ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Первое издание «Статистики для всех» пользовалось оглушительным успехом, однако любую книгу можно улучшить, и я благодарна за предоставленную возможность переработать ее. Мой принцип изложения не изменился: эта книга гораздо больше предназначена тем, кто хочет размышлять и понимать результаты статистической обработки данных, чем тем, кто хочет узнать, как пользоваться конкретным статистическим пакетом программ или углубиться в математические основы при помощи статистических формул. Эта книга также несколько отличается от других изданий в этой серии «Руководств для всех» издательства О'Рейлли – она действительно находится где-то между руководством для тех, кто уже знаком со статистикой, и учебником для людей, которые только начали осваивать этот предмет.

Несмотря на продолжающееся проникновение статистики во многие области нашей жизни, одна вещь осталась неизменной: сказать, что ты работаешь статистиком, – по-прежнему верный способ испортить приятную беседу на вечеринке. Почему-то оказывается, что это побуждает людей рассказать мне, как они ненавидели обязательные занятия по статистике в колледже, или заставляет их процитировать старую шутку, ставшую популярной благодаря Марку Твену, о том, что существует три вида лжецов: простые лжецы, отъявленные лжецы и статистики. Лично я нахожу статистику захватывающей и обожаю работать в этой области. Я также люблю преподавать статистику, и мне нравится думать, что я заражаю своим энтузиазмом окружающих. Хотя часто это превращается в напряженную битву; многие считают, что статистика – это не более чем набор хитростей и подтасовок для искажения реальности, которые нужны, чтобы одурачить других людей. Другие занимают противоположную позицию, полагая, что статистика – это набор волшебных приемов, которые избавят вас от необходимости размышлять над данными.

Ну хорошо, и что же такое статистика?

Прежде чем погрузиться в технические детали изучения и использования статистики, вернемся на минуту назад и обсудим, что можно подразумевать под словом «статистика». Не беспокойтесь, если вы сразу не поймете всю терминологию, она прояснится в ходе чтения этой книги.

Когда люди говорят о статистике, они обычно имеют в виду один или несколько пунктов из приведенного ниже перечня:

1. Числовые данные, такие как уровень безработицы, число людей, умирающих ежегодно от пчелиных укусов, или численность жителей г. Нью-Йорк в 2006 году по сравнению с 1906 годом.
2. Числа, использованные для описания выборок, в противоположность параметрам (числам, характеризующим генеральную совокупность). Например, рекламная компания может интересоваться средним возрастом подписчиков журнала «Спортс Иллюстрейтед» (Sports Illustrated)[1]. Для ответа на этот вопрос компания может создать случайную выборку подписчиков, вычислить среднее значение для этой выборки (статистику) и использовать его как оценку среднего значения для всей генеральной совокупности подписчиков (параметра).
3. Определенные методы анализа данных и результаты такого анализа, такие как *t*-статистика или статистика хи-квадрат.
4. Область науки, которая разрабатывает и использует математические методы для описания данных и формирования суждений о них.

Тот тип статистики, о котором говорится в первом определении, не имеет прямого отношения к этой книге. Если вы просто хотите найти последние данные о безработице, здоровье или о любой из множества других тем, по которым правительство или другие организации регулярно публикуют статистические данные, вам лучше всего проконсультироваться у библиотекаря или у специалиста в данной области. Если же вы хотите узнать, как интерпретировать эти данные (понять, например, почему среднее арифметическое часто бывает плохим показателем средней тенденции, или сравнить исходные и стандартизованные показатели смертности), то «Статистика для всех» точно вам поможет.

Понятия, использованные во втором определении, будут обсуждаться в главе 3, посвященной предсказательным статистикам. Однако эти термины пронизывают всю книгу. Это отчасти терминологические тонкости (статистики – это числа, которые описывают выборки, а параметры характеризуют генеральные совокупности), которые тем не менее подчеркивают ключевой момент применения статистики. Идея использования информации, полученной при изучении выборки, для формирования суждений обо всей генеральной совокупности лежит в основе всей предсказательной статистики, а предсказательная статистика – это основная тема этой книги (как и большинства других книг, посвященных статистике).

