не будут включены также те, кому исполнилось 18 лет после последних выборов. Внезапно ситуация оказывается очень запутанной. Добро пожаловать в мир опросов!

Одно из возможных решений этой проблемы — получить новейшие списки зарегистрированных избирателей, сделать из них выборку и спросить этих людей, собираются ли они голосовать на будущих выборах. Если кто-то не планирует этого делать, не задавайте ему больше никаких вопросов и не включайте этого человека в свой опрос. У тех же, кто собирается идти на выборы, спросите, за кого они хотят проголосовать, и включите их ответы в результаты своего опроса.

У хорошего опроса имеется современная, тщательно продуманная основа выборки, по возможности включающая всех членов целевой совокупности. Если же это невозможно, тогда необходим определенный механизм, благодаря которому у всех членов совокупности будут равные шансы попасть в выборку для проведения опроса. Например, если нужно провести опрос по всем домам в городе, то в качестве основы выборки следует использовать новейшую карту, на которой обозначены все городские строения.

Выборка делается наугад

Еще одна важная черта хорошего опроса — выборка из целевой совокупности делается в случайном порядке. *Случайно* означает, что у всех членов целевой совокупности равные шансы попасть в выборку. Другими словами, процесс, при помощи которого вы делаете выборку, является объективным.

Предположим, есть стадо из 1000 волов, и вам нужно сделать случайную выборку в 50 животных, чтобы проверить их на наличие заболевания. Если взять 50 первых волов, которых вы увидите на поле, это не будет случайной выборкой. Бычки, которые смогли подойти к вам, скорее всего, окажутся здоровыми или же они могут быть более взрослыми, дружелюбными, т.е. более расположенными к болезням. В любом случае, выборка получает смещение. Как же сделать случайную выборку волов? Животные, скорее всего, имеют идентификационные номера, значит, вы берете список этих номеров, делаете из них случайную выборку и находите этих животных. Или же, если бычки находятся в клетках или в закрытых стойлах, пронумеруйте эти помещения и сделайте случайную выборку из номеров. Иногда быть статистиком означает проявлять незаурядную находчивость в том, как сделать действительно случайную выборку!

В случае с опросами с участием людей такие уважаемые организации, как *The Gallup Organization,* используют метод случайного цифрового набора и звонят членам выборки по телефону. Сюда, конечно, не попадают люди, у которых нет телефона, значит, такой опрос тоже не совсем объективен. Однако на сегодня все же телефоны есть у большинства (согласно данным *The Gallup Organization*, у 95% населения), значит, субъективность по отношению к людям без телефона все-таки не слишком серьезна.

Хороший опрос проводится с использованием случайной выборки элементов из целевой совокупности. Если изначально не сообщается, как делалась выборка, обязательно выясните это.

Выборка большого размера

Наверняка вы слышали такое высказывание: “Лучше меньше, да лучше”. В случае с опросами оно звучит слегка по-другому: “Меньше хорошей информации лучше, чем больше плохой, но больше хорошей информации — еще лучше”. (Не совсем доступно, да?)

Разберемся, в чем суть этого высказывания. Если ваша выборка достаточно большая и действительно представляет всю целевую совокупность (т.е. была сделана случайно), то можно рассчитывать, что полученная информация окажется довольно точной. Но вот насколько точной она будет, зависит от размера выборки, т.е. чем больше выборка, тем точнее информация.

Быстрая и обобщенная формула для определения точности опроса такова: разделите единицу на корень квадратный из размера выборки. Например, опрос 1000 человек (выбранных случайно) точен на $\frac{1}{\sqrt{1000}}$, т.е. 0,032 или 3,2%. Этот процент называется *пределом погрешности*.

Остерегайтесь опросов, проведенных с участием большой выборки, которая *не была* случайной. Самый яркий пример — это опросы в сети Интернет. Компания может заявить, что для участия в проводимом ею опросе сайт посе-

тили 50 000 человек, и отсюда следует, что для подведения результатов опроса была собрана масса информации. Однако вся она необъективна, потому что не представляет мнения других людей, помимо тех, что принимали участие в опросе. Другими словами, у них был доступ к сети, они зашли на конкретный Web-сайт и решили ответить на предложенные вопросы. В данном случае меньше было бы лучше: компания могла бы опросить меньше человек, но выбрать их наугад.

Продуманное повторение уменьшает погрешность

После того как был установлен размер выборки, а затем сделана случайная выборка элементов целевой совокупности, вам предстоит получить нужную информацию от членов этой выборки. Если вы когда-либо выбрасывали бланк анкеты или резко отказывались ответить на несколько вопросов по телефону, тогда вам должно быть понятно, что убедить людей принять участие в опросе не так уж и легко. Если исследователь намерен свести смещение к минимуму, то лучше всего это сделать, "надоедая" участникам выборки: повторите опрос два, а то и три раза, предложите вознаграждение за участие, конверты с оплаченной пересылкой, возможность выиграть приз и т.д. Но заметьте, что если перестараться со стимулом к участию в опросе, это тоже может привести к смещению, потому что тогда отвечать будут те люди, которым нужны деньги, а не те, кто действительно хочет высказать свое мнение.

Подумайте, что может лично вас убедить принять участие в опросе. Если стимул, который предлагает исследователь, вас не трогает, тогда, возможно, изучаемый вопрос тоже мало вас волнует. К сожалению, именно здесь и появляется смещение. Если участвовать в опросе будут только те, кто действительно хочет высказаться, тогда будут учитываться только их мнения, ведь остальные люди, которых изучаемая проблема никак не волнует (т.е. каждый ответ "Мне все равно"), не принимаются в расчет. Не учитываются также голоса тех, кто все же имеет, что сказать, но в данный момент у него просто нет времени ответить на ваши вопросы.

Например, представим, 1000 человек должны высказать свое мнение по поводу того, стоит ли отменить запрет на выгул собак в парке. Кто будет отвечать? Скорее всего, респондентами станут те, у кого уже давно сложилось твердое мнение по этому вопросу и кто либо активно поддерживает, либо яростно протестует против изменений. Скажем, по 100 человек из каждой группы решат принять участие в опросе, а оставшиеся 800 анкет так и не будут заполнены. Это означает, что 800 мнений не учитываются. Если ни одному из этих 800 человек действительно нет дела до собак в парке и если бы вы могли учесть и их мнение, тогда результат был бы таким: 800/1000 или 80% ответов "все равно", 10% (100/1000) "за" и 10% (100/1000) "против" предлагаемых изменений. Не учитывая 800 голосов тех, кто отказался от участия в эксперименте, исследователи могут заявить: "Среди респондентов 50% выступили в поддержку предложения, а 50% — против него". Таким образом полученные результаты будут выглядеть уже совсем в ином свете.

Доля ответивших на опрос — это процентное отношение, которое определяется, если разделить количество респондентов на общий размер выборки и умножить частное на 100%. Хорошая доля ответивших — это результат больше 70%. Но в большинстве случаев реальная доля ответивших оказывается

намного меньше этого значения, если только опрос не проводит такая уважаемая организация, как *The Gallup Organization*. Изучая результаты опроса, не забудьте уточнить долю ответивших. Если она будет слишком низкой (намного меньше 70%), то результаты, скорее всего, необъективны и их можно проигнорировать.

Лучше сделать меньшую выборку и добиться высокой доли ответивших, чем остановиться на выборке большого размера, но с очень низкой долей ответивших. Активные повторы опросов уменьшают смещение.

Когда вас в следующий раз попросят принять участие в хорошем опросе (согласно критериям, описанным в этой главе), все же не отказывайтесь. Этим вы внесете свою лепту в дело избавления от необъективности!

Выбран подходящий тип опроса

Существует множество видов опросов: по почте, по телефону, в сети Интернет, поквартирный опрос и опрос на улице (когда к вам подходит человек с папкой в руках и спрашивает: "У вас есть несколько минут для того, чтобы ответить на пару вопросов?"). Один из очень важных, но часто недооцениваемых критериев хорошего опроса — тот факт, правильный ли тип опроса был выбран для конкретной ситуации.

Например, если целевая совокупность — это совокупность людей с плохим зрением, то разослать им по почте анкету, напечатанную мелким шрифтом — это не очень удачная мысль (да, случается и такое). Если вы хотите провести опрос жертв домашнего насилия, то спрашивать их об этом дома тоже не совсем уместно.

