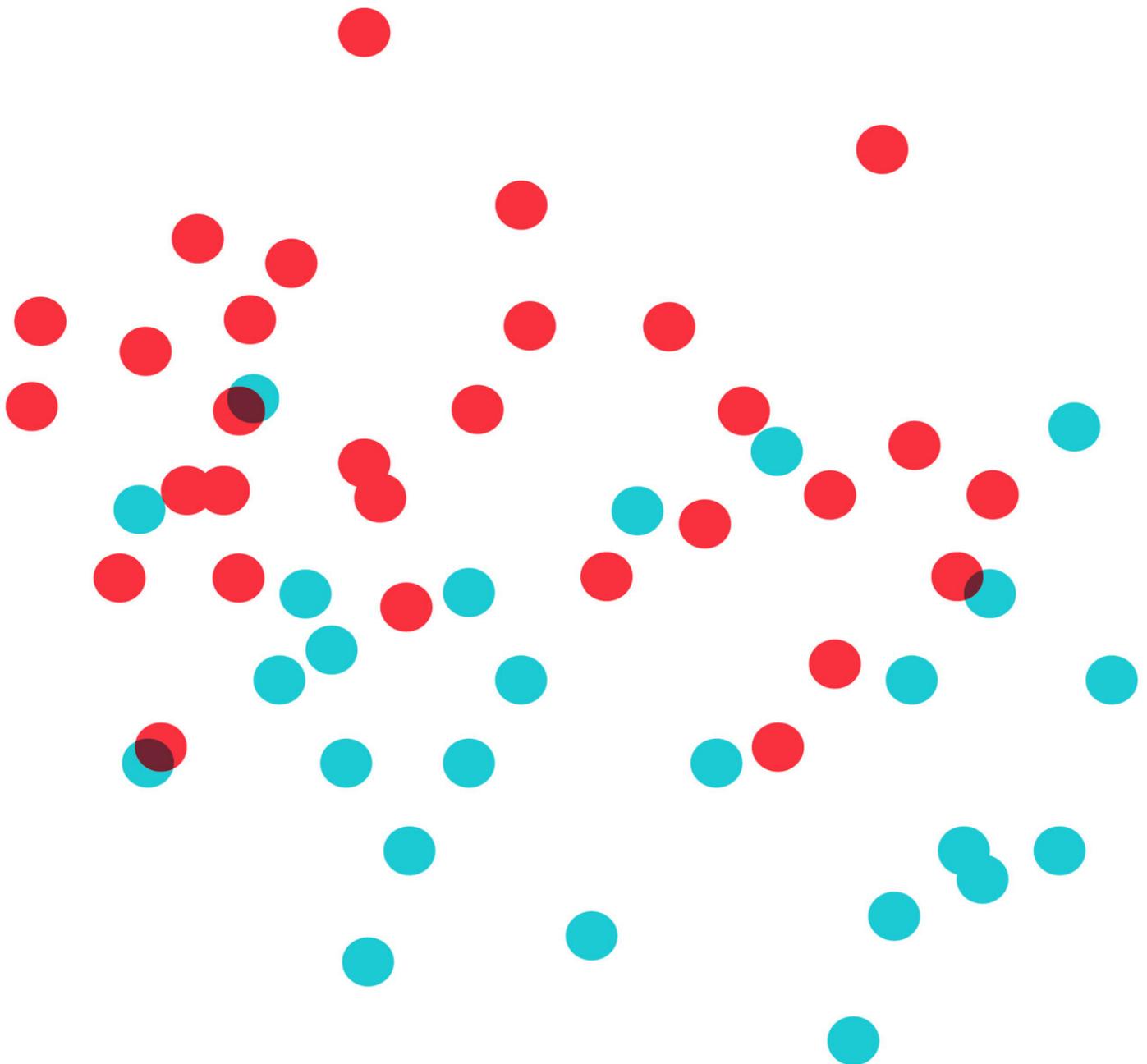


# Искусство статистики

Как находить  
ответы в данных

Дэвид Шпигельхалтер



## Annotation

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.*

- 
- [Дэвид Шпигельхалтер](#)
    - 
    - 
    - [Введение](#)
    - [Глава 1. Расчет долей: качественные данные и проценты](#)
    - [Глава 2. Числовые характеристики выборки и представление данных](#)
    - [Глава 3. Почему мы смотрим на данные? Совокупности и измерение](#)
    - [Глава 4. Причины и следствия](#)
    - [Глава 5. Моделирование зависимости с помощью регрессии](#)
    - [Глава 6. Алгоритмы, аналитика и прогнозирование](#)
    - [Глава 7. Насколько мы можем быть уверены в происходящем? Оценки и интервалы](#)
    - [Глава 8. Вероятность – язык неопределенности и случайности](#)
    - [Глава 9. Объединяем вероятность и статистику](#)
    - [Глава 10. Отвечаем на вопросы и заявляем об открытиях](#)
    - [Глава 11. Учимся на опыте – байесовский путь](#)
    - [Глава 12. Когда дела идут не так](#)
    - [Глава 13. Как можно улучшить статистику](#)

- [Глава 14. В заключение](#)
- [Глоссарий](#)
- [Благодарности](#)
- [Список иллюстраций](#)
- [Список таблиц](#)
- [Эту книгу хорошо дополняют:](#)
- [notes](#)
  - [1](#)
  - [2](#)
  - [3](#)
  - [4](#)
  - [5](#)
  - [6](#)
  - [7](#)
  - [8](#)
  - [9](#)
  - [10](#)
  - [11](#)
  - [12](#)
  - [13](#)
  - [14](#)
  - [15](#)
  - [16](#)
  - [17](#)
  - [18](#)
  - [19](#)
  - [20](#)
  - [21](#)
  - [22](#)
  - [23](#)
  - [24](#)
  - [25](#)
  - [26](#)
  - [27](#)
  - [28](#)
  - [29](#)
  - [30](#)

- [31](#)
- [32](#)
- [33](#)
- [34](#)
- [35](#)
- [36](#)
- [37](#)
- [38](#)
- [39](#)
- [40](#)
- [41](#)
- [42](#)
- [43](#)
- [44](#)
- [45](#)
- [46](#)
- [47](#)
- [48](#)
- [49](#)
- [50](#)
- [51](#)
- [52](#)
- [53](#)
- [54](#)
- [55](#)
- [56](#)
- [57](#)
- [58](#)
- [59](#)
- [60](#)
- [61](#)
- [62](#)
- [63](#)
- [64](#)
- [65](#)
- [66](#)
- [67](#)

- [68](#)
- [69](#)
- [70](#)
- [71](#)
- [72](#)
- [73](#)
- [74](#)
- [75](#)
- [76](#)
- [77](#)
- [78](#)
- [79](#)
- [80](#)
- [81](#)
- [82](#)
- [83](#)
- [84](#)
- [85](#)
- [86](#)
- [87](#)
- [88](#)
- [89](#)
- [90](#)
- [91](#)
- [92](#)
- [93](#)
- [94](#)
- [95](#)
- [96](#)
- [97](#)
- [98](#)
- [99](#)
- [100](#)
- [101](#)
- [102](#)
- [103](#)
- [104](#)

- [105](#)
- [106](#)
- [107](#)
- [108](#)
- [109](#)
- [110](#)
- [111](#)
- [112](#)
- [113](#)
- [114](#)
- [115](#)
- [116](#)
- [117](#)
- [118](#)
- [119](#)
- [120](#)
- [121](#)
- [122](#)
- [123](#)
- [124](#)
- [125](#)
- [126](#)
- [127](#)
- [128](#)
- [129](#)
- [130](#)
- [131](#)
- [132](#)
- [133](#)
- [134](#)
- [135](#)
- [136](#)
- [137](#)
- [138](#)
- [139](#)
- [140](#)
- [141](#)

- [142](#)
- [143](#)
- [144](#)
- [145](#)
- [146](#)
- [147](#)
- [148](#)
- [149](#)
- [150](#)
- [151](#)
- [152](#)
- [153](#)
- [154](#)
- [155](#)
- [156](#)
- [157](#)
- [158](#)
- [159](#)
- [160](#)
- [161](#)
- [162](#)
- [163](#)
- [164](#)
- [165](#)
- [166](#)
- [167](#)
- [168](#)
- [169](#)
- [170](#)
- [171](#)
- [172](#)
- [173](#)
- [174](#)
- [175](#)
- [176](#)
- [177](#)
- [178](#)

- [179](#)
- [180](#)
- [181](#)
- [182](#)
- [183](#)
- [184](#)
- [185](#)
- [186](#)
- [187](#)
- [188](#)
- [189](#)
- [190](#)
- [191](#)
- [192](#)
- [193](#)
- [194](#)
- [195](#)
- [196](#)
- [197](#)
- [198](#)
- [199](#)
- [200](#)
- [201](#)
- [202](#)
- [203](#)
- [204](#)
- [205](#)
- [206](#)
- [207](#)
- [208](#)
- [209](#)
- [210](#)
- [211](#)
- [212](#)
- [213](#)
- [214](#)
- [215](#)

- [216](#)
- [217](#)
- [218](#)
- [219](#)
- [220](#)
- [221](#)
- [222](#)
- [223](#)
- [224](#)
- [225](#)
- [226](#)
- [227](#)
- [228](#)
- [229](#)
- [230](#)
- [231](#)
- [232](#)
- [233](#)
- [234](#)
- [235](#)
- [236](#)
- [237](#)
- [238](#)
- [239](#)
- [240](#)
- [241](#)
- [242](#)
- [243](#)
- [244](#)
- [245](#)
- [246](#)
- [247](#)
- [248](#)
- [249](#)
- [250](#)
- [251](#)
- [252](#)

- [253](#)
  - [254](#)
  - [255](#)
  - [256](#)
  - [257](#)
  - [258](#)
  - [259](#)
  - [260](#)
  - [261](#)
  - [262](#)
  - [263](#)
  - [264](#)
  - [265](#)
  - [266](#)
  - [267](#)
  - [268](#)
  - [269](#)
  - [270](#)
  - [271](#)
  - [272](#)
  - [273](#)
  - [274](#)
  - [275](#)
  - [276](#)
  - [277](#)
  - [278](#)
  - [279](#)
  - [280](#)
  - [281](#)
  - [282](#)
-

# Дэвид Шпигельхалтер

## Искусство статистики. Как находить ответы в данных

*Издано с разрешения Penguin Books Ltd и Andrew Nurnberg Literary Agency*

*Все права защищены.*

*Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.*

Original English language edition first published by Penguin Books Ltd, London

Text copyright © David Spiegelhalter 2019

The author has asserted his moral rights.

All rights reserved.

© Перевод на русский язык, издание на русском языке, оформление. ООО «Манн, Иванов и Фербер», 2021



*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»<sup>[1]</sup>,<sup>[2]</sup>*

### *Зачем нужна статистика?*

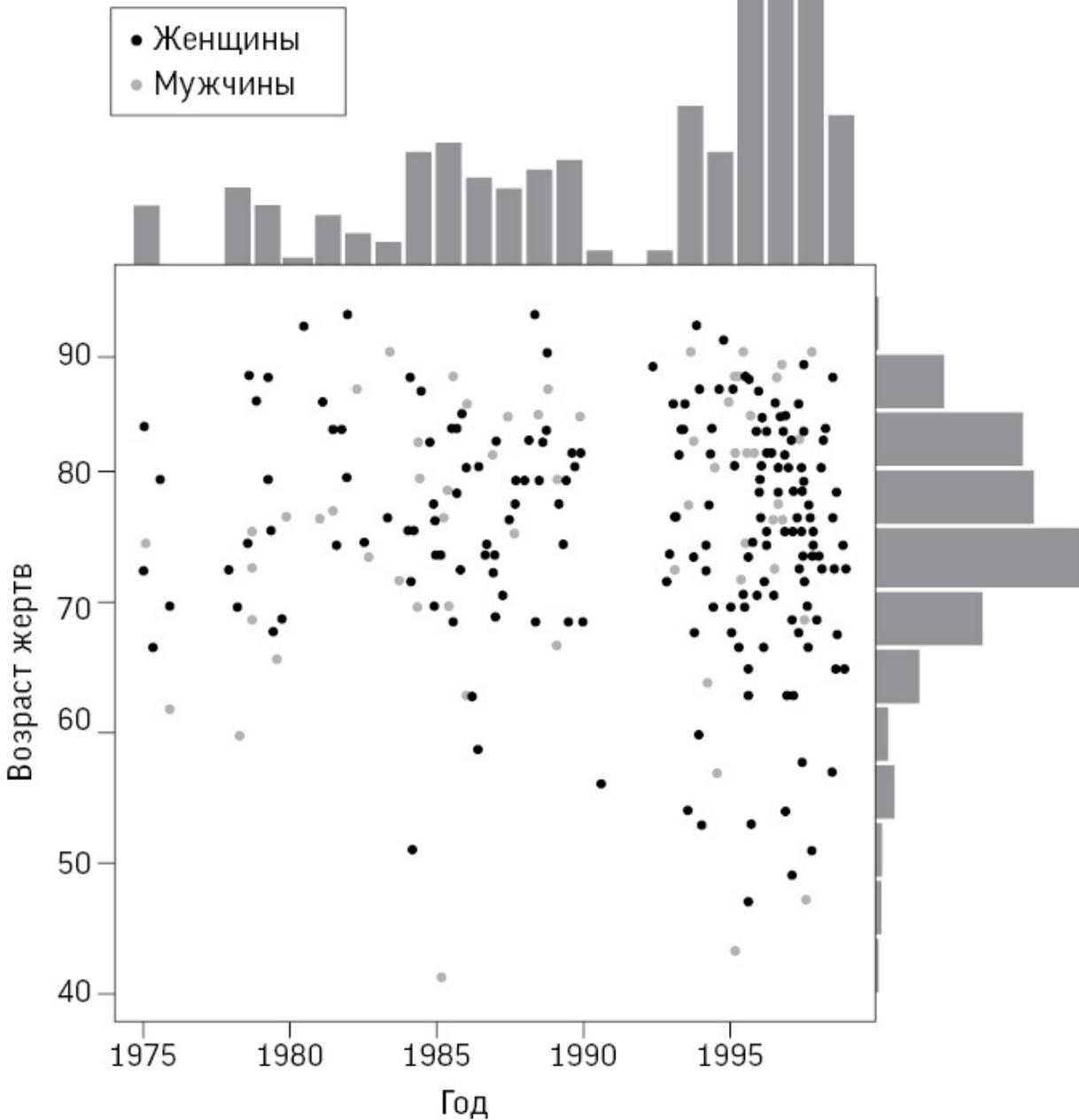
Психологический портрет Гарольда Шипмана, более известного как Доктор Смерть, не похож на серийного убийцу, тем не менее этот человек поставил рекорд по убийствам. Тихий семейный врач, работавший в пригороде Манчестера, в период с 1975 по 1998 год ввел как минимум 215 пожилым пациентам смертельную дозу 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<sup>[3]</sup>.

Эта книга посвящена применению **статистики**<sup>[4]</sup> 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).

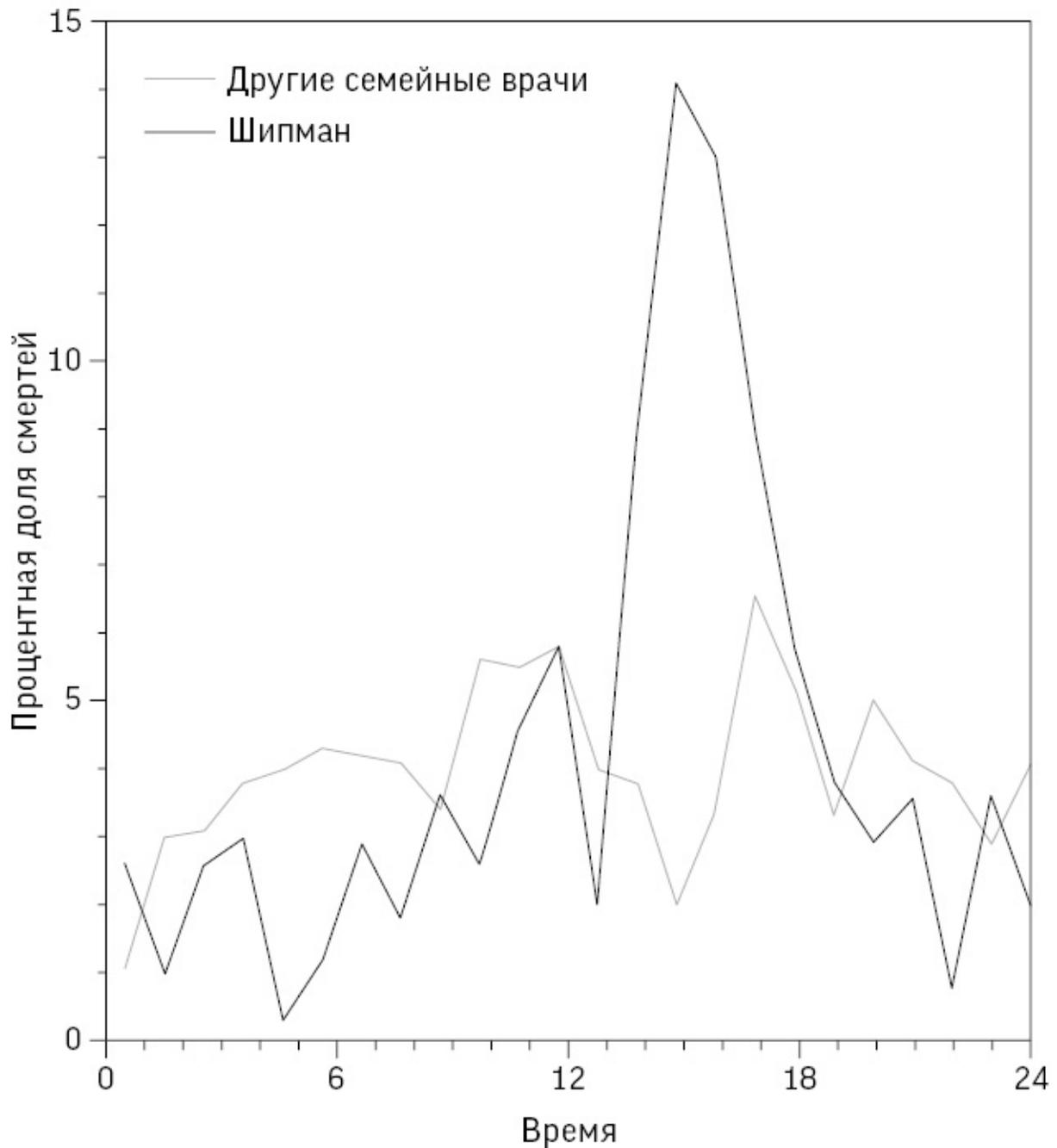


**Рис. 0.1**

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 невыносимо. 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 одноединственное 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<sup>[5]</sup>.

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<sup>[6]</sup>. 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)<sup>[7]</sup>. 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 специалисты, 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<sup>[8] [9]</sup>.

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<sup>[10]</sup>. 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<sup>[11]</sup>. 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<sup>[12]</sup>. 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<sup>[13]</sup>.

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<sup>[14]</sup>.

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 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<sup>[15]</sup> 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 невозпроизводимости многих «открытий» другими исследователями. Как, например, продолжающийся спор, может ли «поза силы» вызвать гормональные и другие изменения у человека<sup>[16]</sup>. На некорректном применении стандартных статистических методов лежит немалая доля вины за то, что известно как кризис воспроизводимости (или репликации) в науке.

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. Большое 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<sup>[17]</sup>. Рис. 0.3 основан на примере Новой Зеландии, которая считается мировым лидером по преподаванию статистики в школах.



**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 двусторчатый шкаф, а ближе всего к «Википедии» было удивительное портативное устройство, описанное в (необычайно прозорливом) путеводителе Дугласа Адамса «Автостопом по галактике»<sup>[18]</sup>. Поэтому для самосовершенствования мы обращались к книгам издательства Pelican, и их легко узнаваемые синие корешки были обычной приметой каждой студенческой полки<sup>[19]</sup>.

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)<sup>[20]</sup>. 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<sup>[21]</sup> или убедительные подтверждения экстрасенсорного восприятия. Другие касаются здравоохранения – например, выше ли показатель выживаемости в более загруженных больницах и полезны ли скрининговые исследования<sup>[22]</sup> 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<sup>[23]</sup>.

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.

# 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<sup>[24]</sup>.

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 (NES), полученная на основе информации, введенной низкооплачиваемыми программистами. У врачей NES 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 NES указывала 62 смерти 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»<sup>[25]</sup>. 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 бристолюбское 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 бристолюбскому 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<sup>[26]</sup>. 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 бристолюбского 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

<b>Больница</b>	<b>Количество прооперированных детей</b>	<b>Количество проживших минимум 30 дней после операции</b>	<b>Количество умерших в течение 30 дней после операции</b>	<b>Процентная доля выживших</b>	<b>Процентная доля умерших</b>
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

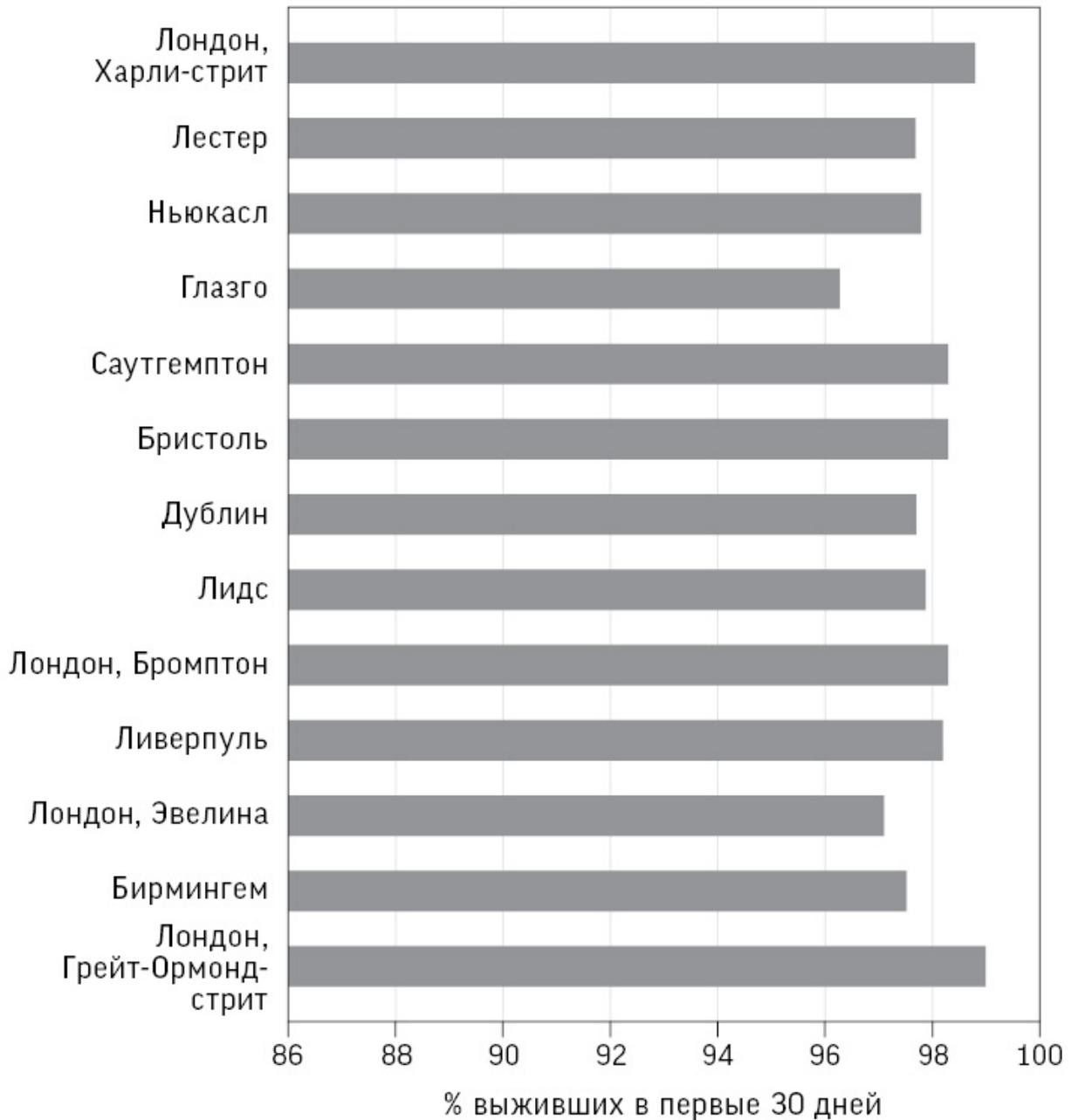
Таблицу можно считать видом графического представления данных, где для привлекательности и удобочитаемости требуется правильно подобрать цвет, шрифт и слова. На эмоциональную реакцию аудитории может также влиять выбор столбцов для отображения. В табл. 1.1 показаны данные об умерших и выживших, однако 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<sup>[27]</sup>.

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.



**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<sup>[28]</sup>, 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, коммуникации – 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)<sup>[29]</sup>.