Третье определение также является ключевым для большинства глав этой книги. Процесс изучения статистики до некоторой степени сводится к освоению определенных статистических методов, включая такие вопросы, как способы вычислений и их интерпретации, выбор подходящей статистики в конкретной ситуации и так далее. На самом деле многие люди, начинающие изучать статистику, держат в голове в основном это определение. Освоить статистику для них означает узнать,

[1] Еженедельный иллюстрированный спортивный журнал, крупнейшее и самое популярное спортивное издание в США. – *Прим. пер.*

как выполнять набор статистических процедур. Это не столько неверный подход к статистике, сколько неполный. Умение применять ряд методов статистической обработки данных – это необходимая составляющая деятельности статистика, но это далеко не все, что нужно. Более того, с тех пор как компьютерные программы сделали применение методов статистического анализа данных существенно проще для всех вне зависимости от уровня математической подготовки, необходимость в понимании и интерпретации результатов статистического анализа значительно превысила необходимость знать, как проводить сами вычисления.

Четвертое определение мне ближе всего, поскольку я избрала статистику своей профессией. Если вы уже студент или закончили вуз, вам, вероятно, знакомо это определение, поскольку в наши дни во многих университетах и колледжах или есть отдельный факультет статистики, или же статистика предлагается как одно из направлений специализации на математическом факультете. Статистика все чаще преподается и в средней школе, а в США число учащихся, выбравших классы с углубленным изучением статистики, быстро растет.

Статистика в университетах – это не только курс для тех, кто решил специализироваться в этой области. На многих факультетах от студентов требуется прослушать один или несколько курсов по статистике, помимо тех предметов, на которых они специализируются. Кроме того, полезно знать, что многие важные методы современной статистики были разработаны людьми, которые изучили и использовали статистику во время своей работы в другой области знаний. Стефан Рауденбуш (Stephen Raudenbush), создатель иерархического линейного моделирования, изучал основы политического анализа и оценочных исследований в Гарварде, а Эдвард Тьюфт (Edward Tufte), наверное, лучший специалист в мире по статистической графике, начинал свою карьеру как политолог: он защитил докторскую диссертацию в Йельском университете по американским движениям в защиту гражданских прав.

Поскольку статистика все чаще применяется во многих специальностях и на всех уровнях от управляющих до рядовых рабочих, базовые знания в этой области необходимо получить многим людям, давно закончившим школу. Они часто недостаточно обеспечены учебниками, предназначенными для вводных университетских курсов, а эти пособия слишком специализированы, слишком много внимания уделяют вычислениям и слишком дороги.

Наконец, статистику нельзя отдать на откуп статистикам, поскольку каждому из нас следует принимать участие в современной общественной жизни, в частности понимать многое из того, что вы прочли в газетах и услышали по радио или телевизору. Рабочие знания по статистике – лучшее противоядие от вводящих в заблуждение или совершенно ложных числовых данных (исходящих или от политиков, или рекламных агентов, или от реформаторов социальной сферы), которые, похоже, составляют постоянно возрастающую часть ежедневно поглощаемой нами информации. Вот почему классическая книга Дэррила Хаффа (Darryl Huff), опубликованная в 1954 г., «Как лгать при помощи статистики» (“How to Lie with Statistics”) до сих пор пользуется спросом. Статистику легко использовать неправильно, стандартные способы искажения статистических данных не меняются на

протяжении десятилетий, а лучшая защита против тех, кто хотел бы солгать при помощи статистики, – стать более образованным, чтобы быть способным выявить лжецов и немедленно остановить их.

Основная цель этой книги

В продаже существует уже столько книг по статистике, что вы могли бы сильно удивиться, почему я чувствую необходимость добавить еще одну книгу к этому множеству. Основная причина заключается в том, что я не нашла ни одной книги по статистике, которая отвечала бы задачам, поставленным мною в «Статистике для всех». На самом деле, если позволите на мгновение впасть в поэтическое настроение, ситуация состоит в том, что, перефразируя состояние старого морехода Кольриджа, «книги, повсюду книги, но ни одной, по которой можно научиться»[2]. Проблемы, которые я постаралась решить в этой книге, таковы:

- нужда в книге, которая была бы посвящена использованию и пониманию статистики в контексте исследований или прикладной науки, не как отдельного набора математических методов, а как части процесса обоснования заключений при помощи цифр;
- необходимость включения таких тем, как теория измерений и управление данными во введение в статистику;
- необходимость в книге по статистике, которая не была бы посвящена одной конкретной области знаний. Простейшая статистика в основном одинакова для всех дисциплин (тест Стьюдента работает одинаково для данных из области медицины, финансов или криминальной юстиции), так что незачем умножать тексты, представляя одну и ту же информацию немного в другом ракурсе;
- нужда во введении в статистику, которое было бы компактным, недорогим и простым для понимания начинающих, избегая снисходительного тона или излишнего упрощения.