Предположим, ваша целевая совокупность — это бездомные и бродяги. Как связаться с ними? У них нет адресов или телефонов, поэтому не подойдет ни один из привычных типов опроса. (Это очень сложная проблема, с которой правительство то и дело вынуждено бороться при проведении очередной переписи.) Вы можете просто ходить и общаться с людьми лично, где бы они ни находились, но вот выяснить, где они находятся, совсем не просто. Начать можно с того, что обратиться в местные приюты, церкви или другие благотворительные организации, помогающие бездомным, и по их подсказке начать поиски.

Изучая результаты опроса, обязательно выясните, какой тип опроса был использован, и подумайте, подходил ли он для конкретной ситуации.

Вопросы сформулированы правильно

Формулировка вопросов тоже может в значительной степени повлиять на результаты опроса. Например, когда президентом был Билл Клинтон и разразился скандал с Моникой Левински, то 21–23 августа 1998 года канал *CNN* и организация *Gallup* попросили респондентов дать оценку президенту, и около 60% высказались положительно. (В этом случае была сделана выборка в 1317 человек, а предел погрешности составил плюс/минус 3%.) Когда же устроители опроса перефразировали вопрос и попросили оценить Клинтона "как человека", результаты изменились: положительно о нем отозвались всего 40% опрошенных.

На следующий день *CNN* и *Gallup* провели еще один опрос по той же теме. Вот некоторые вопросы и ответы на них.

- ✓ Вы одобряете или не одобряете то, как президент Клинтон справляется со своими обязанностями? (62% одобряют, 35% не одобряют работу президента.)
- ✓ Ваше мнение о Клинтоне в целом благосклонное или неблагосклонное? (44% поддерживают Клинтона, 43% не поддерживают.)
- ✓ В сложившейся ситуации рады ли вы, что президентом является Клинтон? (56% сказали "да", 42% — "нет".)
- ✓ Если бы вы могли еще раз проголосовать за кандидатов в президенты от 1996 года, кому бы вы отдали предпочтение? (46% сказали, что проголосовали бы за Билла Клинтона, 34% — за Боба Доула, а 13% — за Росса Перота.)

Все эти вопросы касаются одной и той же темы: что люди думали о президенте Клинтоне во время скандала с Моникой Левински. И хотя вопросы похожи, но все они сформулированы немного по-другому, в зависимости от этого меняются результаты. Следовательно, формулировка вопроса имеет значение.

Наверное, самая серьезная проблема с формулировкой — это использование *наводящих вопросов*. Другими словами, это вопросы, которые сформулированы так, что вы заранее знаете, что от вас хотят услышать в ответ. Отсюда появляются необъективные результаты, придающие слишком большое значение определенному ответу из-за формулировки вопроса, а не потому, что люди действительно придерживаются такого мнения.

Наводящие вопросы встречаются очень часто (то ли случайно, то ли намеренно), что заставляет вас говорить именно то, что от вас хотят услышать. Вот некоторые примеры наводящих вопросов, подобных тем, что я встречала в печати.

- ✓ Какую позицию вы поддерживаете? Демократы выступают за продуманное, реалистичное финансовое планирование, благодаря которому можно сбалансировать бюджет на довольно долгий срок и вместе с тем выполнить предвыборные обещания, данные самым незащищенным слоям населения в США. Республиканцы предлагают заметно сократить расходы на образование и здравоохранение.
- ✓ Согласны ли вы с тем, что у президента должно быть право постатейного вето, что позволит сократить расходы?
- ✓ Как вы думаете, должны ли Конгресс и Белый дом подать хороший пример и сократить надбавки и привилегии, которыми они пользуются за счет налогоплательщиков?

Изучая результаты важного для вас опроса, попросите предоставить вам копию вопросов, которые задавались людям, и проанализируйте их, чтобы убедиться, что они были сформулированы нейтрально и как можно объективнее.

Опрос проводится своевременно

Время проведения опроса решает все. Текущие события влияют на мнения людей, и если одни интервьюеры стремятся выяснить, что же думают люди в действительности, то другие пользуются сложившимися условиями, особенно если они негативны. Например, когда в школах происходят стычки с применением оружия, то в опросах очень часто затрагивается проблема контроля над продажей оружия. Конечно, сразу после трагедии за усиление контроля высказывается больше человек, чем обычно, другими словами, цифры взлетают вверх. Но спустя некоторое время мнения возвращаются к прежнему показателю, а устроители опросов, между тем, сообщают о результатах так, как будто мнение общественности остается неизменным.

Время проведения опроса, независимо от изучаемой темы, тоже может вызвать смещение. Например, предположим, в вашу целевую совокупность попадают люди, работающие полный рабочий день. Если проводить опрос по телефону и при этом звонить им домой днем, с 9 до 17, то результаты будут совершенно смещенными, потому что именно в это время большинство из опрашиваемых находится на работе!

Проверьте, когда проводился опрос, и посмотрите, не было ли в то время каких-либо событий, которые могли бы повлиять на результаты. Также узнайте, в какое время суток проводился опрос: действительно ли это время было самым удобным для общения с представителями целевой совокупности?

Опрос проводят профессионалы

У людей, проводящих опрос, очень непростая работа. Им приходится иметь дело с резкими отказами, грубостью и автоответчиками. И вот, когда наконец на том конце провода или при личной встрече появляется респондент, исследователю предстоит решить основную задачу — точно и объективно собрать данные.

Ниже перечислены проблемы, которые могут возникнуть в процессе опроса.

- ✓ Респондент не понимает вопроса и требует дополнительную информацию. Что можно ему рассказать и при этом сохранить нейтральность?
- ✓ Информация фиксируется неправильно. Например, я говорю, что мне 40 лет, а он по ошибке записывает 60.
- ✓ Человек, проводящий опрос, должен принимать спонтанные решения. К примеру, ему нужно выяснить, сколько человек живет в доме, а респондент спрашивает: “Нужно ли считать кузена Боба, который остановился у нас, пока ищет работу?” Нужно быстро принимать решение.
- ✓ Респонденты могут давать неверную информацию. Например, некоторые люди так сильно ненавидят опросы, что не отказываются от участия в них, а дают неправдивые ответы. Скажем, на вопрос о возрасте женщина может ответить: “101 год”.

Как интервьюеры решают эти и другие проблемы, возникающие в процессе опроса? Главное — ясно и четко представлять, что может произойти, обсудить возможные выходы из ситуации, причем провести подобную дискуссию до того, как вступать в контакт с участниками опроса. Другими словами, персонал должен быть хорошо обучен.

Избежать проблем можно также, если провести *пилотное исследование* (пробный опрос с участием всего нескольких респондентов), которое записывается с тем, чтобы потом можно было попрактиковаться и оценить точность и последовательность персонала при сборе данных. Таким образом исследователи могут предугадать проблемы еще до начала опроса и найти способы их решения.

В опросах не должно быть неясных, двусмысленных или наводящих вопросов. С помощью пилотного исследования можно также обнаружить потенциально сложные вопросы, изучив ответы небольшой группы и те вопросы, которые они попутно задают. Чтобы избежать ошибок при записи информации (т.е. описок и опечаток), нужно, чтобы все возможные ответы были четко обозначены. Например, если активная позиция "против" обозначается 1, а активная позиция "за" — 5 (а не наоборот), то это нужно четко оговорить. Не помешает также записывать процесс опроса на пленку, чтобы позже можно было все перепроверить.

Обязательно заранее решите, как вести себя с *респондентами-шутниками*, т.е. людьми, которые просто развлекаются. Один из вариантов — отметить полученный от такого человека ответ как ненадежный, а затем перезвонить по тому же номеру позже и попробовать все сначала.

Изучение проблемы с помощью опроса

Предположим, исследователь начинает с такого утверждения: "Я хочу выяснить привычки покупателей". На первый взгляд, все ясно. Но затем оказывается, что все вопросы касаются того, как люди относятся к походу в магазин ("Что вам больше/меньше всего нравится в шопинге?" или "Оцените по шкале от 1 до 10, насколько вам нравится ходить по магазинам"). Эти вопросы никак не связаны с привычками покупателей. Хотя отношение к шопингу влияет на привычки, но все же реальным мерилом является поведение покупателей: что они покупают, сколько денег тратят, в какие магазины ходят, с кем, когда, как часто и т.д. Иногда исследователи не осознают, что упустили суть, пока не подведут итоги опроса. Только после этого (когда уже слишком поздно что-либо менять) они понимают, что не могут ответить на вопрос исследования с помощью собранных данных, а это очень плохо!