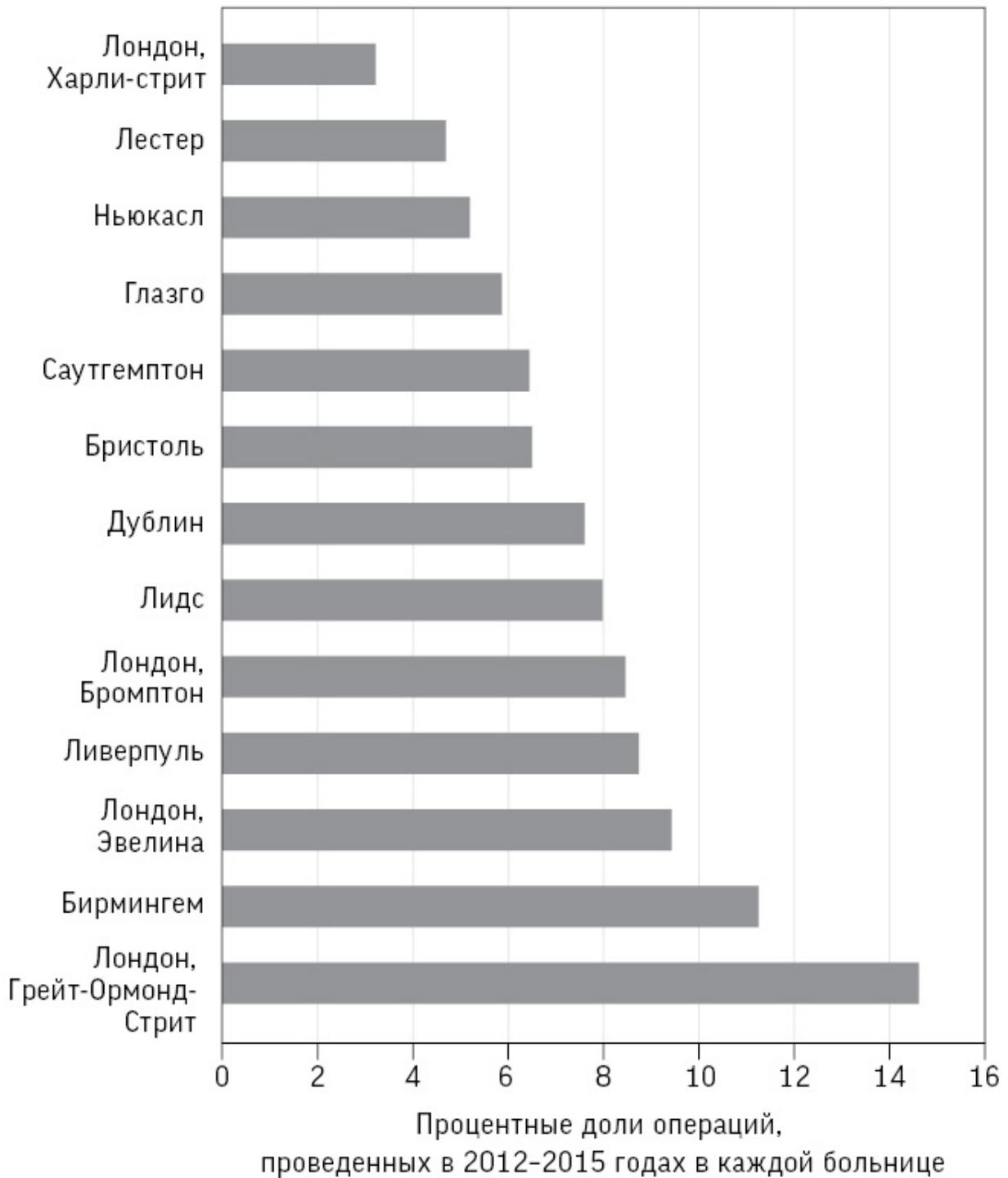
Для отображения качественных данных часто используются круговые диаграммы, что позволяет составить представление о размере каждой категории по занимаемой ею части круга. Однако здесь вероятны проблемы с наглядностью, например при попытке изобразить на одной диаграмме слишком много категорий или использовать 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.



**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»<sup>[30]</sup>.

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<sup>[31]</sup>. 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<sup>[32]</sup>.

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<sup>[33]</sup>. На рис. 1.4 представлена ожидаемая частота случаев рака кишечника в группе из 100 человек в виде **пиктографической диаграммы**.

100 человек, которые не едят бекон



100 человек, которые ежедневно едят бекон



**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)<sup>[34]</sup>

На рис. 1.4 «раковые» пиктограммы случайным образом разбросаны среди 100 изображений. Хотя было продемонстрировано, что такое рассеяние усиливает впечатление непредсказуемости, 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 съесть 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	$(7/93) / (6/94) \approx 1,18$

\* 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»<sup>[35]</sup>. 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:  $0,87 / 0,85 \approx 1,02$ . 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 %<sup>[36]</sup>.

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.
- 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»<sup>[37]</sup>) 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<sup>[38]</sup>. 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.

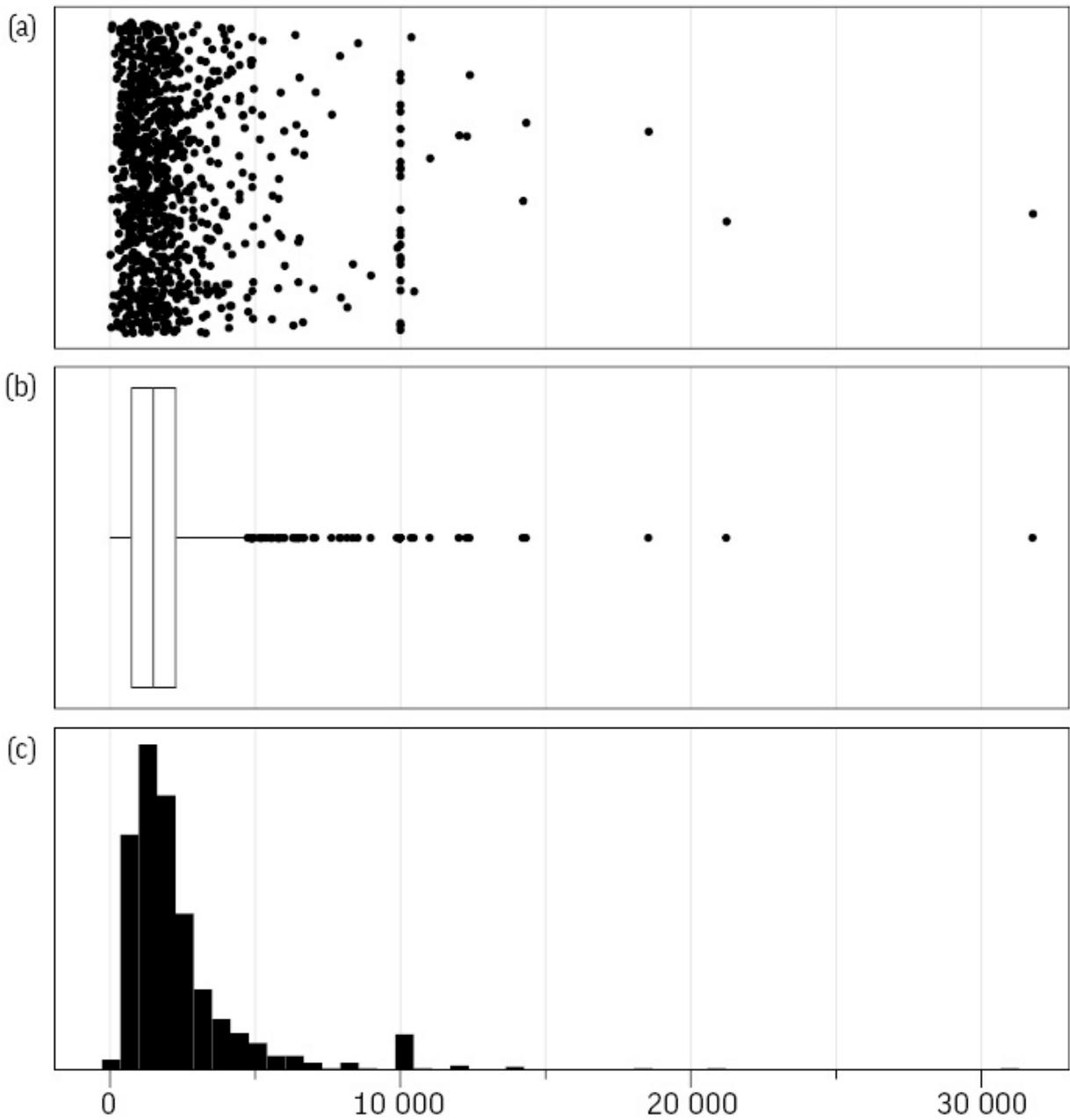
Статистика касается не только таких серьезных вещей, как рак и хирургия. В рамках нашего с популяризатором математики Джеймсом Граймом довольно простого эксперимента мы выложили на YouTube видео и попросили угадать число драже в банке. Вы тоже можете попробовать это сделать, посмотрев на фотографию на 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.



### **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<sup>[39]</sup>.



Высказанные предположения о количестве драже в банке

**Рис. 2.2**

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

(а) 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.

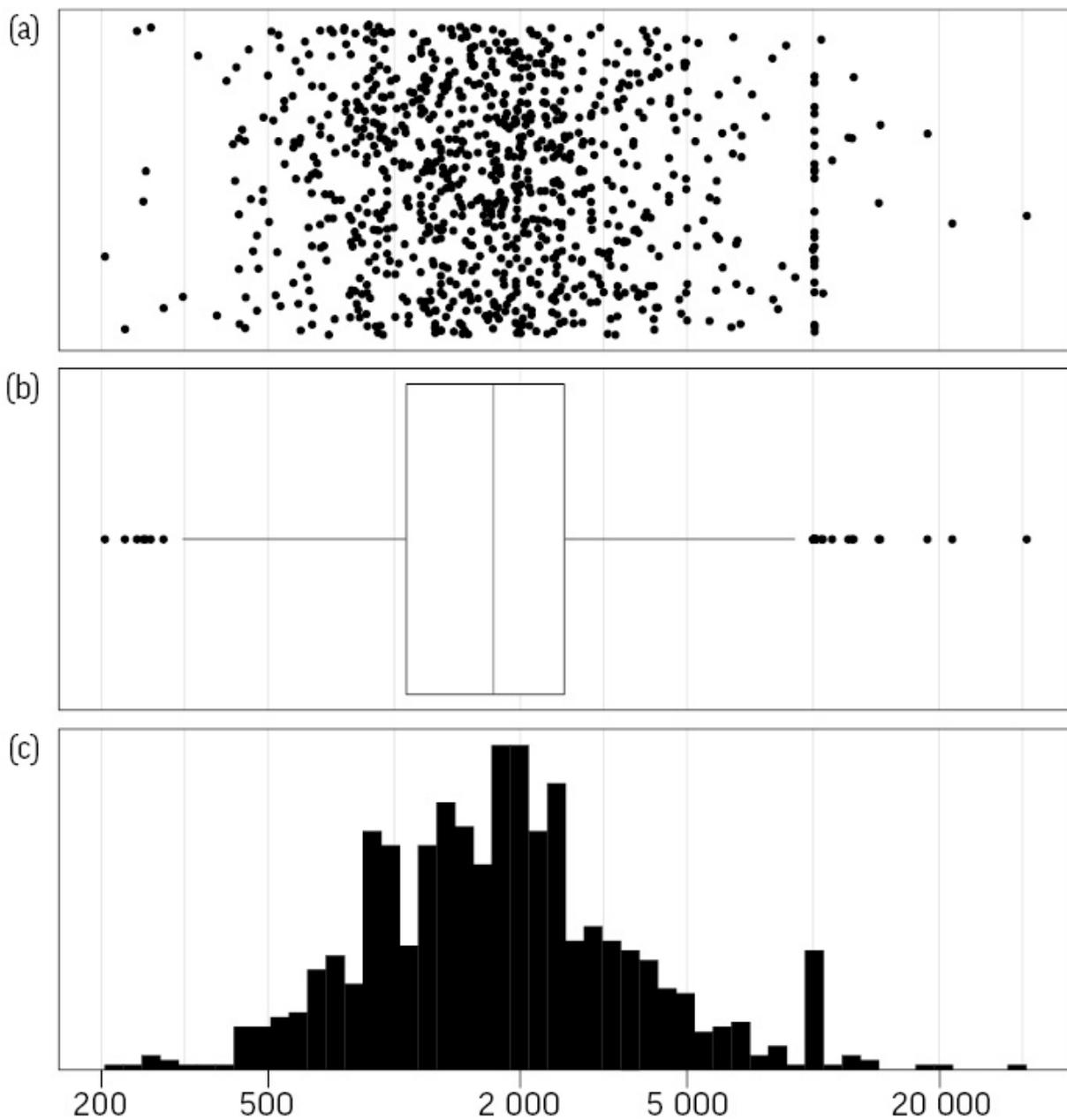
(b) Диаграмма размаха («ящик с усами») показывает некоторые базовые характеристики распределения<sup>[40]</sup>.

(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<sup>[41]</sup>.

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).



Высказанные предположения о количестве драже в банке

**Рис. 2.3**

Графическое отображение догадок о числе драже в банке в логарифмическом масштабе: (а) точечная диаграмма; (б) «ящик с усами»; (с) 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<sup>[42]</sup>.

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)<sup>[43]</sup>. Именно так Гальтон считал голоса толпы<sup>[44]</sup>.

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<sup>[45]</sup>: 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 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 большим 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<sup>[46]</sup>. 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<sup>[47]</sup>. 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**<sup>[48]</sup>, однако он крайне чувствителен к экстремальным значениям, таким как весьма странное предположение о наличии в банке 31 337 драже<sup>[49]</sup>. 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 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<sup>[50]</sup>. Например, удаление из выборки одного (почти гарантированно ошибочного) числа 31 337 приводит к уменьшению среднеквадратичного отклонения с 2422 до 1398<sup>[51]</sup>.

### Таблица 2.1

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

<b>Характеристики выборки при определении количества драже в банке</b>	<b>Значение</b>
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 подкреплён следующим примером.

### *Разница между группами чисел*

#### **Сколько сексуальных партнеров имеют британцы в течение жизни?**

Цель этого вопроса вовсе не любопытство относительно личной жизни людей. Когда в 1980-х годах обозначилась вся серьезность проблемы СПИДа, представители организаций здравоохранения Великобритании осознали, что не располагают достоверными данными о сексуальном поведении в стране, в частности о частоте смены партнеров, количестве людей, имеющих одновременно нескольких партнеров, а также об используемых сексуальных практиках. Такая информация была необходима для прогнозирования распространения болезней, передающихся половым путем, и планирования медицинских услуг. Однако люди все еще пользовались данными Альфреда Кинси для США 1940-х годов, а он не пытался получить репрезентативную выборку.

В конце 1980-х в Великобритании и США, несмотря на противодействие определенных кругов, были проведены масштабные, дорогостоящие и тщательные исследования сексуального поведения. И хотя Маргарет Тэтчер в последний момент отказалась поддержать работы по изучению сексуальных привычек в стране, к счастью, ученые смогли найти благотворительное финансирование, и в результате каждые 10 лет после 1990 года проводят Национальное исследование сексуальных отношений и образа жизни (Natsal).

Третье исследование (Natsal-3) проводилось в 2010 году и обошлось в 7 миллионов фунтов стерлингов<sup>[52]</sup>. В табл. 2.2 представлены сводные данные из Natsal-3 о количестве сексуальных

партнеров (противоположного пола), о которых сообщили люди в возрасте от 35 до 44 лет. Хорошее упражнение – использовать эти сведения, чтобы самостоятельно реконструировать, как могут выглядеть данные. Отметим, что наиболее часто встречающееся значение (мода) – это 1, то есть группа людей, у которых за жизнь был всего один партнер, по-прежнему велика. В таблице также отражены принципиальные различия между средними арифметическими и медианами, что говорит о распределениях с длинным правым хвостом. Среднеквадратичные отклонения велики, и это не лучшая мера разброса из-за неоправданно сильного влияния нескольких чрезвычайно больших значений в выборке.

**Таблица 2.2**

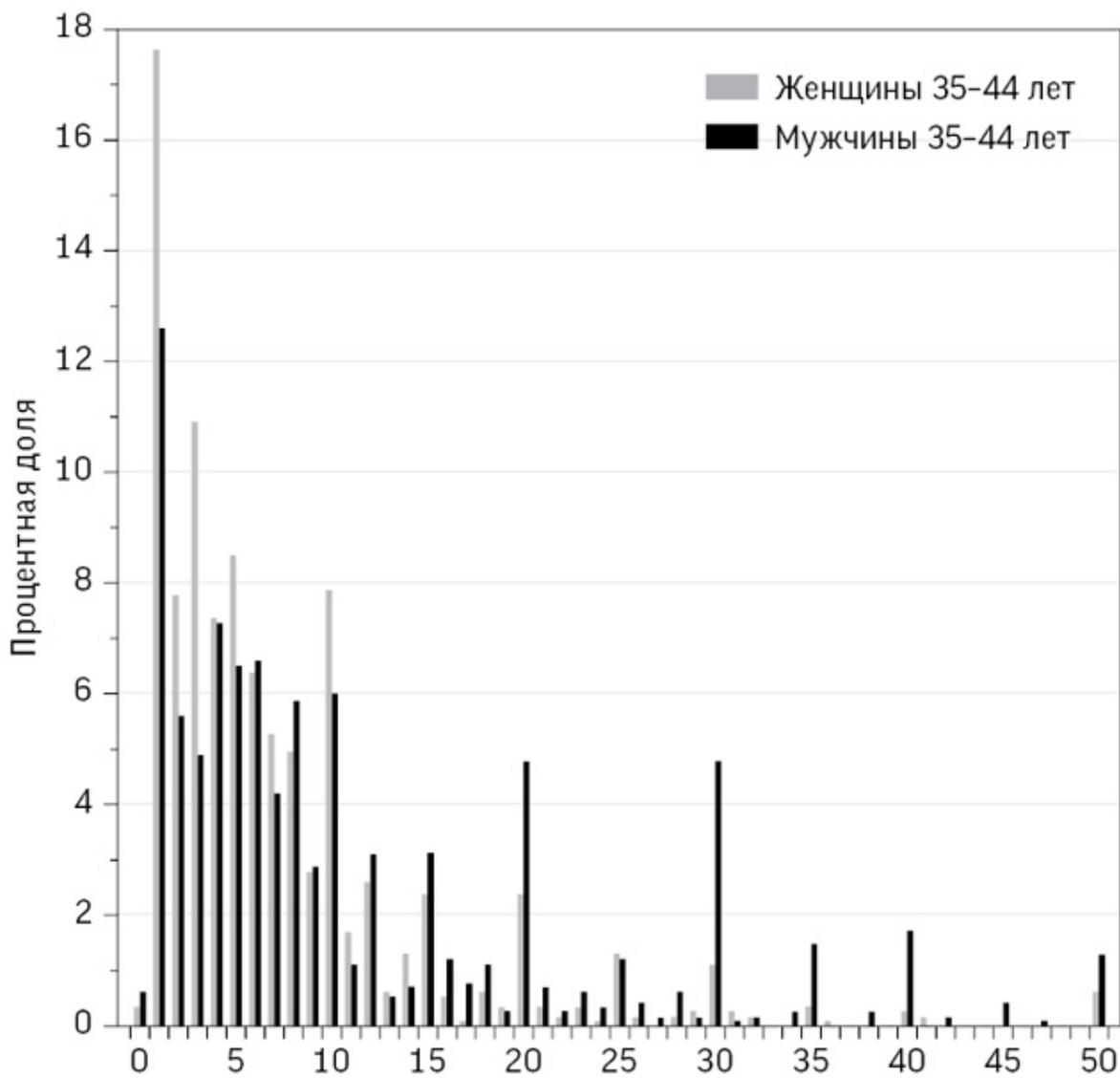
Сводные статистические данные о количестве сексуальных партнеров (противоположного пола) за всю жизнь, согласно ответам 806 мужчин и 1215 женщин в возрасте 35–44 лет, участвовавших в опросе Natsal-3 в период с 2010 по 2012 год. Среднеквадратичное отклонение включено для полноты картины, хотя и не является удачной характеристикой при таком разбросе данных

Указанное число сексуальных партнеров в течение жизни	Мужчины 35–44 лет	Женщины 35–44 лет
Выборочное среднее	14,3	8,5
Медиана	8	5
Мода	1	1
Размах	от 0 до 500	от 0 до 550
Интерквартильный размах	от 4 до 18	от 3 до 10
Стандартное (среднеквадратичное) отклонение	24,2	19,7

При сравнении ответов мужчин и женщин можно отметить, что у мужчин партнеров больше, чем у женщин – как по выборочному среднему (около 6), так и по медиане (3). Или, если воспользоваться относительными показателями, число партнеров, которое сообщают мужчины, примерно на 60 % больше, чем у женщин – как для выборочного среднего, так и для медианы.

Такая разница может вызвать у нас подозрения в отношении данных. В замкнутой **генеральной совокупности (популяции)** с одинаковым количеством мужчин и женщин и примерно одинаковым возрастным профилем среднее (в смысле среднее арифметическое) число партнеров противоположного пола у мужчин и женщин должно быть практически равнозначным!<sup>[53]</sup> Так почему же мужчины в возрастной группе от 35 до 44 лет сообщают о значительно большем количестве партнеров, чем женщины? Отчасти это может объясняться наличием у мужчин более молодых партнерш, которые не попадают в этот возрастной диапазон, а отчасти существованием систематического расхождения между тем, как мужчины и женщины учитывают свою сексуальную историю. Похоже, мужчины склонны преувеличивать число партнеров, а женщины – преуменьшать, или верно и то и другое.

На рис. 2.4 показано реальное распределение, которое подтверждает мнение о тяжелых правых хвостах, сложившееся на основании параметров, представленных в таблице. Кроме того, при взгляде на диаграмму видны и другие важные детали, такие как склонность мужчин и женщин указывать округленные числа при наличии десяти и больше партнеров (за исключением одного педантичного мужчины, возможно, статистика, который точно указал: сорок семь). Конечно, вы можете задуматься о достоверности таких сведений, а возможные искажения в них мы обсудим в следующей главе.



Указанное количество партнеров противоположного пола за всю жизнь

**Рис. 2.4**

Данные, предоставленные Natsal-3 на основе опроса 2010–2012 годов. Из-за экономии места ограничены числом 50, однако общее количество и у мужчин, и у женщин достигало 500. Обратите внимание на склонность мужчин называть большее число партнеров, чем женщины, и указывать круглые числа в случае 10 и более партнеров представителями обоих полов

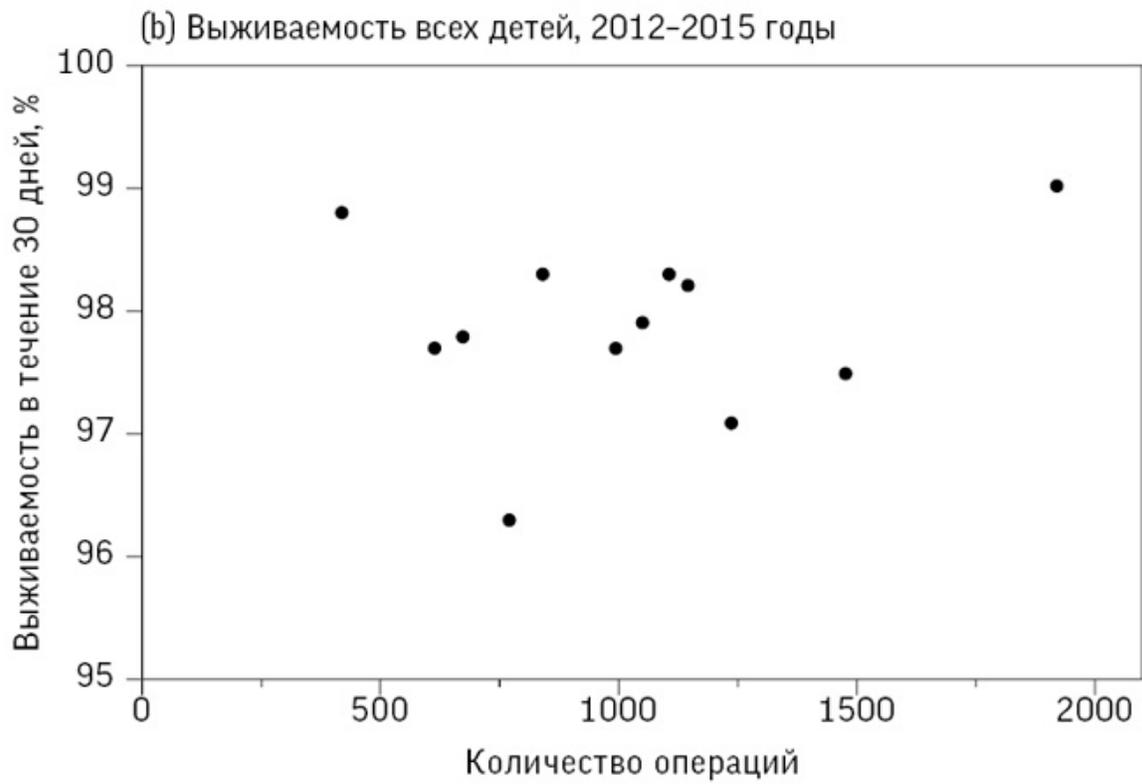
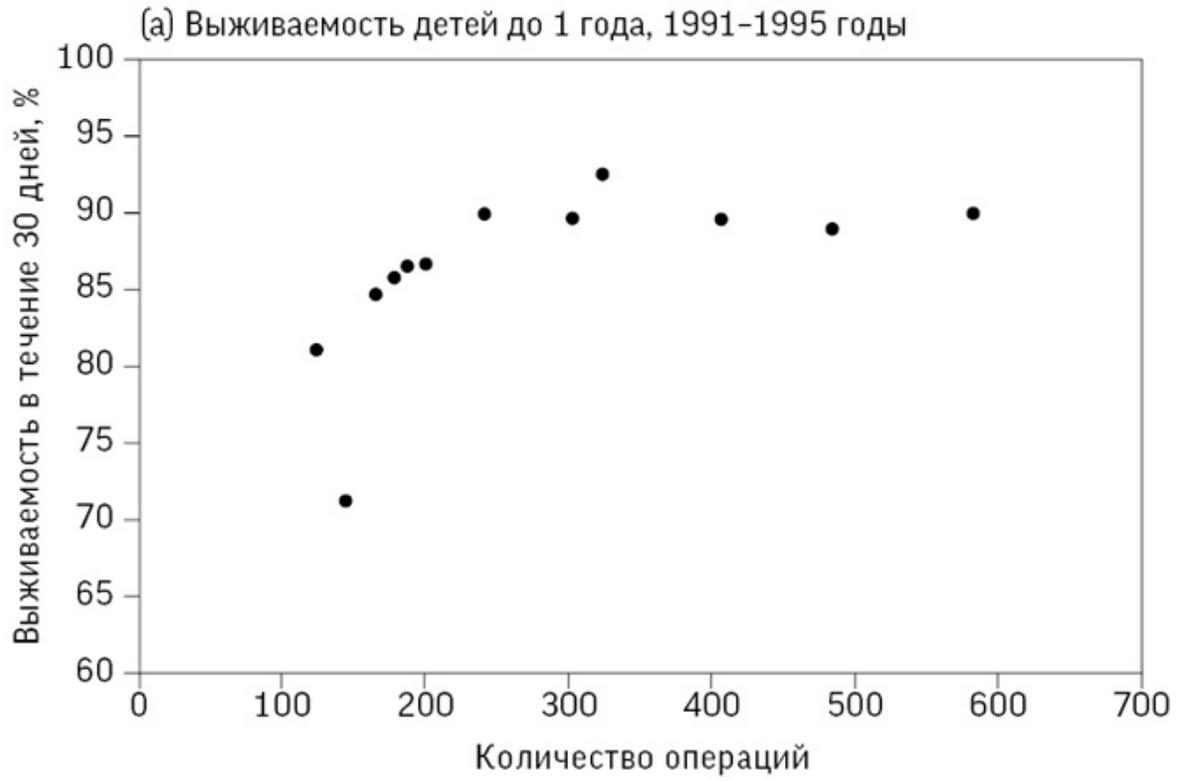
Большие совокупности данных обычно характеризуются несколькими параметрами положения и разброса, а пример с сексуальными партнерами доказал, что эти параметры позволяют

существенно продвинуться в понимании общей картины. Однако ничто не заменит простого внимательного просмотра данных, и следующий пример показывает, что хорошая визуализация особенно полезна при намерении уловить закономерности в большом и сложном наборе чисел.

