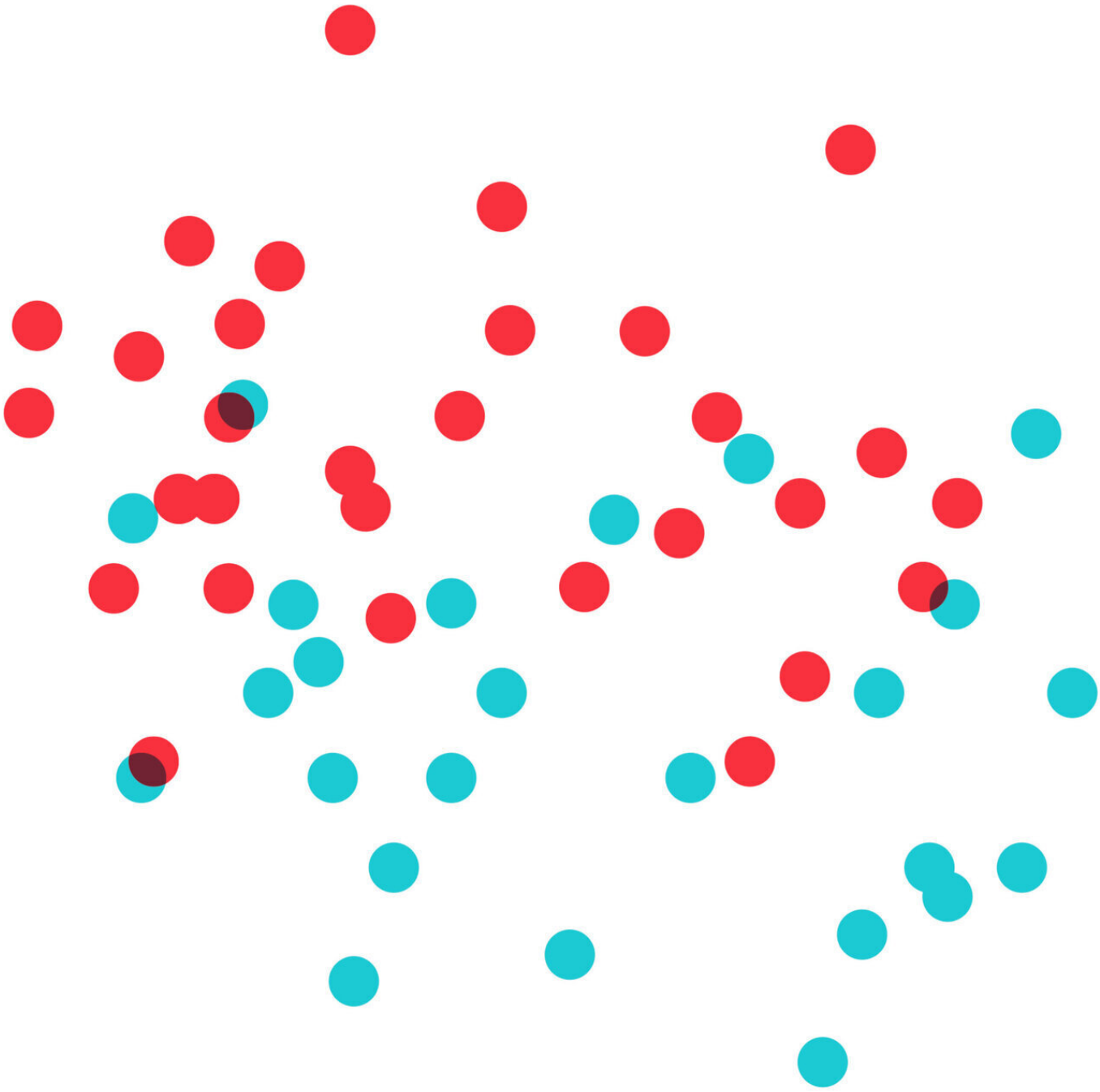
Искусство статистики

Как находить ответы в данных

Дэвид Шпигельхалтер



МИФ Научпоп

Дэвид Шпигельхалтер

Искусство статистики. Как находить ответы в данных

«Манн, Иванов и Фербер»
2019

УДК 311.1
ББК 65.01

Шпигельхалтер Д.

Искусство статистики. Как находить ответы в данных / Д. Шпигельхалтер — «Манн, Иванов и Фербер», 2019 — (МИФ Научпоп)

ISBN 978-5-00-169250-8

Статистика играла ключевую роль в научном познании мира на протяжении веков, а в эпоху больших данных базовое понимание этой дисциплины и статистическая грамотность становятся критически важными. Дэвид Шпигельхалтер приглашает вас в не обремененное техническими деталями увлекательное знакомство с теорией и практикой статистики. Эта книга предназначена как для студентов, которые хотят ознакомиться со статистикой, не углубляясь в технические детали, так и для широкого круга читателей, интересующихся статистикой, с которой они сталкиваются на работе и в повседневной жизни. Но даже опытные аналитики найдут в книге интересные примеры и новые знания для своей практики. На русском языке публикуется впервые.

УДК 311.1
ББК 65.01

ISBN 978-5-00-169250-8

Содержание

Введение 6
Глава 1. Расчет долей: качественные данные и проценты 16
Глава 2. Числовые характеристики выборки и представление данных 29
Конец ознакомительного фрагмента. 38

Дэвид Шпигельхалтер
Искусство статистики. Как находить ответы в данных

Издано с разрешения Penguin Books Ltd и Andrew Nurnberg Literary Agency

⁂

Статистикам всего мира – педантичным, отзывчивым, добросовестным людям, стремящимся использовать данные наилучшим образом

Введение

Цифры сами по себе не умеют говорить. Именно мы говорим за них. Мы наполняем их смыслом.
Нейт Сильвер, «Сигнал и шум»[1],[2]

Зачем нужна статистика?

Психологический портрет Гарольда Шипмана, более известного как Доктор Смерть, не похож на серийного убийцу, тем не менее этот человек поставил рекорд по убийствам. Тихий семейный врач, работавший в пригороде Манчестера, в период с 1975 по 1998 год ввел как минимум 215 пожилым пациентам смертельную дозу опиатов. Но в конце концов он «прокололся», подделав завещание одной из своих жертв, которая якобы оставила ему часть наследства, что весьма насторожило ее дочь-адвоката. Проверка компьютера врача показала, что он задним числом изменял информацию в медицинских картах пациентов, чтобы состояние их здоровья казалось хуже, чем было на самом деле. Он считался увлеченным поборником технологий, но не был достаточно технически подкован, чтобы понимать, что время каждого внесенного изменения фиксируется (кстати, хороший пример метаданных, раскрывающих скрытый смысл данных).

В результате эксгумации пятнадцати тел его пациентов (из тех, которых не кремировали) в них были обнаружены смертельные дозы диаморфина, медицинской формы героина. В 1999 году Шипмана судили за пятнадцать убийств и приговорили к пожизненному заключению. Он не защищался и не произнес на суде ни слова. Впоследствии было инициировано публичное расследование, чтобы определить, какие еще преступления он мог совершить, помимо рассмотренных в суде, и можно ли было разоблачить его раньше. Я был одним из нескольких статистиков, которых тогда привлекали к расследованию. Оно пришло к выводу, что он определенно убил 215 пациентов, а, возможно, и еще 45[3].

Эта книга посвящена применению **статистики**[4] для поиска ответов на вопросы (некоторые из них выделены), которые возникают, когда мы пытаемся лучше понять мир. Чтобы получить представление о мотивах поведения Шипмана, вполне закономерно спросить:

Каких людей убивал Гарольд Шипман, и когда они умирали?

В ходе упомянутого расследования была представлена информация о возрасте, поле и дате смерти каждой жертвы. Рис. 0.1 – довольно сложная визуализация этих данных, отображающая возраст и дату смерти жертвы, при этом цвет точек указывает на пол – мужской или женский. На осях добавлены гистограммы, демонстрирующие распределение по возрасту (с интервалом в пять лет).

---

[1] Издана на русском языке: *Сильвер Н.* Сигнал и шум. Почему одни прогнозы сбываются, а другие – нет. М.: КоЛибри, 2015. *Прим. пер.*

[2] Эта книга Нейта Сильвера – превосходное введение в сферу применения статистики для прогнозов в спорте и других областях.

[3] Подробно данные о Шипмане обсуждаются в работе: D. Spiegelhalter and N. Best, 'Shipman's Statistical Legacy', Significance 1:1 (2004), 10–12. Все документы по этому общественному расследованию находятся на сайте http://www.the-shipman-inquiry.org.uk/reports.asp.

[4] Термины, выделенные **полужирным шрифтом**, включены в глоссарий в конце книги.

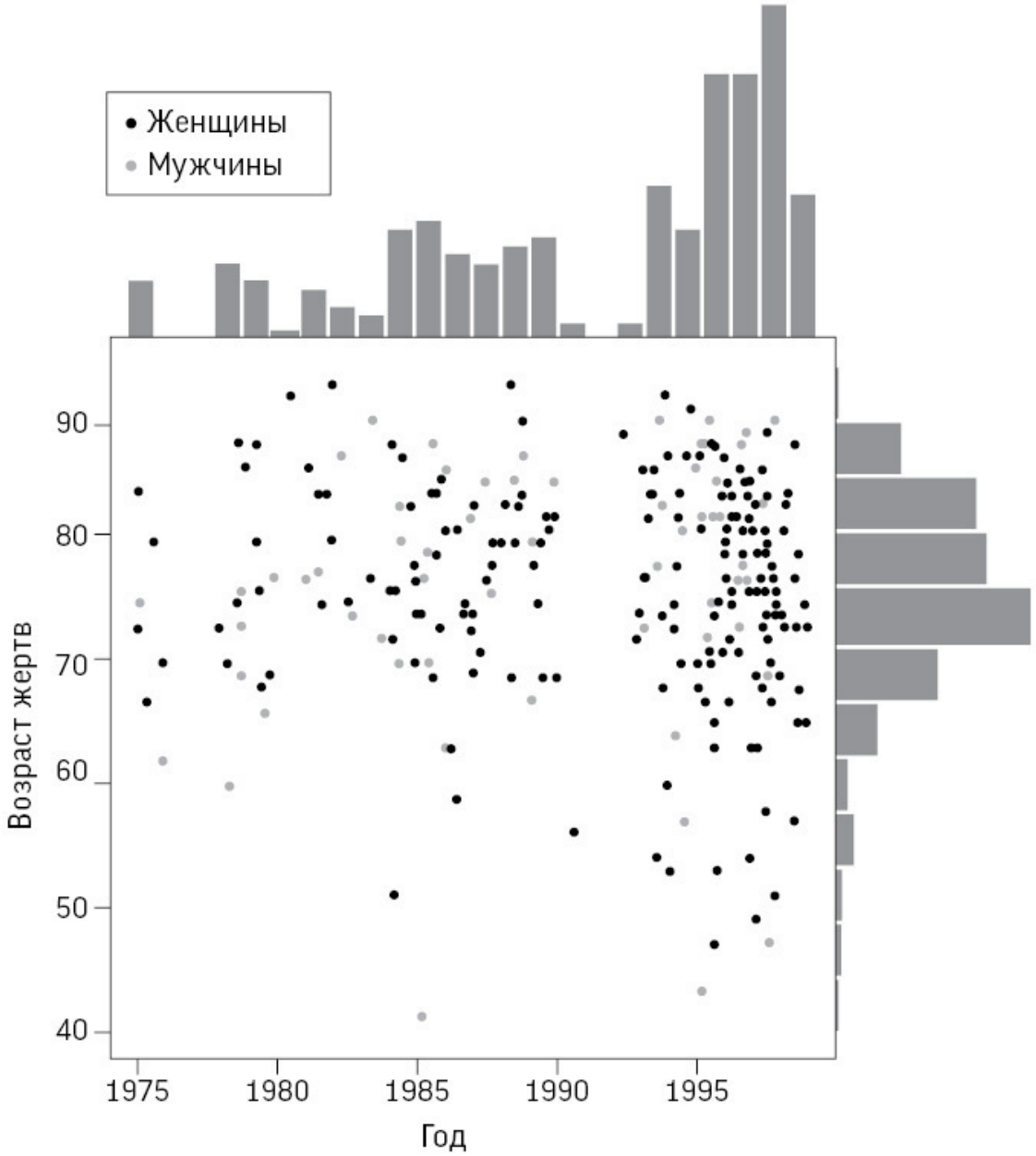


Рис. 0.1

Диаграмма рассеяния, показывающая возраст и год смерти 215 подтвержденных жертв Гарольда Шипмана. По осям добавлены гистограммы, демонстрирующие распределение по возрасту и году совершения убийства

Даже беглый взгляд на рисунок позволяет сделать некоторые выводы. Черных точек больше, чем белых, а значит, жертвами Шипмана в основном были женщины. Гистограмма справа демонстрирует, что возраст большинства жертв – 70–80 лет, но разброс точек показывает, что, хотя изначально все жертвы были пожилыми, впоследствии появилось несколько более молодых пациентов. Гистограмма сверху четко показывает промежуток примерно в 1992 году, когда убийств не происходило. Оказывается, до этого Шипман имел общую практику с другими врачами, но затем – возможно, чтобы избежать подозрений, – стал работать один. После чего его деятельность активизировалась, что и отображено на верхней гистограмме.