Прежде чем участвовать в опросе, выясните у исследователя, что он пытается установить — т.е. какова цель этого опроса. Потом прочтите и выслушайте вопросы, и если окажется, что они уводят вас в совсем другом направлении, я бы посоветовала отказаться от участия и объяснить причину или письменно, или при личной встрече.

Прежде чем разрабатывать опрос, сначала определите его цели. Что вы хотите узнать? Затем сформулируйте вопросы, которые помогут достичь этих целей. Тогда вы обязательно добьетесь своего.

Глава 21

Десять распространенных статистических ошибок

В этой главе ...

- Знакомство с самыми распространенными ошибками в статистике, которые допускают исследователи и журналисты
- Избежание ошибок при работе со статистикой

Эта книга посвящена не только объяснению статистических понятий, с которыми вы сталкиваетесь в средствах массовой информации и в работе, но и позволяет копнуть глубже, чтобы понять, правильны ли, разумны и справедливы эти статистические данные. Чтобы выжить в современном потоке информации, приходится сохранять бдительность — и здоровый скептицизм, — потому что многие статистические данные ошибочны или вводят в заблуждение либо случайно, либо намеренно. Кто же будет критически оценивать потребляемую информацию, если не вы? В этой главе речь пойдет о некоторых самых распространенных статистических ошибках, которые допускают исследователи и журналисты, а также о том, как научиться замечать и избегать таких ошибок.

Графики, вводящие в заблуждение

Во многих графиках и диаграммах содержится неверная информация, или же в них отсутствует как раз та важная информация, которая нужна читателю для принятия взвешенного решения о предложенном материале. На рис. 21.1 приведены примеры четырех типов визуального представления данных: секторные диаграммы, столбиковые диаграммы, временны*е* диаграммы и гистограммы. (Обратите внимание, что *гистограмма* — это, по сути, столбиковая диаграмма для числовых данных.) Для каждого из них существуют самые распространенные способы ввести вас в заблуждение. (Подробнее о диаграммах и графиках, в том числе и тех, что сбивают читателя с толку, см. главу 4.)

Секторные (круговые) диаграммы

Секторные диаграммы по своей сути полностью соответствуют такому названию — это окружность, разделенная на секторы, которые представляют процентное отношение элементов в каждой группе (согласно какой-либо категорийной переменной, например, пол, политическая партия или трудовой статус).

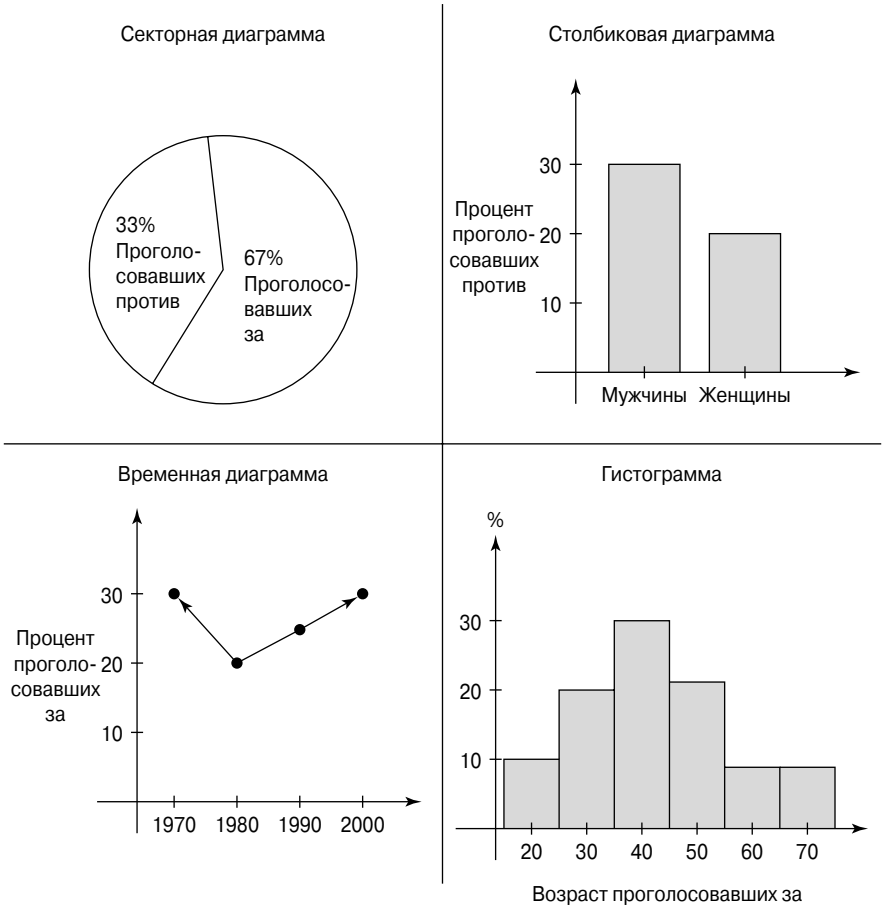


Рис. 21.1. Четыре типа визуального представления данных

Вот как можно проверить качество секторной диаграммы.

- ✓ Убедитесь, что все проценты в сумме дают 100% или около того (любая ошибка при округлении должна быть несущественной).
- ✓ Настороженно относитесь к секторам под названием “Другое”, которые по размеру превосходят остальные секторы. Это значит, что диаграмма слишком расплывчата.
- ✓ Помните об искажениях, которые встречаются в трехмерных диаграммах, когда ближайший к вам сектор кажется больше, чем есть на самом деле, из-за угла изображения.
- ✓ Выясните, указано ли общее число элементов, представленных на секторной диаграмме, чтобы можно было решить, какой была окружность до того, как ее поделили на секторы и представили читателю. Если размер набора данных (количество респондентов) слишком мал, то информация не будет надежной.

Столбиковые диаграммы

Столбиковая диаграмма похожа на секторную с той лишь разницей, что в этом случае каждая группа элементов представлена в виде столбика, высота которого отражает количество или процентное отношение элементов в группе.

Рассматривая столбиковую диаграмму, поступайте следующим образом.

- ✓ Обратите внимание на то, в каких единицах измерения представлена высота столбиков и как эти единицы связаны с окончательными результатами. Например, общее количество преступлений или уровень преступности, который измеряется количеством преступлений *на душу населения*.
- ✓ Оцените обоснованность шкалы, т.е. расстояние между единицами, в которых выражено количество элементов в каждой группе. К примеру, количество жалоб от клиентов-мужчин или от клиентов-женщин можно выразить в единицах 1, 5, 10 или 100. Маленькая цена деления (например, в 10 единиц, если всего на шкале представлены единицы от 1 до 500) приведет к тому, что различие покажется серьезнее, а при большой цене деления (если от 1 до 500 счет идет по сотням) различия будут не столь заметными.

Временны*е* диаграммы

На временн*ой* диаграмме показано, как определенная величина меняется с течением времени (например, цена акций, средний доход семьи или средняя температура).

При изучении временн*ой* диаграммы обратите внимание на следующее.

- ✓ Шк*а*лы как на вертикальной (величина), так и на горизонтальной (промежуток времени) осях. Если всего лишь изменить шкалу, то результаты можно сделать более или менее выразительными, чем они есть на самом деле.
- ✓ Единицы измерения, используемые на диаграмме. Убедитесь, что они подходят для сравнения перемен, произошедших со временем. Например, была ли учтена инфляция, если величины показаны в долларовом эквиваленте?
- ✓ Помните, что люди могут попытаться объяснить замеченные тенденции, не имея для доказательств никаких статистических данных. Как правило, временн*ая* диаграмма показывает, *что* происходит. А вот *почему* это явление наблюдается — это совсем другая история!
- ✓ Остерегайтесь ситуаций, когда ось времени промаркирована неравнозначными отрезками. Такое часто бывает при нехватке данных. К примеру, на оси времени равные отрезки могут быть обозначены годами 1971, 1972, 1975, 1976 и 1978, хотя те годы, для которых нет соответствующих данных, должны также быть показаны в виде пустых промежутков.

Гистограммы

Гистограмма — это график, на котором выборка делится на группы в зависимости от числовой переменной (например, возраст, рост, вес или доход). На таком графике показано либо количество (частота), либо процентное отношение (относительная частота)

элементов в каждой группе. Например, предположим, 20% выборки были представлены людьми в возрасте до 20 лет, 30% — от 20 до 40 лет, 45% — между 40 и 60, а 5% членов выборки были старше 60 (в данном случае числовой переменной является возраст).