Так кто же предполагаемые читатели «Статистики для всех?» Я вижу три группы читателей, для которых эта книга будет наиболее полезной:

- учащиеся, которые посещают вводные курсы по статистике в средней школе, колледжах и университетах;
- взрослые люди, которым нужно освоить статистику для выполнения текущих задач или для карьерного роста;
- те, кому интересно узнать, что такое статистика, из любопытства.

В этой книге я делаю акцент не на конкретные методы, хотя многим из них вы научитесь в процессе чтения, а на обосновании заключений при помощи статистики. Можно сказать, что цель этой книги в меньшей степени заключается в том, чтобы производить статистические вычисления, и в большей степени, – чтобы мыслить статистически. Что это значит? Мышление с использованием чисел тре-

[2] Имеются в виду строки поэмы английского поэта Сэмюэла Кольриджа «Сказание о старом мореходе»: «Вода, вода, одна вода/Мы ничего не пьем» (вольный перевод Н. С. Гумилева). – *Прим. пер.*

бует определенных навыков. В частности, я делаю упор на осмысление данных и использование статистики для облегчения этого процесса. Во многих главах приведены практические задания, которые задуманы как повод пересмотреть представленный материал и подумать о ключевых понятиях, введенных в данной главе, они не требуют бездумных вычислений.

Весь материал «Статистики для всех» был переработан, и многие главы дополнены новыми примерами и упражнениями. В частности, добавлены примеры работы с пропорциями, а также примеры с использованием реальных наборов данных из таких источников, как Проект ООН по развитию человечества (United Nations Human Development Project) и Система слежения за факторами поведенческого риска (Behavioral Risk Factor Surveillance System). Оба этих набора данных можно бесплатно скачать из Интернета, так что студенты могут экспериментировать с ними, а также воспроизвести процедуры, описанные в этой книге. В это издание также добавлена глава 19. Я сделала это, потому что заметила, что умение доводить до сведения окружающих статистическую информацию по меньшей мере так же важно, как и способность выполнять статистические вычисления, в особенности для тех, кто учится статистике для своей профессиональной деятельности. Также добавлено несколько новых приложений, в основном для того, чтобы сделать книгу более самодостаточной и дружественной к читателю. Эти приложения включают вероятностные таблицы для самых распространенных типов распределений, перечень информационных ресурсов Интернета, словарь и таблицу статистических обозначений.

Статистика в информационную эпоху

Стало модным говорить, что мы живем в информационную эпоху, когда люди получают и распространяют столько сведений, что никто не может быть в курсе всего. Это клише основано на правдивом наблюдении; общество «тонет» в данных, и, похожа, эта проблема становится только острее. В этом есть свои плюсы и свои минусы. К положительным моментам можно отнести то, что широкий доступ к компьютерным технологиям и электронным средствам хранения и распространения данных облегчил доступ к информации, так что теперь у исследователей снизилась потребность в поездках в определенную библиотеку или архив для работы с печатными источниками.

Тем не менее данные сами по себе ничего не значат. Они должны быть упорядочены и интерпретированы людьми, чтобы обрести смысл, так что полноценная жизнь в информационную эпоху подразумевает глубокое понимание данных, включая способы их сбора, анализа и интерпретации. И поскольку одни и те же данные могут быть часто интерпретированы разными способами для обоснования совершенно противоположных заключений, даже людям, которые сами не работают в области статистики, нужно понимать, как статистика работает и как выявить безосновательные заявления и аргументы, основанные на неправильном использовании данных.

Структура книги

«Статистика для всех» состоит из трех частей: вводная информация (главы 1–4), где закладывается необходимое основание для понимания последующих глав; методы предсказательной статистики (главы 5–13); специальные методы, которые используются в различных областях науки (главы 14–16), и вспомогательные темы, которые часто являются частью работы статистика, даже если они не относятся к статистике как таковой (главы 17–20). Вот более детальное содержание глав.

Глава 1. Основные понятия, связанные с измерениями

Обсуждаются основополагающие вопросы статистики, включая шкалы измерений, операционализацию, опосредованное измерение, случайные и систематические ошибки, надежность и валидность, а также типы смещения измерений.

Глава 2. Теория вероятности

Описаны основные понятия теории вероятности, включая испытания, события, независимость, взаимное исключение, правила аддитивности и перемножения, комбинации и перестановки, условную вероятность и теорему Байеса.