Анализ случаев, выявленных в ходе расследования, приводит к дальнейшим вопросам о том, как Шипман совершал убийства. Определенная статистическая информация содержится

в данных о времени смерти жертв (указывалось в свидетельстве о смерти). На рис. 0.2 сравниваются два линейных графика: время смерти пациентов Шипмана и пациентов других местных семейных врачей. Здесь не нужен тонкий анализ: разница видна невооруженным глазом. Пациенты Шипмана в подавляющем большинстве умирали вскоре после полудня.

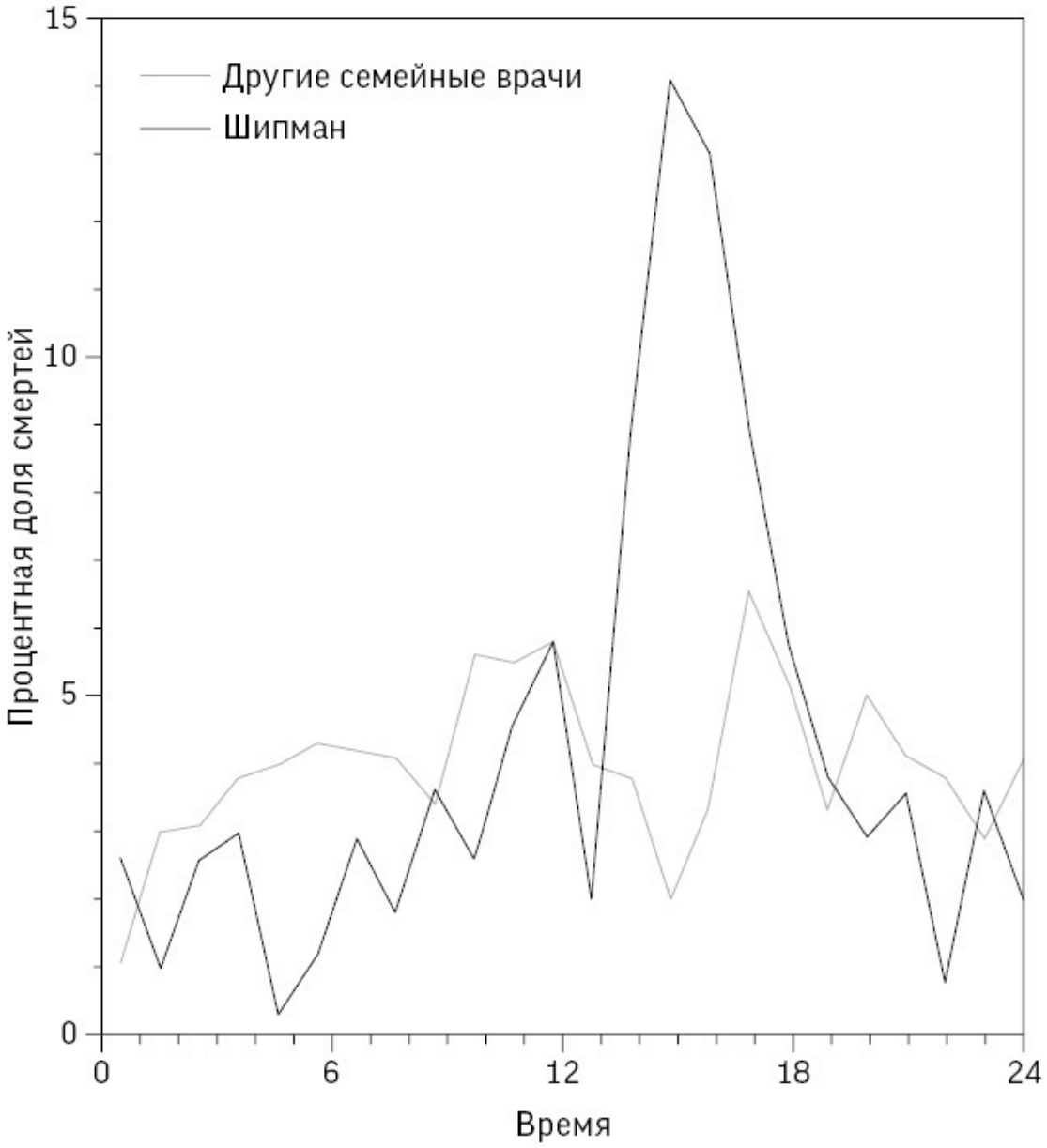


Рис. 0.2

Сравнение времени смерти пациентов Шипмана и пациентов других семейных врачей. Выявление закономерности не требует углубленного статистического анализа

Хотя сами по себе эти данные не объясняют причин такой особенности, дальнейшее расследование обнаружило, что он посещал пожилых больных на дому после обеда, когда, как правило, оставался с ними наедине. Он предлагал им инъекцию якобы для улучшения самочувствия, которая на самом деле была смертельной дозой диаморфина. После того как пациент на его глазах тихо отходил в мир иной, Шипман вносил изменения в медицинскую карту, чтобы смерть выглядела естественной.

Судья Джанет Смит, возглавлявшая публичное расследование, позже говорила: «Я все еще чувствую, насколько это страшно, просто невообразимо и немыслимо. Этот человек изо дня в день ходил к людям, притворяясь на редкость заботливым врачом, неся с собой смертельное оружие, которое он неоднократно хладнокровно использовал».

В определенной степени он рисковал, ведь даже одно-единственное вскрытие могло бы его разоблачить, но, учитывая возраст пациентов и очевидные естественные причины смерти, аутопсию никто не проводил. Мотивы совершения убийств тоже не были установлены: Шипман не давал показаний в суде, никогда ни с кем (включая членов семьи) не говорил на эту тему и окончил жизнь самоубийством в тюрьме в то время, когда жена еще имела право на его пенсию[5].

Мы можем считать такой вид исследовательской работы «криминалистической» статистикой, и в данном случае это название верно буквально. Никакой математики, никакой теории – просто поиск закономерностей, который может привести к более интересным вопросам. Детали злодеяний Шипмана определялись для каждого случая, однако общий анализ данных дает понимание того, как он совершал преступления.

Далее (в главе 10) мы увидим, мог ли формальный статистический анализ помочь поймать Шипмана раньше[6]. Между тем его история достаточно убедительно демонстрирует огромный потенциал использования данных для лучшего понимания мира и вынесения более правильных суждений. Именно для этого и нужна статистика.

Превращение мира в набор данных

Статистический подход к преступлениям Шипмана требует от нас отказаться от перечисления длинного списка отдельных трагедий, за которые он несет ответственность. Все персональные данные о жизни и смерти людей нужно свести к набору фактов и чисел, которые можно подсчитать и отобразить на диаграммах. Каким бы бездушным и бесчеловечным на первый взгляд это ни казалось, но, чтобы использовать статистику для понимания происходящего, наш повседневный опыт следует обратить в данные, а это означает категоризацию и классификацию событий, выполнение измерений, анализ результатов и формулирование выводов. Однако даже простая категоризация и классификация может представлять серьезную проблему. Рассмотрим следующий вопрос, который должен заинтересовать всех, кому небезразличны проблемы окружающей среды.

Сколько деревьев на нашей планете?

Прежде чем задуматься об ответе на этот вопрос, нужно разобраться с простым базовым понятием. Что такое дерево? Возможно, вы посчитаете некий увиденный объект деревом и будете уверены в этом, но другие люди, в отличие от вас, назовут его кустом. Следовательно, чтобы превратить опыт в данные, нужно начинать со строгих определений.

Оказывается, официальное определение дерева звучит так: это многолетнее растение с одревесневшим стеблем (стволом), имеющим довольно большой диаметр на высоте груди (ДВГ)[7]. Лесная служба США считает, что растение можно официально именовать деревом, если его ДВГ не менее 5 дюймов (12,7 сантиметра), но большинство организаций используют значение 10 сантиметров (4 дюйма).

Однако мы не можем бродить по всей планете, измеряя каждое растение с деревянистым стволом, чтобы проверить, удовлетворяет ли оно данному критерию. Поэтому специ-

[5] Шипман повесился в Уэйкфилдской тюрьме за день до своего 58-летия. После этого жена получала деньги от Национальной службы здравоохранения Великобритании, на которые не имела бы права, если бы ее муж умер после 60 лет – возраста выхода на пенсию. *Прим. пер.*

[6] Спойлер: это можно было сделать практически наверняка.

[7] В отечественной практике высотой груди дерева считается расстояние в 1,3 метра от корневой шейки. *Прим. пер.*

алисты, исследовавшие этот вопрос, использовали более прагматичный подход: они взяли несколько участков с общим типом ландшафта (называемый биомом) и подсчитали среднее число деревьев на один квадратный километр. Затем с помощью спутниковой съемки измерили общую площадь поверхности планеты, покрытой каждым типом биома, провели сложное статистическое моделирование и в итоге получили общее число деревьев на планете – примерно 3,04 триллиона (то есть 3 040 000 000 000). Хотя цифра кажется огромной, ученые считают, что когда-то деревьев было вдвое больше[8],[9].

Если разные организации расходятся во мнениях даже относительно того, что следует называть деревом, то стоит ли удивляться, что более сложные понятия поддаются определению еще труднее. Яркий пример – определение безработицы в Великобритании, где за период с 1979 по 1996 год оно менялось по меньшей мере 31 (!) раз[10]. Постоянно пересматривается определение валового внутреннего продукта (ВВП). Так, к ВВП Великобритании в 2014 году были отнесены торговля наркотиками и проституция; для оценок использовались необычные источники данных, например, такие как сайт Punternet, который оценивает услуги проституток. Он-то и предоставил цены различных видов услуг[11]. Даже наши собственные ощущения могут быть систематизированы и подвергнуты статистическому анализу. В рамках проходившего в течение года опроса, закончившегося в сентябре 2017-го, у 150 тысяч человек спросили, насколько счастливыми они себя чувствовали вчера[12]. Средний балл ответов по шкале от 0 до 10 составил 7,5, то есть больше, чем в 2012 году, когда он был 7,3. Это может быть связано с восстановлением экономики после финансового кризиса 2008 года. Самые низкие баллы оказались у людей в возрасте от 50 до 54 лет, а самые высокие – от 70 до 74 лет, что типично для Великобритании[13].

Измерять счастье сложно, тогда как ответить на вопрос, жив человек или мертв, казалось бы, куда проще (как покажут примеры, представленные в книге, рождаемость и смертность – общие проблемы в статистической науке). Однако в США каждый штат может иметь собственное юридическое определение смерти, и, хотя в 1981 году в целях унификации был принят Закон о единообразном определении смерти (Uniform Declaration of Death Act), небольшие расхождения в этом вопросе все же остались. Так, человек, объявленный мертвым в Алабаме, может – по крайней мере, теоретически – перестать быть юридически мертвым при пересечении границы с Флоридой, поскольку там факт смерти должны зарегистрировать два дипломированных врача[14].

Эти примеры показывают, что статистические данные всегда в какой-то степени основаны на суждениях и было бы очевидным заблуждением считать, что всю сложность личного опыта можно однозначно закодировать и записать в электронных таблицах или каких-то компьютерных программах. Все определенные, посчитанные и измеренные характеристики людей

---

[8] T. W. Crowther et al., 'Mapping Tree Density at a Global Scale', Nature 525 (2015), 201–5.

[9] Погрешность для этой величины – 0,1 триллиона, то есть истинное количество деревьев на Земле находится в диапазоне 2,94–3,14 триллиона (я полагаю, что эта величина слишком точна, если учесть большое количество предположений, принятых при моделировании). По оценкам ученых, ежегодно вырубается 15 миллиардов (15 000 000 000) деревьев и с момента возникновения человеческой цивилизации планета уже потеряла 46 % деревьев.

[10] E. J. Evans, Thatcher and Thatcherism (Routledge, 2013), p. 30.

[11] Изменения в национальной статистике: включение незаконных препаратов и проституции в национальную статистику Великобритании [Интернет] (Национальное статистическое управление, 2014).

[12] Национальное статистическое управление Великобритании описывает ряд мер для благосостояния на сайте https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing.

[13] Если бы я был типичным среднестатистическим человеком, этот факт давал бы мне основание заранее чему-то радоваться.

[14] N. T. Nikas, D. C. Bordlee and M. Moreira, 'Determination of Death and the Dead Donor Rule: A Survey of the Current Law on Brain Death', Journal of Medicine and Philosophy 41:3 (2016), 237–56.

и окружающего нас мира – это всего лишь информация и отправная точка к реальному миропониманию.

Как источник таких знаний данные имеют два основных ограничения. Во-первых, это почти всегда несовершенная мера того, что нас действительно интересует: простая просьба оценить, насколько люди были счастливы на прошлой неделе, по шкале от 0 до 10, вряд ли отражает эмоциональное благополучие нации. Во-вторых, все, что мы станем измерять, будет отличаться в разных местах, у разных людей и в разное время, и проблема состоит в умении извлечь осмысленную информацию из этих, на первый взгляд, случайных колебаний.

На протяжении веков статистика сталкивалась с этими двумя задачами и играла ведущую роль в стремлении ученых познать мир. Она дает основу для интерпретации данных (которые всегда несовершенны), чтобы отличить важные взаимосвязи от индивидуальных особенностей, которые делают нас уникальными. Однако мир постоянно меняется, появляются новые вопросы и новые источники данных, поэтому и статистика должна меняться.