### *Взаимосвязи между переменными*

#### **Выше ли показатели выживаемости в более загруженных больницах?**

Отмечается значительный интерес к так называемому эффекту масштаба в хирургии – утверждению, что в более загруженных больницах показатели выживаемости лучше, возможно, потому, что там выше эффективность и врачи имеют шанс приобрести больше опыта. На рис. 2.5 отображены показатели выживаемости детей в течение 30 дней после операций на сердце в больницах Великобритании в зависимости от количества прооперированных детей. На диаграмме 2.5(a) отображены данные о детях до 1 года за 1991–1995 годы (об этом периоде рассказывалось в начале [предыдущей главы](#)), поскольку именно эта возрастная группа отличается повышенным риском и находилась в центре внимания бристольского расследования. На диаграмме 2.5(b) представлены данные обо всех детях до 16 лет за 2012–2015 годы (также указаны в [табл. 1.1](#)); данных о детях до 1 года за этот период нет. По горизонтальной оси откладывается количество операций, а по вертикальной – уровень выживаемости<sup>[54]</sup>.



**Рис. 2.5**

Диаграммы рассеяния показателей выживаемости в зависимости от количества операций на сердце у детей. Для (а) коэффициент корреляции Пирсона равен 0,59, а ранговый коэффициент корреляции – 0,85. Для (б) коэффициент корреляции Пирсона равен 0,17, а ранговый коэффициент корреляции –0,03

Данные за 1991–1995 годы на диаграмме 2.5(а) демонстрируют явный выброс – небольшую больницу с низким показателем выживаемости в 71 %. Это Бристольская больница, низкие показатели которой и последующее расследование мы обсуждали в [главе 1](#). Однако если данные об этой больнице убрать (попробуйте закрыть эту точку пальцем), то вид данных за 1991–1995 годы подтверждает предположение о более высоком уровне выживаемости в больницах, где проводят больше операций.

Прямую или обратную зависимость между величинами на диаграмме рассеяния удобно выражать одним числом. Чаще всего для этого используется **коэффициент корреляции Пирсона** – идея, изначально предложенная Фрэнсисом Гальтоном, но официально закрепленная в работе Карла Пирсона, одного из основоположников современной статистики, в 1895 году [\[55\]](#).

Коэффициент корреляции Пирсона принимает значения от – 1 до 1 и показывает, насколько близко к прямой расположены точки на диаграмме. Коэффициент равен 1, если все точки лежат на прямой с положительным наклоном (чем больше одна величина, тем больше другая), и – 1, если все точки лежат на прямой с отрицательным наклоном (чем больше одна величина, тем меньше другая). Корреляция, близкая к 0, может свидетельствовать о случайном разбросе точек или о какой-либо иной зависимости, при которой отсутствует устойчивый возрастающий или убывающий тренд. Примеры таких случаев приведены на рис. 2.6.



**Рис. 2.6**

Два набора (вымышленных) данных, для которых коэффициент корреляции Пирсона будет примерно равен 0. Совершенно ясно, что это не говорит об отсутствии зависимости между двумя величинами. Из чудесной подборки диаграмм [\[56\]](#) Альберто Каиро [\[57\]](#)

Для данных за 1991–1995 годы, представленных на [диаграмме 2.5\(a\)](#), коэффициент корреляции Пирсона равен 0,59. Это подкрепляет связь между увеличением количества и ростом выживаемости. При удалении данных о Бристольской больнице коэффициент повышается до 0,67, поскольку оставшиеся точки ближе к прямой линии.

Другой критерий – **ранговый коэффициент корреляции Спирмена**, названный в честь английского психолога Чарльза Спирмена (создателя двухфакторной теории интеллекта [\[58\]](#)), – зависит не от конкретных численных значений, а от их рангов, то есть от занимаемых ими мест, если их упорядочить по величине. Это означает, что он может быть близок к 1 или  $-1$ , если точки близки к линии со стабильным подъемом или понижением, даже если эта линия не является прямой. Ранговый коэффициент Спирмена для данных на [диаграмме 2.5\(a\)](#) равен 0,85, что существенно выше, чем коэффициент Пирсона, поскольку точки ближе не к прямой, а к возрастающей кривой.

Для данных за 2012–2015 годы на [диаграмме 2.5\(b\)](#), коэффициент корреляции Пирсона равен 0,17, а ранговый коэффициент Спирмена – 0,03, что говорит об отсутствии четкой связи между количеством операций и уровнем выживаемости. Однако при таком небольшом количестве больниц коэффициент корреляции может быть очень чувствителен к отдельным точкам-данным: если мы уберем самую маленькую больницу с высоким уровнем выживаемости, то коэффициент корреляции Пирсона резко повысится до 0,42.

Коэффициенты корреляции – это просто некоторые характеристики связей, и их нельзя использовать для вывода о наличии взаимозависимости между количеством операций и показателем выживаемости, не говоря уже о том, почему такая связь может существовать<sup>[59]</sup>. Во многих приложениях ось  $x$  представляет **независимую переменную**, и интерес вызывает ее влияние на **зависимую переменную**, которая изображается по оси  $y$ . Однако, как мы увидим далее в главе 4, посвященной причинно-следственным связям, такое предположение заранее фиксирует направление влияния. Даже по [диаграмме 2.5\(a\)](#) мы не можем сделать вывод, что повышение показателя выживаемости в каком-либо смысле вызвано увеличением числа операций, ведь на самом деле все может быть наоборот: лучшие больницы просто привлекают больше пациентов.

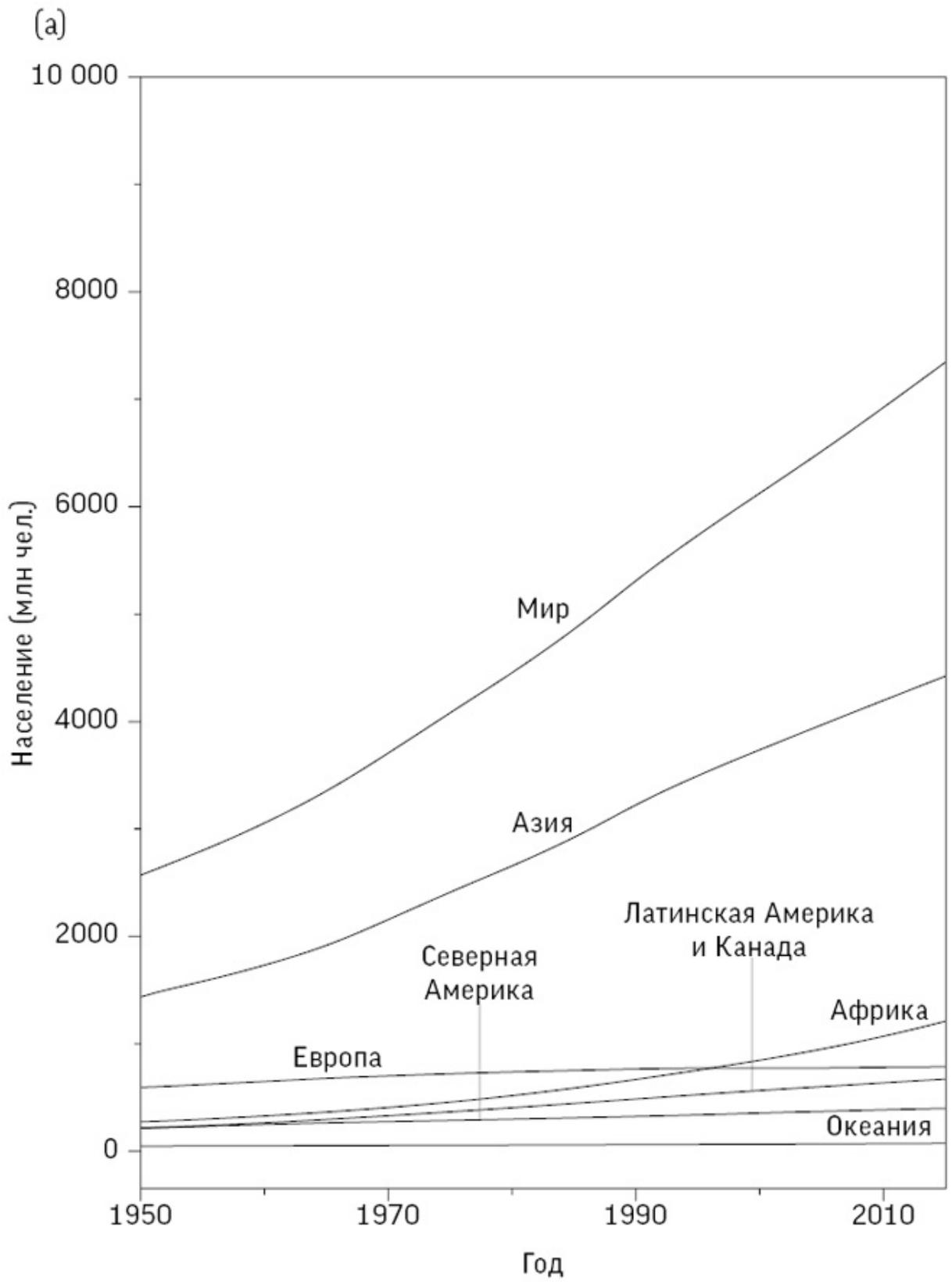
### *Описание трендов*

#### **Каковы закономерности роста мирового населения за последние полвека?**

Население мира растет, и понимание движущих факторов демографических изменений крайне важно для подготовки к вызовам, с которыми разным странам придется столкнуться сейчас или в будущем. Отдел народонаселения ООН дает оценки численности населения для всех стран мира с 1951 года по настоящее время, а также с прогнозом до 2100 года<sup>[60]</sup>. Сейчас мы рассмотрим мировые тенденции, начиная с 1951 года.

На рис. 2.7(a) представлены простые линейные графики для населения начиная с 1951 года. Видно, что за этот период оно утроилось и составляет примерно 7,5 миллиарда. Увеличение произошло в основном за счет стран Азии, однако закономерности для

других континентов на рис. 2.7(a) уловить трудно. Впрочем, использование логарифмической шкалы на рис. 2.7(b) позволяет их разделить, обнаруживая более крутой уклон у Африки и более пологий в других местах, в частности в Европе, где в последнее время численность населения уменьшается.



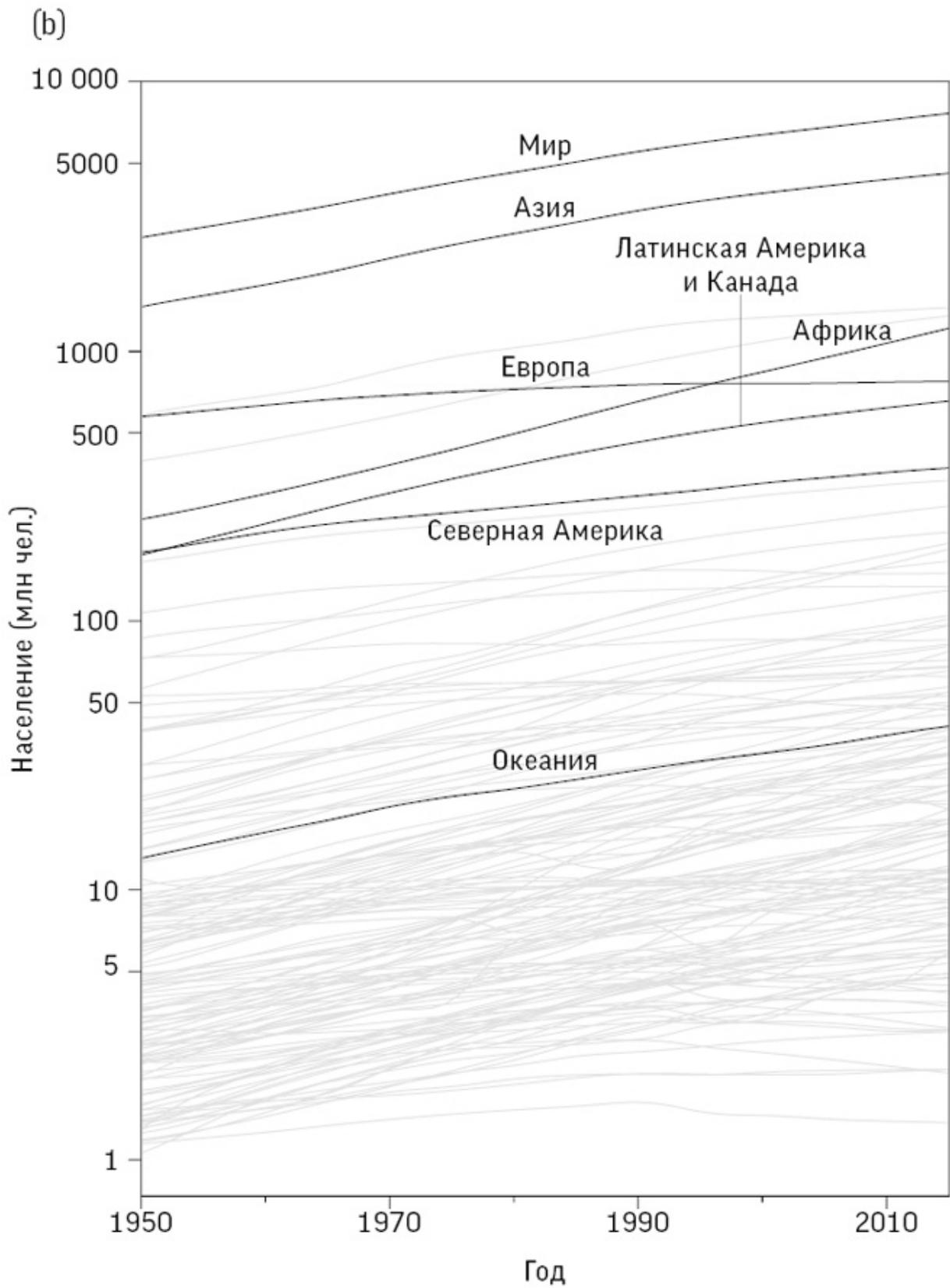
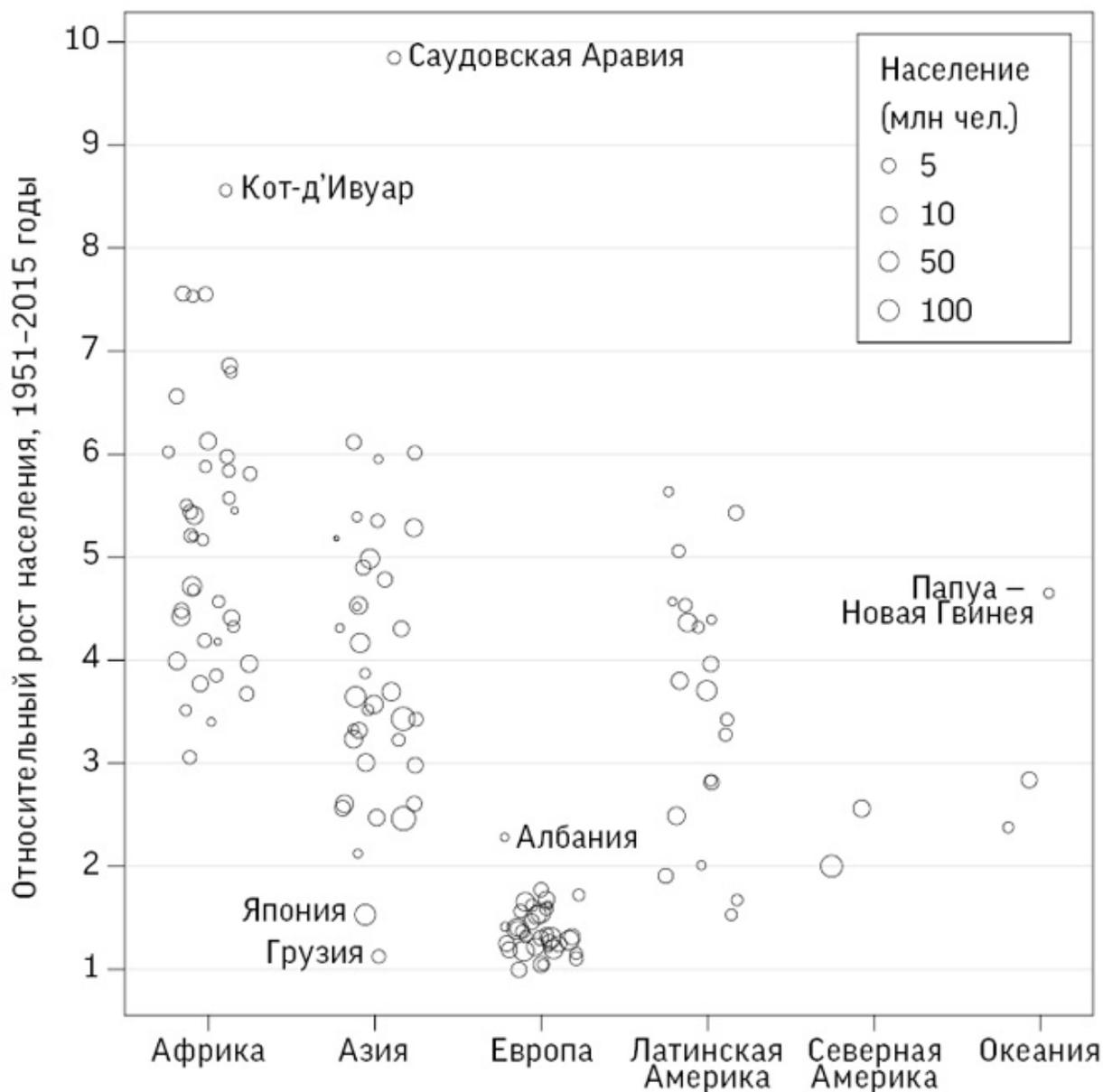


Рис. 2.7

Общая численность населения планеты, отдельных континентов и стран между 1950–2015 годами: (a) показывает тренды на стандартной шкале; (b) – на логарифмической шкале, вместе с линиями трендов для отдельных стран с населением не менее миллиона человек в 1951 году

Серые линии на рис. 2.7(b) отображают изменения в отдельных странах, однако выявить отклонения от общей тенденции к росту невозможно.

На рис. 2.8 представлена простая сводная характеристика тренда для каждой страны – относительный рост населения за период с 1951 по 2015 год. Скажем, относительный рост 4 означает, что в 2015 году в стране жило в четыре раза больше людей, чем в 1951-м (как, например, в Либерии, Камеруне и на Мадагаскаре). Использование значков, пропорциональных размеру страны, привлекает внимание к более крупным государствам, а группировка по частям света позволяет сразу же обнаруживать как общие кластеры, так и выбросы. Всегда полезно разделять данные в соответствии с каким-нибудь фактором (в нашем случае – с континентом), который в какой-то степени объясняет общие изменения.



**Рис. 2.8**

Относительный рост населения с 1951 по 2015 год в странах, население которых в 1951 году составляло не менее миллиона человек

Значительный рост населения наблюдается в Африке, но с большим разбросом и одним экстремальным случаем – Кот-Д’Ивуар. Азия тоже демонстрирует существенные различия, что отражает широкое разнообразие стран этого континента; здесь экстремальные случаи – Грузия и Япония, с одной стороны, и Саудовская Аравия – с

другой (у нее самый высокий показатель относительного роста населения в мире). Рост в Европе относительно низкий.

Как и любая хорошая диаграмма, эта вызывает новые вопросы и побуждает к дальнейшим исследованиям – как с точки зрения идентификации отдельных стран, так и изучения прогнозов будущих тенденций.

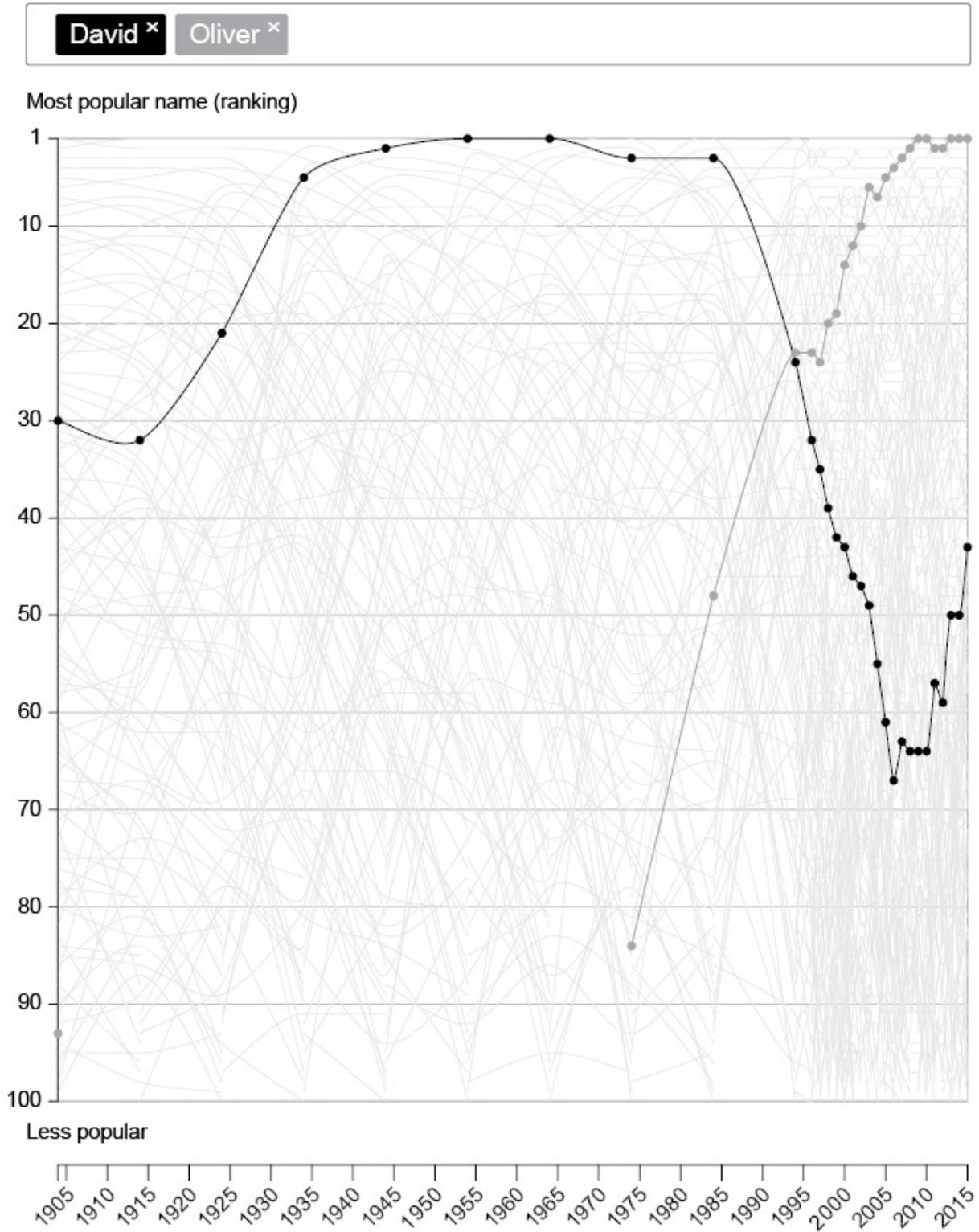
Конечно же, существует множество способов представления таких сложных массивов данных, как данные ООН по народонаселению, но ни один из них нельзя считать правильным. Тем не менее Альберто Каиро определил четыре общих признака хорошей визуализации данных.

1. Содержит достоверную информацию.
2. Схема выбрана так, чтобы соответствующие закономерности были заметны.
3. Выглядит привлекательно, при этом внешний вид не мешает правдивости, ясности и глубине.
4. Когда это уместно, способ организации позволяет проводить некоторые исследования.

Для реализации четвертого признака можно, например, позволить аудитории взаимодействовать с визуализацией. Хотя это трудно реализовать в книге, следующий пример покажет силу персонализации графического представления информации.

### **Как менялась популярность моего имени с течением времени?**

Некоторые графики настолько сложны, что невооруженным взглядом трудно заметить интересные закономерности. Посмотрите на рис. 2.9, где каждая линия показывает рейтинг популярности имен мальчиков, родившихся в Англии и Уэльсе между 1905 и 2016 годами<sup>[61]</sup>. Рисунок отображает замечательную социальную историю, хотя сам по себе всего лишь демонстрирует быстро меняющуюся моду на имена, а уплотнение линий в последние годы говорит о расширении и разнообразии списка имен после середины 1990-х.