Глава 3. Статистический вывод

Введены некоторые базовые понятия статистического вывода, включая распределение вероятностей, зависимые и независимые переменные, генеральные совокупности и выборки, распространенные способы создания выборок, центральную предельную теорему, проверку гипотез, ошибки первого и второго типа, доверительные интервалы и значения p, а также преобразование данных.

Глава 4. Описательные статистики и графическое представление данных

Дана информация о распространенных показателях центральной тенденции и разброса, включая среднее арифметическое, медиану, моду, абсолютный размах, межквартильный размах, дисперсию и стандартное отклонение, а также обсуждаются выбросы. В этой главе рассмотрены наиболее часто используемые графические способы представления статистической информации, включая частотные таблицы, столбчатые и круговые диаграммы, диаграммы Парето, диаграммы типа «стебель с листьями», диаграммы размаха и рассеяния, а также линейные графики.

Глава 5. Категориальные данные

Представлен обзор концепций категориальных и интервальных данных, введено понятие таблицы сопряженности. В этой главе обсуждаются такие статистические методы, как тест хи-квадрат на независимость, тест равенства пропорций, критерий согласия, точный тест Фишера, тест МакНемара, тесты пропорций для больших выборок, а также меры сопряженности для категориальных и порядковых данных.

Глава 6. t-критерий

Обсуждается распределение Стьюдента, теория и применение теста Стьюдента для одной выборки, для двух независимых выборок, для результатов повторных измерений и в случае неравенства дисперсий.

Глава 7. Коэффициент корреляции Пирсона

При помощи диаграмм, демонстрирующих разную силу связи между двумя переменными, вводится понятие связи, также обсуждается коэффициент корреляции Пирсона и коэффициент детерминации.

Глава 8. Введение в регрессию и дисперсионный анализ

Показано отношение линейной регрессии и дисперсионного анализа к концепции обобщенной линейной модели, и обсуждаются допущения, которые принимаются при использовании этих видов анализа данных. Обсуждается и на примерах разбирается применение простой регрессии (для двух переменных), однофакторного дисперсионного анализа и апостериорного тестирования гипотез.

Глава 9. Многофакторный дисперсионный анализ и ковариационный анализ

Обсуждаются более сложные схемы дисперсионного анализа, включая двух- и трехфакторный дисперсионный анализ и ковариационный анализ, а также поднимается тема взаимодействия переменных.

Глава 10. Множественная линейная регрессия

Регрессионная модель расширяется за счет включения множественных независимых переменных. Рассмотрены связи между независимыми переменными, стандартизованные и нестандартизованные коэффициенты, фиктивные переменные, способы построения моделей, а также отклонения от допущений, принимаемых при линейной регрессии, включая нелинейность, автокорреляцию и гетероскедатичность.

Глава 11. Логистическая, мультиномиальная и полиномиальная регрессия

Расширяет применение регрессионного анализа до бинарных данных (логистическая регрессия), категориальных данных (мультиномиальная регрессия) и нелинейных моделей (полиномиальная регрессия), также обсуждается проблема избыточной подгонки модели.

Глава 12. Факторный, кластерный и дискриминантный анализ

Описаны три сложные статистические процедуры: факторный, кластерный и дискриминантный анализ, обсуждаются группы задач, для решения которых эти методы могут быть полезны.

Глава 13. Непараметрическая статистика

Обсуждается, когда нужно использовать непараметрическую статистику вместо параметрической, а также описаны методы для внутри- и межгрупповых сравнений, включая тесты Вилкоксона, Манна–Уитни, Краскел–Уоллиса, Фридмана, критерий знаков и медианный критерий.

Глава 14. Статистика для бизнеса и контроля качества

Приведены статистические методы, которые часто используются в бизнесе

и при контроле качества. Описанные аналитические и статистические процедуры включают в себя индексы, временны́е серии, критерии принятия решений минимакс, максимакс и максимин, принятие решений в условиях риска, деревья решений и контрольные карты.

Глава 15. Статистика в медицине и эпидемиологии

Вводятся понятия и демонстрируются статистические методы, которые особенно актуальны для медицины и эпидемиологии. В главу вошли такие темы, как определение и использование отношений, пропорций и долей, показатели заболеваемости и распространения, исходные и стандартизованные данные, прямая и непрямая стандартизация, меры риска, искажающие факторы, коэффициент несогласия (простой и Мантеля–Гензеля), а также вычисления точности, мощности и объема выборок.

Глава 16. Статистика в образовании и психологии

Обсуждаются концепции и статистические методы, наиболее часто используемые в образовании и психологии, такие как перцентили, стандартизованные баллы, методы создания тестов, классическая теория тестов, надежность комбинированного теста, меры внутренней согласованности, включая коэффициент альфа, а также методы анализа заданий. Также приводится обзор современной теории тестирования.