Люди считали и измеряли всегда. Однако современная статистика как наука фактически зародилась в 1650-х годах, когда, как мы увидим в главе 8, понятие вероятности впервые было правильно представлено Блезом Паскалем и Пьером Ферма. С такой прочной математической основой прогресс заметно ускорился. В сочетании с данными о возрасте смерти людей теория вероятностей позволила рассчитывать пенсии и годовые платежи. Когда ученые поняли, как работать с разбросами в измерениях, это революционизировало астрономию. Энтузиасты Викторианской эпохи[15] были одержимы сбором сведений о человеческом теле (и о многом другом) и установили прочную связь между статистическим анализом и генетикой, биологией и медициной. Позже, в XX веке, статистика приблизилась к математике, и, к сожалению, для многих студентов и практиков эта область стала синонимом механического приложения определенных статистических инструментов, многие из которых были названы в честь эксцентричных статистиков – с ними мы познакомимся далее в книге.

Этот распространенный взгляд на статистику как на базовый «набор инструментов» в настоящее время сталкивается с серьезными проблемами. Во-первых, мы живем в век **науки о данных**, когда большие и сложные массивы данных собираются из самых обычных источников, таких как мониторинг дорожного движения, социальных сетей и покупок онлайн, а затем используются в качестве основы для технологических инноваций – например, оптимизации движения транспорта, целевой рекламы или систем рекомендации покупок. **Алгоритмы,** основанные на **больших данных**, мы рассмотрим в главе 6. Сегодня, чтобы стать специалистом по обработке данных, нужно не только изучать статистику, но и обладать навыками программирования, разработки алгоритмов, управления данными, а также разбираться в самом предмете.

Еще одну реальную угрозу традиционному взгляду на статистику представляет колоссальный рост количества проводимых исследований, особенно в биомедицине и социальных науках, в сочетании с требованием публикаций в высокорейтинговых журналах. Это привело к сомнениям в надежности определенной части научной литературы и утверждениям о невоспроизводимости многих «открытий» другими исследователями. Как, например, продолжающийся спор, может ли «поза силы» вызвать гормональные и другие изменения у человека[16]. На некорректном применении стандартных статистических методов лежит немалая доля вины за то, что известно как кризис воспроизводимости (или репликации) в науке.

---

[15] Викторианская эпоха – время правления королевы Виктории (1837–1901). *Прим. пер.*

[16] J. P. Simmons and U. Simonsohn, 'Power Posing: P-Curving the Evidence', Psychological Science 28 (2017), 687–93. Возражения смотрите в работе: A. J. C. Cuddy, S. J. Schultz and N. E. Fosse, 'P-Curving a More Comprehensive Body of Research on Postural Feedback Reveals Clear Evidential Value for Power-Posing Effects: Reply to Simmons and Simonsohn (2017)', Psychological Science 29 (2018), 656–66.

В связи с растущей доступностью больших массивов данных и удобного программного обеспечения для их анализа может показаться, что необходимость в изучении статистических методов снижается. Однако крайне наивно так думать. Увеличение объема данных, рост количества и сложности научных исследований еще больше затрудняют процесс формулирования соответствующих выводов. Большее количество данных означает, что нам надо еще лучше осознавать, чего на самом деле стоят такие доказательства.

Например, интенсивный анализ массивов данных может повысить вероятность ложных открытий – как вследствие систематической ошибки, присущей источнику, так и в результате выполнения множества тестов, но сообщения только о тех из них, которые выглядят интересными, то есть так называемого слепого прочесывания данных. Чтобы иметь возможность критически относиться к опубликованным научным работам, а тем более к ежедневным сообщениям СМИ, нужно четко осознавать опасность такого избирательного подхода, понимать необходимость проверки утверждений независимыми специалистами и осознавать риск неправильной интерпретации результатов одного исследования вне контекста.

Все это можно объединить под термином «**грамотность в работе с данными**», который описывает не только способность проводить статистический анализ реальных проблем, но и умение понять и критически проанализировать любые выводы, сделанные другими на основе статистики. Повышение такой грамотности предполагает изменение методики обучения статистике.

Преподавание статистики

Целые поколения студентов страдали от сухих курсов статистики, основанных на изучении набора методов, применяемых в различных ситуациях, причем больше внимания в них уделялось математической теории, чем пониманию причин применения той или иной формулы, или проблемам, возникающим при попытке использовать данные для ответа на вопросы.

К счастью, все меняется. Наука о данных и грамотность в работе с ними требуют подхода, направленного на решение основных проблем, где применение конкретных статистических инструментов рассматривается лишь как один из компонентов цикла исследований. Цикл **PPDAC** (Problem, Plan, Data, Analysis, Conclusion) был предложен как модель решения проблем, которую мы будем использовать в этой книге[17]. Рис. 0.3 основан на примере Новой Зеландии, которая считается мировым лидером по преподаванию статистики в школах.

[17] Основная рекомендация Американской статистической ассоциации (ASA) – «Преподавать статистику как исследовательский процесс решения проблем и принятия решений». См. https://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx. Цикл PPDAC был представлен в работе: R. J. MacKay and R. W. Oldford, 'Scientific Method, Statistical Method and the Speed of Light', Statistical Science 15 (2000), 254–78. Его активно поддерживает школьная система Новой Зеландии, которая обеспечивает хорошее статистическое образование. См. C. J. Wild and M. Pfannkuch, 'Statistical Thinking in Empirical Enquiry', International Statistical Review 67 (1999), 223–265, и онлайн-курс «Данные для идей», https://www.futurelearn.com/courses/data-to-insight.

Рис. 0.3

Цикл решения проблем PPDAC (от проблемы, плана, данных, анализа к заключению и коммуникации), начинающийся заново в другом цикле

Первая стадия цикла – определение *проблемы*: статистическое исследование всегда начинается с вопроса, например, с такого как наш вопрос о закономерностях убийств Гарольда Шипмана или о количестве деревьев в мире. Далее мы рассмотрим самые разные проблемы – от ожидаемой пользы различных методов послеоперационного лечения рака молочной железы до вопроса, почему у стариков большие уши.

Искушение пренебречь необходимостью в хорошем *плане* довольно велико. В случае с Шипманом требовалось просто собрать как можно больше данных о жертвах. Однако люди, считавшие деревья, уделили пристальное внимание точным определениям и методам измерения, поскольку надежные заключения можно сделать только на основе тщательно спланированного исследования. К сожалению, желание быстрее получить данные и приступить к их анализу приводит к тому, что эта стадия часто игнорируется.

Сбор *данных* требует определенных организаторских навыков и навыков кодирования, наличие которых все больше ценится в науке о данных, особенно потому, что данные из некоторых источников могут нуждаться в тщательной очистке перед их анализом. Системы сбора данных со временем меняются, там могут быть выявлены ошибки – само выражение «найти

данные» четко указывает на то, что они бывают довольно грязными, как нечто, подобранное на улице.

В курсах статистики основной упор делается на стадию *анализа*, и мы рассмотрим в книге ряд аналитических методов; однако иногда все, что необходимо сделать на данном этапе, – это наглядная визуализация, как на рис. 0.1.

Наконец, главное в статистической науке – сделать соответствующие *заключения*, которые полностью признают и четко показывают ограничения в доказательствах, как на графических иллюстрациях данных Шипмана. Любые заключения, как правило, приводят к новым вопросам, поэтому цикл начинается заново – как в случае, когда мы стали анализировать время смерти пациентов Шипмана.

Хотя на практике цикл PPDAC, представленный на рис. 0.3, может не соблюдаться с абсолютной точностью, он подчеркивает, что формальные методы статистического анализа – это только часть работы статистика или специалиста по обработке данных. Статистика – нечто гораздо большее, чем область математики, содержащая заумные формулы, с которыми пытались совладать (нередко против своего желания) поколения учащихся.

Эта книга

В 1970-е годы, когда я был студентом, в Великобритании работало всего три телеканала, компьютеры напоминали огромный двустворчатый шкаф, а ближе всего к «Википедии» было удивительное портативное устройство, описанное в (необычайно прозорливом) путеводителе Дугласа Адамса «Автостопом по галактике»[18]. Поэтому для самосовершенствования мы обращались к книгам издательства Pelican, и их легко узнаваемые синие корешки были обычной приметой каждой студенческой полки[19].

Поскольку я изучал статистику, моя коллекция Pelican включала Facts from Figures («Факты из цифр») Майкла Морони (1951) и How to Lie with Statistics Дарелла Хаффа (1954)[20]. Тираж этих почтенных трудов составлял сотни тысяч экземпляров, что отражало как степень интереса к статистике, так и удручающее отсутствие выбора в те времена. Эти классики прекрасно продержались 65 лет, однако нынешнее время требует других подходов к преподаванию статистики, основанных на вышеизложенных принципах. Поэтому решение проблем реального мира используется в книге в качестве отправной точки для представления статистических идей. Некоторые из этих идей могут показаться очевидными, тогда как другие, более тонкие, требуют определенных умственных усилий, хотя математические знания даже в этом случае не понадобятся. В отличие от традиционных текстов эта книга сосредоточена на концептуальных вопросах, а не на технических аспектах, и содержит лишь несколько вполне безобидных уравнений, а также глоссарий с объяснениями. Хотя программное обеспечение – важная часть любой работы в науке о данных и статистике, эта книга на нем не фокусируется – вы и так без труда найдете руководства по таким языкам, как R или Python.

На все выделенные в книге вопросы можно в какой-то степени ответить с помощью статистического анализа, хотя они и сильно отличаются по масштабности. Одни – важные научные гипотезы, например, существует ли бозон Хиггса[21] или убедительные подтверждения экстрасенсорного восприятия. Другие касаются здравоохранения – например, выше ли показатель

---

[18] Книга Дугласа Адамса вышла в 1979 году, когда он уже получил степень и преподавал. *Прим. пер.*Издана на русском языке: *Адамс Д.* Автостопом по галактике. М.: АСТ, 2014. *Прим. ред.*

[19] Издательство (дочернее предприятие (Penguin Books) было основано в 1937 году и выпускало недорогие научно-популярные (и другие нехудожественные) книги в мягких обложках. Они активно использовались для самообразования после войны, а газета The Guardian даже назвала эти книги «неформальным университетом для британцев 1950-х». *Прим. пер.*

[20] Издана на русском языке: *Хафф Д.* Как лгать при помощи статистики. М.: Альпина Паблишер, 2015. *Прим. пер.*

[21] Питер Хиггс (род. 1929) – британский физик, предложивший в 1964 году идею нового поля и соответствующей частицы (бозона), которые сейчас носят его имя. *Прим. пер.*

выживаемости в более загруженных больницах и полезны ли скрининговые исследования[22] для обнаружения рака яичников. Иногда мы просто хотим оценить некоторые величины, такие как риск развития рака от употребления сэндвичей с беконом, количество сексуальных партнеров британцев в течение жизни и пользу от ежедневного употребления статинов[23].

Многие вопросы просто интересны: скажем, определение самого счастливого выжившего при крушении «Титаника»; мог ли Гарольд Шипман быть разоблачен раньше; какова вероятность того, что скелет, найденный под автостоянкой в Лестере, действительно принадлежит Ричарду III.

Эта книга предназначена как для студентов-статистиков, которые хотят ознакомиться с предметом, не углубляясь в технические детали, так и для обычных читателей, интересующихся статистикой, с которой они сталкиваются на работе и в повседневной жизни. Я делаю акцент на осторожном обращении со статистическими данными: числа могут казаться сухими фактами, однако описанные выше попытки измерить деревья, счастье и смерть уже показали, что с ними нужно обращаться очень осторожно.

Статистика помогает прояснить стоящие перед нами вопросы, но при этом мы прекрасно знаем, что данными можно злоупотреблять – часто для навязывания чужого мнения или просто для привлечения внимания. Умение оценивать истинность статистических утверждений становится ключевым навыком в современном мире, и я надеюсь, что эта книга научит людей ставить под сомнение достоверность цифр, с которыми они сталкиваются в повседневной жизни.

Выводы

- Превращение опыта в данные – непростое дело, а способность данных описывать мир, безусловно, ограничена.
- У статистики как науки долгая, вполне успешная история, однако сейчас она меняется вследствие повышения доступности данных.
- Владение статистическими методами – важный навык специалиста по обработке данных.
- Преподавание статистики сегодня сосредоточивается не на математических методах, а на полном цикле решения задачи.
- Цикл PPDAC предоставляет удобный алгоритм поиска ответа на вопросы: проблема → план → данные → анализ → заключение и коммуникация.
- Грамотность в использовании данных – ключевой навык в современном мире.

[22] Скрининговые исследования – обследование людей, не имеющих симптомов, с целью выявить какое-нибудь заболевание. *Прим. пер.*

[23] Статины – препараты, которые применяются для снижения уровня холестерина в крови. *Прим. пер.*

Глава 1. Расчет долей: качественные данные и проценты

Что происходило с детьми, которым делали операции на сердце в Бристоле между 1984 и 1995 годами?