Далее перечислены аспекты, о которых стоит помнить при изучении гистограммы.

- ✓ Обратите внимание на шкалу на вертикальной (частота/относительная частота) оси, с помощью которой можно намеренно преувеличить или приуменьшить результаты, если выбрать неподходящую цену деления.
- ✓ Проверьте единицы на вертикальной оси в зависимости от того, о чем идет речь — о частоте или относительной частоте. При изучении информации не забывайте о единицах измерения.
- ✓ Изучите шкалу, которая используется для групп числовых переменных на горизонтальной оси. Если группы основываются на небольших промежутках (например, 0–2, 2–4 и т.д.), то данные могут показаться очень неустойчивыми. Если же за основу взяты большие промежутки (0–100, 100–200 и т.п.), данные будут выглядеть стабильнее, чем есть в действительности.

Смещенные данные

Смещение в статистике — это результат систематической ошибки, которая приводит или к преувеличению, или к приуменьшению истинного значения. Например, если я измеряю высоту растений с помощью линейки, которая короче, чем нужно, на 1 см, то все полученные результаты будут смещенными — они систематически будут оказываться меньше истинного значения.

Вот некоторые самые распространенные источники смещенных данных.

- ✓ **Измерительные приборы показывают неверный результат**. Например, радар у полицейского фиксирует, что вы ехали со скоростью 76 миль в час, а вы *знаете*, что скорость была всего 72 миль в час. Или вспомните весы, которые всегда добавляют 2 кг к вашему весу.
- ✓ **На участников влияет сам процесс сбора данных**. Например, вопрос, сформулированный так: “Вы *когда-нибудь* не соглашались с правительством?”, приведет к тому, что процентное отношение людей, недовольных правительством, заметно увеличится.
- ✓ **Выборка не представляет изучаемую совокупность**. К примеру, если изучать привычки студентов, посетив только университетскую библиотеку, то в выборке будет представлено больше прилежных учащихся. (Подробнее об этом см. в разделе “Неслучайные выборки” далее в этой главе.)
- ✓ **Исследователи необъективны**. Например, предположим, одной группе пациентов дают сахарные пилюли, а другой — настоящее лекарство. Исследователи не должны знать, кто какой препарат принимает, потому что если им будет это известно, то они могут спроецировать ожидаемый результат на пациентов (например, сказать: “Вам ведь уже лучше, правда?”) или уделять больше внимания тем, кто получает настоящее лекарство.

Чтобы заметить смещенные данные, выясните, как они собирались. Поинтересуйтесь у некоторых участников, как проводилось исследование, какие вопросы задавались, какое лечение (лекарства, процедуры, терапия и т.п.) предлагалось (если оно вообще было) и кто знал об этом, какие измерительные приборы использовались, как они калибровались и т.п. Обращайте внимание на повторяющиеся ошибки, и если их будет слишком много, игнорируйте результат.

Предел погрешности отсутствует

У слова "погрешность" чувствуется какая-то негативная окраска, как будто погрешность — это явление, которого всегда можно избежать. В статистике такое бывает не каждый раз. Например, определенная степень так называемой ошибки выборки будет присутствовать всякий раз при попытке оценить совокупность с помощью любого другого явления, нежели вся эта совокупность целиком. Сам факт выборки элементов из совокупности означает, что вы не включаете в исследование каких-то представителей этой совокупности, а значит, и не получите точного, достоверного значения для всей совокупности. Но не волнуйтесь. Помните, что в статистике никогда не приходится говорить, что вы в чем-то уверены наверняка — нужно лишь приблизиться к истине. И если выборка достаточно большая и сделана случайно, то ошибка выборки будет незначительной.

Для оценки статистического результата нужно определить его точность — обычно это делается с помощью предела погрешности. *Предел погрешности* сообщает, в какой степени исследователь допускает, что полученные результаты будут варьироваться от выборки к выборке. (Подробнее о пределе погрешности см. главу 10.) Если исследователь или журналисты не могут назвать предел погрешности, то вам остается только догадываться, насколько точны результаты, или еще хуже — вы просто предполагаете, что все в порядке, хотя во многих случаях это не так. Результаты опросов, о которых ранее сообщалось по телевидению, очень редко содержали упоминание о пределе погрешности, но теперь такая информация приводится часто. Хотя многие опросы в газетах, журналах и в Интернете проводятся без указания предела погрешности, или же предел погрешности указывается совершенно бессмысленный, потому что данные изначально были смещены (см. главу 10).

Изучая статистические результаты, в которых дается предположение о каком-либо количестве (например, процентное отношение людей, полагающих, что президент хорошо справляется со своими обязанностями), всегда выясняйте предел погрешности. Если он не указывается, требуйте! (Или же, имея достаточно информации, вы сможете подсчитать предел погрешности самостоятельно, используя для этого формулы, которые приведены в главе 10.)

Неслучайные выборки

Все статистические показатели, от данных опросов до результатов медицинских исследований, основываются на сведениях, собранных у выборки элементов, а не у всей совокупности, что объясняется затратами времени и средств. Далее, для того, чтобы получить крайне точные результаты, выборка не должна быть очень большой, главное, чтобы она представляла всю изучаемую совокупность. Например, у хорошо разработан-

ного и проведенного опроса 2500 человек предел погрешности равен плюс/минус 2% (см. главу 10). Для получения точной информации в ходе эксперимента с изучаемой и контрольной группами желательно, чтобы в этих группах было как минимум по 30 человек в каждой.

Но как можно добиться того, чтобы выборка представляла всю совокупность? Лучший способ — в *случайном* порядке выбрать представителей совокупности. Случайная выборка — это представители совокупности, выбранные таким образом, что у всех членов совокупности остаются равные шансы попасть в эту группу (например, вытягивание бумажки с именем из шапки). При случайной выборке никому не отдается предпочтение и никто не исключается из группы намеренно.

Многие исследования и опросы проводятся на основе неслучайных выборок элементов. К примеру, очень часто в медицинских экспериментах участвуют добровольцы, которые сами изъявили желание принять участие — значит, они не были выбраны случайно. Нельзя было бы позвонить человеку домой и сказать: "Вас наугад выбрали для участия в нашем эксперименте по изучению сна. Вам нужно прийти к нам в лабораторию и пробыть здесь четыре ночи". В такой ситуации лучшее, что можно сделать — это изучить добровольцев и выяснить, наглядно ли они представляют совокупность, а только потом обнародовать результаты. Можно также подбирать определенные типы добровольцев.

Опросы должны проводиться с участниками, выбранными случайно, и сделать это намного проще, чем в случае с медицинскими исследованиями. Но очень часто опросы основываются на неслучайных выборках. Например, телеопрос, при котором зрителям предлагают "позвонить и высказать свое мнение", — это неслучайная выборка. В таких опросах не у всех представителей совокупности есть равные шансы быть услышанными (более того, в подобных ситуациях люди сами изъявляют желание высказаться).

Прежде чем принимать решение о статистических результатах опроса или исследования, узнайте, как делалась выборка участников. Если она не была случайной, не воспринимайте результаты слишком серьезно.

Выборки неправильного размера

Количество информации всегда важно для точности статистического показателя. Чем больше информации получено, тем точнее будет результат — конечно, пока эта информация остается объективной (см. раздел "Смещенные данные" чуть выше в этой главе). Потребителю статистической информации нужно оценить ее точность, а для этого вы должны выяснить, как информация собиралась (см. главу 16 об опросах и главу 17 об экспериментах) и сколько данных было получено (т.е. вам нужно знать размер выборки).

Многие диаграммы и графики, появляющиеся в средствах массовой информации, не указывают размер выборки. Можно также заметить, что многие заголовки "не совсем" соответствуют тому, о чем идет речь в статье, потому что оказывается, что либо выборки была слишком маленького размера (что снижает надежность результатов), либо вообще никакой информации о размере выборки найти не удается. (Например, вы, возможно, видели рекламу жевательной резинки, в которой говорится: "Четыре из пяти опрошенных стоматологов рекомендуют эту жевательную резинку своим пациентам". А что, если и было опрошено всего пять стоматологов?)