Глава 17. Управление данными

Обсуждаются практические вопросы управления данными, включая кодификацию, группировку данных, методы устранения ошибок в файлах, методы хранения данных в цифровом виде, текстовые и числовые данные и пропущенные значения.

Глава 18. Планирование исследования

Обсуждаются наблюдения и эксперименты, слагаемые хорошего планирования исследований, этапы сбора данных, типы валидности и способы ограничить или предотвратить искажение результатов.

Глава 19. Представление статистических материалов

Рассмотрены основные проблемы представления статистической информации различной аудитории, затем более детально обсуждается изложение результатов для специализированных журналов, для общественности и для коллег по работе.

Глава 20. Оценка работ по статистике других авторов

Содержит руководство по проверке правильности использования статистики, включая список контрольных вопросов, которые помогут оценить представление статистических данных, и примеры манипуляций с корректными статистическими методами для подтверждения спорных заключений.

В шести приложениях приведены сведения, которые лежат в основе материала, изложенного в основной части книги, а также указаны источники дополнительной информации:

Приложение A. Обзор основных математических понятий

Содержит материалы для самопроверки и обзор основ арифметики и алгебры для тех, у кого остались лишь ускользающие воспоминания о последнем курсе по математике. Обсуждаются арифметические правила, экспоненты, корни и логарифмы, методы решения уравнений и систем уравнений, дроби, факториалы, перестановки и комбинации.

Приложение B. Краткий обзор статистических пакетов

Представлен обзор некоторых наиболее распространенных компьютерных программ, используемых для статистических вычислений, приведены примеры простейшего анализа данных в каждой из программ, обсуждаются сильные и слабые стороны каждой из них. Рассмотрены такие программы, как Minitab, SPSS, SAS и R; также обсуждается использование Microsoft Excel (это не статистический пакет) для статистического анализа.

Приложение C. Ссылки

Аннотированный список литературы к каждой главе включает бумажные публикации и сайты в Интернете, которые упоминаются в тексте, и прочие источники, с которых хорошо начать углубленное изучение соответствующей темы.

Приложение D. Таблицы вероятностей для распространенных типов распределений

Приведены таблицы для большинства широко используемых статистических распределений – нормальное, Стьюдента, биномиальное и хи-квадрат. Даже в эпоху компьютера и Интернета сто́ит знать, как читать таблицы распределений, и удобно иметь их под рукой в печатном виде.

Приложение E. Интернет-ресурсы

Приведен перечень лучших сайтов в Интернете, которые пригодятся тем, кто учит, использует или преподает статистику. Источники разделены на общие руководства, словари, вероятностные таблицы, калькуляторы и учебники.

Приложение F. Словарь статистических терминов

Сюда вошли греческий алфавит (проклятие многих начинающих статистиков), расшифровка статистических обозначений и краткий словарь для большинства статистических терминов, использованных в этой книге.

Эта книга – руководство, которое можно приспосабливать к имеющимся знаниям и нуждам отдельных читателей. Некоторые главы посвящены темам, которые часто отсутствуют в вводных книгах по статистике, однако я считаю их важными. Это касается управления данными, изложения статистических результатов и чтения статистических статей, написанных другими людьми. Эти главы также послужат полезным справочным материалом для людей, которые внезапно обнаружат, что их назначили разбираться с данными по проекту, или которым было поручено, более или менее неожиданно, представить статистические данные о работе их команды. Ни один из этих сценариев, к сожалению, не слишком редок.

Классификация сведений на элементарные и сложные зависит от личных знаний и задач. Я написала «Статистику для всех» так, чтобы она отвечала задачам многих категорий читателей. Из-за этого невозможно расположить материал в идеальной последовательности, так, чтобы это удовлетворяло запросам каждого. Это соображение приводит нас к важному заключению: нет никакой необходимости читать главы в том порядке, в каком они представлены здесь. В статистике есть много дилемм типа «что было раньше, яйцо или курица?». К примеру, вы не можете спланировать эксперименты, не зная, какие типы статистической обработки данных вам доступны, при этом вы не сможете понять, как применяется статистика, без каких-либо знаний о планировании исследований. Сходным образом может казаться логичным, что тот, кто занялся управлением данными, уже имеет опыт статистического анализа, однако я консультировала многих лаборантов и руководителей проектов, которым было поручено разобраться с объемными наборами данных до того, как они прослушали хотя бы один курс по статистике. Так что читайте эти главы в том порядке, который облегчает выполнение стоя́щих перед вами задач, и не стесняйтесь пропустить что-то и сосредоточиться на том, что отвечает вашим конкретным потребностям.