У 16-месячного Джошуа Л. была транспозиция магистральных сосудов – тяжелая форма врожденного порока сердца, при котором крупные артерии, отходящие от сердца, присоединены к неправильному желудочку. Ему требовалась операция по «переключению» сосудов. В 7 утра 12 января 1995 года родители пожелали Джошуа удачи, и медики увезли его на операцию в Королевскую больницу Бристоля. Но родители малыша не знали, что слухи о невысоком уровне выживаемости после хирургических операций в Бристоле ходили с начала 1990-х. Никто не сказал им и того, что медсестры увольнялись, чтобы избежать тех непростых моментов, когда приходится сообщать родителям, что их ребенок умер, или что накануне вечером проходил консилиум, где обсуждался вопрос об отмене операции Джошуа[24].

Ребенок умер на операционном столе. А в следующем году Генеральный медицинский совет (регулирующий орган) начал расследование после жалобы родителей Джошуа и родителей других умерших детей, и в 1998-м два хирурга и бывший руководитель отделения были признаны виновными в ненадлежащем исполнении профессиональных обязанностей. Волнения в обществе не утихали, поэтому было инициировано еще одно официальное расследование: группе статистиков поручили сравнить показатели выживаемости в Бристоле с другими больницами Соединенного Королевства в период с 1984 по 1995 год. Я возглавлял эту группу.

Сначала нам предстояло выяснить, сколько детей перенесли операцию и сколько умерли. Звучит вроде бы незамысловато, но, как мы убедились в предыдущей главе, даже простой подсчет событий может вызывать сложности. Что значит ребенок? Что считается операцией на сердце? Когда можно утверждать, что смерть наступила в результате операции? И даже если вопрос со всеми этими понятиями урегулирован, можно ли определить количество таких событий?

Мы решили считать ребенком любого человека до 16 лет и сосредоточились на открытых операциях с подключением к аппарату искусственного кровообращения. За один раз на сердце могло проводиться несколько операций, но они рассматривались нами как одно событие. Случаи смерти учитывались, если она наступала в течение 30 дней после операции, будь то в больнице или нет, вследствие хирургического вмешательства. Мы понимали, что смерть – несовершенная мера качества операции, поскольку не учитывались дети, которые в результате ее проведения получили повреждение мозга или другие виды инвалидности, однако сведениями о таких долгосрочных последствиях мы не располагали.

Основным источником данных стала Национальная статистика эпизодов в больницах (HES), полученная на основе информации, введенной низкооплачиваемыми программистами. У врачей HES пользовалась плохой репутацией, но гигантским преимуществом этого источника было то, что его можно было связать с национальными данными о смертности. Существовала также параллельная система данных, вносимых непосредственно в Реестр операций на сердце (CSR), созданный профессиональным сообществом хирургов.

Хотя оба источника, по логике, должны быть примерно одинаковыми, на практике они демонстрировали существенное расхождение: за 1991–1995 годы HES указывала 62 смерти

[24] См. 'History of Scandal', Daily Telegraph, 18 July 2001, and D. J. Spiegelhalter et al., 'Commissioned Analysis of Surgical Performance Using Routine Data: Lessons from the Bristol Inquiry', Journal of the Royal Statistical Society: Series A (Statistics in Society) 165 (2002), 191–221.

при 505 операциях на открытом сердце (14 %), а CSR – 71 смерть при 563 операциях (13 %). В нашем распоряжении было еще не менее пяти дополнительных местных источников сведений – от анестезиологической документации до собственных журналов хирургов. Бристоль располагал множеством данных, но ни один из источников не мог считаться истинным и никто не брал ответственность за анализ результатов хирургических вмешательств и принятие мер.

Мы подсчитали, что если бы в бристольской больнице средний риск для пациентов был таким же, как в целом по Великобритании, то за указанный период было бы зафиксировано 32 смерти, а не 62 фактических, что мы определили как «30 избыточных смертей в период с 1991 по 1995 год»[25]. Цифры менялись в зависимости от источников данных, и может показаться необычным, что мы даже не смогли установить основные факты о количестве операций и их результатах, хотя нынешние системы регистрации стоило бы улучшить.

Наши выводы широко освещались в прессе, и бристольское расследование привело к значительному изменению отношения к отслеживанию ситуации в здравоохранении: контроль над медициной больше не доверяли ей самой. Появились механизмы для публичного представления данных о выживаемости в больницах, хотя, как мы сейчас увидим, даже способ отображения может влиять на их восприятие аудиторией.

Представление результатов

Данные, фиксирующие, произошли какие-то события или нет, известны как **бинарные (двоичные) данные,** поскольку они могут выражаться только двумя значениями, например да или нет, болен или здоров. Из набора бинарных данных можно извлечь обобщенную информацию – общее количество и доля случаев, когда событие произошло.

В этой главе подчеркивается важность способа представления статистических данных. В каком-то смысле мы переходим к последней стадии цикла PPDAC, на которой делаются заключения; и хотя форма их подачи традиционно не считается значимой темой в статистике, растущий интерес к визуализации данных отражает изменения в данном вопросе. Поэтому в этой и следующей главах мы сосредоточимся на способах отображения данных, позволяющих быстро уловить суть происходящего без детального анализа. И начнем с рассмотрения альтернативных способов их представления, которые – во многом благодаря бристольскому расследованию – теперь стали общедоступны.

В табл. 1.1 отображены результаты лечения примерно 13 тысяч детей, перенесших операцию на сердце в Соединенном Королевстве Великобритании и Северной Ирландии в 2012–2015 годах[26]. В течение 30 дней после операции умерли 263 ребенка, и, безусловно, каждая из смертей – трагедия для семьи. Для них будет слабым утешением то, что со времени бристольского расследования показатель выживаемости значительно повысился и теперь составляет 98 %, поэтому у семей с детьми, нуждающимися в операции на сердце, более обнадеживающие перспективы.

Таблица 1.1

Результаты операций на сердце у детей в больницах Соединенного Королевства Великобритании и Северной Ирландии за 2012–2015 годы с точки зрения выживаемости в течение 30 дней после операции

[25] Сейчас я сожалею об использовании выражения «избыточные смерти», поскольку газеты потом интерпретировали его как «предотвратимые случаи смерти». На деле просто по вероятностным соображениям примерно в половине больниц количество смертей будет больше ожидаемого, и лишь некоторых из них можно было бы избежать.

[26] Данные о результатах выживания детей, перенесших операции на сердце, в Соединенном Королевстве Великобритании и Северной Ирландии можно получить на сайте http://childrensheartsurgery.info/.

Больница	Количество прооперированных детей	Количество проживших минимум 30 дней после операции	Количество умерших в течение 30 дней после операции	Процентная доля выживших	Процентная доля умерших
Лондон, Харли-стрит	418	413	5	98,8	1,2
Лестер	607	593	14	97,7	2,3
Ньюкасл	668	653	15	97,8	2,2
Глазго	760	733	27	96,3	3,7
Саутгемптон	829	815	14	98,3	1,7
Бристоль	835	821	14	98,3	1,7
Дублин	983	960	23	97,7	2,3
Лидс	1038	1016	22	97,9	2,1
Лондон, Бромптон	1094	1075	19	98,3	1,7
Ливерпуль	1132	1112	20	98,2	1,8
Лондон, Эвелина	1220	1185	35	97,1	2,9
Бирмингем	1457	1421	36	97,5	2,5
Лондон, Грейт-Ормонд-стрит	1892	1873	19	99,0	1,0
Всего	12 933	12 670	263	98,0	2,0

Таблицу можно считать видом графического представления данных, где для привлекательности и удобочитаемости требуется правильно подобрать цвет, шрифт и слова. На эмоциональную реакцию аудитории может также влиять выбор столбцов для отображения. В табл. 1.1 показаны данные об умерших и выживших, однако в США сведения о результатах операций представлены в виде показателя *смертности*, а в Великобритании – в виде показателя *выживаемости.* Такая форма подачи называется эффектом **фрейминга**, и он интуитивно понятен и хорошо документирован: например, «смертность – 5 %» звучит и воспринимается хуже, чем «выживаемость – 95 %». Указание фактического количества смертей и их процентной доли также может создать впечатление о повышении риска, поскольку эту величину можно представить как группу реальных людей.

Классическим примером того, как фрейминг меняет эмоциональное восприятие какого-нибудь показателя, стали плакаты, появившиеся в 2011 году в лондонском метро, которые гласили, что «99 % молодых лондонцев не совершают серьезных насильственных преступлений». Предполагалось, что такие заявления будут способствовать спокойствию пассажиров. Однако мы могли бы изменить их эмоциональное воздействие с помощью двух простых вещей. Во-первых, с помощью заявления, что 1 % молодых лондонцев совершают серьезные насильственные преступления. Во-вторых, учитывая, что население Лондона составляет около 9 миллионов человек, возраст примерно 1 миллиона из них – от 15 до 25 лет, и если считать эту категорию молодежью, то получается, что в городе проживает 1 % от миллиона, или 10 тысяч агрессивно настроенных молодых людей. А такая цифра звучит удручающе и уж вовсе не ободряет. Обратите внимание на две хитрости, используемые для манипулирования воздействием таких статистических данных: переход от позитива к негативу и превращение процентной доли в фактическое количество людей.

В идеале – если мы хотим беспристрастной подачи информации – нужно давать как положительные, так и отрицательные значения, хотя даже порядок столбцов в таблице может влиять на интерпретацию. Необходимо тщательно продумывать и порядок строк. Например, в табл. 1.1 больницы распределены в порядке увеличения количества проведенных операций, но если их упорядочить, например, в порядке убывания смертности (с наибольшим значением в верхней части таблицы), то это может создать впечатление, что перед нами правильный и важный способ сравнения больниц. Такие рейтинговые таблицы любят средства массовой информации и некоторые политики, однако они могут вводить в заблуждение, причем не только потому, что различия бывают вызваны случайными отклонениями, но и потому, что больницы принимают пациентов с заболеваниями разной степени тяжести. Например, по данным табл. 1.1 можно заподозрить, что больница в Бирмингеме – одна из крупнейших и наиболее известных детских больниц – берет наиболее тяжелые случаи. Поэтому было бы несправедливо говорить, что у нее не самые впечатляющие показатели выживаемости[27].

Показатели выживаемости можно представить и в виде горизонтальной столбчатой диаграммы, как на рис. 1.1. Главное – решить, где начинать горизонтальную ось: если с 0 %, то полосы займут практически всю ширину диаграммы, что покажет необычайно высокий уровень выживаемости во всех больницах, но полосы между собой будет трудно различить. Гораздо хуже старый трюк, использующийся для обмана, – начать, например, с 95 %. Тогда все больницы будут резко отличаться, даже если на самом деле разница в показателях объясняется чистой случайностью.

[27] Оказывается, нет никаких веских доказательств каких-либо принципиальных различий между этими больницами, если учитывать степень серьезности случаев.

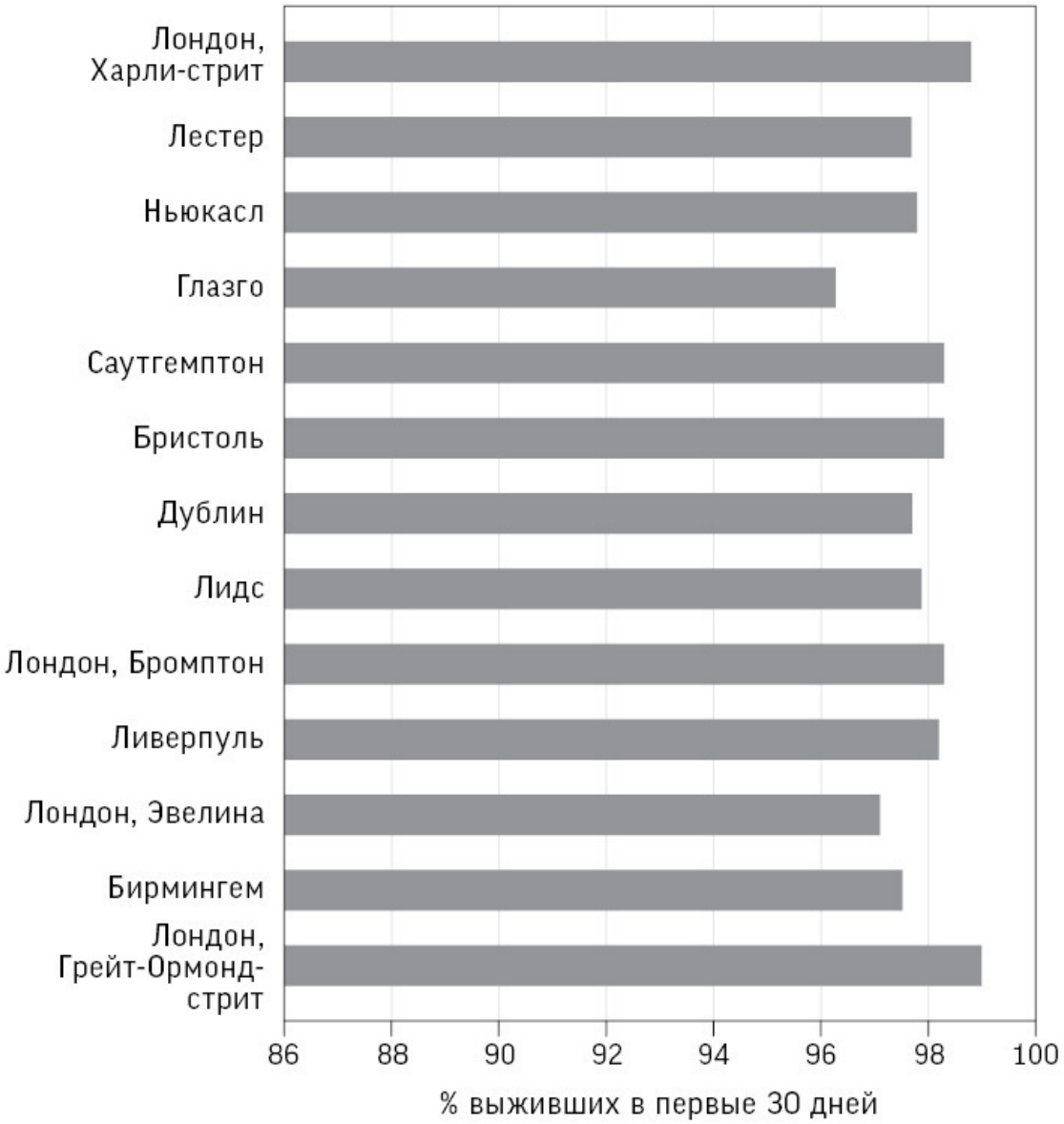


Рис. 1.1

Горизонтальная гистограмма уровня выживаемости за 30 дней в тринадцати больницах. Выбор начала горизонтальной оси (в данном случае 86 %) может существенно сказаться на впечатлении, вызываемом графиком. Если ось начинается с 0 %, все больницы выглядят неразличимыми; если же начать с 95 %, разница будет обманчиво драматичной