Прежде чем принимать решение о статистической информации, всегда выясняйте размер выборки. Чем меньше выборка, тем менее надежна информация. Если в статье ничего не говорится о размере выборки, получите копию полного отчета об исследовании, свяжитесь с исследователем или журналистом, автором статьи.

Неверно истолкованные корреляции

В статистике *корреляция* понимается как прочность и направление линейной взаимосвязи между двумя числовыми переменными. Другими словами, это степень, на которую должна увеличиться или уменьшиться одна числовая переменная (например, вес) при увеличении/уменьшении другой числовой переменной (например, рост). Корреляция — одно из самых неверно истолковываемых и используемых статистических понятий, которое употребляют исследователи, журналисты и широкая общественность. С корреляцией связаны три следующих важных момента.

- ✓ **Корреляцию нельзя относить к двум *дискретным переменным*, таким как политическая партия или пол. Она используется только по отношению к двум *числовым переменным*, например, рост или вес.** Значит, если вы слышите, как кто-то говорит: "По-видимому, между политическими предпочтениями и полом существует корреляция", знайте, что это неправильно. Между политическими предпочтениями и полом может наблюдаться только ассоциация, а не корреляция, что и подразумевается в статистическом определении этого понятия.
- ✓ **Корреляция оценивает прочность и направление *линейной* зависимости между двумя числовыми переменными**. Другими словами, если вы соберете данные о двух числовых переменных (например, рост и вес) и нанесете все точки на график, то при наличии корреляции вы сможете провести через эти точки линию (направленную вверх или вниз), которая будет иметь вид прямой. Если прямая не получается, то между переменными корреляции нет. Однако это не обязательно означает, что переменные никак не связаны. Между ними может существовать и другая взаимосвязь, просто эта зависимость нелинейная. Например, количество бактерий за определенное время возрастает в *геометрической* прогрессии (удваиваясь все быстрее и быстрее), а не *линейно* (т.е. постепенно).
- ✓ **Корреляция не означает причину и следствие**. К примеру, предположим, вам говорят, что люди, которые пьют диетическую соду, страдают от опухоли мозга чаще, чем те, кто не потребляет соду. Если вы любите содовую, не стоит паниковать. Это может оказаться загадкой природы, которую кто-то заметил. И как минимум необходимо провести еще множество исследований (помимо простого наблюдения), чтобы утверждать, что содовая *вызывает* опухоль мозга.

Внешние переменные

Внешняя переменная — это переменная, не включенная в исследование, хотя она может в значительной степени повлиять на результат, т.е. оказать внешнее воздействие.

Например, предположим, что исследователь пытается доказать, что если употреблять в пищу морские водоросли, то вы проживете дольше. Ближе знакомясь с исследованием, вы выясняете, что оно основано на выборке людей, которые регулярно едят морские водоросли и перешагнули столетний рубеж. А затем вы читаете интервью с этими людьми и узнаете некоторые из их секретов долголетия (помимо употребления в пищу морских водорослей): они ели очень здоровую пищу, спали в среднем по 8 часов в день, пили много воды и ежедневно занимались спортом. Так действительно ли морские водоросли помогли им прожить дольше? Возможно, но наверняка сказать нельзя, потому что внешние переменные (упражнения, потребление воды, диета и достаточный сон) тоже могли повлиять на продолжительность жизни.

Распространенная ошибка, которую часто допускают в исследованиях, — это невнимание к внешним переменным, в связи с чем результаты начинают вызывать сомнение (а вы и должны в них усомниться)! Лучший способ контролировать внешние переменные — провести хорошо спланированный *эксперимент*, для которого нужно собрать две группы, максимально похожие друг на друга, при этом одна группа (которая называется *исследуемой*) принимает лекарство, а другая получает *плацебо* (ненастоящее лекарство; такая группа называется *контрольной*). Затем вы сравниваете результаты двух групп, после чего можете отнести любую существенную разницу на счет лекарства (и, в идеале, ничего другого).

Исследование влияния морских водорослей не было спланированным экспериментом — это было *исследование по результатам наблюдений*. В таких случаях не существует никакого контроля внешних переменных. Людей просто наблюдают и записывают информацию о них.

Когда вы видите результаты исследования, в котором вас пытаются убедить в наличии причинно-следственной взаимосвязи или существенной разницы между группами, выясните, было ли это исследование хорошо спланированным экспериментом и проводился ли контроль над внешними переменными. Если проводить эксперимент неэтично (например, чтобы доказать, что курение вызывает рак легких, нельзя заставить половину испытуемых выкуривать по 10 пачек в день в течение 20 лет, а вторую половину — не курить вообще), то приходится полагаться на данные многочисленных исследований по результатам наблюдений, которые в самых разных условиях показали тот же результат.

Исследования по результатам наблюдений прекрасно подходят для опросов, но не для доказательства причинно-следственной взаимосвязи, потому что с их помощью невозможно проконтролировать внешние переменные. Намного более убедительные доказательства вы получите в ходе хорошо спланированного эксперимента.

Ошибки в цифрах

Тот факт, что статистический показатель был опубликован в средствах массовой информации, еще не означает, что он правильный. Более того, ошибки происходят постоянно (случайно или намеренно), поэтому не забывайте о них. Вот несколько советов, как заметить ошибки в цифрах.

- ✓ **Убедитесь, что в сумме все числа дадут ту величину, которая указана**. В случае с секторными диаграммами обязательно проверьте, чтобы сумма всех процентов равнялась 100%.
- ✓ **Перепроверьте даже простейшие вычисления**. Например, в диаграмме сказано, что 83% американцев поддерживают определенный вопрос, а в отчете говорится, что вопрос поддерживают 7 американцев из 8. Это одно и то же? (Нет, 7, деленное на 8, дает 87,5%. А вот 5 из 6 — это около 83%.)
- ✓ **Выясните долю ответивших на опрос, не довольствуйтесь только количеством участников**. (*Доля ответивших* — это количество респондентов, деленное на общее число опрошенных и умноженное на 100%.) Если доля ответивших намного меньше 70%, результаты могут оказаться смещенными, потому что вы не знаете, что бы сказали те люди, которые отказались отвечать на вопросы.
- ✓ **Узнайте, какой статистический показатель использовался и подходил ли он для конкретной ситуации**. К примеру, количество преступлений увеличилось, но вместе с ним выросла и численность населения. Исследователи должны были вместо этого указать *уровень преступности* (количество преступления на душу населения).

Статистические показатели получаются благодаря формулам и вычислениям, которые равнодушны по своей природе — люди, подставляющие числа в эти формулы, должны знать больше, а они или не знают, или не хотят ставить в известность вас. Вы же, как потребитель информации (а также убежденный скептик), должны действовать. И лучший способ для этого — задавать вопросы.

Результаты сообщаются выборочно

Еще один плохой вариант — когда исследователь сообщает полученный им *статистически значимый результат* (т.е. результат, который не мог быть получен случайно), но не упоминает о том, что на самом деле он провел сотни других, не зарегистрированных тестов, и результаты всех их оказались *не*значительными. Если бы вы знали обо всех эти тестах, то могли бы усомниться, действительно ли этот заявленный результат настолько значим, а может, он объясняется простой случайностью, связанной с большим количеством проведенных испытаний. В статистике это принято называть *вылавливанием данных*.

Как защититься от вводящих в заблуждение результатов, которые получаются в ходе вылавливания данных? Узнайте больше подробностей исследования: сколько тестов было проведено, какое количество результатов признано незначительными, а что было сочтено значимым. Другими словами, по возможности получите всю картину, чтобы можно было оценить заявленные результаты под другим углом.

Чтобы заметить подтасованные числа и упущения, лучше всего помнить о том, что если результат слишком хорош, чтобы быть правдивым, то иногда это действительно так. Не верьте сразу тому, что услышите, особенно если на этом делают сенсацию. Подождите, смогут ли другие люди проверить и повторить такой результат.

Всемогущие случаи из жизни

Ах, эти случаи из жизни — они оказывают невероятно сильное влияние на общественное мнение и модели поведения. И вместе с тем их практически нельзя проверить с точки зрения статистики. Случай из жизни — это история, которая основана на опыте всего одного человека. Например:

- ✓ официантка, которая выиграла в лотерею;
- ✓ кот, который научился ездить на велосипеде;
- ✓ женщина, которая похудела на 45 кг за два дня, попробовав чудодейственную картофельную диету;
- ✓ звезда, которая утверждает, что пользуется известной краской для волос, выпускаемой компанией, лицом которой она является (ну да, а как иначе?).