Не весь материал этой книге и актуален для каждого, это наиболее очевидно для глав 14–16, которые посвящены определенным областям науки (бизнес и контроль качества, медицина и эпидемиология, образование и психология соответственно). Однако полезно быть открытым всему новому, если дело касается знания статистических методов. В данный момент вы можете быть уверенным, что вам никогда не понадобится проводить непараметрический тест или логистический регрессионный анализ, но вы никогда не знаете, что пригодится в будущем. Также неправильно слишком четко делить методы по областям знаний; поскольку статистические методы в конечном счете имеют дело с числами, а не с содержанием; методы, разработанные в одной области знаний, часто пригождаются в другой. Например, контрольные карты (обсуждаемые в главе 14) были разработаны для производственных нужд, а теперь широко используются во многих областях от медицины до образования, тогда как коэффициент несогласия (глава 15), разработанный в эпидемиологии, теперь применяется ко всем типам данных.

Условные обозначения, используемые в этой книге

В этой книге принята следующая система обозначений:

Обычный текст
: Обозначает названия пунктов меню, опций, кнопок на экране и клавишей клавиатуры (таких как **Alt** и **Ctrl**).

Курсив
: Обозначает новые термины, названия файлов и их расширения, путь к файлам, директории и утилиты Unix.

Нижнее подчеркивание

Ссылки на страницы в Интернете, адреса электронной почты.

Эта пиктограмма обозначает совет, предложение или общее замечание.

Эта пиктограмма обозначает предостережение.

Благодарности

На обложке указан только один автор, однако многие люди приложили руку к созданию этой книги.

Я хотела бы поблагодарить моего агента Нейла Залкинда (Neil Salkind) за постоянные советы и поддержку; команду О'Рейлли, включая Мэри Трезелер (Mary Treseler), Сару Шнейдер (Sarah Schneider) и Меган Бланше (Meghan Blanchette), а также всех статистиков, которые помогали при техническом рецензировании текста. Я бы также хотела поблагодарить моих далеких от статистики друзей, которые постоянно требовали от меня объяснять им статистические концепции, что подтолкнуло меня к написанию этой книги, и моих коллег из центра устойчивой журналистики в государственном университете Кеннесо (Center for Sustainable Journalism at Kennesaw State University) за их терпение и снисходительность во время моего труда над переработкой этой книги. От всей души хочу поблагодарить мою бывшую коллегу Ранд Росс (Rand Ross) из университета Вашингтона в Сент-Луисе (Washington University in St. Louis) за то, что она помогала мне не сойти с ума во время написания первого издания этой книги, и моего мужа Дэна Пека (Dan Peck) за то, что он был воплощением современного супруга, готового всегда оказать поддержку.

Об авторе

Сара Бослаф (Sarah Boslaugh) получила докторскую степень по исследованиям и оцениванию в городском университете Нью-Йорка. В течение 20 лет она работала как статистический аналитик в различных профессиональных организациях, включая городской совет Нью-Йорка по образованию (New York City Board of Education), исследовательское отделение (Institutional Research Office) городского университета Нью-Йорка, медицинский центр Монтефиоре (Montefiore Medical Center), отдел социального обеспечения в Вирджинии (Virginia Department of Social Services), медицинская организация Магеллан (Magellan Health Services), медицинская школа при университете г. Вашингтон (Washington University School of Medicine) и организации BJC HealthCare. Она преподавала статистику в разных

аудиториях, а сейчас работает составителем заявок на гранты в государственном университете Кеннесоу (Kennesaw).

Сара Бослаф уже опубликовала две книги: «Справочник по программированию в SPSS средней сложности: использование программного кода для управления данными» (“An Intermediate Guide to SPSS Programming: Using Syntax for Data Management”, SAGE Publications, 2004) и «Вторичные источники данных в здравоохранении» (“Secondary Data Sources for Public Health”, Cambridge University Press, 2007), а также редактировала «Энциклопедию эпидемиологии» (“Encyclopedia of Epidemiology” for SAGE Publications, 2007).

В 2013 году издательством SAGE опубликована её новая книга, – «Системы здравоохранения во всем мире: сравнительный справочник» (“Healthcare Systems Around the World: A Comparative Guide”).