Следовательно, выбор начала оси представляет собой дилемму. Альберто Каиро, автор авторитетных книг по визуализации данных[28], предлагает всегда начинать с «логической и взвешенной точки отсчета», которую в нашем случае трудно определить. Мой собственный произвольный выбор – 86 %, что примерно отражает недопустимо низкий уровень выживаемости в Бристольской больнице двадцатью годами ранее.

Я начал книгу цитатой Нейта Сильвера, основателя цифровой платформы FiveThirtyEight и автора точного прогноза президентских выборов 2008 года в США. Он красноречиво высказал идею, что цифры не говорят сами за себя – это мы наполняем их смыслом. А значит, ком-

[28] См. A. Cairo, The Truthful Art: Data, Charts, and Maps for Communication (New Riders, 2016), и The Functional Art: An Introduction to Information Graphics and Visualization (New Riders, 2012).

муникации – ключевая часть цикла решения проблем, и в этом разделе я показал, как способ представления данных может влиять на наше восприятие.

Теперь нам нужно ввести важное и удобное понятие, которое поможет выйти за рамки простых вопросов типа «да/нет».

Качественные переменные

Переменной называется любая величина, которая может принимать различные значения в разных обстоятельствах; это очень полезный сокращенный термин для всех видов наблюдений, содержащих данные. Бинарные переменные могут принимать только два значения (да/нет) – например, жив человек или мертв, женщина он или мужчина. Значения могут отличаться у разных людей и даже у одного человека в разные моменты жизни. **Качественная (или категорийная) переменная** – это переменная, которая может принимать одно, два или более значений, попадающих в ту или иную категорию. При этом категории могут быть:

• *неупорядоченными:* страна рождения человека, цвет автомобиля или больница, где делали операцию;

• *упорядоченными*: воинские звания;

• *сгруппированными числами:* степени ожирения, которые часто определяются в терминах пороговых значений по индексу массы тела (ИМТ)[29].

Для отображения качественных данных часто используются круговые диаграммы, что позволяет составить представление о размере каждой категории по занимаемой ею части круга. Однако здесь вероятны проблемы с наглядностью, например при попытке изобразить на одной диаграмме слишком много категорий или использовать трехмерное представление, искажающее площади. Рис. 1.2 показывает весьма уродливый пример, смоделированный с помощью Microsoft Excel, где представлены данные из табл. 1.1 о результатах операций на сердце для 12 933 детей.

[29] Индекс массы тела разработан бельгийским статистиком и социологом Адольфом Кетле в 1830-х годах. Он определяется так: ИМТ = масса (кг) / рост^2 (м). Используются самые разные способы группирования людей по этому параметру; в настоящее время в Великобритании применяются такие категории: недостаточная масса (ИМТ < 18,5), нормальная масса (ИМТ от 18,5 до 25), избыточная масса (от 25 до 30), ожирение (от 30 до 35), болезненное ожирение (свыше 35).Сам термин «индекс массы тела» появился намного позднее, в статье Анселя Киза с соавторами, опубликованной в 1972 году в Journal of Chronic Diseases. *Прим. пер.*

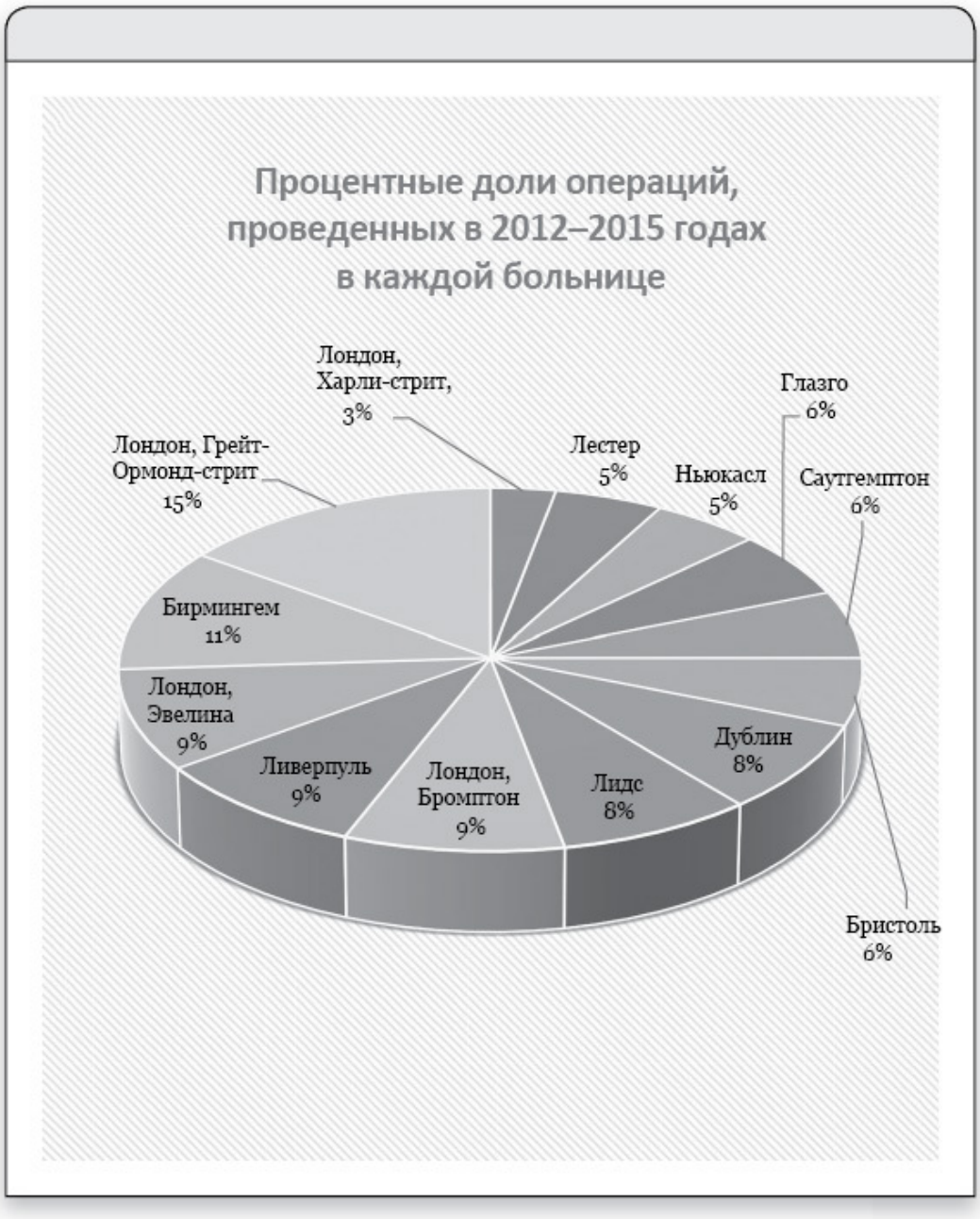


Рис. 1.2

Процентные доли операций на сердце у детей в каждой больнице, отображенные на круговой 3D-диаграмме из Excel. Это крайне неудачное представление данных зрительно увеличивает категории на переднем плане, делая невозможным визуальное сравнение между больницами

Использование сразу нескольких круговых диаграмм, как правило, не очень хорошая идея, поскольку это затрудняет сравнение относительных размеров областей разной формы. Сравнения лучше проводить с помощью гистограмм (столбчатых диаграмм) – при этом хорошо видна разница в высоте или длине. Рис. 1.3 – более простой и понятный пример горизонтальной гистограммы, где длина горизонтальной полосы отражает долю операций каждой больницы.

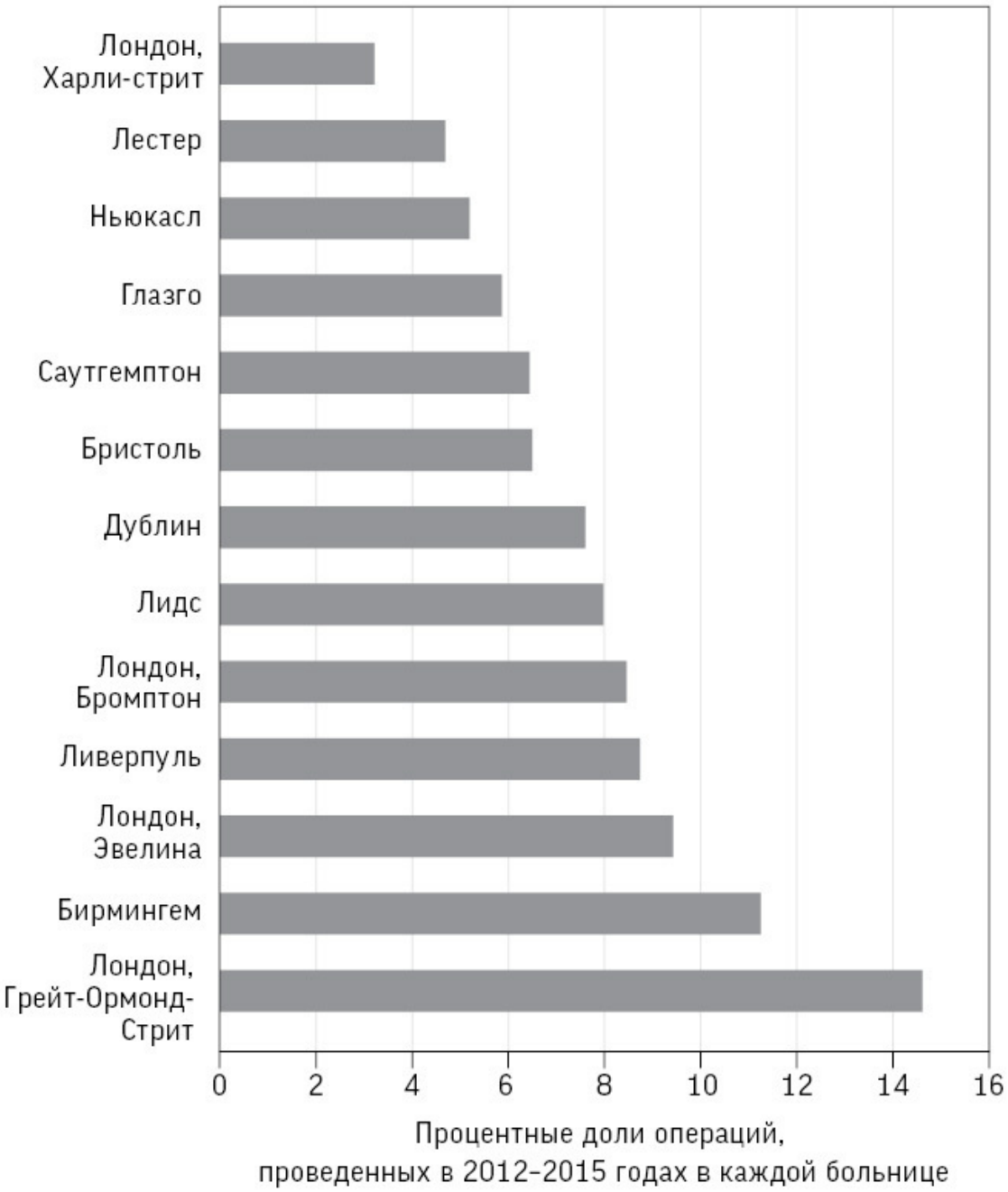


Рис. 1.3

Процентные доли всех операций на сердце у детей, проведенных в каждой больнице: более четкое представление с помощью горизонтальной гистограммы

Сравнение двух долей

Итак, увидев, как с помощью гистограммы можно элегантно сравнить несколько пропорциональных долей, было бы логично полагать, что сравнение двух долей вообще тривиальное дело. Однако когда эти доли представляют собой оценку рисков причинения какого-либо вреда, метод их сравнения становится серьезным, дискуссионным вопросом. Типичный пример:

Каков риск развития рака от употребления сэндвичей с беконом?