**Рис. 2.9**  
 Скриншот интерактивной диаграммы, предоставленный Национальным статистическим управлением Великобритании, где

показаны тенденции изменения популярности имен мальчиков. Мои лишённые воображения родители дали мне в 1953 году самое популярное на то время имя, но с тех пор оно вышло из моды, в отличие от Оливера. Однако в последние годы имя Дэвид снова демонстрирует некоторые признаки повышения востребованности, возможно, благодаря Дэвиду Бекхэму

Только добавив интерактивность, мы можем выделить линии, представляющие для нас интерес. Например, мне интересен тренд для имени Дэвид, которое было особенно популярно в 1920-х и 1930-х годах, возможно, потому, что Дэвидом звали принца Уэльского (будущего короля Эдуарда VIII)<sup>[62]</sup>. Но затем оно резко утратило популярность – и если в 1953 году я был одним из десятков тысяч Дэвидов, то в 2016-м этим именем назвали всего 1461 ребенка, при этом больше сорока имен оказались гораздо популярнее.

### ***Коммуникация***

В этой главе мы старались обобщить и обнародовать данные открытым неманипулятивным способом, чтобы избежать влияния на эмоции и отношение аудитории и не навязывать ей определенную точку зрения. Мы просто хотим рассказать все как есть или по крайней мере как должно быть. Хотя мы не вправе претендовать на то, что излагаем абсолютную истину, мы пытались быть максимально правдивыми.

Конечно, о такой научной объективности проще говорить, чем реализовывать на практике. Когда в 1834 году Чарльз Бэббидж, Томас Мальтус и другие ученые создали Лондонское статистическое общество (впоследствии Королевское статистическое общество), они помпезно заявили, что «статистическое общество будет считать первым важнейшим правилом своей деятельности тщательное исключение всех частных мнений из своих протоколов и публикаций и основываться исключительно на фактах, причем – насколько это вообще возможно – на тех, которые могут быть записаны в численном виде и зафиксированы в таблицах»<sup>[63]</sup>. Увы, на это ограничение никто не обращал внимания с самого начала: авторы работ стали вставлять свои мнения о данных относительно преступлений, здоровья и экономики и советовать, что с этим делать. Возможно, лучшее, что мы можем

сейчас, – признать это искушение и всячески стараться держать свое мнение при себе.

Первое правило коммуникации – закрыть рот и слушать, чтобы лучше познакомиться с аудиторией, будь то политики, профессионалы или широкие массы. Мы должны понимать их неизбежные ограничения и любые возможные недоразумения и бороться с искушением казаться слишком умными или чрезмерно вдаваться в детали.

Второе правило коммуникации – знать, чего вы хотите добиться. Будем надеяться, что цель – способствовать открытым обсуждениям и принятию взвешенных решений. Однако, похоже, нелишне повторить еще раз, что цифры не говорят сами за себя: контекст, язык и графический вид способствуют коммуникации. Нужно признать, что мы рассказываем историю, а люди неизбежно станут сравнивать и выносить суждения, даже если мы всего лишь хотели информировать, а не убеждать. Все, что мы можем, – это постараться предотвратить неуместные инстинктивные реакции с помощью предупреждений или системы представления данных.

### *Изложение с использованием статистики*

В этой главе мы ввели понятие визуализации данных. Соответствующие методы часто используются для исследователей или достаточно подготовленной аудитории благодаря арсеналу средств, выбранных исходя из их ценности, чтобы обеспечить понимание и изучение данных, а не по причине их визуальной привлекательности. Но когда мы хотим донести до аудитории важное сообщение, содержащееся в данных, мы можем применить инфографику или визуализацию, чтобы привлечь внимание людей и рассказать хорошую историю.

Сложная инфографика регулярно появляется в СМИ, однако на рис. 2.10 представлен довольно простой пример, который говорит о социальных тенденциях, объединяя ответы на три вопроса из Национального исследования сексуальных отношений и образа жизни (Natsal-3) 2010 года: в каком возрасте мужчины и женщины впервые занялись сексом, когда они начали вместе жить и завели первого ребенка<sup>[64]</sup>. Медианный возраст для каждого из этих трех событий нанесен на график в зависимости от года рождения женщин, и три точки соединены жирной вертикальной линией. Устойчивое

удлинение этой линии для диапазона между 1930 и 1970 годами демонстрирует увеличение периода, когда необходима эффективная контрацепция.

За прошедшие 60 лет промежуток между возрастом, когда люди впервые начинают заниматься сексом, и возрастом, когда они начинают жить вместе и заводят первого ребенка, увеличился – соответственно, увеличился и период, когда необходимо предотвращать незапланированную беременность

Медианный возраст при первом половом акте, начале совместной жизни и рождении первого ребенка

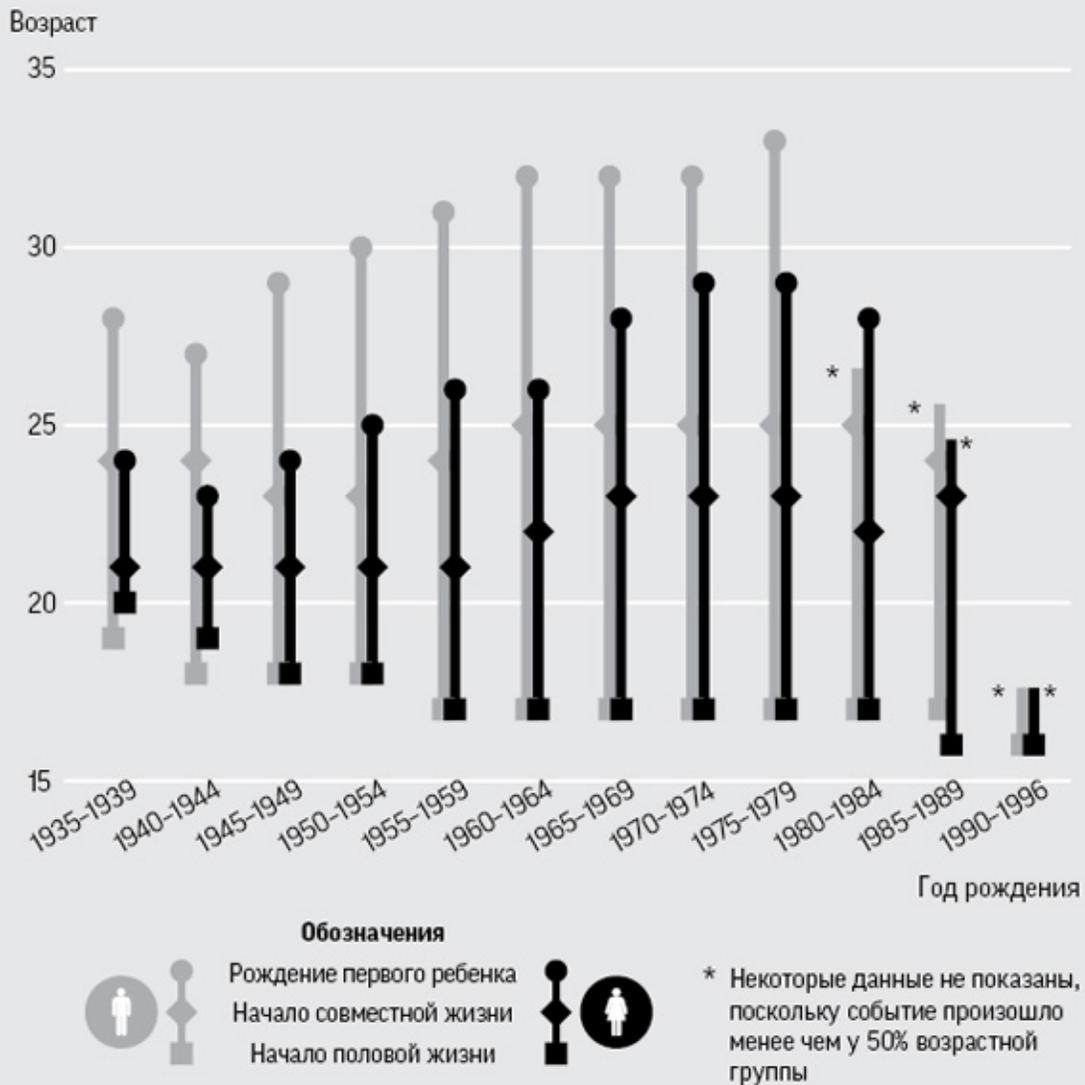


Рис. 2.10

Инфографика на основании данных Национального исследования сексуальных отношений и образа жизни (Natsal-3); выводы представлены как визуально, так и словесно

Еще более продвинутой является динамическая графика, где движение используется для выявления закономерностей изменений с течением времени. Специалистом по такой методике был Ханс Рослинг, чьи выступления на конференции TED<sup>[65]</sup> и видеоролики установили новый стандарт для выступлений с применением статистики, например демонстрация взаимосвязи между изменениями благосостояния и здоровья с помощью перемещения пузырьков, отражающих прогресс в каждой стране с 1800 года до наших дней. Рослинг использовал графику, чтобы исправить ошибочное представление о различии между развитыми и слаборазвитыми странами: динамические графики показывали, что со временем почти все страны стабильно двигались по одному и тому же пути в сторону улучшения благосостояния и процветания<sup>[66]</sup>, <sup>[67]</sup>.

В этой главе продемонстрирован весь диапазон представления информации – от простых описаний и изображения необработанных данных до сложных примеров изложения с применением статистики. Современные вычисления делают визуализацию данных проще и гибче. А поскольку характеристики выборки могут как скрывать, так и подчеркивать существенные особенности, важно наглядное графическое представление. Тем не менее выделение сводных характеристик выборки – только первый этап в процессе изучения данных. Чтобы продвинуться дальше по этому пути, нужно обратиться к фундаментальной идее того, чего мы намерены достичь в первую очередь.

### **Выводы**

- При анализе эмпирических распределений данных (в частности, определения среднего и разброса) применяются различные числовые характеристики.
- Часто встречаются асимметричные распределения, а некоторые показатели крайне чувствительны к выбросам.

- Сводные характеристики выборки всегда скрывают какие-то детали, поэтому нужно проявлять осторожность, чтобы не потерять важную информацию.
- Наглядно эмпирические данные можно представить в виде точечной диаграммы, диаграммы типа «ящик с усами» или гистограмм.
- Для лучшего выявления закономерностей используйте преобразования. Для обнаружения закономерностей, выбросов, сходств и кластеров используйте глаза.
- Рассматривайте пары чисел как точки на плоскости, а динамические (изменяющиеся во времени) величины – как линии на графике.
- При исследовании данных основная цель – поиск факторов, объясняющих изменчивость.
- Графика может быть интерактивной и анимированной.
- Инфографика выделяет интересные особенности и помогает читателям погружаться в повествование, но она должна использоваться с осознанием ее цели и воздействия на аудиторию.

## Глава 3. Почему мы смотрим на данные? Совокупности и измерение

### Сколько сексуальных партнеров у британцев на самом деле?

В предыдущей главе мы рассмотрели несколько примечательных результатов недавнего британского исследования, в рамках которого люди сообщали о количестве своих сексуальных партнеров за всю жизнь. Графические методы анализа этих ответов выявили определенные особенности, включая очень длинный хвост, склонность указывать круглые числа (например, 10 и 20) и тот факт, что мужчины называют большее число партнеров, чем женщины. Но исследователей, потративших миллионы фунтов на сбор таких данных, на самом деле интересовали не ответы конкретных респондентов (в конце концов, всем им гарантировалась полная анонимность), а общие закономерности сексуального поведения британцев, которые они на основе этих ответов хотели обнаружить.

На самом деле переход от реальных ответов, собранных в исследовании, к выводам обо всей Великобритании нельзя считать тривиальным. Было бы неправильно просто заявить, что ответы респондентов точно отражают ситуацию в стране. Опросы в СМИ о сексе, где добровольцы заполняют анкеты на сайтах, сообщая о том, чем они занимаются за закрытыми дверями, грешат этим постоянно.

Процесс перехода от сырых данных к утверждениям о поведении жителей всей страны можно разбить на несколько этапов.

1. Записанные *первичные данные* о числе сексуальных партнеров, которое указали участники исследования, говорят нам кое-что об...

2. *Истинном количестве* партнеров у людей в нашей *выборке*, что расскажет нам кое-что о...

3. *Количестве партнеров* у людей в *исследуемой совокупности* – тех, кто мог бы потенциально стать участником исследования. Это говорит нам кое-что о...

4. Числе сексуальных партнеров у всех британцев, которые и являются нашей *целевой совокупностью*.

Где самые слабые места в этой цепочке рассуждений? Переход от первоначальных данных (этап 1) к правде о нашей выборке (этап 2) означает наличие определенных предположений о том, насколько точно респонденты указали количество своих партнеров и насколько обоснованы причины для сомнений в их ответах. Мы уже наблюдали явную склонность мужчин преувеличивать, а женщин – преуменьшать количество своих связей, возможно, из-за того, что женщины не включают в них те, о которых предпочли бы забыть, из-за различных склонностей к округлению вверх или вниз, плохой памяти или просто вследствие «искажений из-за социальной приемлемости»<sup>[68]</sup>.

Переход от нашей выборки (этап 2) ко всей исследуемой совокупности, пожалуй, самый сложный шаг. Прежде всего мы должны быть уверены, что участники исследования представляют собой случайную выборку из тех, кто подходит для целей такого хорошо организованного исследования, как Natsal. Но мы также должны предположить, что люди, согласившиеся участвовать, составляют репрезентативную выборку, а это не просто. Доля отвечавших респондентов составила 66 %, что на удивление хорошо, учитывая характер вопросов. Однако существуют определенные доказательства того, что процент участия тех, кто менее сексуально активен, несколько ниже, что, впрочем, в какой-то степени уравнивается сложностью опроса членов общества с нетрадиционной сексуальной ориентацией.

Наконец, переход от исследуемой (этап 3) к целевой (этап 4) выборке упрощается, когда мы можем предположить, что потенциальные участники надлежащим образом представляют взрослое население. В случае Natsal это обеспечивается путем тщательного отбора на основании случайной выборки домохозяйств, хотя и означает, что люди из таких мест, как тюрьмы или женские монастыри, не учтены.

К тому времени, когда мы проработаем все, что может пойти не так, этого, вероятно, окажется достаточно, чтобы кто угодно скептически относился к любым общим утверждениям о сексуальном поведении страны на основании ответов респондентов. Однако весь смысл статистики – сгладить процесс на всех этапах и в конце с

должным смирением сказать, что мы можем (а что не можем) узнать из этих данных.

### ***Выводы из данных – процесс «индуктивного умозаключения»***

В предыдущих главах предполагалось, что у вас есть какая-то проблема, вы получаете какие-то данные, смотрите на них и находите их сводные характеристики. Иногда ответ уже заключен в подсчете, измерении или описании. Например, если мы хотим знать, сколько людей в прошлом году обращалось в службу экстренной медицинской помощи, то данные дадут нам ответ.

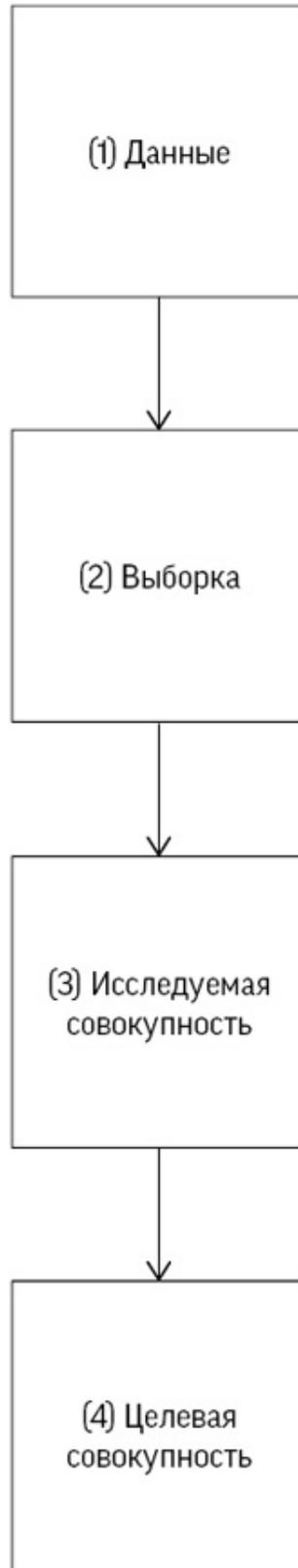
Однако часто вопрос выходит за рамки обычного описания данных: мы стремимся узнать нечто большее, чем просто набор имеющихся у нас наблюдений, например, хотим делать прогнозы (что будет происходить с показателями в следующем году?) или сообщить о причинах (почему цифры растут?)

Прежде чем приступить к обобщению на основе данных, чтобы узнать что-то о мире за пределами непосредственных наблюдений, нужно задать себе вопрос: «Узнать о чем?». А это требует обращения к сложной идее **индуктивного умозаключения**.

Многие люди имеют некоторое смутное представление о *дедукции* благодаря Шерлоку Холмсу, использовавшему ее при поиске преступников<sup>[69]</sup>. В реальной жизни дедукция – это процесс применения правил логики для перехода от общего к частному. Если согласно законодательству в стране установлено правостороннее движение, то мы можем прийти к дедуктивному заключению, что в любой ситуации лучше ехать по правой стороне. *Индукция* работает наоборот: на основании частных случаев предпринимаются попытки сделать общие заключения. Например, мы не знаем, принято ли в каком-то сообществе целовать подруг в щеку, и пробуем это выяснить, наблюдая, целуют ли женщины друг друга один, два, три раза или не целуют вовсе. Принципиальное отличие индукции от дедукции состоит в том, что дедукция дает истинные заключения, а индукция – в общем случае нет<sup>[70]</sup>.

На рис. 3.1 индуктивное умозаключение представлено в виде диаграммы, показывающей шаги, связанные с переходом от данных к

конечной цели нашего исследования. Как мы увидели, данные, собранные в ходе опроса, рассказывают нам о поведении людей в выборке; эту информацию мы используем для изучения поведения людей, которые могли бы стать участниками опроса, а уже из этого делаем некоторые предварительные выводы о сексуальном поведении в масштабе страны.



**Рис. 3.1**

Процесс индуктивного умозаключения: каждую стрелку можно истолковать как «говорит нам кое-что о...»<sup>[71]</sup>

Конечно, было бы идеально, если бы мы могли сразу перейти от просмотра первоначальных данных к общим утверждениям о целевой совокупности. В стандартных курсах статистики предполагается, что наблюдения извлекаются совершенно случайно и непосредственно из интересующей нас совокупности.

Однако в реальной жизни так бывает редко, поэтому нам приходится рассматривать всю процедуру перехода от первичных данных к конечной цели. При этом, как мы увидели на примере с исследованием Natsal, проблемы могут возникать на каждом этапе.

*Переход от данных (этап 1) к выборке (этап 2)* – это проблемы измерения. Является ли то, что мы фиксируем в своих данных, точным отражением того, что нас интересует? Мы хотим, чтобы наши данные были:

- надежными – в том смысле, что у них низкая изменчивость от случая к случаю и их можно считать воспроизводимыми и точными;
- достоверными – в том смысле, что вы измеряете именно то, что хотите, без какой-либо систематической ошибки.

Например, адекватность в опросе о сексе основывается на том, что люди на один и тот же вопрос каждый раз, когда их об этом спрашивают, отвечают практически одинаково, причем вне зависимости от интервьюера, настроения респондента или его памяти. Это в какой-то степени можно проверять, задавая в начале и в конце специальные вопросы. Качество исследования также требует, чтобы участники описывали свою сексуальную активность честно, а не систематически преувеличивая или преуменьшая свой опыт. Это довольно строгие требования.

Исследование станет недостоверным, если сами вопросы демонстрируют предвзятость в пользу конкретного ответа. Например, в 2017 году авиакомпания Ryanair объявила, что 92 % ее пассажиров довольны предоставляемым сервисом во время перелетов. Но, как оказалось на самом деле, опрос об уровне удовлетворенности

предусматривал только ответы *отлично, очень хорошо, хорошо, удовлетворительно и окей*<sup>[72]</sup>.

Мы уже видели, как форма подачи чисел (в положительном или отрицательном ключе) влияет на восприятие; точно так же формулировка вопроса может влиять на ответ. Например, в ходе опроса, проведенного в Великобритании в 2015 году, людей спрашивали, поддерживают ли они предоставление 16- и 17-летним подросткам права голосовать на референдуме о выходе из Евросоюза. Оказалось, что 52 % выступают за и 41 % – против. Таким образом, большинство людей поддержали это предложение, поскольку оно сформулировано с позиции признания и расширения прав молодежи.

Но когда тем же респондентам задали вопрос (логически идентичный предыдущему), поддерживают ли они уменьшение возрастного ценза для голосования на референдуме с 18 до 16 лет, доля сторонников этой идеи снизилась до 37 %, а против высказались 56 %. Таким образом, когда то же самое предложение было сформулировано в терминах более рискованной либерализации, большинство оказалось против. Мнение изменилось из-за простой переформулировки вопроса<sup>[73]</sup>.

На ответы также может влиять то, что спрашивалось ранее, – механизм, известный в психологии как прайминг (или фиксирование установки, или эффект предшествования). Согласно официальным исследованиям благосостояния, 10 % молодых британцев считают себя одиночками, при этом в ходе онлайн-опроса службы «Би-би-си» этот ответ выбрало гораздо больше участников – 42 %. Возможно, такое повышение показателя обусловлено двумя факторами: 1) самооценкой при добровольном «исследовании» и 2) тем, что вопросу об одиночестве предшествовал длинный ряд вопросов о том, испытывал ли респондент в целом недостаток дружеского общения, чувство брошенности, отстраненности и так далее. Возможно, все эти вопросы и побудили его дать положительный ответ на ключевой вопрос об одиночестве<sup>[74]</sup>.

*Переход от выборки (этап 2) к исследуемой совокупности (этап 3)* зависит от фундаментального качества исследования, называемого **внутренней валидностью**: отражает ли наблюдаемая выборка то свойство группы, которое мы изучаем? Именно здесь мы приходим к

ключевому способу для избежания искажений – случайной выборке. Даже дети понимают, что значит выбирать что-нибудь случайным образом – с закрытыми глазами сунуть руку в мешок с конфетами и посмотреть, какого цвета будет фантик у той, которую ты вытащил, или извлечь наугад номер из шапки, чтобы определить, кому достанется (или не достанется) приз или угощение. Этот метод тысячами использовался для обеспечения справедливости – определения вознаграждения<sup>[75]</sup>, проведения лотерей, назначения присяжных заседателей и прочего – и именовался жеребьевкой<sup>[76]</sup>. Применялся он и в более серьезных случаях – при выборе, кому идти на войну или кого съезить в спасательной шлюпке, затерявшейся в море.

Джордж Гэллап, фактически разработавший в 1930-е годы научные методы исследования общественного мнения, предложил изящную аналогию для понимания ценности случайной выборки, сказав, что, если вы сварили большую кастрюлю супа, вам не нужно съедать его весь, чтобы узнать, достаточно ли в нем приправы. Хватит и одной ложки, но *при условии, что вы хорошо все перемешали*. Буквальное доказательство это утверждение получило в 1969 году во время лотереи, определявшей порядок призыва на войну во Вьетнаме. Сначала в рамках лотереи создавался упорядоченный список дней рождения, а затем те, чья дата рождения оказывалась в верхних строках списка, отправлялись во Вьетнам, и так далее. В попытке сделать эту процедуру справедливой было подготовлено 366 капсул с уникальной датой рождения в каждой. Предполагалось, что капсулы будут извлекаться из ящика наугад. Однако складывали их в коробку в соответствии с месяцем рождения и не удосужились должным образом перемешать. Это не привело бы к проблемам, если бы люди, доставая капсулы, запускали руку поглубже в коробку, но, как показывает видеозапись, они, как правило, брали капсулы сверху<sup>[77]</sup>. В результате меньше всего повезло тем, кто родился в конце года: из 31 дня декабря были выбраны 26, в то время как из января – только 14 дней.

Идея надлежащего «перемешивания» имеет решающее значение: если вы хотите перейти от выборки ко всей генеральной совокупности, вы должны убедиться, что выборка репрезентативна. Наличие большого массива данных вовсе не гарантирует хорошую выборку и даже может вселить ложную уверенность. Например, на всеобщих выборах в Великобритании в 2015 году компании, проводящие опросы,

с треском провалились, хотя их выборки включали тысячи потенциальных избирателей. Последующее расследование обвинило нерепрезентативную выборку, особенно для телефонных опросов, так как в большинстве случаев звонили только на стационарные номера и фактически на эти звонки ответили менее 10 % абонентов. Вряд ли такую выборку можно считать репрезентативной.

*Переход от исследуемой (этап 3) к целевой (этап 4) совокупности.* Наконец, даже при превосходных измерениях и хорошей случайной выборке результаты по-прежнему могут не отражать того, что мы хотим исследовать, если нам не удалось опросить людей, в которых мы особенно заинтересованы. Мы хотим, чтобы наше исследование имело **внешнюю валидность** <sup>[78]</sup>.

Крайнее проявление – это ситуация, в которой целевая совокупность состоит из людей, тогда как изучать мы можем только животных, например при анализе воздействия какого-то химического вещества на мышей. Не столь кардинальная разница будет в случае, если клинические испытания нового препарата проводились исключительно на взрослых мужчинах, а затем он использовался для женщин и детей. Мы хотели бы знать влияние на всех людей, но одним статистическим анализом тут не обойтись – мы неизбежно должны делать предположения и проявлять осторожность.