Случаи из жизни делают сенсации, и чем они сенсационнее, тем лучше. Но сенсационные истории — это аномалии в реальной жизни. С большинством людей они не происходят.

Возможно, вы подумаете, что случаи из жизни на вас не влияют. Но что вы скажете о всех тех ситуациях, когда оказались под влиянием опыта всего одного человека? Ваш сосед обожает своего поставщика услуг Интернет, и вы тоже подключаетесь к нему. Вашему другу не повезло с автомобилем определенной марки, поэтому вы даже не хотите попробовать сесть за руль такой машины. Ваш отец знал человека, который погиб в автомобильной аварии, потому что оказался зажат ремнем безопасности, поэтому он никогда не пристегивается.

Если некоторые решения вполне можно принять, основываясь только на случаях из жизни, то другие, более важные, нужно принимать, исходя из реальных статистических показателей и данных, которые получены в ходе хорошо спланированных и тщательно проведенных исследований.

Случай из жизни — это набор данных с выборкой всего из одного элемента. У вас нет информации, с которой можно сравнить эту историю, никаких статистических данных для анализа, никаких возможных объяснений, от которых можно отталкиваться — это всего лишь одна история. Не позволяйте случаям из жизни оказывать на вас слишком большое влияние. Лучше полагайтесь на научные исследования и статистическую информацию, которая получена на основе больших случайных выборок элементов, представляющих целевую совокупность (а не просто отдельную ситуацию).

Если вас пытаются убедить, рассказывая случай из жизни, лучшее, что можно сделать — это сказать: “Покажите мне данные!”.

Приложение

Источники

В приложении содержатся источники примеров, которые были использованы в книге. Поскольку вы — восходящая звезда статистики, то, возможно, вам потребуется более подробная информация по описанным ситуациям, позволяющая принять более информированные решения.

Глава 1

1. Все упомянутые статьи взяты из газет *The Cincinnati Enquirer* и *The Columbus Dispatch*, January 26, 2003.

Глава 2

1. Опрос о микроволновках: *USA Today,* September 6, 2001.
2. Web-сайт Trident Gum: `www.tridentgum.com/consumer/html/c0000.html`.
3. Количество преступлений в США с 1990 по 1998 гг. взято из отчетов ФБР: `www.fbi.gov/ucr/ucr.htm`.
4. Web-сайт лотереи Канзаса: `www.kslottery.com`.
5. Врачебный такт спасает от судебных исков: *USA Today,* February 19, 1997.
6. Исследование Росса Перота: *TV Guide,* March, 21, 1993; подробнее см. на сайте движения United We Stand America `www.uwsa.com`.
7. *Journal of the American Medical Association:* `jama.ama-assn.org`.
8. *The New England Journal of Medicine:* `content.nejm.org`.
9. *The Lancet:* `www.thelancet.com`.
10. *British Medical Journal:* `bmj.com`.
11. The Gallup Organization: `www.gallup.com`.

Глава 3

1. The Gallup Organization: `www.gallup.com`.
2. Бюро переписей США: `www.census.gov`.
3. Цинк при простуде, положение подушки и сон, высота каблука и комфорт при ходьбе — все эти исследования цитируются по статье “Healthy Habits — that Aren’t,” *Womans Day,* February 11, 2003.
4. Трели сверчка и температура: “Cricket thermometers,” *Field & Stream* July 1993, Vol. 98, Issue 3, p. 21; подробнее см. *The Songs of the Insects* (1949). by George W. Pierce, Harvard University Press, pp. 12–21.
5. Преступность и полиция, Министерство юстиции США: `www.ojp.usdoj.gov/bjs/lawenf.htm`.
6. Мороженое и убийства (Нью-Йорк): хорошая статья для того, чтобы приступить к изучению этого и других вопросов — Spellman, B. A., & Mandel, D. R. (2003). Подробнее о психологии причинности см. Nadel, L. (Ed.) *Encyclopedia of Cognitive Science* (vol. 1, pp. 461–466).

Глава 4

1. Изучение потребительских расходов, Бюро трудовой статистики www.bls.gov/cex.
2. Лотереи: (Огайо) www.ohiolottery.com; (Флорида) www.flalottery.com/lottery/edu/edu.html; (Мичиган) www.michigan.gov/lottery; (Нью-Йорк) www.nylottery.org.
3. Доллары налогообложения: www.irs.gov/app/cgi-bin/slices.cgi.
4. Тенденции в населении, этническом составе, рабочей силе, Министерство труда США, Herman Report: "Futurework: Trends and Challenges for Work in the 21st Century": www.dol.gov/asp/programs/history/herman/repports/futurework.
5. Расходы на дорогу, Бюро транспортной статистики: www.bts.gov/publications/transportation_in_the_united_states/pdf/teconomy.pdf.
6. Статистика рождаемости, Департамент здравоохранения и окружающей среды штата Колорадо: www.cdphe.state.co.us/../cohid/birthdata.html.
7. Сайт Налогового управления: www.irs.gov.
8. Бюро трудовой статистики, занятость работников и оплата их труда: www.wa.gov/esd/lmea/occdata/oeswage/Page2067.htm.
9. Оценка изменений в численности населения штата, Бюро переписей США: eire.census.gov/popest/data/states/tables/ST-EST2002-01.php.

Глава 5

1. Данные о населении США, Бюро переписей США: http://www.census.gov/population/, www/documentation/twps0038.pdf.
2. Оплата игроков NBA: www.hoopsworld.com/article_21.shtml.
3. Знакомства в киберпространстве: Parks and Floyd (1996), *Journal of Communication* 46(1), 80–97.
4. Доход семей, Бюро переписей США, *Money Income in the United States,* 2001, pp. 26-27.

Глава 6

1. Опрос американских общин, Коламбус, штат Огайо, 2001 г.: www.census.gov/acs/www/Products/Profiles/Single/2001/ACS/Narrative/155/NP15500US3918000049.htm.
2. Бюро переписей США: www.census.gov.

Глава 7

1. Лотерея Connecticut Powerball: www.ctlottery.org/powerball.htm.

Глава 8

1. В этой главе не цитировались никакие источники.

Глава 9

1. Стандартные ошибки для опроса расходов общины: www.bls.gov/cex/2001/stnderror/age.pdf.
2. Бюро трудовой статистики: www.bls.gov.

3. Баллы и среднеквадратичное отклонение для теста ACT (2002 г.): www.act.org/news/data/02/pdf/t6-7-8.pdf.
4. Таблицы для теста ACT 2002 г.: www.act.org/news/data/02/pdf/data.pdf.

Глава 10

1. *The Gallup Organization*: www.gallup.com.

Глава 11

1. Доверительные интервалы для медианы дохода семьи: www.census.gov/hhes/income/income01/statemhi.html; полный отчет: www.census.gov/prod/2002pubs/p60-218.pdf.

Глава 12

1. Исследование потребления наркотиков подростками: "Monitoring the Future," при поддержке Национального института по изучению наркомании и Университета Мичигана: monitoringthefuture.org/data/2002data-drugs.

Глава 13

В этой главе не цитировались никакие источники.

Глава 14

1. Варикозное расширение вен: *Woman's Day,* Feb 11, 2003 p. 28.
2. Сон младенцев: *Woman's Day,* Feb 11, 2003, p. 120 "And So, to Bed" by Loraine Stern.
3. Исследование потребления наркотиков подростками: "Monitoring the Future," при поддержке Национального института по изучению наркомании и Университета Мичигана: monitoringthefuture.org/data/2002data-drugs.
4. *The Gallup Organization:* www.gallup.com.
5. Институт страхования безопасности на дорогах (проверки аварий): www.hwysafety.org/default.htm.
6. *Consumer Reports:* www.consumerreports.org/main/home.jsp.
7. Институт правильного домашнего хозяйства (одобрение): magazines.ivillage.com/goodhousekeeping.

Глава 15

1. Доктор Рут: *Family Circle,* 2/1/97 page 102. "Full Circle."
2. Adderall: *Woman's Day,* Feb 11, 2003, Advertisement following page 110. (Shire U.S. Inc.)

Глава 16

1. *The Gallup Organization*: www.gallup.com.
2. Gallup's Pollwatcher's Guide: www.gallup.com/Publications/pollwatcher.asp.
3. The Harris Poll: vr.harrispollonline.com/register.
4. Zogby International: www.zogby.com.
5. Американская медицинская ассоциация: www.ama-assn.org.
6. Национальная оружейная ассоциация: www.nra.org.
7. Бюро переписей США: www.census.gov.