Каждому из нас знакомы громкие заголовки в СМИ, предупреждающие о том, что какая-то вполне обыденная вещь увеличивает риск возникновения чего-нибудь плохого. Я обычно называю такие истории «кошки вызывают рак». Например, в ноябре 2015 года Международное агентство по изучению рака (МАИР) Всемирной организации здравоохранения объявило обработанное мясо «канцерогеном группы I», то есть отнесло его к той же категории, что сигареты и асбест. Естественно, это привело к появлению устрашающих заголовков. Так, Daily Record написала, что «по мнению экспертов, бекон, ветчина и сосиски подвергают такому же риску развития рака, как и сигареты»[30].

МАИР попыталось подавить панику, подчеркнув, что попадание в группу I всего лишь говорит о существовании повышенного риска рака, а не о реальной величине самого риска. В пресс-релизе МАИР сообщалось, что ежедневное употребление 50 граммов обработанного мяса связано с повышением риска развития рака кишечника на 18 %. Звучит тревожно, но так ли это на самом деле?

Величина 18 % известна как **относительный риск**, который отражает разницу в опасности развития рака кишечника (колоректального рака) у двух групп людей: ежедневно употребляющих 50 граммов обработанного мяса (например, сэндвич с двумя ломтиками бекона) и тех, кто его не ест. Статистики наложили этот относительный показатель на каждую отдельную группу риска и посмотрели, какие абсолютные значения он принимает в каждом случае, что позволило выявить **абсолютный риск** этого исхода для каждой группы. Они пришли к выводу, что при нормальном ходе вещей примерно 6 из каждых 100 человек, которые не едят бекон ежедневно, заболеют раком кишечника. Если же 100 таких человек ели бы бекон ежедневно всю жизнь, то, согласно отчету МАИР, можно было бы ожидать, что больных будет на 18 % больше, то есть не 6, а 7 человек из 100[31]. Один дополнительный случай рака кишечника на 100 человек, ежедневно употреблявших бекон в течение жизни, звучит вовсе не так впечатляюще, как относительный риск (увеличение на 18 %), и позволяет оценивать риски более объективно. Нужно отличать то, что действительно опасно, от того, что только выглядит пугающе[32].

Пример с сэндвичем показывает, что риски полезно выражать в **ожидаемых частотах**, то есть вместо того, чтобы обсуждать доли или вероятности, просто спросить: «А что это означает для группы в 100 (или 1000) человек?» Психологические исследования продемонстрировали, что такой метод улучшает понимание: утверждение, что потребление мяса приводит к «18-процентному повышению риска», можно считать манипулятивным, поскольку мы знаем, что такая форма подачи информации создает преувеличенное впечатление о степени опасности[33]. На рис. 1.4 представлена ожидаемая частота случаев рака кишечника в группе из 100 человек в виде **пиктографической диаграммы.**

---

[30] Информацию Всемирной организации здравоохранения о канцерогенности потребления красного мяса и обработанного мяса см. http://www.who.int/features/qa/cancer-red-meat/en/. ‘Bacon, Ham and Sausages Have the Same Cancer Risk as Cigarettes Warn Experts’, Daily Record, 23 October 2015.

[31] Строго говоря, относительное увеличение на 18 % дает 6 × 1,18 = 7,08 процента, но для наших целей округления до 7 % вполне достаточно.

[32] Это было любимое наблюдение Ханса Рослинга, см. следующую главу.

[33] E. A. Akl et al., ‘Using Alternative Statistical Formats for Presenting Risks and Risk Reductions’, Cochrane Database of Systematic Reviews 3 (2011).

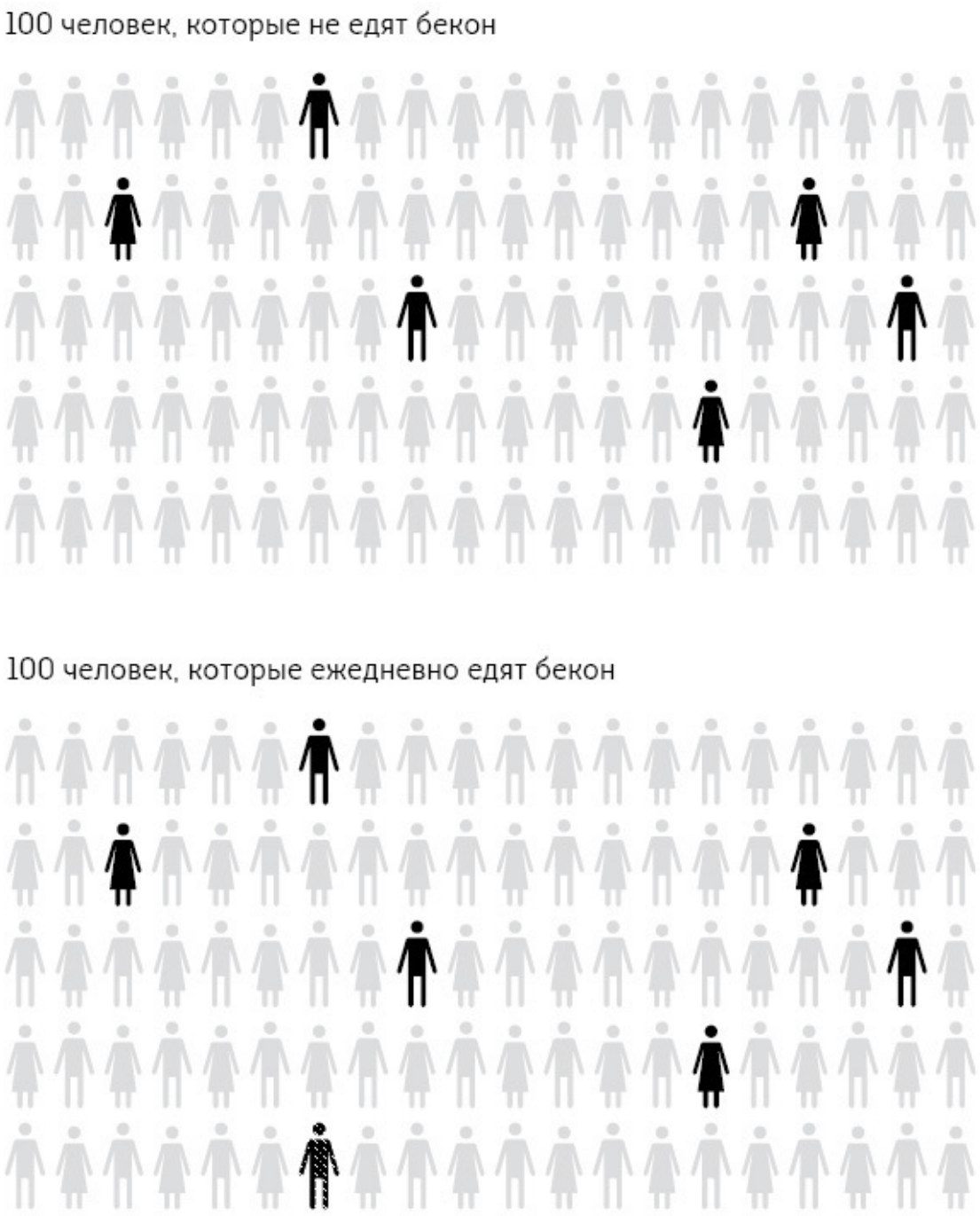


Рис. 1.4

Пример с сэндвичем в виде двух пиктографических диаграмм, где люди с раком кишечника случайно рассеяны в общей группе. При нормальных обстоятельствах в группе из 100 человек, не употребляющих бекон, рак кишечника развивается у 6 человек (выделены темным на первой диаграмме). В группе из 100 человек, которые ежедневно едят бекон (вторая диаграмма), выявляется один дополнительный случай заболевания (заштрихованная пиктограмма)[34]

На рис. 1.4 «раковые» пиктограммы случайным образом разбросаны среди 100 изображений. Хотя было продемонстрировано, что такое рассеяние усиливает впечатление непред-

[34] Строго говоря, шесть темных фигурок в обеих частях рисунка следовало бы разместить по-разному, поскольку диаграммы представляют разные группы из 100 человек. Но это затруднило бы их сравнение.

сказуемости, его следует использовать только в случае одной дополнительной выделенной пиктограммы, тогда для быстрого визуального сравнения не нужно будет их считать.

Еще несколько способов сравнить две доли представлены в табл. 1.2, отражающей те же риски для людей, которые едят и не едят бекон.

Таблица 1.2

Примеры способов информирования о риске развития рака кишечника при ежедневном употреблении сэндвича с беконом и без него. «Число больных, которых нужно лечить», – это число людей, которые должны всю жизнь ежедневно съедать сэндвич с беконом, чтобы можно было ожидать один дополнительный случай рака кишечника (поэтому, пожалуй, этот параметр лучше назвать «числом людей, которые должны есть»)

Метод	Не употреблявшие бекон	Ежедневно употреблявшие бекон
Частота события	6%	7%
Ожидаемая встречаемость	6 из 100	7 из 100
	1 из 16	1 из 14
Шансы	6/94 (6 к 94)	7/93 (7 к 93)

Показатели сравнения

Разница в абсолютных рисках	1%, или 1 из 100
Относительный риск	1,18, или увеличение на 18%
«Число больных, которых нужно лечить»*	100
Отношение шансов	(7/93) / (6/94) ≈ 1,18

* Число больных, которых нужно лечить (ЧБНЛ), – один из важных параметров в здравоохранении. В обычном смысле это среднее число пациентов, которых необходимо лечить, чтобы предотвратить один неблагоприятный исход или добиться какого-то благоприятного исхода, по сравнению с контрольной группой. Автор использует понятие в более широком смысле. *Прим. пер.*

Обычно риск выражают фразой «1 из *x*», то есть «1 из 16 человек» означает 6-процентный риск. Однако использовать несколько выражений «1 из…» не рекомендуется, потому что многим людям трудно их сравнивать. Например, на вопрос «Какой риск больше – 1 из 100, 1 из 10 или 1 из 1000?» около четверти людей ответили неверно: проблема в том, что большее число здесь связывается с меньшим риском, поэтому для правильного ответа требуется некоторая сообразительность.

Под **шансами** на событие понимается отношение вероятности его наступления к вероятности того, что оно не произойдет. Например, из 100 человек, не употребляющих бекон, у 6

будет выявлен колоректальный рак, а у 94 – нет, а значит, шансы заболеть раком у людей в этой группе составляют 6/94, что читается как «6 к 94»[35]. Шансы обычно используют в различных ставках, но они также широко применяются в статистическом моделировании долей, а это означает, что медицинские исследования обычно выражают эффекты, связанные с лечением или поведением, именно в **отношении шансов**.

Несмотря на то что отношение шансов часто встречается в исследовательской литературе, это не всегда подходящий способ показать разницу в рисках. Если события происходят достаточно редко, то такие отношения будут численно близки к относительным рискам, как в случае сэндвичей с беконом, но для распространенных событий отношения шансов могут сильно отличаться от относительных рисков, и следующий пример показывает, как это может запутать журналистов (и остальных людей).

Как можно рост с 85 до 87 % назвать 20-процентным повышением?

Статины широко используются для снижения уровня холестерина и риска инфарктов и инсультов, однако некоторых врачей беспокоят побочные эффекты их применения. Исследование, опубликованное в 2013 году, установило, что 87 % людей, принимавших статины, сообщали о мышечных болях – по сравнению с 85 % тех, кто их не принимал. Если посмотреть на способы сравнения рисков, представленные в табл. 1.2, то можно сказать либо об увеличении абсолютного риска на 2 %, либо о примерно таком же увеличении относительного риска: 0,87 / 0,85 ≈ 1,02. Шансы для обеих групп равны, соответственно 0,87 / 0,13 = 6,7 и 0,85 / 0,15 = 5,7, а значит, их отношение составляет 6,7 / 5,7 = 1,18. Получилось такое же значение, как и у сэндвичей с беконом, хотя при совершенно других абсолютных рисках.

Газета Daily Mail неправильно интерпретировала это отношение шансов 1,18 как относительный риск и напечатала статью под заголовком: «Статины повышают риск на 20 %», что является серьезным искажением результатов исследования. Однако винить надо не только журналистов: в кратком содержании статьи было указано лишь отношение шансов – без упоминания о том, что оно соответствует разнице между абсолютными рисками в 87 и 85 %[36].

Это подчеркивает опасность применения отношения шансов в любом контексте, кроме научного. Всегда лучше сообщать аудитории о понятных ей абсолютных рисках вне зависимости от того, касаются они бекона, статинов или чего-то другого.

Примеры в этой главе продемонстрировали, как кажущаяся простой задача по вычислению и выражению величины долей может превратиться в довольно сложную, и здесь нужно проявлять осторожность. Психологи все активнее изучают воздействие различных форматов числовых и графических данных на наше восприятие. Коммуникации – важная часть цикла решения проблем, и она не должна зависеть от личных предпочтений.

Выводы

• Бинарные переменные принимают только два значения: да и нет. Информацию о нескольких таких переменных можно выражать в виде доли случаев, которую составляет какая-то из них.