Об иллюстрации на обложке

На обложке книги «Статистика для всех» изображен колючий краб-паук (*Maja squinado*, *Maja brachydactyla*). Этот краб обитает в северо-восточной части Атлантического океана и в Средиземном море. Это самый крупный краб в Европе, диаметр его карапакса колеблется от 5 до 17 см. Его легко отличить от других крабов по двум похожим на рога шипам между глаз и шести, или около того, шипикам расположеным на каждой стороне панциря. Панцирь краба-паука красноватый с розовыми, коричневыми или желтыми отметинами и вся его поверхность покрыта мелкими шипами, как следует из названия животного.

Крабы-пауки иногда выползают на берег, но предпочитают глубины от 30 до 180 м. Это одиночные животные, за исключением периода спаривания, когда они образуют большие скопления. В годы, когда эти крабы особенно многочисленны, они могут досаждать ловцам омаров, поскольку могут разорять ловушки. Крабы-пауки сами являются объектом промысла из-за вкусного мяса конечностей.

Самцы крабов-пауков – активные хищники; их, кажущиеся слабыми конечности, на самом деле довольно мощные и могут открывать раковины небольших моллюсков, которых крабы поедают. Их конечности имеют два сочленения, так что крабы-пауки способны достать клешнями до своей спины, чтобы ущипнуть обидчика, хотя в целом безопаснее его держать за створки панциря. Клешни самок мельче и менее подвижные, поэтому они более уязвимы для нападения. Для защиты от врагов, к которым относятся омары, рыбы-губаны и каракатицы, многие виды крабов-пауков украшают свои колючие панцири водорослями, губками или грунтом, чтобы лучше замаскироваться на фоне дна.

Изображение на обложке предоставлено естественно-научной библиотекой Лидеккера (Lydekker’s Library of Natural History).

ГЛАВА 1. Основные понятия, связанные с измерениями

Для использования статистики при решении определенной задачи необходимо преобразовать информацию об этой задаче в данные. Это значит, что вы должны разработать или применить систему присвоения значений, чаще всего чисел, ключевым для рассматриваемой проблемы объектам или понятиям. Это не скрытый от понимания непосвященных процесс, а то, что люди делают ежедневно. Например, когда вы покупаете что-нибудь в магазине, сумма, которую вы платите, – это измерение: она выражает количество денег, которое вы должны заплатить, чтобы купить что-то. Аналогичным образом, когда вы утром становитесь на весы, число, которое вы видите, – это измерение вашего веса. В зависимости от места вашего проживания это число может быть выражено в фунтах или килограммах, но принцип присвоения числа физической величине (весу) сохраняется в любом случае.

Подходящие для анализа данные не обязательно должны быть числовыми. Например, понятия *мужчина* и *женщина* обычно используются в науке и повседневной жизни для классификации людей, и за этими категориями не стоит никаких чисел. Аналогично мы часто говорим о цветах объектов, таких как *красный* и *синий*, и к этим категориям также не привязано никаких чисел. (Хотя вы можете сказать, что этим цветам свойственны разные длины волны света, это знание не нужно для классификации объектов по цветам.)

Этот тип категориального мышления – привычный ежедневный опыт, и нас редко раздражает тот факт, что разные категории используются в разных ситуациях. Например, художник может различать *карминовый*, *малиновый* и *гранатовый*, тогда как неспециалисту достаточно называть их все *красным*. Сходным образом социолог, собирающий информацию о семейном статусе людей, будет различать *никогда не состоявших в браке*, *разведенных* и *вдовцов*, тогда как для кого-нибудь человек, относящийся к любой из этих трех категорий, будет просто *холостым*. Здесь важно понять, что уровень детализации, используемый при классификации, должен соответствовать ситуации, исходить из цели классификации и назначения собранной информации.

Измерение

Измерение – это процесс систематичного присвоения чисел объектам и их свойствам для облегчения использования математического аппарата при изучении и описании объектов и их взаимосвязей. Некоторые типы измерений абсолютно конкретны: например, измерения веса человека в фунтах или килограммах или его роста в футах и дюймах или метрах. Обратите внимание, что определенная система единиц измерения не так важна, как применение определенного набора правил: мы можем легко преобразовать вес, выраженный в килограммах, в вес, выраженный в фунтах, например. Хотя любая система единиц измерения может показаться необоснованной (попробуйте защитить футы и дюймы от нападок того, кто вырос, используя метрическую систему!), пока система остается постоянной по отношению к измеряемым признакам, мы можем использовать полученные результаты для вычислений.