8. Опрос канала CBS о деятельности знаменитостей: www.cbsnews.com/stories/2003/03/06/eveningnews/main543046.shtml.
9. Опрос канала CBS о знакомствах в Интернете: www.cbsnews.com/stories/2003/02/17/opinion/polls/main540870.shtml.
10. Опрос канала CBS о боли: www.cbsnews.com/stories/2003/01/28/opinion/polls/main538259.shtml.
11. Вопросы по медицине в Сети: www.cnn.com/2000/HEALTH/08/18/watercooler.ap/.
12. Мнение инвесторов: www.gallup.com/poll/releases/pr030224.asp.
13. Худший автомобиль тысячелетия: cartalk.cars.com/About/Worst-Cars/results5.html.
14. Езда в нетрезвом виде: www.reuters.com/newsArticle.jhtml?type=healthNews&storyID=2345367.
15. Детское здравоохранение: my.webmd.com/content/article/60/67060.htm?lastselectedguid={5FE84E90-BC77-4056-A91C-9531713CA348).
16. Опрос жертв насилия: www.ojp.usdoj.gov/bjs/pub/pdf/cvus01.pdf.
17. Рак груди: my.webmd.com/content/article/36/1728_50583.htm?lastselectedguid={5FE84E90-BC77-4056-A91C-9531713CA348}.
18. Использование сотовых телефонов: www.consumerreports.org/main/detailv2.jsp?C0NTENT%3C%3Ecnt_id=23371&F0LDER%3C%3Efolder_id=23051&bmUID=1047262402709.
19. Преступления в киберпространстве: www.gocsi.com/press/20020407.html.
20. Сексуальные домогательства: womensissues.about.com/library/blsexharassmentstats.htm.
21. Ложь: *Journal of Applied Social Psychology,* 1997, v. 27, 12 pp. 1048–1062, by Ester Backbier, Johan Hoogstraten, and Katharina Meerum Terwogt-Kouwenhoven.

Глава 17

1. Национальный институт по изучению раковых заболеваний: cancer.gov.
2. Марихуана и химиотерапия: *The New York Times,* September 16,197S
3. Информация о клинических испытаниях. Национальный институт здравоохранения США: www.ClinicalTrials.gov.
4. ВИЧ плацебо: *The Manhattan Mercury* (Manhattan, KS), September 21, 1997.
5. Исследование связи между поздними родами и продолжительностью жизни: *Kansas City Star,* September 11, 1997.
6. Исследование потребления наркотиков подростками: “Monitoring the Future,” при поддержке Национального института по изучению наркомании и Университета Мичигана: monitoringthefuture.org/data/2002data-drugs.

Глава 18

1. Национальный институт здравоохранения: www.nih.gov.

2. Связь между просмотром телепередач и другими видами пассивной деятельности с риском ожирения и диабета типа 2 у женщин: *Journal of the American Medical Association* (2003) Volume 289, pp. 1785–1791, (NIH News Release).
3. Обратная взаимосвязь между выражением гнева и сердечными приступами/ударами: *Psychosomatic Medicine* (2003), 65(1), pp. 100–110, (NIH News Release).
4. Умеренное потребление алкоголя и снижение риска сердечных приступов у мужчин: `www.nih.gov/news/pr/jan2003/niaaa-08.htm` (пресс-релиз Национального института здравоохранения).
5. Лечение на ранней стадии останавливает развитие глаукомы: `www.nih.gov/news/pr/oct2002/nei-14.htm`, (пресс-релиз Национального института здравоохранения).
6. Цинк при лечении простуды — это неразумно: "Healthy Habits — that Aren't," *Woman's Day,* February 11, 2003.
7. Сверчки и температура: журнальная статья, *Garden Gate,* Issue Number 5: `www.gardengatemagazine.com/tips/25tip13.html`.
8. "Cricket Chirp Converter": Национальное бюро метеопрогнозов, `www.srh.noaa.gov/elp/wxcalc/cricketconvert.html`.
9. Данные о трелях сверчков: для этого случая существует множество наборов данных, вот источник, которым пользуется большинство статистиков: Pierce, George W, *The Songs of the Insects,* (1949), Harvard University Press, pp. 12–21. Примечание: в своем примере я использовала только часть этих данных (исключительно для наглядности).
10. Подробнее о связи между трелями сверчков и температурой см. `www.dartmouth.edu/~genchem/0102/spring/6winn/cricket.html`.
11. Аспирин останавливает образование полипов у больных раком толстой кишки: Национальный институт по изучению раковых заболеваний, группа больных раком и лейкемией под наблюдением Электры Паскетт, Государственный университет Огайо: `www.osu.edu/researchnews/archive/aspirin.htm`.
12. Мороженое и убийства (Нью-Йорк): хорошая статья, с которой можно начать изучение этого вопроса и других, связанных с ним — это Spellman, B.A., & Mandel, D.R. (2003). Подробнее о психологии причинности см. Nadel, L. (Ed.) *Encyclopedia of Cognitive Science* (vol. 1, pp. 461–466).

Глава 19

1. Управление по контролю над лекарственными препаратами и пищевыми продуктами США: `www.fda.gov`.
2. Тотальное управление качеством: `www.6sigma.us`.
3. W. Edwards Deming: *Out of the Crisis* (Center for Advanced Engineering Study. Massachusetts Institute of Technology, 1986) `www.deming.org`.
4. Аппараты для наполнения тюбиков зубной пастой: `www.packagingdigest.com/articles/199710/52.html`.
5. Популярные вопросы о наполнении тюбиков: `www.keyinternational.com/FAQ_Tube_Filling.html#Q4`.
6. Совет по наполнению тюбиков: `www.tube.org`.

Предметный указатель

Научно-популярное издание

Дебора Рамси

Статистика для "чайников"

В издании использованы карикатуры американского художника Рича Теннанта

Литературный редактор *П.Н. Мачуга*
Верстка *Т.Н. Артеменко*
Художественный редактор *Е.П. Дынник*
Корректоры *В.В. Смоляр, Л.В. Чернокозинская*

Издательский дом "Вильямс"
127055, г. Москва, ул. Лесная, д. 43, стр. 1

Подписано в печать 25.01.2008. Формат 70×100/16.
Гарнитура Times. Печать офсетная.
Усл. печ. л. 25,80. Уч.-изд. л. 21,00.
Тираж 2000 экз. Заказ № 0000.

Отпечатано по технологии CtP
в ОАО "Печатный двор" им. А. М. Горького
197110, Санкт-Петербург, Чкаловский пр., 15.

Шпаргалка

Основные статистические показатели

Ниже приводятся самые распространенные статистические показатели, с которыми вы можете столкнуться, а также формулы для их вычисления и краткое описание.

Статистический показатель	Формула	Использование
Среднее выборки	$\bar{x} = \frac{\sum x}{n}$	Взвешенный центр всех данных (с учетом выбросов)
Медиана	Среднее значение из упорядоченного набора данных	Числовой центр данных; выбросы не учитываются
Стандартное отклонение выборки	$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$	Мера изменчивости; типичное расстояние до среднего
Коэффициент корреляции	$r = \frac{1}{n-1}\sum \frac{(x-\bar{x})(y-\bar{y})}{s_x s_y}$	Модуль и направление линейной зависимости между x и y

Доверительные интервалы

Доверительный интервал — это обоснованное предположение о какой-либо характеристике совокупности (например, какой процент людей в стране имеют сотовый телефон). Доверительный интервал содержит исходное предположение (скажем, средняя начальная зарплата на основе выборки из 1000 выпускников вузов) плюс/минус предел погрешности (степень, на которую, как вы считаете, могут варьироваться полученные результаты, если будут сделаны другие выборки). Ниже показаны формулы самых распространенных доверительных интервалов. Подробнее см. главу 13.