---

[35] Подчеркиваем, что в данном случае вовсе не подразумевается, что вероятность рака равна 6/94. Объясним это на простом примере. Когда говорят о «шансах 1 к 2», то вероятность не равна 1/2. Это означает, что в вашу пользу один возможный исход, а против вас – два исхода. Следовательно, «шансы 1 к 2» означают один удачный исход из трех возможных, то есть вероятность успеха равна 1/3. Аналогично, в нашем случае вероятность рака равна 6/100, а число 6/94 – это отношение вероятности рака к вероятности его отсутствия: (6/100) / (94/100) = 6/94. *Прим. пер.*

[36] 'Statins Can Weaken Muscles and Joints: Cholesterol Drug Raises Risk of Problems by up to 20 per cent', Mail Online, 3 June 2013. Исходная работа: I. Mansi et al., 'Statins and Musculoskeletal Conditions, Arthropathies, and Injuries', JAMA Internal Medicine 173 (2013), 1318–26.

• Положительный или отрицательный фрейминг может повлиять на эмоциональное восприятие данных.

• Относительные риски склонны преувеличивать важность, поэтому для полноты картины следует предоставлять информацию об абсолютных рисках.

• Ожидаемая частота обеспечивает понимание и правильное представление о важности.

• Отношения шансов можно оценивать в научных работах, но их не стоит использовать в обычных публикациях.

• Визуальное представление информации должно быть тщательно продумано с учетом особенностей его восприятия.

Глава 2. Числовые характеристики выборки и представление данных

Можно ли доверять мудрости толпы?

В 1907 году Фрэнсис Гальтон (двоюродный брат Чарльза Дарвина, эрудит, создатель метода идентификации отпечатков пальцев, метеоролог и автор термина «евгеника» [37]) написал письмо в престижный научный журнал Nature о своем посещении выставки животноводства и птицеводства в Плимуте. Там он увидел необычный конкурс: участникам, заплатившим по 6 пенсов, предлагалось угадать вес выставленного напоказ большого откормленного быка, после того как его забьют и освежуют. По окончании конкурса ученый взял 787 заполненных билетов и выбрал из них в качестве среднего значения 1207 фунтов (547 килограммов). «Любая иная оценка рассматривалась большинством голосовавших как слишком высокая или слишком низкая», – пояснил он. Реальный вес животного составил 1198 фунтов (543 килограмма), что оказалось на удивление близко к выбранному числу[38]. Гальтон назвал свое письмо Vox Populi («Глас народа»), хотя сегодня такой процесс принятия решений более известен как **мудрость толпы**.

Гальтон выполнил то, что сегодня мы назвали бы сводкой данных: он взял множество чисел на билетах и свел их к одному весу в 1207 фунтов. В этой главе мы рассмотрим методы, разработанные в последующем столетии для получения сводной информации из имеющейся массы данных. Мы увидим, что числовые характеристики выборки (показатели положения, распространения, разброса, тренды и корреляция) тесно связаны со способом их представления на бумаге или экране. Мы также поговорим о переходе от простого описания данных к сторителлингу с помощью инфографики.

Начнем с моей собственной попытки экспериментировать с мудростью толпы, которая выявляет многие из проблем, возникающих, когда в качестве источника данных используется реальный мир, со всей его склонностью к странностям и ошибкам.

Статистика касается не только таких серьезных вещей, как рак и хирургия. В рамках нашего с популяризатором математики Джеймсом Граймом довольно простого эксперимента мы выложили на YouTube видео и попросили угадать число драже в банке. Вы тоже можете попробовать это сделать, посмотрев на фотографию на рис. 2.1 (истинное число станет известно позже). Свои предположения высказали 915 человек, их ответы варьировались от 219 до 31 337. В этой главе мы увидим, как такие переменные можно изображать графически и обрабатывать численно.

---

[37] Евгеника (др.-греч. εὐγενής – хорошего рода) – это учение о том, что человеческую расу можно улучшать путем селекции либо путем поощрения деторождения у «подходящих» людей (например, с помощью финансовых стимулов), либо препятствуя размножению «неподходящих» (скажем, за счет принудительной стерилизации). Многие из первых создателей статистических методов были увлеченными евгениками. Однако опыт нацистской Германии положил конец этой концепции, хотя академический журнал Annals of Eugenics поменял свое название на Annals of Genetics только в 1955 году.

[38] F. Galton, 'Vox Populi', Nature (1907); доступно по адресу: https://www.nature.com/articles/075450a0.

Рис. 2.1
Сколько драже в банке? Мы спросили об этом в ролике на YouTube и получили 915 ответов. Ответ будет дан позже

Начнем с того, что на рис. 2.2 отображены три способа представления чисел, указанных 915 участниками. Их можно назвать по-разному: распределение данных, **выборочное распределение** или эмпирическое распределение[39].

[39] Слово «распределение» широко используется в статистике, но может иметь разные смыслы, поэтому я постараюсь объяснить, что оно означает в каждой ситуации. Диаграммы построены с помощью программного обеспечения для языка R.

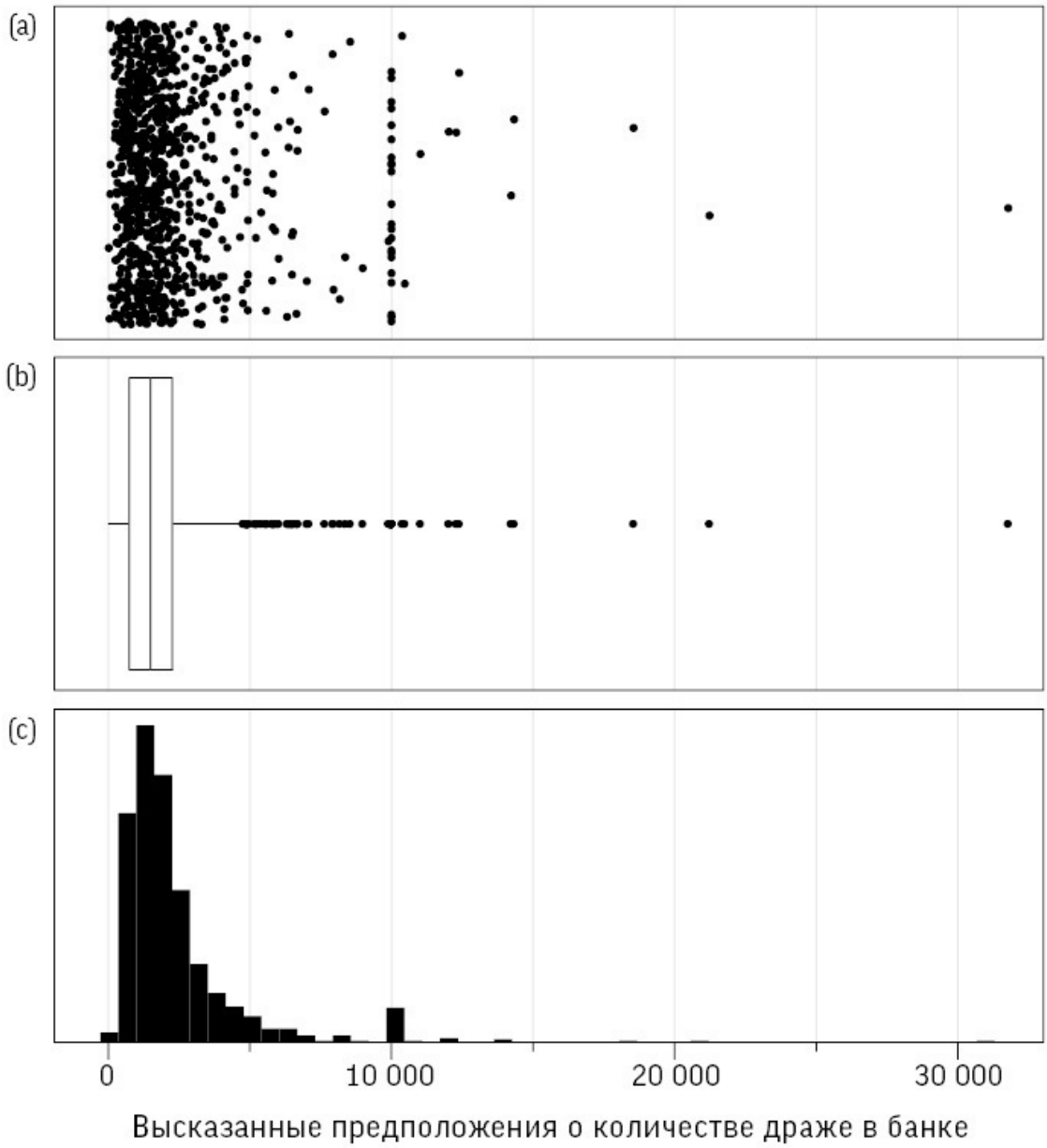


Рис. 2.2

Различные способы отображения 915 предположений о количестве драже в банке: (a) точечная диаграмма с разбросом, чтобы точки не перекрывали друг друга; (b) диаграмма размаха, или «ящик с усами»; (c) гистограмма

(a) Точечная диаграмма просто показывает все значения в виде отдельных точек, но для каждой добавлено случайное отклонение по вертикали, чтобы точки не перекрывали друг друга, поскольку некоторые догадки были высказаны по несколько раз. Четко видна концентрация большого количества значений в диапазоне примерно до 3000, а затем длинный «хвост» тянется более чем за 30 000, причем в точке 10 000 наблюдается всплеск.

(b) Диаграмма размаха («ящик с усами») показывает некоторые базовые характеристики распределения[40].

[40] На диаграмме размаха центральная вертикальная линия в прямоугольнике представляет собой медиану (серединное значение), сам ящик-прямоугольник включает основную часть точек, расположенную близко к медиане [обычно в ящик включают половину наблюдений, то есть границами ящика являются первый и третий квартили, и, соответственно, ширина ящика отражает интерквартильный размах; *Прим. пер.*], а горизонтальные линии-«усы» показывают наименьшее и наибольшее зна-

(c) На гистограмме просто учитывается, сколько точек данных попало в тот или иной интервал. Она дает очень приблизительное представление о форме распределения.

Эти способы отображения сразу же позволяют выделить некоторые особенности распределения. Видно, что оно сильно скошено, то есть **асимметрично (**отсутствует даже приблизительная симметрия относительно какой-нибудь центральной точки) и из-за наличия нескольких очень больших чисел имеет длинный «правый хвост». Вертикальные ряды точек на точечной диаграмме (изображающие повторяющиеся числа) также указывают на некоторое предпочтение круглых чисел.

Однако у всех диаграмм есть общая проблема. Внимание сосредоточено на самых больших значениях, причем основная часть чисел сконцентрирована в левой части. Можно ли представить эти данные более информативно? Мы могли бы отбросить самые большие числа как нелепые (когда я первоначально анализировал полученные величины, я сознательно исключил все, превышающие 9000). Кроме того, мы можем уменьшить влияние экстремальных наблюдений, скажем, отобразив данные в **логарифмическом масштабе**, когда интервал от 100 до 1000 имеет такую же длину, что и интервал от 1000 до 10 000[41].

На рис. 2.3 представлена более понятная структура с вполне симметричным распределением и отсутствием значительных выбросов. Это избавляет нас от исключения каких-либо значений наблюдений, что обычно не считается хорошей идеей (если, конечно, речь не идет о явных ошибках).

чение, либо доходят только до краев статистически значимой выборки, а выбросы изображаются отдельно.

[41] Десятичный логарифм числа x – это такое число y, что $10^y = x$. Например, десятичный логарифм 1000 равен 3, потому что $10^3 = 1000$. Логарифмические преобразования особенно уместны, когда есть основания полагать, что люди совершают скорее относительные, а не абсолютные ошибки. Скажем, если мы ожидаем, что люди получают неверный ответ, ошибаясь на 20 % в ту или иную сторону, а не на 200 драже в банке.