### *Когда есть все данные*

Хотя вышеописанные исследования хорошо иллюстрируют идею извлечения информации из данных, фактически большая часть используемых сегодня данных не основывается на случайной или вообще на какой-либо выборке. Регулярно собираемые данные, скажем об онлайн-покупках или социальных взаимодействиях, а также об администрировании образовательных или правоохранительных систем, можно переориентировать, чтобы лучше понять происходящее в мире. В таких ситуациях у нас есть полные данные. С точки зрения индуктивного процесса, показанного на [рис. 3.1](#), между этапами 2 и 3 нет разрыва – выборка и исследуемая совокупность, по сути, совпадают. Это избавляет от беспокойства по поводу малого размера выборки, однако многие другие проблемы все же могут оставаться.

Рассмотрим вопрос об уровне преступности в Великобритании и его важный политический аспект: растет он или снижается.

Существуют два ключевых источника данных: один – на основе опросов, второй – официальный. Первый, «Исследование преступности в Англии и Уэльсе» – классический пример опроса, в рамках которого примерно 38 тысяч человек ежегодно рассказывают о своем опыте соприкосновения с криминалом. Как и в случае исследования Natsal о сексе, здесь могут возникать проблемы при переходе между этапами. Во-первых, приходится использовать самоотчеты (этап 1) для оценивания реального опыта людей (этап 2), поскольку они могут скрывать правду, например, о том, что сами замешаны в незаконных делах, связанных с наркотиками. Во-вторых, мы вынуждены предположить, что выборка репрезентативна для соответствующей совокупности, и учесть ее ограниченный размер (переход от этапа 2 к этапу 3). В-третьих, нам нужно признать, что план исследования не охватывает какой-то части общей целевой совокупности, скажем подростков младше 16 лет или людей в местах совместного проживания (переход от этапа 3 к этапу 4). Тем не менее «Исследование преступности в Англии и Уэльсе» с определенными оговорками считается официально признанной национальной статистикой и применяться для отслеживания долгосрочных тенденций<sup>[79]</sup>.

Второй источник данных – сообщения о преступлениях, зарегистрированных полицией. Это делается для официальных целей и не является выборкой: поскольку можно учесть каждое преступление, зарегистрированное в стране, «исследуемая совокупность» совпадает с выборкой. Конечно, мы по-прежнему должны предполагать, что записанные данные действительно отображают то, что случилось с жертвами преступлений (переход от этапа 1 к этапу 2), но главная проблема возникает при утверждении, что данные об исследуемой совокупности (люди, которые сообщают о преступлениях) представляют целевую совокупность по всем преступлениям, совершенным в Англии и Уэльсе. К сожалению, полицейская статистика систематически упускает случаи, которые полиция не зарегистрировала как преступления или о которых жертвы предпочли умолчать, такие как незаконное употребление наркотиков или отказ людей сообщать о краже или вандализме, когда из-за этого падают цены на недвижимость в месте их проживания. Вот яркий пример: когда в ноябре 2014 года полицейские методы регистрации подверглись суровой критике, число зафиксированных преступлений на сексуальной

почве возросло с 64 тысяч в 2014 году до 121 тысячи в 2017-м, то есть почти удвоилось за три года.

Неудивительно, что эти два разных источника данных могут приводить к различным выводам о наблюдаемых тенденциях. Например, согласно «Исследованию преступности», между 2016 и 2017 годами уровень преступности снизился на 9 %, в то время как полиция зарегистрировала на 13 % больше правонарушений. Чему тут верить? Статистики больше доверяют опросу, а сомнения в достоверности предоставляемых полицией данных привели к тому, что в 2014 году они перестали использоваться в качестве национальной статистики.

Располагая полными данными, нетрудно получить статистику, описывающую то, что было измерено. Но если мы хотим применять их для более масштабных заключений о происходящем вокруг, качество данных приобретает первостепенное значение. И мы должны быть внимательны к систематическим ошибкам любого рода, которые могут поставить под угрозу надежность этих заключений.

Целые сайты посвящены перечислению возможных ошибок в статистике – от ошибки распределения (ошибка при распределении пациентов по группам) до ошибки добровольного участия (люди, добровольно участвующие в исследованиях, систематически отличаются от людей в генеральной совокупности). Хотя причины возникновения многих из них очевидны, в главе 12 мы узнаем и о более завуалированных причинах появления плохих статистических данных. Но сначала мы должны рассмотреть способы описания нашей конечной цели – целевой совокупности.

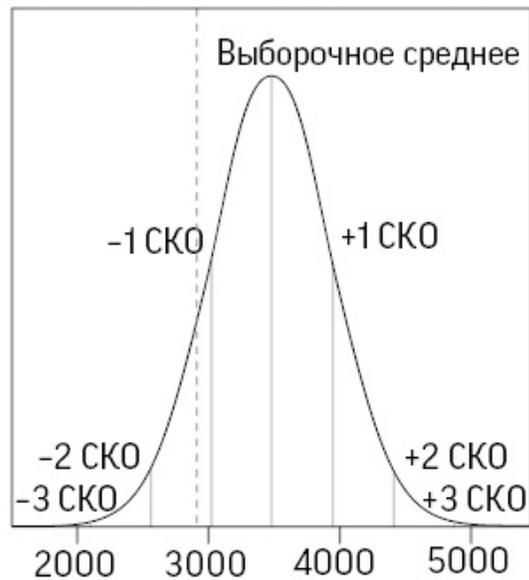
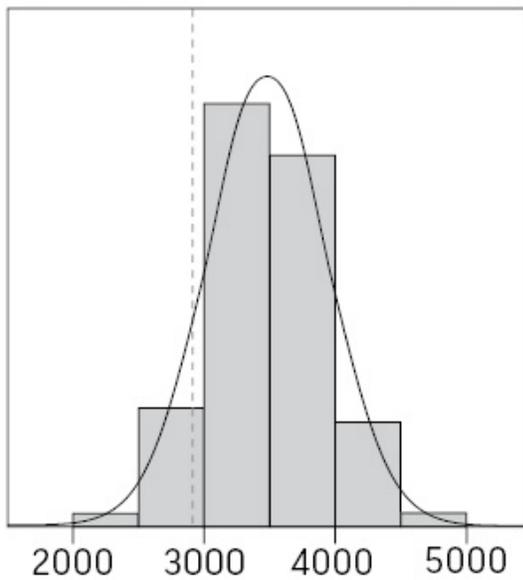
### *Колоколообразная кривая*

**Подруга в США родила доношенного ребенка весом 2910 граммов. Ей сказали, что это ниже среднего, и она обеспокоена. Действительно ли этот вес недостаточен?**

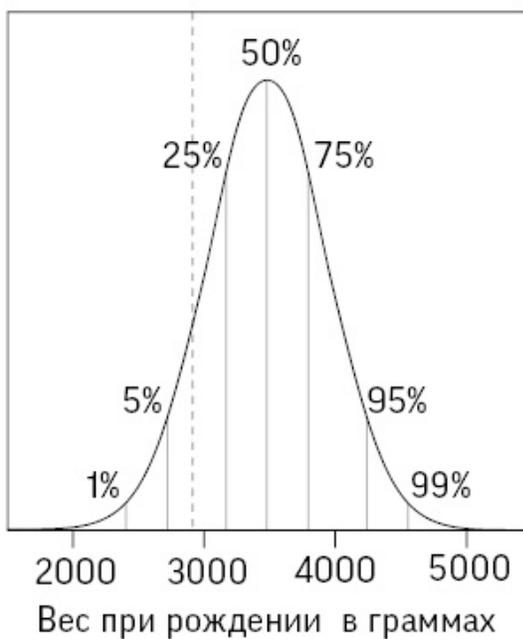
Мы уже обсуждали понятие распределения данных (эмпирическое или выборочное распределение) – закономерность, которой подчинены данные в выборке. Теперь нам нужно рассмотреть концепцию **распределения генеральной совокупности**, то есть модель во всей интересующей нас группе.

Вернемся к нашей роженице. Будем думать о ее ребенке как о своего рода выборке из одного человека, взятой из генеральной совокупности всех детей, недавно родившихся в США у неиспаноязычных белых женщин (указание расы важно, поскольку вес новорожденных сообщается для различных рас). Распределение генеральной совокупности определяется по весу при рождении для всех таких младенцев; эти данные можно получить из Национальной системы статистического учета естественного движения населения США, в которой зарегистрировано свыше миллиона доношенных детей, родившихся в США в 2013 году у белых неиспаноязычных женщин. Хотя это не все множество рождений, тем не менее выборка настолько велика, что ее можно рассматривать как генеральную совокупность<sup>[80]</sup>. Новорожденные распределяются по группам в соответствии с их весом при рождении (с шагом 500 граммов); эти данные представлены на рис. 3.2(a).

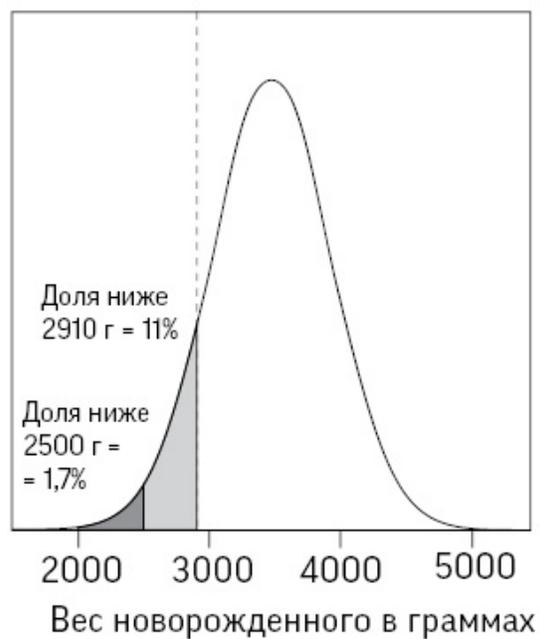
(a) Распределение веса при рождении (b) Выборочное среднее  $\pm 1, 2, 3$  СКО



(c) Процентили



(d) Низкий вес при рождении



**Рис. 3.2**

(a) Распределение веса при рождении для 1 096 277 детей, родившихся в США у белых неиспаноязычных женщин в 2013 году на 39–40 неделе беременности, а также кривая нормального распределения с теми же значениями среднего и среднеквадратичного отклонения (СКО), что и регистрируемый вес детей в этой генеральной

совокупности. Ребенок весом 2910 граммов отображен пунктирной линией. (b) Значения среднего  $\pm 1, 2, 3$  СКО для нормального распределения. (c) Перцентили для нормального распределения. (d) Доля новорожденных с низкой массой тела (темно-серая область) и с массой менее 2910 г (серая область)

Вес ребенка вашей подруги (2910 граммов) указан в виде пунктирной линии, положение которой относительно всего распределения можно использовать для оценки того, насколько он «необычен». Важна форма этого распределения. Такие измерения, как вес, доход, рост и другие аналогичные величины, можно, по крайней мере теоретически, производить с любой желаемой точностью. Поэтому для них можно использовать непрерывные распределения, отображаемые не ступенчатыми, а плавными линиями<sup>[81]</sup>. Классический пример – колоколообразная кривая, или **нормальное (гауссовское) распределение**, которое впервые было подробно исследовано Карлом Фридрихом Гауссом в 1809 году в контексте анализа ошибок измерений в астрономии и геодезии<sup>[82]</sup>.

Как показывает теория, нормальное распределение случайной величины можно встретить в ситуациях, обусловленных влиянием на нее большого количества мелких факторов, – например, когда на какую-нибудь физическую характеристику нашего тела влияет большое количество генов. Массу тела при рождении (для одной этнической группы и сходного срока беременности) вполне можно считать такой характеристикой, и на [рис. 3.2\(a\)](#) представлена теоретическая кривая нормального распределения с теми же значениями среднего и среднеквадратичного отклонения, что и вся совокупность зарегистрированного веса у детей. Гладкая теоретическая кривая и гистограмма, отображающая реальные данные, удовлетворительно близки<sup>[83]</sup>. Аналогично и другие характеристики человека, такие как рост или когнитивные навыки, также имеют распределение, близкое к нормальному. Однако существуют и величины, распределение которых далеко от гауссовского и часто имеет длинный правый хвост. Классический пример – доход.

Нормальное распределение случайной величины характеризуется двумя параметрами – своим **средним** (или **математическим ожиданием**) и стандартным отклонением (которое является мерой

разброса или отклонения от среднего); кривая на [рис. 3.2\(a\)](#) имеет среднее на уровне 3480 граммов и стандартное отклонение 462 грамма. Мы видим, что величины, используемые в [главе 2](#) для характеристики выборки, можно также применять для описания всей генеральной совокупности. Разница лишь в том, что термины *среднее* и *стандартное отклонение* в контексте выборки называются **статистиками**, а в контексте генеральной совокупности в целом – **параметрами**. Это впечатляющая возможность – описать больше миллиона измерений (то есть больше миллиона рождений) только этими двумя величинами.

Огромное преимущество использования нормального распределения – в его изученности и возможности взять все его важные характеристики из таблиц или программ. На [рис. 3.2\(b\)](#) показано положение среднего и 1, 2 и 3 среднеквадратичных отклонения в обе стороны от него. Из математических свойств нормального распределения мы знаем, что примерно 95 % всей генеральной совокупности содержится в промежутке [среднее  $\pm 2$  СКО], а примерно 99,8 % всей генеральной совокупности – в промежутке [среднее  $\pm 3$  СКО]. Ребенок вашей подруги находится приблизительно на 1,2 СКО ниже среднего – параметр, известный как **Z-оценка** (или просто число, показывающее, на сколько СКО данное значение отличается от среднего).

Среднее и стандартное отклонение могут также использоваться в качестве кратких описаний (большинства) других распределений, однако полезными могут быть и другие характеристики. На [рис. 3.2\(c\)](#) показаны выбранные **процентили** для нормальной кривой: например, 50-й процентиль – это медиана, которая делит генеральную совокупность пополам. Можно сказать, что медианное значение – это вес «среднего» ребенка. В случае симметричных распределений (каким и есть нормальное) медиана совпадает со средним значением. 25-й процентиль (3167 граммов) – это вес, меньше которого имеют 25 % родившихся детей. 25-й и 75-й процентиль (3791 граммов) называются **квартелями**, а расстояние между ними (в нашем случае 624 грамма), или интерквартильный размах – мерой разброса для распределения. И снова те же характеристики, которые в [главе 2](#) мы относили к выборке, здесь применяются ко всей совокупности в целом.

Ребенок вашей подруги находится в 11-м процентиле, а значит, 11 % всех доношенных детей у белых неиспаноговорящих женщин будут весить меньше. На [рис. 3.2\(d\)](#) эта 11-процентная область выделена серым цветом. Перцентили веса ребенка важны на практике, поскольку изменения массы его тела будут отслеживаться по отношению к росту, ожидаемому у малышек в 11-м процентиле<sup>[84]</sup>, и низкое значение перцентиле может стать причиной для беспокойства.

По медицинским, а не статистическим причинам дети с весом ниже 2500 граммов считаются «имеющими низкую массу тела при рождении», а с весом меньше 1500 граммов – «очень низкую массу тела при рождении». [Рис. 3.2\(d\)](#) показывает, что, согласно ожиданиям, 1,7 % младенцев в этой генеральной совокупности будут иметь низкую массу тела при рождении. Фактическое число таких детей составило 14 170 (1,3 %) – хорошее соотношение с прогнозом, который дает нормальная кривая. Следует отметить, что в этой группе (доношенные дети у белых неиспаноязычных женщин) уровень детей с низкой массой тела очень небольшой, в то время как общий уровень в 2013 году по всей стране составил 8 %, а у черных женщин – 13 %; как видите, разница между расами существенна.

Возможно, самый важный урок, извлеченный из этого примера, состоит в том, что темно-серая закрашенная область на [рис. 3.2\(d\)](#) выполняет две функции:

1. Отображает *долю* детей с низкой массой тела при рождении в генеральной совокупности.
2. Демонстрирует *вероятность* того, что вес случайно выбранного ребенка, родившегося в 2013 году, будет меньше 2500 граммов.

Таким образом, генеральную совокупность можно рассматривать не только как группу реальных людей, но и как представление **вероятностного распределения** для случайных наблюдений. Эта двойная интерпретация будет иметь фундаментальное значение, когда мы перейдем к более формальным статистическим заключениям.

Конечно, в этом случае мы знаем форму и параметры генеральной совокупности, поэтому можем что-то сказать и о долях, и о вероятностях различных событий, которые могут наступать при случайных наблюдениях. Но суть этой главы в том, что мы, как

правило, не знаем параметры генеральной совокупности, а потому хотим с помощью индукции переходить от данных выборки ко всей совокупности. Мы видели, что стандартные измерения выборочного среднего, медианы, моды и так далее, которые мы создали для выборки, распространяются на всю генеральную совокупность. Но разница в том, что мы не знаем, что это такое. Именно с этой проблемой мы и столкнемся в следующей главе.

### ***Что такое генеральная совокупность?***

Рассмотренные выше индуктивные этапы хорошо работают с плановыми исследованиями, однако значительная часть статистических анализов не так легко вписывается в эту структуру. Мы видели, что иногда (например, при использовании полицейской документации о преступлениях) у нас могут быть все доступные данные. И хотя это не выборка, идея лежащей в их основе какой-то генеральной совокупности все же имеет ценность.

Вернемся к данным об операциях на сердце у детей из [главы 1](#). Мы сделали довольно смелое предположение, что проблем с измерениями не было – иными словами, что у нас есть полный набор операций и всех выживших детей в течение 30 дней во всех больницах, то есть идеальное знание выборки (этап 2).

Но что такое изучаемая совокупность? Мы располагаем данными обо всех больницах и всех детях, поэтому нет большей группы, из которой они могут быть взяты. Хотя идея генеральной совокупности обычно вводится в курсах статистики довольно буднично и вскользь, наш пример показывает, что это сложное и запутанное понятие, требующее подробного изучения, поскольку на нем основаны многие важные идеи.

Существуют три вида генеральных совокупностей, из которых мы можем делать выборки – вне зависимости от того, являются ли источниками данных люди, сделки, деревья или что-либо другое.

- *Буквальная* совокупность. Это идентифицируемая группа, откуда мы, к примеру, выбираем случайным образом человека при опросе. Или группа людей, для которых можно провести измерения, и, хотя мы на самом деле не выбираем наугад, у нас есть данные от добровольцев. Например, мы можем рассматривать людей, угадавших число даже в

банке, как выборку из совокупности всех любителей математики, которые смотрят видеоролики на YouTube.

- *Виртуальная* совокупность. Мы часто проводим измерения с помощью каких-либо устройств, скажем, измеряем кровяное давление или уровень загрязнения воздуха. Мы знаем, что всегда можем сделать еще несколько измерений и получить немного другие результаты – вам это прекрасно известно, если вы когда-нибудь повторно измеряли артериальное давление. Близость полученных результатов зависит от точности прибора и неизменности обстановки. Мы могли бы думать об этом как о получении наблюдений из некой виртуальной совокупности всех измерений, которые могли бы сделать, если бы имели достаточно времени.

*Метафорическая* совокупность. В этом случае никакой большей совокупности нет вообще. Это необычное понятие. Мы действуем так, будто наши данные получены случайным образом из какой-то большей совокупности, хотя это не так. Например, в случае детей, перенесших операцию на сердце, у нас не было никакой выборки, а были полные данные, и ничего сверх них мы собрать уже не могли. Подумайте о количестве ежегодно совершаемых убийств, результатах экзаменов для определенного класса или данных обо всех странах мира – ни в одном из этих случаев мы не можем считать имеющиеся данные выборкой из какой-то фактической совокупности.

Идея метафорической совокупности требует осмысления: возможно, предпочтительнее думать, что наши наблюдения берутся из некоего воображаемого пространства возможностей. Например, мировая история такая, какая есть, но мы можем представить, что она развивалась по совершенно иному сценарию, а мы просто оказались в одном из ее возможных состояний. Это множество альтернативных историй можно считать метафорической совокупностью. А если конкретнее, то, когда мы рассматривали детские операции в Соединенном Королевстве за 2012–2015 годы, у нас были полные данные о детях за этот период: мы знали и число смертей, и число выживших. Однако мы можем себе представить гипотетические истории, в которых выжили бы другие дети вследствие непредвиденных обстоятельств, которые мы склонны именовать «случайностью».

Должно быть очевидно, что в статистике выборка редко составляется буквально наугад и что более распространены ситуации, когда потенциально доступны полные данные. Тем не менее крайне полезно придерживаться концепции воображаемой генеральной совокупности, из которой взята наша «выборка», поскольку в этом случае мы можем использовать все математические методы, разработанные для составления выборок из реальных генеральных совокупностей.

Лично мне больше нравится действовать так, будто происходящее вокруг – результат случайного выбора из всех вероятных сценариев. От нас зависит, будем ли мы верить, что это действительно случайность, или воля Божья или богов, или какая-то иная теория причинности: для математики разницы нет. Это всего лишь одно из расширяющих кругозор требований при работе с данными.

### **Выводы**

- Для перехода от данных к выборке, а затем к изучаемой и далее к целевой совокупности требуются индуктивные умозаключения.
- На каждом из этапов могут возникать ошибки и проблемы.
- Лучший способ перейти от выборки к исследуемой совокупности – обеспечить случайность выборки.
- Генеральную совокупность можно представлять и как группу объектов, и как отображение вероятностного распределения для случайного наблюдения, полученного из этой совокупности.
- Описывать совокупности можно с помощью тех же характеристик, что и выборки.
- Часто данные не являются выборкой из буквальной совокупности. Когда в выборку входят все данные, мы можем вообразить, что они взяты из метафорической совокупности событий, которые могли бы случиться, но не произошли.

## Глава 4. Причины и следствия

### Повышает ли поступление в университет риск развития опухоли мозга?

Эпидемиология изучает, как и почему возникают и распространяются заболевания, при этом скандинавские страны – мечта эпидемиолога. А все потому, что в них каждый человек имеет личный идентификационный номер, который используется при регистрации во всех сферах: здравоохранение, образование, налогообложение и прочие. Это позволяет исследователям комплексно изучать различные аспекты жизни людей, что невозможно сделать (и, наверное, не всегда целесообразно) в других государствах.

Одно масштабное исследование, проведенное более чем на 4 миллионах шведов и шведок, в рамках которого связывались сведения о налогообложении и здоровье за 18 лет, установило, что у людей с более высоким социально-экономическим положением чаще диагностировали опухоль головного мозга. Это было одно из тех солидных, но весьма неинтересных исследований, которые обычно не привлекают особого внимания, поэтому специалист по связям с общественностью посчитал, что в пресс-релизе гораздо лучше написать так: «Высокий уровень образования связан с повышенным риском развития опухоли головного мозга», хотя работа посвящалась скорее социально-экономическому положению, чем образованию. Однако к тому времени, когда результаты были представлены широкой публике, помощник редактора одной из газет выдал классический заголовок: «Почему поступление в университет повышает риск развития опухоли мозга»<sup>[85]</sup>.

Такой заголовок встревожил бы любого, кто имеет высшее академическое образование. Но стоит ли на самом деле беспокоиться? Исследование основывалось на всей доступной генеральной совокупности, а не на выборке, поэтому мы с уверенностью можем заключить, что у более образованных людей действительно немного чаще выявляли опухоль головного мозга. Но неужели интенсивные нагрузки в библиотеке действительно перегревали мозг и вели к неблагоприятным мутациям клеток? Несмотря на газетный заголовок, я

в этом сомневаюсь. Как, собственно, и авторы статьи, которые добавили: «Потенциальным объяснением такого результата могут быть полнота регистрации рака и ошибка выявления». Другими словами, люди с более высоким уровнем образования с большей вероятностью пройдут обследование, а значит, опухоли будут регистрироваться чаще (пример того, что в эпидемиологии называется **ошибкой обращаемости**<sup>[86]</sup>).

### *Корреляция не означает причинность*

Из [главы 2](#) мы узнали, что коэффициент корреляции Пирсона показывает, насколько близко к прямой расположены точки на диаграмме рассеяния. Когда мы рассматривали английские больницы, проводившие в 1990-х операции на сердце у детей, и отображали на диаграмме точки, отражавшие число операций и уровень выживаемости, высокая корреляция демонстрировала, что более крупные больницы ассоциировались с более низким уровнем смертности. Однако мы не могли сделать вывод, что более крупные больницы и есть *причина* более низкой смертности.

У такого осторожного отношения солидная родословная. Когда в журнале Nature в 1900 году обсуждали предложенный Карлом Пирсоном коэффициент корреляции, один комментатор предупредил, что «корреляция не означает причинно-следственной связи». В течение следующего столетия эта фраза стала мантрой, постоянно повторяемой статистиками при столкновении с заявлениями, основанными на простом наблюдении, что какие-то две вещи имеют тенденцию изменяться вместе. Существует даже специальный сайт, который автоматически находит невероятные связи: например, очаровательную корреляцию 0,96 между ежегодным потреблением сыра моцарелла в США за 2000–2009 годы и количеством докторских степеней по гражданскому строительству, полученных за этот период<sup>[87]</sup>.

Похоже, у людей есть глубокая внутренняя потребность объяснять происходящее в виде простейшей зависимости «причина → следствие». Уверен, что каждый из нас мог бы придумать увлекательную историю обо всех этих остепененных инженерах, поглощающих пиццу с сыром. Существует даже специальное слово для склонности конструировать

связи между событиями, которые в реальности не связаны, – апофения, причем ее крайнее проявление – объяснять простую случайность или невезение злонамеренностью других и даже колдовством.

К сожалению (а, возможно, к счастью), мир несколько сложнее, чем колдовство. И первая сложность появляется при попытке понять, что подразумевается под «причиной».

### ***Что такое причинность?***

Причинность – это довольно спорный и активно обсуждаемый вопрос, что, вероятно, кажется удивительным, поскольку в реальной жизни все выглядит просто: мы что-то делаем, и это к чему-то приводит. Дверь машины зажала мой большой палец, и теперь он болит.

Но откуда мне знать, что большой палец не заболел бы в любом случае? Возможно, мы могли бы обратиться к тому, что называется **контрфактуальным** мышлением<sup>[88]</sup>. Если бы мой палец не зажала дверь, то он бы не болел. Но это всегда будет предположением, требующим переписывания истории, поскольку мы никогда точно не узнаем, что я мог бы почувствовать (хотя в данном случае я могу быть вполне уверен, что мой палец не заболит внезапно сам по себе).