Измерения не ограничены физическими величинами, такими как рост и вес. Тесты для измерения абстрактных величин, таких как интеллект или академическая успеваемость, широко используются в образовании и психологии, а разработкой и улучшением методов исследований этих типов абстрактных конструктов занимается специальная дисциплина – психометрика. Утверждать, что определенное измерение точно и осмысленно, более трудно, если его нельзя напрямую наблюдать. Однако вы можете оценить точность одной шкалы измерений, сравнивая результаты, которые были получены при помощи другой шкалы, точность которой известна. Применимость такого подхода при измерении веса не вызывает сомнений, дело обстоит сложнее, когда вам нужно измерить такой параметр, как интеллект. В данном случае не только не существует общепризнанных метрик интеллекта, с которыми можно сравнить новую шкалу, нет даже общего согласия по поводу того, что подразумевается под интеллектом. Иными словами, трудно уверенно судить о чьем-нибудь интеллекте, поскольку не существует ясного способа его измерения и, строго говоря, нет общепринятого определения интеллекта. Эти вопросы особенно актуальны в социологии и образовании, в которых основная часть исследований сосредоточена на таких абстрактных понятиях.

Типы измерений

В статистике обычно выделяют четыре типа, или уровня, измерений, эти же термины могут быть отнесены и к самим данным. Уровни измерений различаются и по смыслу чисел, используемых в системе измерений, и по типу статистических процедур, которые корректно применять для обработки данных.

Номинальные данные

Для *номинальных данных* числа выступают в виде имени или ярлыка и не имеют смысла как числа. Например, вы можете создать переменную для пола, которая принимает значение 1 для мужчин и 0 для женщин. Эти 0 и 1 не имеют смысла как

числа, а выступают в роли «ярлыков», сходным образом вы можете закодировать эти значения как *М* и *Ж*. Однако исследователи часто предпочитают числовую кодировку значений по нескольким причинам. Во-первых, это упрощает анализ данных, поскольку некоторые статистические программы не допускают использования нечисловых значений при определенных типах обработки данных. (Так что любые нечисловые данные придется перекодировать перед анализом.) Во-вторых, кодирование данных при помощи чисел позволяет избежать некоторых проблем при вводе данных, таких как конфликт между прописными и строчными буквами (для компьютера *М* и *м* – разные значения, однако тому, кто вводит данные, они могут показаться одинаковыми).

Номинальные данные могут иметь больше двух значений. Например, если вы изучаете связь между опытом игроков в бейсбол и их зарплатой, вы можете классифицировать игроков по их основной роли, используя традиционную систему: 1 – подающий, 2 – принимающий, 3 – первый полевой игрок и так далее.

Если вы не можете решить, относятся ли ваши данные к номинальному типу, задайте себе вопрос: отражают ли числа некоторое свойство так, что более высокое значение означает наличие большего количества этого свойства? Рассмотрим пример с кодировкой пола, где 0 обозначает женщину, а 1 – мужчину. Есть ли некоторое свойство пола, которым мужчина обладает в большей степени, чем женщина?[1] Конечно нет, и кодировка будет работать, если обозначать женщин 1, а мужчин 0. Тот же принцип применим и к бейсбольным игрокам: нет такого качества, как «бейсбольность», которое свойственно в большей степени полевым игрокам, по сравнению с подающими. Числа – всего лишь удобный способ обозначения объектов исследования, и наиболее важно то, что каждому состоянию признака соответствует свое значение. Другое название номинальных данных – *категориальные*, что отражает тот факт, что измерения скорее разделяют объекты на категории (мужчина или женщина, подающий или полевой игрок), а не измеряют некоторые присущие им свойства. В пятой главе обсуждаются методы анализа, подходящие для этого типа данных, и некоторые из разобранных в главе 13 непараметрических методов также подходят для категориальных данных.

Когда данные принимают только два значения, как в случае с женщинами и мужчинами, их называют *бинарными*. Этот тип данных настолько распространен, что для его анализа разработаны специальные методы, включая логистическую регрессию (обсуждается в главе 11), которая применяется во многих областях науки. Многие используемые в медицине статистики, такие как отношение шансов и отношение рисков (обсуждаются в главе 15), были разработаны для описания взаимосвязи между двумя бинарными переменными, поскольку они очень часто используются в медицинских исследованиях.

Порядковые данные

Порядковые данные – это данные, которые можно расположить в каком-либо осмысленном *порядке*, так что бóльшие значения соответствуют бóльшему про-

[1] Неудачный пример с точки зрения биолога. – *Прим. пер.*