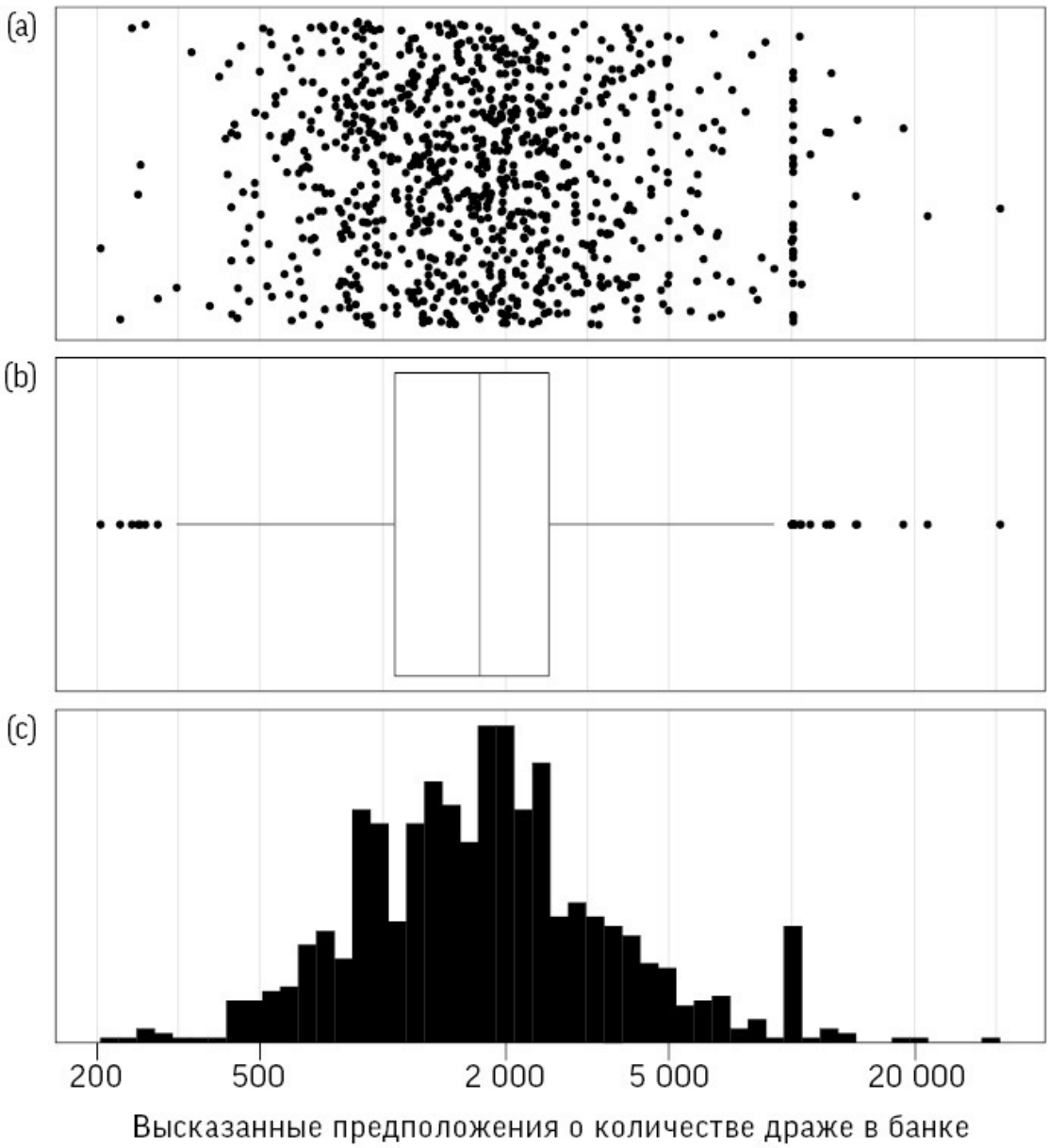


Рис. 2.3

Графическое отображение догадок о числе драже в банке в логарифмическом масштабе: (a) точечная диаграмма; (b) «ящик с усами»; (c) гистограмма – на всех заметна достаточная степень симметрии

Единственно правильного способа отображения чисел нет, у каждого из способов свои преимущества: на точечной диаграмме показаны все отдельные точки, «ящик с усами» дает визуальное представление, а гистограмма помогает полнее понять вид исходного распределения.

Переменные, которые записываются в виде чисел, могут быть разного типа:

• **Счетные переменные:** могут принимать целочисленные значения 0, 1, 2, 3… Например, ежегодное число самоубийств или предположения о количестве драже в банке.

• **Непрерывные переменные:** могут принимать любые значения. Например, некоторые вещи теоретически можно измерять с любой точностью и получать любые числа. Скажем, вес и рост, которые отличаются как у разных людей, так и у одного человека в зависимости от

времени. Разумеется, эти значения можно округлить до целого числа сантиметров или килограммов[42].

Когда набор наблюдений (выборка) сводится к одному числу, мы, как правило, называем его **средним значением.** Все знакомы с понятием средней зарплаты, средней оценки на экзамене или средней температуры, но часто не знают, как интерпретировать эти величины (особенно если человек, который о них говорит, сам не понимает, о чем речь).

Чаще всего встречаются три толкования термина «среднее значение»:

1. **Среднее арифметическое** (или **выборочное среднее):** сумма всех величин, деленная на их количество.
2. **Медиана:** среднее по величине число ранжированного ряда (то есть слева и справа от него будет поровну чисел)[43]. Именно так Гальтон считал голоса толпы[44].
3. **Мода:** чаще всего встречающееся значение в выборке.

Эти параметры также называются показателями положения центра распределения.

Интерпретация термина «среднее» как «среднее арифметическое» дает повод для старых шуток о том, что почти у всех людей число ног превышает среднее (которое, по оценкам, примерно равно 1,99999) и что у человека в среднем одно яичко. Однако среднее арифметическое может не подходить не только при измерении ног и яичек. Вычисленное таким образом среднее число сексуальных партнеров или средний доход по стране может иметь крайне мало общего с представлением большинства людей из-за сильного влияния больших значений в выборке, которые тянут среднее арифметическое вверх[45]: подумайте об Уоррене Битти или Билле Гейтсе (в отношении числа сексуальных партнеров и дохода соответственно).

Средние значения способны сильно вводить в заблуждение, когда исходные данные имеют не симметричное распределение, а сильно перекошенное в какую-либо сторону (как при догадках о количестве драже). Как правило, так происходит при наличии большой группы стандартных случаев и хвоста из нескольких высоких (скажем, величина дохода) или низких (число ног) значений. Я могу практически гарантированно утверждать, что вы гораздо меньше рискуете умереть в следующем году по сравнению с людьми вашего возраста и пола (если средний риск вычислять как среднее арифметическое). Например, согласно таблицам смертности для Соединенного Королевства, 1 % 63-летних мужчин не доживают до 64-летия. Однако многие из тех, кто умрет, уже серьезно больны, а потому риск для подавляющего большинства (тех, кто относительно здоров) меньше, чем средний.

К сожалению, когда в СМИ пишут о *среднем*, часто непонятно, следует это толковать как среднее арифметическое или как медиану. Например, Национальная статистическая служба Великобритании вычисляет средний недельный заработок (который рассчитывается как среднее арифметическое), а также публикует *медианные* заработки, предоставляемые местными органами. Это позволяет отличить «средний доход» (среднее арифметическое) от «дохода среднего человека» (медиана). Цены на дома имеют крайне асимметричное распределение с

---

[42] Вообще говоря, непрерывным переменным противопоставляются дискретные, которые необязательно принимают неотрицательные целые значения, а могут принимать значения в произвольном конечном или счетном множестве. *Прим. пер.*

[43] Это определение удобно для нечетного количества элементов в выборке. Если число элементов четное, то обычно медианой считают полусумму двух средних элементов ряда. *Прим. пер.*

[44] Хотя в 1907 году в Nature оспаривали выбор Гальтоном медианы, считая, что среднее арифметическое дало бы лучшую оценку.

[45] Представьте, что в комнате сидят три человека, которые зарабатывают 400, 500 и 600 фунтов в неделю. В таком случае выборочное среднее для их зарплат составляет 1500 / 3 = 500 фунтов. Медианное значение тоже 500 фунтов. Затем в комнату заходят два человека, зарабатывающие по 5000 фунтов, и выборочное среднее взлетает до 11 500 / 5 = 2300 фунтов, в то время как медиана поднялась только до 600.

длинным правым хвостом элитной недвижимости, поэтому официальные индексы для цен на жилье указываются в виде медианных значений. Однако обычно пишут о «цене в среднем», что является весьма неоднозначным термином. Это «цена среднего дома» (то есть медиана)? Или «средняя цена дома» (то есть среднее арифметическое)? Как видите, перестановка слов имеет большое значение.

А теперь пришло время обнародовать результаты нашего эксперимента с мудростью толпы; может, он не такой захватывающий, как определение веса быка, зато с чуть б*о*льшим количеством голосов, чем у Гальтона.

Из-за наличия длинного правого хвоста среднее арифметическое 2408 было бы плохой оценкой, а мода (чаще других названное значение) 10 000, похоже, отражает склонность людей выбирать круглые числа. Поэтому предпочтительнее последовать примеру Гальтона и использовать в качестве общей оценки медиану. Она равна 1775, хотя на самом деле в банке находилось 1616 драже[46]. Правильно это число угадал только один человек, 45 % дали оценки ниже этого значения, а 55 % – выше. Поэтому наблюдается небольшая асимметрия, и мы говорим, что истинное значение находится на 45-м процентиле[47]. Медиана, которая является 50-м процентилем, дала избыточную оценку: 1775–1616 = 159 и оказалась примерно на 10 % больше правильного ответа. Только каждый десятый человек указывал оценку лучше, чем полученное медианное значение. Таким образом, мудрость толпы оказалась вполне на уровне, а именно гораздо ближе к истине, чем 90 % отдельных людей.

Разброс распределения данных

Свести распределение к единственному числу недостаточно – нужно иметь представление о разбросе данных (рассеивании, отклонении от среднего). Например, знание среднего размера обуви взрослого мужчины никак не поможет обувной фабрике определить, сколько пар обуви каждого размера производить. Один размер не годится для всех, что прекрасно иллюстрируют пассажирские кресла в самолетах.

В табл. 2.1 приведены статистические данные для выборки по драже. Она предлагает три способа демонстрации разброса. Естественный вариант – **размах**[48], однако он крайне чувствителен к экстремальным значениям, таким как весьма странное предположение о наличии в банке 31 337 драже[49]. Напротив, на **интерквартильный размах** такие выбросы не очень влияют. Интерквартильный размах – это разность между третьим и первым квартилем (то есть 75-м и 25-м процентилем); иными словами, сюда входит «центральная половина» всех чисел, в нашем случае – от 1109 до 2599 драже. Ящик на диаграмме типа «ящик с усами» как раз и

---

[46] В ролике о нашем эксперименте (https://www.youtube.com/watch?v=n98BhnwWmsc) я принудительно убрал 33 максимальных числа (9999 и выше), взял логарифм для получения симметричного распределения, вычислил среднее арифметическое для такого преобразованного распределения, а затем произвел обратное преобразование, чтобы получить оценку в первоначальном масштабе. Это дало число 1680, которое оказалось самой близкой оценкой к истинному значению 1616. Описанный процесс (взять логарифм, вычислить среднее арифметическое, вернуться обратно) дает то, что известно как среднее геометрическое. Это эквивалентно такой процедуре: перемножить все N чисел и извлечь корень N-й степени. Среднее геометрическое используется при создании некоторых экономических индексов, в частности основанных на отношениях. Причина в том, что у него есть «устойчивость к переворачиванию отношения»: если стоимость апельсинов измерять в килограммах на апельсин или в апельсинах на килограмм, то это даст одно и то же геометрическое среднее. В то же время среднее арифметическое может давать большой разброс.

[47] Если не вдаваться в тонкости, то N-й процентиль – значение, которое не превышает N% наблюдений. 25-й процентиль называют первым квартилем, 50-й процентиль – вторым квартилем (или медианой), 75-й процентиль – третьим квартилем. В общем случае, когда доля наблюдений не превосходит числа α, то говорят об α-квантиле. *Прим. пер.*

[48] Размах – это разность между наибольшим и наименьшим значением в выборке. Впрочем, у автора в таблице указываются только границы диапазона – как для размаха, так и для интерквартильного размаха. *Прим. пер.*

[49] Почти наверняка это опечатка при наборе числа 1137, которое является числовым изображением слова leet, что на сетевом сленге означает «элитный» [Leet – это язык интернета, где латинские буквы заменяются похожими символами. *Прим. пер.*]*; среди ответов было девять чисел 1337.*

включает интерквартильный размах. Наконец, в качестве меры разброса широко используется **стандартное (среднеквадратичное) отклонение.** Но поскольку его сложнее вычислять и оно сильно подвержено влиянию выбросов, оно лучше всего подходит для симметричных и хорошо себя ведущих данных[50]. Например, удаление из выборки одного (почти гарантированно ошибочного) числа 31 337 приводит к уменьшению среднеквадратичного отклонения с 2422 до 1398[51].

Таблица 2.1

Характеристики выборки для 915 предположений о количестве драже в банке. Истинное число равно 1616

Характеристики выборки при определении количества драже в банке	Значение
Выборочное среднее	2408
Медиана	1775
Мода	10 000
Размах	от 219 до 31 337
Интерквартильный размах	от 1109 до 2599
Среднеквадратичное отклонение	2422

Толпа в нашем маленьком эксперименте продемонстрировала значительную мудрость, даже несмотря на несколько странных ответов. Это показывает, что, хотя данные часто включают ошибки, выбросы и другие странные величины, их вовсе не обязательно выискивать и исключать. Кроме того, это указывает на полезность использования характеристик выборки, на которые не влияют даже столь эксцентричные наблюдения, как 31 337. Такие характеристики называются робастными (то есть устойчивыми) и включают медиану и интерквартильный размах. Наконец, эксперимент подчеркивает ценность обычного просмотра данных – урок, который будет подкреплен следующим примером.

[50] В качестве меры неравенства для сильно асимметричных распределений (например, доходов) используется коэффициент Джини, однако он сложен и не всегда интуитивно понятен.

[51] Квадрат среднеквадратичного отклонения называется **дисперсия:** его трудно интерпретировать прямо, но с математической точки зрения это очень полезное понятие. [Дисперсия интерпретируется вполне естественно – это средний квадрат отклонения наблюдений от выборочного среднего. *Прим. пер.*].

Разница между группами чисел

Конец ознакомительного фрагмента.