Ситуация осложняется еще больше, когда мы начинаем учитывать неизбежную изменчивость, лежащую в основе событий в реальной жизни. Например, медицинское сообщество сейчас соглашается с тем, что курение вызывает рак легких, однако врачам потребовались десятилетия, чтобы прийти к такому заключению. Почему так долго? Потому что большинство курильщиков не заболевают раком легких, в то время как некоторые некурящие заболевают. Все, что мы можем сказать, – это то, что у вас выше риск заболеть раком легких, если вы курите, чем если не курите; и это одна из причин того, почему для принятия законов об ограничении курения понадобилось столько времени.

Таким образом, наша «статистическая» идея причинности не будет строго детерминистской. Когда мы говорим, что X обуславливает Y, мы не имеем в виду, что каждый раз, когда наступает X, наступает и Y. Мы всего лишь подразумеваем, что если вмешаемся и заставим X происходить чаще, то и Y будет случаться чаще. Соответственно, мы никогда не сможем сказать, что X вызывает Y в данном случае, а можем

лишь утверждать, что X увеличивает долю случаев, когда происходит Y. Из этого вытекают два важнейших следствия относительно того, что нам нужно делать при намерении понять причинно-следственную связь. Во-первых, чтобы вывести причинно-следственную связь с полной уверенностью, в идеале нам нужно вмешаться и провести эксперименты. Во-вторых, поскольку мир статистический и стохастический, вмешаться нужно не один раз, чтобы собрать доказательства.

Все это естественным образом подводит нас к очень деликатной теме – проведению клинических испытаний на больших группах людей. Мало кому понравится идея экспериментов над собой, особенно если речь идет о жизни и смерти. Это тем более примечательно, что тысячи людей изъявляли желание участвовать в масштабных исследованиях, в которых ни они, ни врачи не знали, какое лечение в итоге будет применено.

### **Уменьшают ли статины риск инфарктов и инсультов?**

Каждый день я принимаю маленькую белую таблетку – статин, потому что мне сказали, что он понижает уровень холестерина и тем самым уменьшает риск инфарктов и инсультов. Но как это сказывается на мне? Я почти уверен, что эти таблетки снижают уровень холестерина липопротеинов низкой плотности (ЛПНП)<sup>[89]</sup>, поскольку мне сообщили, что он упал вскоре после того, как я начал их принимать. Снижение ЛПНП – непосредственный, по сути, детерминированный эффект, который, как я полагаю, вызван приемом статина.

Однако я никогда не узнаю, принесет ли мне этот ежедневный ритуал пользу в долгосрочной перспективе; все зависит от того, какой из многочисленных сценариев моей дальнейшей жизни будет на самом деле разыгран. Если инфаркта или инсульта у меня никогда не будет, то я так и не узнаю, в какой-то степени это результат приема таблеток, или их многолетнее глотание здесь ни при чем и просто оказалось напрасной тратой времени. Если инфаркт или инсульт все же случится, то я не узнаю, было ли это событие отложено благодаря приему статина. Все, что мне дано знать, – это то, что в среднем препарат

приносит пользу большой группе похожих на меня людей и что это знание основано на масштабных клинических испытаниях.

Цель клинических испытаний – провести «правильный тест», который верно определяет причинность и оценивает средний эффект нового медицинского метода лечения, и при этом избежать ошибок, которые могли бы дать ложное представление о его эффективности.

Правильное клиническое исследование в идеале должно соответствовать следующим принципам:

1. *Контроль.* При намерении изучить влияние статинов на популяцию мы не можем просто дать их нескольким добровольцам, а затем, если инфаркта не будет, заявить, что его удалось избежать благодаря приему таблеток (несмотря на наличие сайтов, которые используют подобные смехотворные рассуждения для продвижения своей продукции). Нам нужна **экспериментальная группа**, которой будут давать статины, и **контрольная группа**, принимающая сахарные таблетки или **плацебо**.

2. *Распределение при лечении.* Важно сравнивать подобное с подобным, поэтому и лечение, и группы сравнения должны быть максимально похожи. Лучший способ этого добиться – случайно распределить участников по группам, а потом наблюдать, что с ними происходит. Такой метод называется **рандомизированным контролируемым исследованием (РКИ)**. В тестировании статинов задействуется значительное количество людей, поэтому обе группы должны быть сходны по всем факторам, которые могли бы повлиять на результат, включая (что критически важно) *те, о которых мы не знаем*. Такие исследования могут быть весьма масштабными: в исследовании по защите сердца (HPS), проведенном в Великобритании в конце 1990-х годов, 20 536 человек с повышенным риском инфаркта или инсульта были случайным образом распределены на две группы: одним ежедневно давали 40 мг симвастатина, а другим – пустую таблетку<sup>[90]</sup>.

3. *Подсчет количества людей в обеих группах.* Люди, попавшие в группу «статинов» в исследовании по защите сердца, включались в итоговый анализ, *даже если не принимали свои таблетки*. Такой принцип называется «**анализ по назначенному лечению**» и может показаться довольно странным. Это означает, что итоговая оценка эффекта статинов в действительности измеряет эффект прописанных

статинов, а не фактически принимаемых. На практике, конечно, людям настоятельно рекомендовали пить таблетки в течение всего исследования, хотя через пять лет HPS 18 % тех, кому были прописаны статины, прекратили их принимать, в то время как целых 32 % тех, кому было назначено плацебо, в действительности начали принимать статины в ходе испытаний. Поскольку люди, изменяя лечение, как правило, размывают различия между группами, мы можем ожидать, что видимый эффект в анализе по назначенному лечению может быть меньше, чем эффект от реального приема препарата.

4. *Если возможно, люди не должны знать, в какую из двух групп входят.* В испытаниях статинов и настоящие препараты, и плацебо выглядели одинаково, поэтому участники не знали, что именно принимают<sup>[91]</sup>.

5. *Процедуры для групп должны быть одинаковыми.* Если бы группу, которая употребляла статины, чаще приглашали в больницу или более тщательно обследовали, то было бы невозможно разграничить пользу от применения препарата и от улучшенного ухода. В HPS персонал, наблюдавший за пациентами, не знал, кто из них принимает статины, а кто – плацебо.

6. *По возможности те, кто оценивает итоговые результаты, не должны знать, к какой группе относятся испытуемые:* полагая, что лечение помогает, врач может преувеличить пользу для экспериментальной группы, то есть допустить неосознанную ошибку.

7. *Измеряйте всех.* Нужно приложить максимум усилий, чтобы отследить всех участников, поскольку люди, бросившие исследование, могли, например, это сделать из-за побочных эффектов препарата. У HPS были замечательные 99,6 % полного наблюдения за всеми в течение пяти лет – эти результаты приведены в табл. 4.1.

#### **Таблица 4.1**

Результаты пяти лет исследования защиты сердца в соответствии с лечением, назначенным пациентам. Абсолютное снижение риска инфаркта составило  $11,8 - 8,7 = 3,1$  %. Таким образом, в группе из 1000 человек, принимавших статины, был предотвращен примерно 31 инфаркт. Это означает, что для предотвращения одного инфаркта примерно 30 человек должны принимать статины в течение пяти лет

Событие	Процентная доля в группе из 10 267 человек, которым назначили плацебо	Процентная доля в группе из 10 269 человек, которым назначили статины	Процентное (относительное) снижение риска у людей, которым назначили статины
Инфаркт	11,8	8,7	27%
Инсульт	5,7	4,3	25%
Смерть по любой причине	14,7	12,9	13%

Те, кто попал в группу, принимавшую статины, явно в среднем имели лучшие показатели здоровья, а поскольку пациенты распределялись случайным образом и в остальном лечились одинаково, результат можно считать следствием приема статинов. Однако мы видели, что многие люди на самом деле не придерживались назначенного лечения, и это приводит к некоторому размыванию разницы между группами: специалисты, проводившие HPS, оценивают реальный эффект от приема статинов примерно на 50 % выше, чем показано в табл. 4.1.

Два важных итоговых замечания:

1. *Не полагайтесь на одно исследование.* Один отдельный эксперимент может нам сказать, что лекарство работало в определенной группе в определенном месте, но надежные выводы требуют нескольких исследований.

2. *Систематически проверяйте доказательства.* При рассмотрении нескольких испытаний обязательно включайте каждое

проведенное исследование, создавая таким образом систематический обзор. Затем результаты можно формально объединить в **метаанализ**.

Например, недавний систематический обзор собрал данные двадцати семи рандомизированных контролируемых исследований статинов, в которых участвовало более 170 тысяч человек с пониженным риском сердечно-сосудистых заболеваний<sup>[92]</sup>. Но вместо того чтобы фокусироваться на разнице между экспериментальной (принимающей статины) и контрольной группами, оценивался эффект от снижения уровня ЛПНП. По сути, исследователи предположили, что эффект статинов достигается посредством изменения липидов в крови, и основывали свои расчеты на среднем уменьшении ЛПНП, установленном в каждом из испытаний, учитывающем любое несоблюдение назначенного лечения. Такое дополнительное предположение позволило оценить эффект от фактического приема статинов. Ученые пришли к выводу, что снижение уровня ЛПНП на 1 ммоль/л (миллимоль на литр) уменьшает риск серьезных проблем (в том числе преждевременную смерть) с сердечно-сосудистой системой на 21 %. Мне, например, этого достаточно, чтобы продолжать принимать такие таблетки<sup>[93]</sup>.

Мы проигнорировали вероятность того, что любая наблюдаемая связь необязательно является причинно-следственной, а может быть просто результатом случайности. Большинство лекарственных препаратов на рынке обладают лишь умеренным воздействием и помогают только меньшинству принимающих их людей; их общую полезность можно точно выявить исключительно в рамках крупных тщательных рандомизированных исследований. Испытания статинов довольно-таки масштабны, особенно когда они объединены в метаанализ, а значит, полученные результаты нельзя объяснить простым случайным отклонением. (Мы узнаем, как это проверить, из [главы 10](#).)

### **Эффективна ли молитва?**

Список принципов РКИ не нов: почти все они были введены в 1948 году в эксперименте, который считается первым правильным клиническим испытанием. Тогда исследовался стрептомицин — лекарство, предназначенное для борьбы с туберкулезом. Конечно, было

бы слишком безнравственно случайным образом определять, кого лечить, а кого оставить без потенциально спасающего жизнь препарата. Однако принятию столь трудного решения способствовал тот факт, что имеющегося в то время в Великобритании лекарства в любом случае не хватило бы на всех, поэтому случайный выбор казался вполне справедливым и этически обоснованным. Но даже по прошествии стольких лет и тысяч проведенных РКИ общество все еще может удивиться, узнав, что медицинские решения о том, какое лечение рекомендовать человеку (даже такие драматичные, как радикальная мастэктомия или лампэктомия при раке молочной железы<sup>[94]</sup>), фактически принимаются путем подбрасывания монеты (пусть это и метафорическая монета, воплощенная в генераторе случайных чисел в компьютере)<sup>[95]</sup>.

На практике процесс назначения лечения в испытаниях гораздо сложнее, чем простая рандомизация в каждом случае, так как мы хотим убедиться, что все типы людей одинаково представлены в группах, получающих различные виды лечения. Например, мы можем захотеть, чтобы количество пожилых людей с повышенным риском, принимающих статины и плацебо, распределялось примерно поровну. Эта схема позаимствована из сельскохозяйственных экспериментов, где многие идеи рандомизированных исследований возникли в основном благодаря работе Рональда Фишера (о котором мы расскажем чуть позже). Например, большое поле делится на отдельные участки, а затем для каждого участка случайным образом выбирается удобрение – так же как люди случайным образом получают назначение на лечение. Но части поля могут различаться по дренажу, затененности и другим признакам, поэтому предварительно поле нужно разделить на блоки, содержащие примерно сходные участки, а уже затем осуществить рандомизацию, чтобы в каждом блоке было равное число участков с тем или иным удобрением. В таком случае способы обработки земли будут сбалансированными, скажем на заболоченных участках одинаково применяют все виды удобрений.

Например, однажды я работал над рандомизированным испытанием, где сравнивались два альтернативных метода лечения грыжи: стандартная «открытая» операция и лапароскопия (операция с минимальным вмешательством). Предполагалось, что мастерство хирургической бригады во время испытаний может возрастать, поэтому

было важно, чтобы в течение всего исследования эти два метода были сбалансированы. Поэтому я разбил цепочку пациентов на блоки по 4 и 6 человек, а затем случайно распределял их внутри каждого блока по методам операции. Используемые методы были напечатаны на листочках бумаги, которые я сложил и поместил в пронумерованные непрозрачные коричневые конверты. Помню, как я наблюдал за больными, лежащими на предоперационной каталке, понятия не имея, какую именно операцию им будут делать, в то время как анестезиолог открывал конверт и узнавал, что с ними случится дальше, в частности вернутся ли они домой с одним большим шрамом или несколькими точечными проколами.

Рандомизированные испытания стали золотым стандартом тестирования новых медицинских методов, а теперь все чаще используются и при оценке эффективности новых методик в сфере образования или правоохранительной деятельности. Например, британская организация Behavioural Insights Team<sup>[96]</sup> случайным образом отобрала половину школьников, пересдающих экзамены по математике и английскому языку, и регулярно отправляла им поощряющие текстовые сообщения, чтобы поддержать в учебе. В результате доля сдавших экзамены среди тех, кто имел такую поддержку, оказалась на 27 % выше. Эта же группа исследователей наблюдала и ряд положительных эффектов в рандомизированном испытании видеокамер, закрепленных на теле полицейских, – к примеру, снижение количества остановленных и безосновательно обысканных<sup>[97]</sup>.

Проводились даже эксперименты для определения эффективности молитвы. Например, в рамках исследования терапевтических эффектов ходатайственной молитвы (STEP) свыше 1800 пациентов с шунтированием сердца случайным образом разделили на три группы: за пациентов 1-й и 2-й групп, соответственно, молились и не молились, но при этом они не знали, молятся за них или нет, а вот члены 3-й группы знали, что за них молятся. Единственным заметным эффектом было незначительное *увеличение* осложнений в группе, где знали, что за них молятся. Один из исследователей прокомментировал это так: «Возможно, это заставило их сомневаться и задаться вопросом: “Неужели я настолько болен, что им пришлось вызвать свою молитвенную команду?”»<sup>[98]</sup>

Основное из последних нововведений в рандомизированных экспериментах – А/В-тестирование в веб-дизайне<sup>[99]</sup>, при котором пользователей направляют на различные варианты веб-страницы (о чем они не знают). Далее измеряется количество времени, проведенного на том или ином варианте страницы, переходов по рекламным объявлениям и так далее. Серия А/В-тестов может быстро привести к оптимальному дизайну, а огромные размеры выборки означают, что даже небольшие, но потенциально выгодные эффекты гарантированно обнаружатся. Следовательно, совершенно новое сообщество людей должно было узнать о тонкостях пробных испытаний, в том числе о рисках при проведении множественных сравнений, которые мы рассмотрим в главе 10.

### ***Что делать, если рандомизация невозможна?***

#### **Почему у стариков большие уши?**

Легко провести рандомизацию, когда нужно, скажем, изменить сайт: можно без проблем найти участников, поскольку они даже не знают, что участвуют в эксперименте, и нет никаких этических проблем в использовании их в качестве подопытных кроликов. Однако иногда осуществить рандомизацию не просто трудно, а невозможно: мы не можем проверять влияние привычек, например, заставляя людей в рамках исследования курить или употреблять нездоровую пищу (даже если такие эксперименты проводятся на животных). Когда данные появляются не в результате эксперимента, а просто из наблюдений, их называют наблюдательными (а соответствующие исследования – наблюдательными, или обсервационными). Поэтому часто наша задача – постараться как можно лучше отделить корреляцию от причинно-следственной связи, применяя к наблюдательным данным статистические принципы и хороший план исследования в сочетании со здоровой дозой скептицизма.

Вопрос об ушах стариков, возможно, не так важен, как многие другие темы в этой книге, но он иллюстрирует необходимость выбора плана исследования, который подойдет для ответа на вопросы. Если мы обратимся к подходу на основе цикла РРДАС, то *проблема* строится на моем личном наблюдении, что у стариков, похоже, действительно слишком большие уши. Но почему? Очевидный *план* – посмотреть,

коррелирует ли в генеральной совокупности возраст с длиной ушей взрослых людей. Как оказалось, группа медиков-исследователей в Великобритании и Японии собрала *данные* в таком **поперечном исследовании**: их *анализ* показал явную положительную корреляцию, и они пришли к *заклучению*, что длина ушей связана с возрастом<sup>[100]</sup>.

Теперь наша задача – попытаться объяснить такую связь. Уши продолжают расти с возрастом? Или у нынешних пожилых людей они всегда были большими, а из-за каких-то событий, произошедших за последние десятилетия, у предыдущих поколений уши меньше? Или же люди с ушами меньшего размера просто умирают раньше по каким-то причинам, ведь существует же у китайцев поверье, что большие уши предсказывают долгую жизнь. Чтобы придумать, какие исследования могли бы проверить такие идеи, нужно определенное воображение. В **перспективном когортном исследовании** участники измеряли бы уши всю свою жизнь, проверяя, не растут ли они, или не умирают ли раньше люди с небольшими ушами. Но это требует много времени, поэтому можно применить альтернативу – **ретроспективное когортное исследование**, то есть взять нынешних стариков и попытаться выяснить, выросли ли у них уши, например, с помощью старых фотографий. **Исследование типа «случай-контроль»** могло бы к уже умершим людям подобрать живущих, которые соответствуют им по возрасту и прочим факторам, связанным (по нашим сведениям) с долголетием, и посмотреть, больше ли уши у тех, кто прожил дольше<sup>[101]</sup>.

А затем цикл решения задачи запустится снова.

### ***Что мы можем сделать, наблюдая какую-то связь?***

Именно здесь требуется определенное статистическое воображение, и попытка догадаться о причинах того, почему наблюдаемая корреляция может быть ложной, обещает стать приятным упражнением. Некоторые причины довольно просты: значительная корреляция между потреблением моцареллы и числом инженеров, по-видимому, обусловлена тем, что обе категории увеличиваются со временем. Точно так же любые корреляции между продажами мороженого и числом утонувших зависят от погоды. Когда видимую связь между двумя величинами можно объяснить наличием какого-то наблюдаемого внешнего фактора, влияющего на обе величины, его

называют **возмущающим, или искажающим фактором**. И год, и погода – это потенциальные возмущающие факторы, которые можно регистрировать и учитывать при анализе.

Простейший метод работы с возмущающим фактором – посмотреть на видимые связи при каждом его уровне. Это называется **поправкой**, или стратификацией. Например, мы могли бы изучить связь между продажами мороженого и числом утонувших в дни с примерно одинаковой температурой воздуха.

Однако поправка может привести к некоторым парадоксальным результатам, как показал анализ процента зачисления абитуриентов в Кембриджский университет для обоих полов в 1996 году. Общая доля поступивших на пять учебных дисциплин в Кембридже была чуть выше у мужчин (24 % из 2470 абитуриентов), чем у женщин (23 % из 1184 абитуриенток). Это те дисциплины, которые сегодня принято обозначать аббревиатурой НТИМ (STEM) – наука, технологии, инженерия и медицина<sup>[102]</sup>, то есть предметы, исторически изучаемые преимущественно мужчинами. Была ли тут гендерная дискриминация?

Внимательно посмотрите на табл. 4.2. Хотя общий процент зачисления выше у мужчин, на каждую отдельную дисциплину он выше у женщин. Как мог возникнуть такой парадокс? Объяснение заключается в том, что женщины чаще подавали заявления на более популярные, конкурентные дисциплины – медицину и ветеринарию, и реже – на инженерию, где у них более высокий процент поступления. Поэтому мы можем заключить, что никаких подтверждений дискриминации нет.

### **Таблица 4.2**

Иллюстрация парадокса Симпсона на примере данных о поступлении в Кембриджский университет в 1996 году. Общий процент зачисленных абитуриентов выше у мужчин, однако процент зачисления на каждую дисциплину отдельно выше у женщин

	Женщины			Мужчины		
	Подали	Поступили	%	Подали	Поступили	%
<b>Информатика</b>	26	7	27	228	58	25
<b>Экономика</b>	240	63	26	512	112	22
<b>Инженерия</b>	164	52	32	972	252	26
<b>Медицина</b>	416	99	24	578	140	24
<b>Ветеринария</b>	338	53	16	180	22	12
<b>ВСЕГО</b>	1184	274	23	2470	584	24

Описанная ситуация известна как **парадокс Симпсона**<sup>[103]</sup>, который возникает, когда видимое направление взаимосвязи становится обратным с учетом возмущающего фактора. В результате вывод, извлеченный из данных, становится противоположным. Статистики наслаждаются поиском подобных примеров в реальной жизни, так как каждый из них подчеркивает, насколько осторожно нужно обращаться с наблюдательными данными. Тем не менее такие случаи показывают идеи, возникающие при разделении данных по факторам, которые могут объяснить наблюдаемые связи.

### **Добавляет ли близость к супермаркету Waitrose 36 тысяч фунтов к стоимости вашего дома?**

В 2017 году британские СМИ опрометчиво опубликовали заявление, что соседство с супермаркетом сети Waitrose «добавляет 36 тысяч фунтов к цене дома»<sup>[104]</sup>. Однако это было не исследование

изменения цен на жилье в связи с открытием нового супермаркета, и Waitrose, конечно же, не размещает свои магазины случайным образом: эти данные всего лишь корреляция между ценами на жилье и близостью супермаркетов, особенно таких высококлассных, как Waitrose.

Эта корреляция практически наверняка отражает политику Waitrose по открытию магазинов в более богатых районах, а потому представляет собой прекрасный пример того, что фактическая причинно-следственная связь является полной противоположностью заявлению в газете. Неудивительно, что она называется **обратной причинной зависимостью**. Более серьезные примеры встречаются в исследованиях, изучающих взаимосвязь между употреблением алкоголя и состоянием здоровья: как правило, показатель смертности у непьющих людей в целом существенно выше, чем у умеренно пьющих. Как это понимать, учитывая, что мы знаем о влиянии алкоголя на печень? Частично это объясняется обратной причинной зависимостью: люди, которые умирают с более высокой вероятностью, не пьют, потому что уже больны (возможно, из-за чрезмерного употребления алкоголя в прошлом). Сегодня более тщательный анализ исключает бывших алкоголиков, а также игнорирует неблагоприятные для здоровья события, происходящие в первые несколько лет исследования, поскольку они могут быть результатом предыдущих условий. Однако даже при таких исключениях некоторая общая польза для здоровья от умеренного употребления алкоголя, похоже, остается, хотя и активно оспаривается.

Еще одно забавное упражнение – попробовать сочинить историю с обратной причинной зависимостью для любого статистического заявления, основанного исключительно на корреляции. Моя любимая история – о корреляции между потреблением безалкогольных газированных напитков американскими подростками и их склонностью к насилию. Хотя одна газета преподнесла это так: «Газированные напитки делают подростков жестокими»<sup>[105]</sup>, что, скорее всего, так же правдоподобно, как и утверждение, что насилие вызывает жажду. Или, что более правдоподобно, мы могли бы придумать некие общие факторы, влияющие на обе величины, например принадлежность к какой-то группе сверстников. Потенциальные причины, которые мы не измеряем, называются **скрытыми факторами**, поскольку они остаются

на заднем плане, не входят в поправки и только и ждут подходящего момента, чтобы опровергнуть наивные выводы из наблюдательных данных.

Вот еще несколько примеров того, как легко поверить в наличие причинно-следственной связи, хотя на самом деле на события влияет посторонний фактор.

- У многих детей диагностируется аутизм после вакцинации. Вызывает ли вакцинация аутизм? Нет, но эти события возникают примерно в одном возрасте, а потому неизбежны случайные совпадения.

- Среди ежегодно умирающих людей доля левшей меньше, чем во всей популяции. Означает ли это, что левши живут дольше? Нет, это происходит потому, что те, кто умирает сейчас, родились во времена, когда детей насильственно переучивали пользоваться правой рукой, поэтому пожилых левшей меньше<sup>[106]</sup>.

- Средний возраст смерти римских пап выше, чем в среднем в популяции. Означает ли это, что избрание папой помогает жить дольше? Нет, просто пап выбирают из группы людей, которые не умерли молодыми (в противном случае они не были бы кандидатами)<sup>[107]</sup>.

Миллионы способов, которыми нас можно поймать в ловушку, могут подтолкнуть к мысли, что причинно-следственную связь можно выявить только в рандомизированном эксперименте. Но, по иронии судьбы, эту точку зрения опроверг человек, отвечавший за первое современное рандомизированное клиническое исследование.

### ***Как сделать заключение о причинной связи по наблюдательным данным?***

Остин Брэдфорд Хилл был блестящим британским прикладным статистиком, находящимся в авангарде двух изменивших мир научных достижений: он разработал упомянутые ранее клинические испытания стрептомицина, которые фактически установили стандарты для всех последующих РКИ, и провел вместе с Ричардом Доллом в 1950-х годах исследование, по сути, подтвердившее связь между курением и раком легких. В 1965 году он изложил список критериев, которые необходимо

учесть, прежде чем делать вывод о том, что наблюдаемая связь между **воздействием** и результатом является причинной. Под воздействием здесь понимается что угодно – от химических веществ в окружающей среде до таких привычек, как курение или недостаточное количество физических упражнений.

Впоследствии эти критерии широко обсуждались. Представленная ниже версия разработана Джереми Хоуиком с коллегами, которые выделили в ней прямые, механистические и параллельные доказательства<sup>[108]</sup>.

Прямое доказательство:

1. Масштаб эффекта настолько велик, что его нельзя объяснить разумными возмущающими факторами.

2. Существует соответствующая временная и/или пространственная близость, когда причина предшествует эффекту, а эффект возникает через разумный интервал, и/или причина происходит в том же месте, что и эффект.