Объект	Статистический показатель	Предел погрешности	Использование
Среднее совокупности (μ)	$\bar{x}$	$\pm Z \times \frac{s}{\sqrt{n}}$	n не меньше 30
Среднее совокупности (μ)	$\bar{x}$	$\pm t_{n-1} \times \frac{s}{\sqrt{n}}$	n меньше 30
Доля совокупности (p)	$\bar{p}$	$\pm Z \times \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$	$n \times \bar{p}$ и $n(1-\bar{p})$ не меньше 5
Разность средних двух совокупностей ($\mu_x - \mu_y$)	$(\bar{x}-\bar{y})$	$\pm Z \times \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	n_1 и n_2 не меньше 30
Разность долей двух совокупностей ($p_1 - p_2$)	$(\bar{p}_1-\bar{p}_2)$	$\pm Z \times \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}+\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$	$1 \times \bar{p}$ и $n(1-\bar{p})$ для каждой группы не меньше 5

Доверительные коэффициенты (Z-значения, коэффициенты Стьюдента)

Доверительные коэффициенты (Z-значения) — это важный компонент доверительных интервалов. Z-значение — одна из составляющих предела погрешности, т.е. того числа, которое вы добавляете или отнимаете, чтобы получить определенный доверительный уровень своих результатов. Чем больше Z-значение, тем более достоверным является результат. Подробнее см. главу 9.

Доверительный уровень	Z-значение	Доверительный уровень	Z-значение
80 %	1,28	95 %	1,96
85 %	1,44	98 %	2,33
90 %	1,64	99 %	2,58

Проверка гипотез

При проверке гипотезы, с помощью определенных данных можно выяснить, справедливо ли какое-либо утверждение относительно совокупности. (Например, кто-то может заявить, что у 40 % американцев есть сотовый телефон. Правда ли это?). Чтобы проверить гипотезу, вы делаете выборку, собираете данные, находите статистический показатель, стандартизируете его и получаете тестовую статистику (которую можно нанести на стандартную шкалу), после чего определяете, доказывает ли эта тестовая статистика сделанное утверждение. (Подробнее см. главы 14 и 15.) Ниже приводятся формулы самых распространенных критериев проверки гипотезы.

Проверяется	***Нулевая гипотеза (H_0)***	***Тестовая статистика***	***Распределение***	***Использование***
Среднее совокупности (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_0)}{s/\sqrt{n}}$	Стандартное нормальное (Z)	n не меньше 30
Среднее совокупности (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_0)}{s/\sqrt{n}}$	t_{n-1}	n меньше 30
Доля совокупности (p)	$p = p_0$	$\frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Стандартное нормальное (Z)	$n \times p_0$ и $n(1 - p_0)$ не меньше 5
Разность средних двух совокупностей ($\mu_x - \mu_y$)	$\mu_x - \mu_y = 0$	$\frac{(\bar{x} - \bar{y}) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Стандартное нормальное (Z	n_1 и n_2 не меньше 30
Среднее разности (“до” – “после”)	$\mu_d = 0$	$\frac{\bar{d} - 0}{s/\sqrt{n}}$	Стандартное нормальное (Z)	30 и больше пар данных
Среднее разности (“до” – “после”)	$\mu_d = 0$	$\frac{\bar{d} - 0}{s/\sqrt{n}}$	t_{n-1}	Меньше 30 пар данных
Разность долей двух совокупностей ($p_1 - p_2$)	$p_1 - p_2 = 0$	$\frac{(\bar{p}_1 - \bar{p}_2) - 0}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$	Стандартное нормальное (Z)	$n \times \bar{p}$ и $n(1 - \bar{p})$ для каждой группы не меньше 5

ПРАКТИЧЕСКАЯ БИЗНЕС-СТАТИСТИКА

Эндрю Ф. Сигел

www.williamspublishing.com

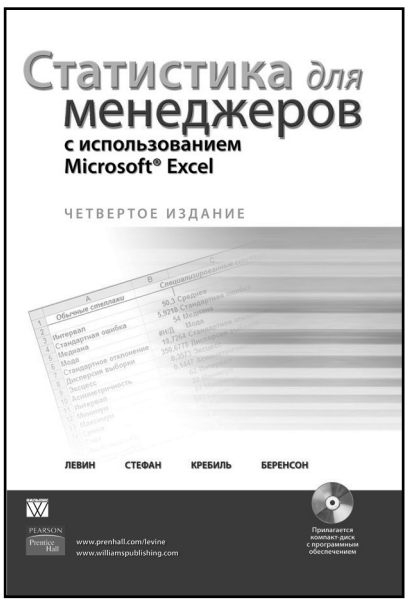
Эта книга представляет собой прекрасно организованный курс статистических методов анализа данных. Теоретический материал не перегружен математическими подробностями и дополняется большим количеством тщательно отобранных примеров с использованием Microsoft Excel. Здесь есть анализ финансового состояния предприятий и конъюнктуры фондового рынка, прогнозирование уровня продаж и результатов избирательных кампаний, анализ качества продукции и эффективности рекламы, изучение аудитории средств массовой информации и много других непростых и практически важных задач. Книга может быть полезна преподавателям, студентам, научным сотрудникам, аналитикам консалтинговых фирм и рекламных агентств, а также практикующим менеджерам, желающим использовать прикладной статический анализ в своей работе.

ISBN 978-5-8459-1367-8

в продаже

СТАТИСТИКА ДЛЯ МЕНЕДЖЕРОВ С ИСПОЛЬЗОВАНИЕМ EXCEL 4-е издание

Д. Левин, Д. Стефан, Т. Кребиль, М. Беренсон



www.williamspublishing.com

Книга представляет собой вводный курс бизнес-статистики. В ней рассмотрены практически все традиционные темы, касающиеся анализа данных — от описательных статистик до регрессионного анализа и карт контроля. Особенную ценность книге придает огромное количество примеров, почерпнутых из практики, а также компакт-диск с большим количеством приложений, иллюстрирующих методы статистического анализа данных с помощью программы Microsoft Excel. Книга предназначена для студентов, изучающих основы менеджмента, преподавателей бизнес-школ, а также менеджеров, желающих повысить качество своей работы.

ISBN 5-8459-0607-5

в продаже

ПРИКЛАДНОЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Н. Дрейпер, Г. Смит

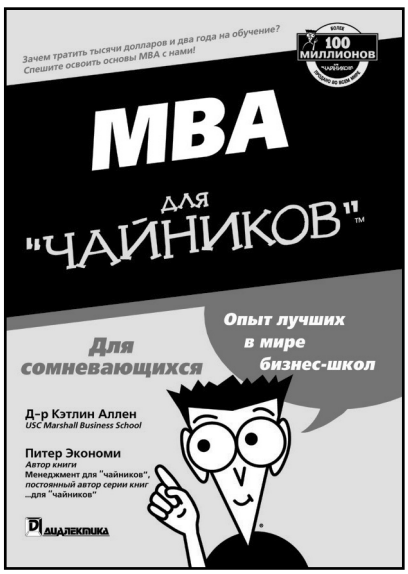
www.dialektika.com

Полное классическое введение в фундаментальные основы регрессионного анализа. В книге описываются методы подбора и исследования линейных и нелинейных регрессионных моделей различной степени сложности, а также рассматриваются практические аспекты их применения, в том числе с использованием специальных компьютерных программ. Помимо стандартного набора тем, составляющих ядро курса регрессионного анализа, в это издание включены отдельные главы, посвященные мультиколлинеарности, обобщенным линейным моделям, геометрическим свойствам регрессии, робастной регрессии и процедурам тиражирования выборки (бутстрепа). Содержит множество примеров и упражнений (с полными или частичными решениями), а также вопросы для самоконтроля. Предназначена для аналитиков, экспериментаторов и студентов высших учебных заведений.

ISBN 978-5-8459-0963-3 **в продаже**

MBA
ДЛЯ "ЧАЙНИКОВ"

Д-р Кэтлин Аллен,
Питер Экономи



www.dialektika.com

Книга *MBA для "чайников* содержит много полезной информации и подсказок для тех, кто стремится эффективно управлять или руководить компанией, развивать собственный бизнес. Она поможет читателю овладеть предметами, изучаемыми студентами университетских программ MBA, претендующими на степень Мастера бизнес-администрирования. Шесть разделов книги посвящены основным вопросам, свободно ориентироваться в которых обязан современный профессиональный руководитель или менеджер: стратегическому планированию, менеджменту, маркетингу, финансам, умению вести переговоры, организовывать взаимоотношения с клиентами, осуществлять бухгалтерский учет и многому другому. Написанная в доступной и интересной форме, эта книга станет прекрасным учебным пособием для тех, кто стремится овладеть основами современного ведения бизнеса и управления основными бизнес-процессами, но не располагает временем или средствами для изучения курсов MBA в университетах.

ISBN 5-8459-0803-5

в продаже