3. *Чувствительность к дозе и обратимость*: при увеличении воздействия эффект увеличивается. Подтверждение еще сильнее, если эффект уменьшается при уменьшении дозы.

Механистическое доказательство:

4. Существует правдоподобный механизм действия, который может быть биологическим, химическим или механическим, с внешним подтверждением «причинно-следственной цепочки».

Параллельное доказательство:

5. Эффект соответствует тому, что уже известно.

6. Эффект обнаруживается при повторном воспроизведении исследования.

7. Эффект выявляется в аналогичных, но не идентичных исследованиях.

Такие принципы позволяют выявить причинно-следственную связь в наборе разрозненных данных даже при отсутствии рандомизированных испытаний. Например, установлено, что при втирании аспирина в ротовой полости (например, для облегчения

зубной боли) образуются язвочки. Эффект силен (удовлетворяет критерию 1), происходит при втирании (2), является правдоподобной реакцией на кислотный компонент препарата (3), не противоречит современным научным данным и аналогичен известному эффекту, при котором аспирин вызывает язву желудка (4), а также регулярно наблюдается у различных пациентов (5). Итого соблюдены пять из семи критериев, оставшиеся два не проверялись, поэтому вполне резонно заключить, что мы имеем дело с истинной побочной реакцией на препарат.

Критерии Брэдфорда Хилла применяются к общим научным заключениям, касающимся генеральных совокупностей. Но нас могут интересовать и отдельные случаи – например, в гражданской тяжбе, когда суду нужно решить, привело ли определенное воздействие (скажем, наличие асбеста на работе) к негативному результату для конкретного лица (например, к раку легких у Джона Смита). Никогда нельзя установить с полной уверенностью, что асбест стал причиной рака, поскольку невозможно доказать, что без асбеста рак не развился бы. Тем не менее некоторые суды признают, что по «принципу большей вероятности» прямая причинная связь установлена, если показатель относительного риска, связанного с воздействием, превосходит 2. Но почему именно 2?

Предположительно аргументация этого решения такова:

1. Допустим, при обычных обстоятельствах из 1000 человек, подобных Джону Смигу, раком легких заболеют 10. Если асбест повышает риск более чем вдвое, то при его воздействии на 1000 человек наблюдалось бы, возможно, 25 случаев рака.

2. Таким образом, среди тех, кто подвергался воздействию асбеста и получил рак легких, меньше половины заболели бы раком при отсутствии асбеста.

3. Следовательно, более половины случаев рака в этой группе вызваны асбестом.

4. А поскольку Джон Смит принадлежит к этой группе, по принципу большей вероятности его рак вызван асбестом.

Подобные рассуждения привели к появлению новой области науки – **судебной эпидемиологии**, которая пытается использовать сведения, полученные из общей популяции, для заключения о том, что могло стать причиной конкретных событий. По сути, эта дисциплина обязана своим появлением людям, требующим компенсаций, но это очень интересная область для статистических рассуждений о причинно-следственных связях.

Правильная трактовка причинно-следственной связи по-прежнему остается предметом жарких споров в статистике, неважно, касается это фармацевтических препаратов или больших ушей, и без рандомизации редко удается сделать надежные выводы. Один творческий подход основывается на том, что многие гены распределяются по популяции фактически случайным образом, поэтому мы как будто рандомизированы при зачатии – получилась вот такая версия. Этот подход известен как менделевская рандомизация, названная так в честь Грегора Менделя, автора учения о наследственности<sup>[109]</sup>.

Чтобы максимально учесть все возмущающие факторы и тем самым приблизиться к оценке реального эффекта воздействия, были разработаны и другие сложные статистические методы, в значительной степени основанные на важной идее регрессионного анализа. И за это мы снова должны поблагодарить богатое воображение Фрэнсиса Гальтона.

### **Выводы**

- Причинность в статистическом анализе означает, что при нашем вмешательстве шансы различных исходов изменяются по определенной системе.
- Причинно-следственную связь трудно установить статистически, однако хорошо спланированные рандомизированные исследования – наилучшая возможная схема.
- Принципы слепоты, анализа по назначенному лечению и прочего позволяют проводить масштабные клинические исследования для выявления умеренных, но важных эффектов.
- Наблюдательные данные могут отличаться фоновыми факторами, которые влияют на видимые наблюдаемые взаимосвязи между воздействием и результатом. Они могут

оказаться либо наблюдаемыми возмущающими, либо скрытыми факторами.

- Для поправки на прочие факторы существуют специальные статистические методы, однако всегда требуется уточнение о степени уверенности, с которой можно говорить о предполагаемой причинно-следственной связи.

## Глава 5. Моделирование зависимости с помощью регрессии

Идеи, изложенные в предыдущих главах, позволяют визуализировать и находить характеристики данных, а также рассматривать зависимости между парами переменных. Эти базовые методы могут помочь нам проделать будущий довольно долгий путь, хотя в целом современные данные намного сложнее. Часто появляется список переменных, возможно, имеющих отношение к вопросу, одна из них нам особенно интересна при объяснении или прогнозировании, будь то риска развития рака для одного человека или будущего жителей целой страны. В этой главе мы познакомимся с идеей **статистической модели** – формальным представлением взаимоотношений между переменными, – которую сможем использовать для желаемого объяснения или прогноза. Это означает неизбежное появление определенных математических идей, однако базовые концепции должны быть понятны без применения алгебры.

Но прежде вернемся к Фрэнсису Гальтону. Он был одержим идеей сбора данных (что характерно для классического джентльмена-ученого Викторианской эпохи), и обращение к мудрости толпы для выяснения массы быка – это всего лишь один из примеров. Он использовал свои наблюдения для составления прогнозов погоды, оценки эффективности молитвы и даже для сравнения относительной красоты молодых женщин в различных частях страны<sup>[110]</sup>. Он также разделял интерес своего двоюродного брата Чарльза Дарвина к наследственности и намеревался изучить способы изменения личных характеристик людей от поколения к поколению. В частности, его заинтересовал такой вопрос:

**Как предсказать будущий рост детей по росту их родителей?**

В 1886 году Гальтон опубликовал данные о росте большой группы родителей и их взрослых детей. Характеристики выборки приведены в табл. 5.1<sup>[111]</sup>. Рост людей в выборке Гальтона близок к росту

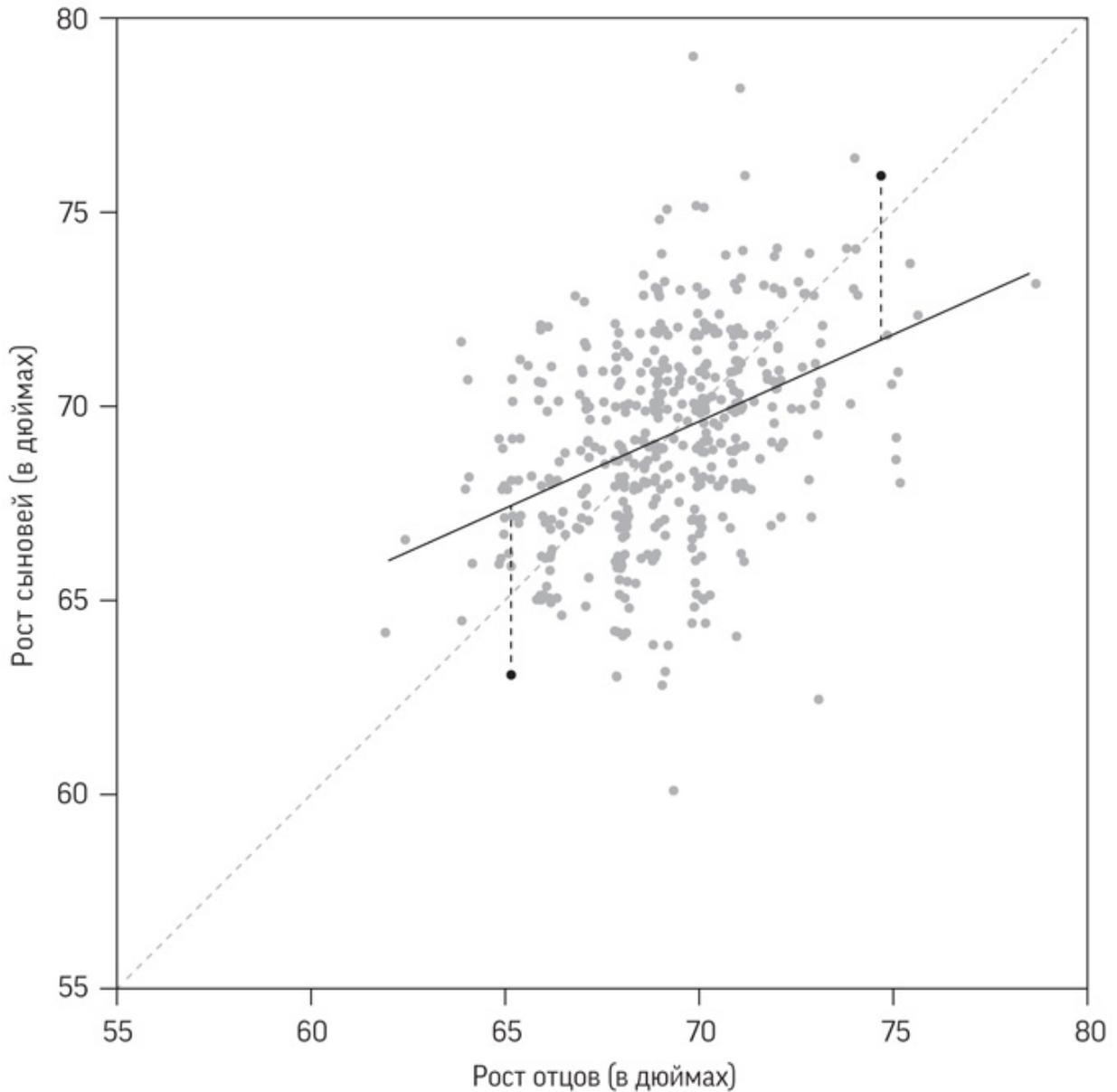
современных взрослых (как сообщалось, средний рост взрослых женщин и мужчин в Соединенном Королевстве в 2010 году составлял примерно 63 и 69 дюймов [160 и 175 сантиметров соответственно]), что говорит о том, что эти люди хорошо питались и обладали высоким социально-экономическим статусом<sup>[112]</sup>.

### Таблица 5.1

Характеристики выборки роста (в дюймах) 197 родительских пар и их взрослых детей, по данным Гальтона 1886 года. Для справки: 64 дюйма = 163 см, 69 дюймов = 175 см. Даже без построения диаграммы близость выборочных средних и медианы позволяет предположить, что распределение симметрично

	Количество	Выборочное среднее	Медиана	Стандартное отклонение
Матери	197	64,0	64,0	2,4
Отцы	197	69,3	69,5	2,6
Дочери	433	64,1	64,0	2,4
Сыновья	465	69,2	69,2	2,6

Рис. 5.1 – это точечная диаграмма, где отображен рост 465 сыновей в зависимости от роста их отцов. Между ростом отцов и сыновей четко прослеживается корреляция, при этом коэффициент корреляции Пирсона равен 0,39. Как нам поступить, если мы хотим предсказать рост сына по росту его отца? Начать можно с построения прямой линии для прогноза: она позволит указать рост сына по росту отца. Первая мысль – провести линию точно «по диагонали», то есть при таком прогнозе рост сына будет точно таким же, как и у отца. Однако, оказывается, есть способ лучше.



**Рис. 5.1**

Точки отображают рост отцов и их сыновей, по данным Гальтона (многие отцы повторяются, потому что у них несколько сыновей). Для разделения точек добавлен случайный разброс, а диагональная пунктирная линия демонстрирует точное равенство между ростом отцов и сыновей. Сплошная линия – стандартная «прямая наилучшего соответствия» (регрессионная прямая). У каждой точки есть «остаток» (вертикальные пунктирные линии) – разность между наблюдаемым значением и значением, которое предсказывает регрессионная модель

Какую бы прямую мы ни выбрали, у любой точки данных будет **остаток** (вертикальные пунктирные линии на диаграмме), который представляет собой величину допустимой ошибки при использовании для прогноза этой линии. Нам нужна линия, которая делает эти остатки маленькими, и стандартный способ ее провести – это выбор прямой по **методу наименьших квадратов**, то есть прямой, для которой сумма квадратов всех остатков будет наименьшей<sup>[113]</sup>. Уравнение для такой прямой получить несложно (см. [гlossарий](#)); этот метод разработан одновременно, но независимо друг от друга французскими математиками Адриеном-Мари Лежандром и Карлом Фридрихом Гауссом в конце XVIII века. Прямую часто называют прямой наилучшего соответствия, и с ее помощью определяется лучший прогноз, который мы можем сделать для роста сына, зная рост его отца.

Линия, построенная по методу наименьших квадратов на рис. 5.1, проходит через середину облака точек, отражая средние значения роста для отцов и сыновей, но не совпадая с диагональю, отображающей «равенство». Она ниже диагонали у отцов выше среднего и выше диагонали у отцов ниже среднего роста. Это означает, что у высоких отцов сыновья в среднем ниже их, а у низкорослых – в среднем выше их. Гальтон назвал это явление «регрессией<sup>[114]</sup> к посредственности», а позднее оно стало именоваться «**регрессией к среднему значению**», или «**регрессом к среднему**». Аналогичный феномен отмечается и для матерей и дочерей: дочери более высоких матерей в среднем ниже их, а низкорослых – в среднем выше. Это объясняет происхождение термина в названии главы: со временем любая стохастическая зависимость, определяемая по данным, стала называться регрессией.

В регрессионном анализе зависимой переменной (или переменной отклика) называется величина, которую мы хотим предсказать или объяснить; обычно ее откладывают по вертикальной оси  $y$ . Независимая переменная (или объясняющая переменная) – это величина, которую мы используем для прогноза или объяснения; обычно она откладывается по горизонтальной оси  $x$ . Наклон (точнее, угловой коэффициент) регрессионной прямой называется **коэффициентом регрессии**.

Табл. 5.2 показывает корреляцию между ростом родителей и потомков, а также наклон для регрессионных прямых<sup>[115]</sup>. Существует

простая зависимость между угловыми коэффициентами, коэффициентом корреляции Пирсона и стандартными отклонениями и переменными<sup>[116]</sup>. В реальности если у зависимой и независимой переменной среднеквадратичные отклонения одинаковы, то угловой коэффициент просто совпадает с коэффициентом корреляции Пирсона, что и объясняет их сходство в табл. 5.2.

**Таблица 5.2**

Коэффициенты корреляции между ростом взрослых детей и родителей того же пола, а также коэффициенты регрессии для роста детей по отношению к росту родителей

	<b>Коэффициент корреляции Пирсона</b>	<b>Коэффициент регрессии для детей и родителей</b>
Матери и дочери	0,31	0,33
Отцы и сыновья	0,39	0,45

Смысл углового коэффициента полностью зависит от наших предположений о взаимосвязи между изучаемыми переменными. Для корреляционных данных угловой коэффициент показывает, какое среднее изменение зависимой переменной можно ожидать, если значение независимой переменной изменится на единицу. Например, если Алиса на 1 дюйм выше Бетти, то мы можем предсказать, что взрослая дочь Алисы будет на 0,33 дюйма выше, чем взрослая дочь Бетти. Конечно, мы не ожидаем, что этот прогноз будет соответствовать их истинной разнице в росте, но это наилучшее предположение, которое мы можем сделать исходя из имеющихся данных.

Однако если мы предполагаем *причинно-следственную* связь, то у углового коэффициента будет совершенно иная интерпретация: это изменение, которого мы можем ожидать в зависимой переменной, если вмешаемся и изменим значение независимой переменной на единицу. Это точно не относится к примеру с ростом, так как рост нельзя

изменить экспериментальным путем (по крайней мере, для взрослых). Даже с учетом описанных выше критериев Брэдфорда Хилла статистики, как правило, неохотно признают причинно-следственную связь без проведения эксперимента. Впрочем, некоторые исследователи, включая Джуда Перла, добились значительного прогресса в построении моделей причинной регрессии для наблюдательных данных [\[117\]](#).

### *Линии регрессии – это модели*

Линия регрессии для роста отцов и сыновей – очень простой пример статистической модели. Федеральная резервная система США определяет модель как «представление некоторого аспекта мира, основанное на упрощающих предположениях»: по сути, какое-нибудь явление представляется в математической форме, встраивается в программное обеспечение, а затем создается упрощенная «воображаемая» версия реальности [\[118\]](#).

У статистических моделей есть два основных компонента. Первый – это математическая формула, которая выражает детерминистский, предсказуемый компонент, например формула прямой линии, позволяющая нам делать прогноз о росте сына, зная рост его отца. Однако такая детерминистская часть модели не будет идеальным отображением реального мира. Как мы видели на [рис. 5.1](#), рост весьма сильно рассеян вокруг регрессионной прямой. Разница между тем, что предсказывает модель, и тем, что происходит на самом деле, – второй компонент модели, известный как **остаточная ошибка**, хотя важно помнить, что в статистическом моделировании термин «ошибка» означает не какой-то просчет, а неизбежную неспособность модели точно представить наблюдаемый мир. Поэтому в целом мы можем считать, что

наблюдение = детерминистская модель + остаточная  
ошибка.

Эта формула может быть истолкована как утверждение, что в статистическом мире то, что мы видим и измеряем, можно рассматривать как сумму систематической математической

идеализированной формы и некоего случайного компонента, который пока нельзя объяснить. Классическая идея сигнала и шума.

### **Уменьшают ли камеры контроля скорости количество ДТП?**

Этот раздел содержит простой урок: тот факт, что мы что-то делаем и что-то меняется, сам по себе не означает, что мы несем ответственность за результат. Похоже, людям трудно уловить эту простую истину, мы всегда стараемся придумать какое-то объяснение, причем гораздо охотнее, если лично находимся в центре ситуации. Конечно, иногда такая интерпретация верна: если вы щелкнете выключателем и зажжется свет, то обычно ответственны вы. Но иногда ваши действия определенно не отвечают за результат: если вы не взяли зонтик, а пошел дождь, в том нет вашей вины (хотя может показаться, что это так). Однако порой последствия наших действий менее ясны. Предположим, у вас болит голова, вы приняли аспирин, и боль прошла. Но откуда вам знать, что она не прекратилась бы, если бы вы не принимали таблетку?

У нас сильная психологическая склонность приписывать перемены какому-нибудь вмешательству, и это делает сравнения «до и после» ненадежными. Классический пример относится к камерам контроля скорости, которые обычно размещают в местах повышенной аварийности. Последующее снижение аварийности приписывают наличию камер. Но разве не понизился бы этот уровень в любом случае?

Полосы удач и неудач не бесконечны, и в конце концов все возвращается на круги своя – это тоже можно воспринимать как регресс к среднему, как у высоких отцов, имеющих более низких в среднем сыновей. Но когда мы убеждены, что полосы везения-невезения отражают постоянное состояние дел, мы ошибочно будем рассматривать возврат к нормальному состоянию как следствие какого-либо нашего вмешательства. Возможно, все это кажется вам очевидным, но эта простая идея имеет примечательные последствия:

- Преемникам футбольных тренеров, уволенных после череды неудач, лавры достаются всего лишь за возврат к нормальному состоянию.

- Управляющие фондами ухудшают показатели эффективности работы после нескольких успешных лет (и вероятного получения хороших бонусов).
- «Проклятие обложки Sport Illustrated»: как только спортсмены за свои достижения попадают на обложку этого известного журнала, их результаты резко ухудшаются.

Удача играет немалую роль в положении спортивных команд в турнирной таблице, а регресс к среднему означает, что можно ожидать того, что команды, преуспевающие в этом году, ухудшат игру в следующем, а плохо играющие сегодня улучшат ситуацию в будущем, особенно если команды примерно равны по силам. Напротив, видя такие изменения, мы можем подозревать, что работает регресс к среднему, и не стоит слишком обращать внимание на заявления о влиянии, скажем, новых методов тренировок.

В турнирных таблицах ранжируются не только спортивные команды. Рассмотрим таблицы Международной программы по оценке образовательных достижений учащихся (PISA), где сравниваются результаты преподавания математики в школах разных стран. Изменение положения страны в сводной таблице за 2003–2012 годы отрицательно коррелирует с начальным ее положением в списке. Иными словами, страны из верхней части списка имеют тенденцию опускаться, а из нижней части – подниматься: коэффициент корреляции равен  $-0,60$ . Однако определенные теоретические выкладки показывают, что если бы ранжирование было полностью случайным и действовал исключительно регресс к среднему, то для коэффициента корреляции можно было бы ожидать значения  $-0,71$ , которое не очень отличается от реально наблюдаемого<sup>[119]</sup>. Это говорит о том, что различия между странами меньше, чем заявляется, и что у изменений в таблицах мало общего с изменениями в методиках преподавания.

Регресс к среднему проявляется и в клинических испытаниях. В предыдущей главе мы узнали, что для правильной оценки нового лекарственного препарата нужно провести рандомизированные испытания, поскольку даже люди из контрольной группы демонстрируют улучшение состояния – так называемый эффект плацебо. Это часто истолковывают так: простой прием сахарной пилюли (предпочтительно красной) на самом деле благотворно влияет

на здоровье человека. Но значительная часть улучшений у людей, не проходивших активного лечения, может приходиться на регресс к среднему, поскольку пациентов включают в испытания, когда у них есть определенные симптомы, а многие из симптомов пропали бы в любом случае. Так что если мы хотим знать подлинный эффект установки камер контроля скорости в местах повышенной аварийности, нам нужно подойти к этому вопросу так же, как к проверке фармацевтического препарата, и распределить камеры случайным образом. (После проведения таких исследований оказалось, что примерно две трети пользы от камер приходится на регресс к среднему<sup>[120]</sup>.)

### *Если у нас есть несколько независимых переменных*

Со времен первой работы Гальтона появилось множество расширений базовой идеи регрессии, в значительной степени подкрепленных современными вычислительными возможностями. Такие обобщения включают:

- несколько независимых (объясняющих) переменных;
- независимые переменные, которые не являются числами;
- зависимости, которые отображаются не прямыми, а кривыми, и гибко подстраиваются под закономерность в данных;
- зависимые переменные, которые не являются непрерывными.

В качестве примера наличия более одной независимой переменной рассмотрим, как рост сына или дочери зависит от роста их отца и матери. Теперь точки на диаграмме расположены в трех измерениях, и на странице книги изобразить это сложнее. Однако мы по-прежнему можем применить метод наименьших квадратов, чтобы вывести формулу, которая наилучшим образом предскажет рост потомства. В этом случае говорят о **множественной линейной регрессии**<sup>[121]</sup>. Когда у нас была всего одна независимая переменная, связь с зависимой переменной показывал наклон (угловой коэффициент) прямой, который одновременно интерпретировался как коэффициент в уравнении регрессии. Эту идею можно распространить на несколько независимых переменных.

В табл. 5.3 приведены результаты для семей Гальтона. Как можно интерпретировать показанные здесь коэффициенты? Прежде всего они являются частью формулы, которую можно использовать для прогнозирования роста взрослого потомства у конкретных отца и матери [122]. Одновременно они также иллюстрируют идею поправки для наблюдаемого отношения, учитывая третий, возмущающий фактор.

**Таблица 5.3**

Результаты множественной линейной регрессии для роста взрослого потомства в зависимости от роста их отцов и матерей. Отсекаемый на прямой отрезок (сдвиг) – это средний рост потомства (табл. 5.1). Коэффициенты линейной регрессии показывают прогнозируемое изменение в росте потомков при изменении среднего роста родителей на 1 дюйм

Зависимая переменная	Отсекаемый отрезок (средний рост потомства)	Коэффициент множественной регрессии для роста матери	Коэффициент множественной регрессии для роста отца
Рост дочери	64,1	0,30	0,40
Рост сына	69,2	0,33	0,41

Например, мы видели в табл. 5.2, что угловой коэффициент регрессионной прямой, связывающей рост дочерей и матерей, равнялся 0,33 (вспомните, что угловой коэффициент прямой наилучшего приближения для точечной диаграммы – это всего лишь другое название коэффициента регрессии). Табл. 5.3 показывает, что если мы учтем еще и влияние роста отца, то этот коэффициент уменьшится до 0,30. Аналогично, если мы учтем при прогнозировании роста сына еще и рост матери, то коэффициент регрессии для отца снизится с 0,45 в табл. 5.2 до 0,41 в табл. 5.3. Таким образом, влияние роста одного родителя слегка снижается, если учитывать рост второго. Причиной может быть тот факт, что рост родителей не совсем независимая величина, ведь высокие женщины чаще выходят замуж за высоких мужчин. В целом эти данные говорят о том, что изменение роста отца

на 1 дюйм сильнее сказывается на росте взрослого ребенка, чем изменение роста матери на 1 дюйм. Множественная регрессия часто используется, когда исследователей интересует одна объясняющая переменная, а остальные служат в качестве поправок для учета дисбаланса.

Давайте вернемся к шведскому исследованию опухолей мозга, о котором мы упоминали в [главе 4](#), иллюстрируя то, как СМИ неверно трактуют причинность. В регрессионном анализе количество опухолей рассматривалось как зависимая переменная (переменная отклика), а образование как независимая (объясняющей) переменная. В регрессионную модель включались и другие факторы: возраст при диагностике, календарный год, регион Швеции, семейное положение и доход; все это считалось потенциальными возмущающими переменными. Поправка на возмущения была попыткой выделить чистую зависимость между образованием и опухолями мозга, однако полной адекватности здесь все равно никогда не добиться. Всегда будет оставаться подозрение, что могут срабатывать какие-то скрытые факторы, например, такой: более образованные люди больше заботятся о здоровье, поэтому активнее занимаются диагностикой.

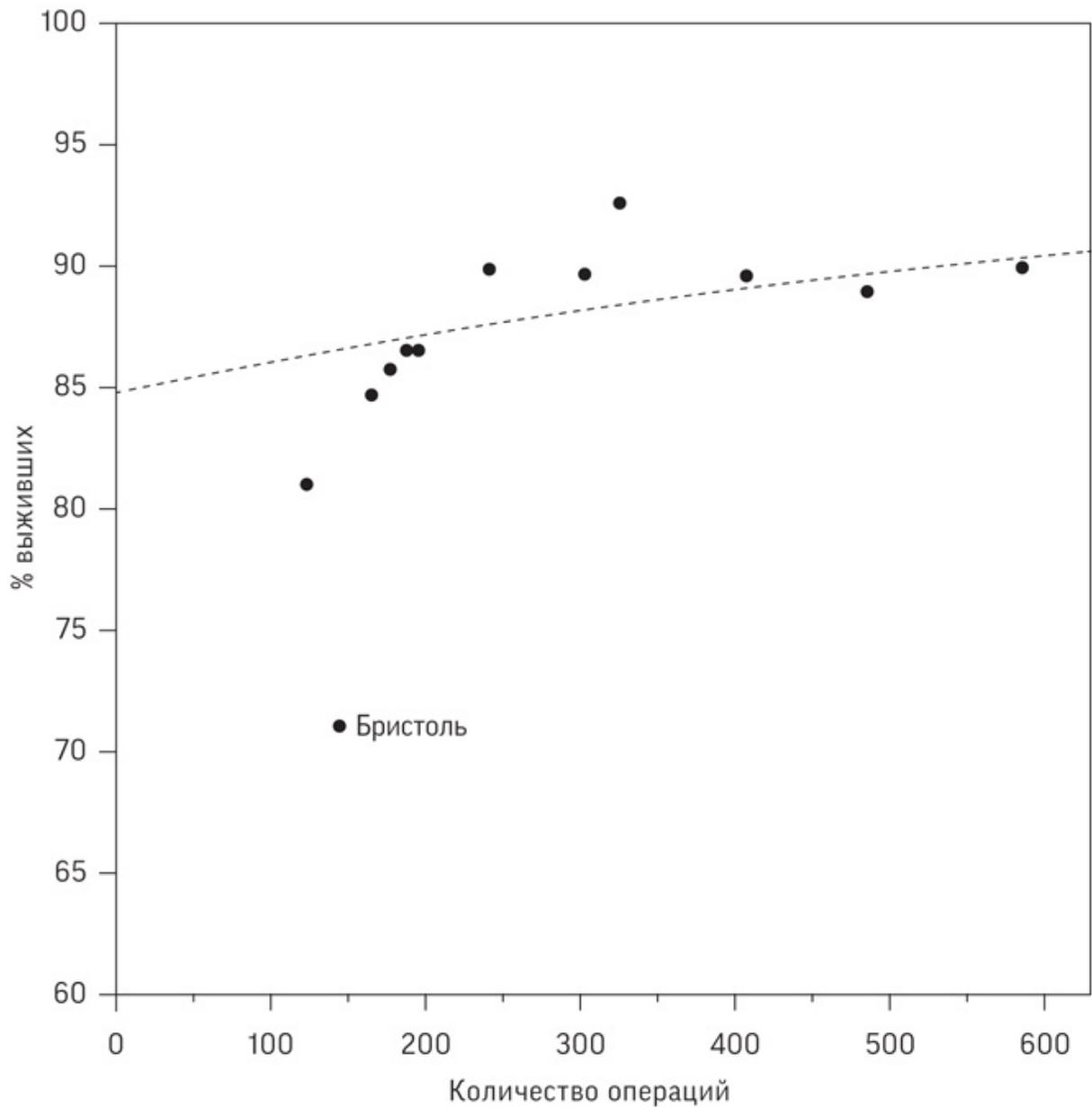
В рандомизированном испытании нет необходимости вносить поправки из-за возмущающих факторов, поскольку случайное распределение по группам гарантирует, что все факторы, кроме изучаемого, будут равномерно сбалансированы между группами. Однако исследователи часто все равно проводят регрессионный анализ – на случай, если вкрадется какой-либо дисбаланс.

### ***Различные виды зависимых переменных***

Не все данные являются непрерывными измерениями, такими как рост. В статистическом анализе зависимые переменные часто могут иметь другой вид: доля случаев, когда произошло какое-нибудь событие (например, доля людей, переживших операцию), количество каких-нибудь событий (например, число выявленных случаев рака в год в определенном регионе) или продолжительность времени до определенного события (например, количество лет, которое пациент прожил после операции). Для каждого из таких видов зависимых переменных существуют собственные формы множественной

регрессии, и соответственно меняется интерпретация получающихся коэффициентов [123].

Рассмотрим данные об операциях на сердце у детей, которые обсуждались в главе 2, где на [рис. 2.5\(a\)](#) показаны доли пациентов, переживших операцию, и количество операций, проведенных в каждой из больниц в 1991–1995 годах. На рис. 5.2 снова представлена точечная диаграмма и линия регрессии, которая построена без учета точки-выброса, соответствующей бристольской больнице.



**Рис. 5.2**

Модель логистической регрессии для данных об операциях на сердце у детей в возрасте до 1 года в больницах Соединенного Королевства в период с 1991 по 1995 год. В больницах, где больше пациентов, показатель выживаемости выше. Линия является частью кривой, которая никогда не достигнет 100 %, и не учитывает выброс, соответствующий бристольской больнице

Мы могли бы провести через эти точки прямую линейной регрессии, но тогда наивная экстраполяция говорила бы, что при очень большом количестве случаев выживаемость превысит 100 %, а это полный абсурд. Поэтому для показа долей была разработана **логистическая регрессия**, где кривая не выходит за рамки диапазона от 0 % до 100 %.

Даже без учета Бристоля в больницах с большим количеством пациентов выше показатели выживаемости, а коэффициент логистической регрессии (0,001) означает, что ожидаемый уровень смертности будет примерно на 10 % (относительно) ниже на каждые дополнительные сто операций, которые проводила больница детям до 1 года за четырехлетний период<sup>[124]</sup>. Конечно, еще раз повторим клише, что корреляция не означает причинно-следственной связи, и мы не можем заключить, что увеличение нагрузки приводит к повышению качества операций. Как мы уже упоминали, причинность может быть обратной: больницы с хорошей репутацией привлекают больше пациентов.

Этот спорный вывод, опубликованный в 2001 году, внес свою лепту в длительные, до сих пор продолжающиеся дискуссии о том, сколько больниц в Великобритании должны проводить подобные операции.

### ***Более сложные модели регрессии***

Методы, описанные в этой главе, прекрасно работали с момента их появления более века назад. Однако доступность огромных объемов данных и колоссальное увеличение вычислительных мощностей позволили создать более сложные модели. В широком смысле различные группы исследователей используют четыре основные стратегии моделирования:

- Достаточно простые математические представления зависимостей, такие как описанные в этой главе линейные регрессии. Статистики, как правило, предпочитают именно их.

- Сложные детерминистские модели, основанные на научном понимании физических процессов, например, используемые при прогнозировании погоды. Они предназначены для реалистичного воспроизведения механизмов, лежащих в их основе, и разрабатываются, как правило, прикладными математиками.

- Сложные алгоритмы, используемые для принятия решений и прогнозов, основанных на анализе большого количества прошлых случаев – например, для рекомендации книг, которые вы, возможно, хотели бы купить в сетевом магазине. Создаются в мире компьютерных наук и **машинного обучения**. Они часто будут «черными ящиками» в том смысле, что могут делать хорошие прогнозы, но их внутренняя структура в какой-то степени непостижима (см. [следующую главу](#)).

- Регрессионные модели, которые делают заключения о причинно-следственных связях; за них выступают экономисты.

Это значительные обобщения. К счастью, профессиональные барьеры рушатся, и, как мы увидим позже, формируется все более универсальный подход к моделированию. Но какая бы стратегия ни была принята, при создании и использовании модели возникают общие проблемы.

Хорошая аналогия состоит в том, что модель похожа на карту, а не на саму территорию. Все мы знаем, что одни карты лучше, чем другие: простой карты может быть достаточно для поездки из одного города в другой, но для прогулки в сельской местности нужно что-то более подробное. Британский статистик Джордж Бокс прославился бесценным афоризмом: «Все модели неверны, но некоторые полезны». Это поучительное заявление основывалось на опыте применения статистики в промышленных процессах, который позволял Боксу оценивать и силу моделей, и опасности излишней веры в них.

Но такие предостережения легко забываются. Как только какая-то модель принимается (и особенно тогда, когда она уходит из рук создателей, понимающих ее ограничения), она может превратиться в своего рода оракула. Финансовый кризис 2007–2008 годов в значительной степени был вызван чрезмерным доверием к сложным

финансовым моделям, которые использовались для определения рисков, например ипотечных пакетов. Эти модели предполагали лишь умеренную корреляцию между неисполнением обязательств по ипотеке и успешно работали, пока рынок недвижимости процветал. Но когда условия изменились и возникли проблемы с выплатами, оказалось, что проблемы начались по всем фронтам: модели сильно недооценили риски из-за корреляций, которые оказались намного выше, чем предполагалось. Руководители просто не поняли хрупкости фундамента, на котором строились эти модели, упустив из виду тот факт, что модели всего лишь упрощение реального мира, то есть *карты, а не территории*. Результатом стал один из тяжелейших мировых кризисов в истории.

### **Выводы**

- Регрессионные модели обеспечивают математическое представление отношений между набором независимых (объясняющих) переменных и зависимой переменной (переменной отклика).
- Коэффициенты в регрессионной модели показывают, какое изменение в отклике мы можем ожидать при изменении независимой переменной.
- Регресс к среднему наблюдается, когда отклонения возвращаются к долговременному среднему значению, поскольку такие выбросы были чисто случайными.
- Регрессионные модели могут включать различные виды зависимой и независимых переменных, а также нелинейные взаимосвязи.
- При интерпретации моделей нужно быть осторожным. Помните: «Все модели неверны, но некоторые полезны».

## Глава 6. Алгоритмы, аналитика и прогнозирование

До сих пор акцент в этой книге делался на том, как статистика может помочь нам лучше понять, как устроен мир, будь то потенциальный вред от поедания сэндвичей с беконом или взаимосвязь между ростом родителей и их детей. По сути, это научное исследование, где выясняется, что происходит на самом деле, а что (если пользоваться терминами, введенными в предыдущей главе) – просто остаточная ошибка, к которой нужно относиться как к неизбежной изменчивости, поскольку ее нельзя смоделировать.

Однако основные идеи статистической науки сохраняются, когда мы пытаемся решить не научную, а практическую задачу. Желание найти сигнал в шуме уместно и тогда, когда нам просто нужен метод для конкретного случая в повседневной жизни. Тема этой главы – решение практических задач с помощью имеющихся данных, то есть использование какого-то алгоритма, механической формулы, которая будет автоматически выдавать ответ для каждого нового случая, возникающего без или с минимальным вмешательством человека. Фактически это уже не наука, а «технология».

У такого алгоритма есть два широких класса задач:

- Классификация (также известна как разделение, контролируемое обучение, **обучение с учителем**): сообщить, с какой ситуацией мы столкнулись. Например, пристрастия и предубеждения онлайн-покупателя или является ли объект в поле зрения робота ребенком или собакой.

- Прогнозирование: сообщить, что будет дальше. Например, какая погода будет на следующей неделе, какая может быть завтра цена акций, какие продукты может купить этот клиент и не выбежит ли тот ребенок перед нашим самоуправляемым автомобилем.

Хотя эти задачи и различаются в том смысле, что одна касается настоящего, а вторая – будущего, обе имеют одинаковую природу: взять набор наблюдений, относящихся к текущей ситуации, и сделать

соответствующее заключение. Такой метод называется **предсказательной аналитикой**, но здесь мы уже вторгаемся на территорию **искусственного интеллекта (ИИ)**, когда воплощенные в машинах алгоритмы используются либо для выполнения задач, которые обычно требуют участия человека, либо для предоставления людям советов экспертного уровня.

В узком смысле термин ИИ относится к системам, которые могут выполнять строго предписанные задачи, причем есть ряд крайне успешных примеров, основанных на машинном обучении, которые включают разработку алгоритмов посредством статистического анализа больших массивов данных, взятых из прошлого. Среди заметных успехов – системы распознавания речи, встроенные в телефоны, планшеты и компьютеры; программы типа Google Translate, которые не знакомы с грамматикой, зато научились подбирать тексты из необъятного опубликованного архива; программное обеспечение для компьютерного зрения, использующее прошлые изображения, чтобы «научиться» распознавать, например, лица на фотографиях или другие автомобили, попадающие в поле зрения самоуправляемого автомобиля. Наблюдается значительный прогресс и в системах, играющих в различные игры, таких как программное обеспечение DeepMind. Они изучают правила компьютерных игр и становятся опытными игроками, обыгрывающими чемпионов мира по шахматам и «Го», пока компьютер IBM Watson обыграл людей в викторине, где требуются общие знания. Эти системы начинались не с попыток закодировать человеческий опыт и знания, а с огромного числа примеров, и обучались методом проб и ошибок, как обычный ребенок, играя в игры сам с собой.

Опять же, подчеркнем, что это технологические системы, использующие прошлые данные для ответа на возникающие практические вопросы, а не научные системы, которые стремятся понять, как устроен мир: их следует оценивать исключительно по тому, насколько хорошо они выполняют ограниченную задачу, и, хотя форма изученных алгоритмов может натолкнуть на какие-то идеи, от них не ждут воображения или сверхчеловеческих способностей в обычной жизни. Здесь требуется ИИ «общего назначения», который выходит за рамки этой книги и (по крайней мере на данный момент) за рамки возможностей компьютеров.

С тех пор как в 1690-х годах Эдмунд Галлей вывел формулы для расчета страховых взносов и платежей, статистика имеет дело с алгоритмами, помогающими в принятии решений. Современное развитие науки о данных продолжает эту традицию, однако за последние годы масштабы собираемых данных и разрабатываемых продуктов изменились: появились так называемые большие данные.

Данные могут быть «большими» в двух разных смыслах. Во-первых, по числу примеров в базе данных: это могут быть отдельные люди, звезды в небе, школы, поездки на автомобиле или посты в социальных сетях. Количество примеров часто обозначают буквой  $n$ , и в начале моей деятельности  $n$  считалось «большим», когда превосходило 100. Но сегодня данные могут включать миллионы и миллиарды случаев.

Второй смысл термина «большие данные» – это измерение в каждом примере многих характеристик или признаков. Они часто обозначаются буквой  $p$  (возможно, от слова *parameter*). Если снова обратиться ко временам моей статистической юности, то обычно  $p$  не превышало 10 – мы знали не так много пунктов в медицинской карте отдельного человека. Но с получением доступа к миллионам генов человека проблемы геномики оказались в малых значениях  $n$ , но больших значениях  $p$ , то есть в наличии колоссального количества информации об относительно небольшом числе случаев.

А теперь мы вступили в эру задач с большими  $n$  и большими  $p$ , когда имеется огромное количество случаев, каждый из которых может быть очень сложным, – подумайте об алгоритмах, анализирующих все посты, лайки и дизлайки каждого из миллиардов подписчиков Facebook, чтобы определить, какие новости и рекламу показывать.

Такие новые захватывающие задачи привели в науку о данных массу новых людей. Но давайте еще раз вспомним утверждение, приведенное в начале книги: данные не говорят сами по себе. Это мы наполняем их смыслом, а потому с ними нужно обращаться умело и с осторожностью, чтобы избежать многих потенциальных ловушек при наивном использовании алгоритмов. В этой главе мы приведем примеры некоторых классических катастроф, но сначала рассмотрим фундаментальную задачу по сведению данных в нечто полезное.

### ***Выявление закономерностей***

Одна из стратегий при работе с чрезмерным количеством случаев – формирование групп по схожести – процедура, известная как кластеризация, или неконтролируемое обучение (**обучение без учителя**), поскольку мы должны определить эти группы сами, нас не предупреждают заранее об их существовании. Нахождение таких однородных кластеров может быть и конечной целью. Например, можно определить группы людей с одинаковыми предпочтениями или предубеждениями, установить их характеристики, дать название, придумать алгоритмы для классификации будущих случаев. А затем давать таким выявленным кластерам соответствующие рекомендации фильмов, политическую, коммерческую и социальную рекламу – в зависимости от мотивации людей, построивших алгоритм.

Прежде чем конструировать алгоритм для классификации или прогнозирования, возможно, придется сократить объем исходных данных по величине  $p$  до приемлемого размера, поскольку изначально она может быть очень большой (в каждом конкретном случае измерялось слишком много характеристик). Этот процесс называется **конструированием признаков**. Просто представьте, сколько измерений можно сделать на человеческом лице. Чтобы разрабатывать программное обеспечение для распознавания лиц и сравнения их с базой данных, можно ограничиться измерением нескольких важных признаков. Те измерения, которые не нужны для прогноза или классификации, можно определить с помощью визуализации данных или методов регрессионного анализа, а затем отбросить. Кроме того, уменьшить число признаков можно с помощью «составных» измерений, которые включают большую часть информации.

Последние разработки в области чрезвычайно сложных моделей (вроде тех, что относятся к так называемому **глубокому обучению**) предполагают, что необходимости в первоначальном этапе сокращения данных может и не быть, то есть один алгоритм способен обработать все исходные данные.

### ***Классификация и прогнозирование***

Сегодня доступно ошеломляющее количество различных методов для построения алгоритмов классификации и прогнозирования. Исследователи обычно используют то, к чему привыкли в ходе своей профессиональной деятельности: например, статистики предпочитают

регрессионные модели, а специалисты по теории вычислительных машин и систем – логику на основе правил и нейронные сети, которые были альтернативными способами имитации человеческого познания. Реализация любого из этих методов требует специальных навыков и программного обеспечения, но сейчас появились удобные программы, которые позволяют выбирать методы с помощью меню и тем самым поощряют менее однобокий подход, когда эффективность важнее, чем философия моделирования.

Начав измерять и сравнивать практическую эффективность алгоритмов, люди неизбежно стали соревноваться, и сейчас такие платформы, как [Kaggle.com](https://www.kaggle.com), проводят конкурсы в сфере науки о данных. Какая-нибудь коммерческая или академическая организация предоставляет участникам набор данных: это может быть задача по обнаружению китов по зафиксированным звукам, регистрация темной материи по астрономическим данным или прогнозирование числа госпитализированных больных. В каждом случае конкурсантам предоставляется тренировочный (обучающий) набор данных для конструирования алгоритма, а также тестовый набор для определения его эффективности. Особенно популярен конкурс (привлекающий тысячи команд) по созданию алгоритма для следующей задачи.

### **Можно ли сказать, какие пассажиры выжили после гибели «Титаника»?**

Во время своего первого рейса «Титаник» столкнулся с айсбергом и медленно затонул в ночь с 14 на 15 апреля 1912 года. Только около 700 из 2200 пассажиров и членов экипажа оказались в спасательных шлюпках и выжили, и последующие исследования и расчеты сосредоточились на том, что шансы попасть в шлюпку и выжить критически зависели от того, билет какого класса у вас был.

Алгоритм, прогнозирующий выживание, может на первый взгляд показаться странным выбором *проблемы* в рамках стандартного цикла RPDAC, поскольку такая ситуация вряд ли возникнет снова, поэтому не представляет никакой ценности для будущего. Но один человек помог мне найти мотивацию. В 1912 году Фрэнсис Сомертон уехал из Илфракомба в Северном Девоне, расположенного недалеко от того места, где я родился и вырос. Отправившись искать счастье в США, он купил билет третьего класса за 8 фунтов и 1 шиллинг на новенький

«Титаник», оставив в Европе жену и маленькую дочь. Однако так и не добрался до Нью-Йорка – его надгробие находится на церковном кладбище в Илфракомбе (рис. 6.1). Точный прогнозирующий алгоритм сможет сказать нам: Фрэнсису Сомертону действительно просто не повезло или его шансы на самом деле были невелики.



ALSO OF  
FRANCIS WILLIAM  
SON OF THE ABOVE  
WHO PERISHED  
IN THE "TITANIC" DISASTER  
APRIL 14 1912, AGED 30 YEARS

Рис. 6.1

Надгробие Фрэнсиса Сомертона на кладбище в Илфракомбе. Надпись гласит: «Также Фрэнсис Уильям, сын вышеуказанного, который погиб при катастрофе “Титаника” 14 апреля 1912 года в возрасте 30 лет»

План – собрать имеющиеся данные и попробовать ряд различных методов для создания алгоритмов, предсказывающих тех, кто выжил. Это можно считать скорее проблемой классификации, чем прогнозирования, поскольку все события уже случились. Данные – это открытая информация о 1309 пассажирах «Титаника»: потенциальные предикторные (предсказывающие) переменные включают их полное имя, форму обращения, пол, возраст, класс на судне (первый, второй, третий), сумму, уплаченную за билет, были ли они частью семьи, место посадки на судно (Саутгемптон, Шербур, Куинстаун), а также неполные данные о некоторых номерах кают<sup>[125]</sup>. Зависимая переменная – это указатель, выжил человек (1) или нет (0).

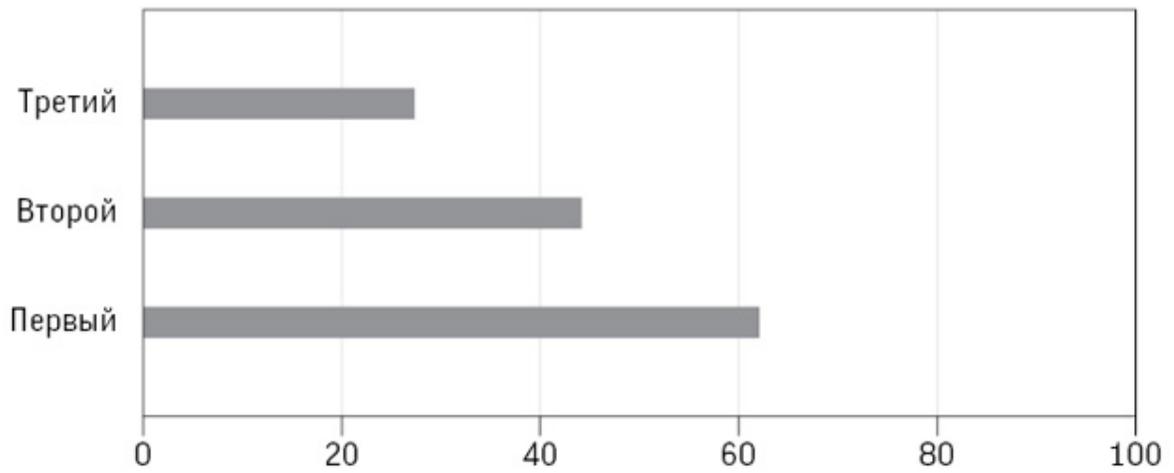
На этапе *анализа* важно разделить данные на две части: тренировочный набор, используемый для создания алгоритма, и тестовый набор, который служит только для оценки эффективности – смотреть на тестовый набор до готовности алгоритма было бы серьезным жульничеством. Как и в конкурсе Kaggle, мы возьмем в качестве тренировочного набора случайную выборку из 897 случаев, а оставшиеся 412 человек составят тестовый набор.

Это реальные, а потому довольно загрязненные данные, требующие определенной предварительной обработки. У восемнадцати пассажиров отсутствует информация о плате за поездку, так что будем считать, что они заплатили медианную стоимость для своего класса. Были добавлены некоторые родители, братья и сестры для создания единой переменной, характеризующей размер семьи. Упростились обращения: «мадемуазель» было объединено с «мисс», «мадам» – с «миссис», и целый ряд обращений был закодирован как «редкие формы обращений»<sup>[126]</sup>.

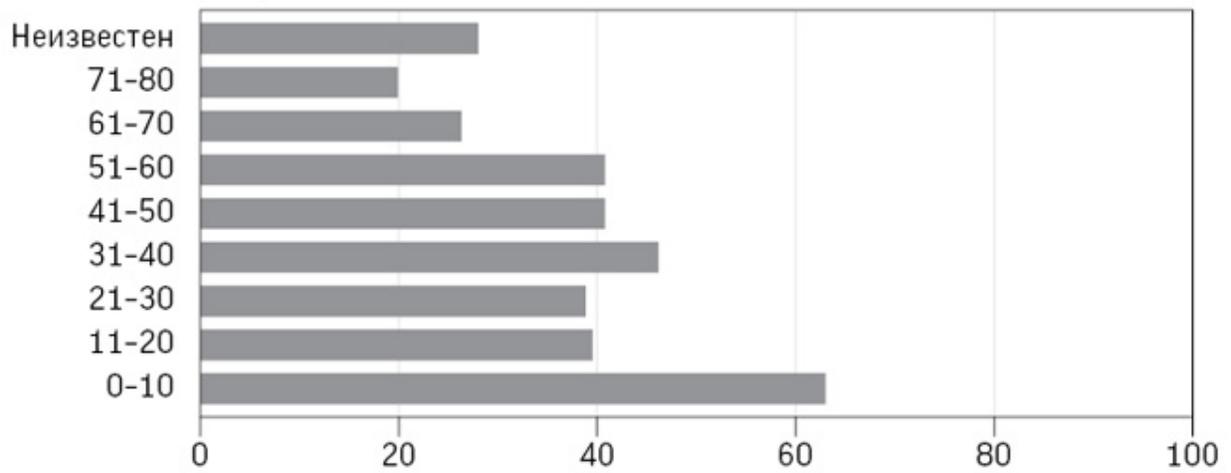
Следует пояснить, что даже для простой подготовки данных к анализу, кроме требуемых навыков кодирования, могут понадобиться серьезные знания и рассуждения – например, об использовании доступной информации о каютах для определения их положения на судне. Несомненно, я мог бы сделать эту работу лучше.

На рис. 6.2 показаны доли выживших для разных категорий из 897 пассажиров, выбранных в качестве тренировочного (обучающего) набора. Все эти признаки сами по себе обладают прогностической способностью: видно, что уровень выживаемости выше среди тех, кто путешествовал более высоким классом; среди женщин и детей; тех, кто больше заплатил за билет; среди имевших небольшую семью и тех, к кому обращались *миссис*, *мисс* или *мастер*<sup>[127]</sup>. Все это соответствует нашим предположениям.

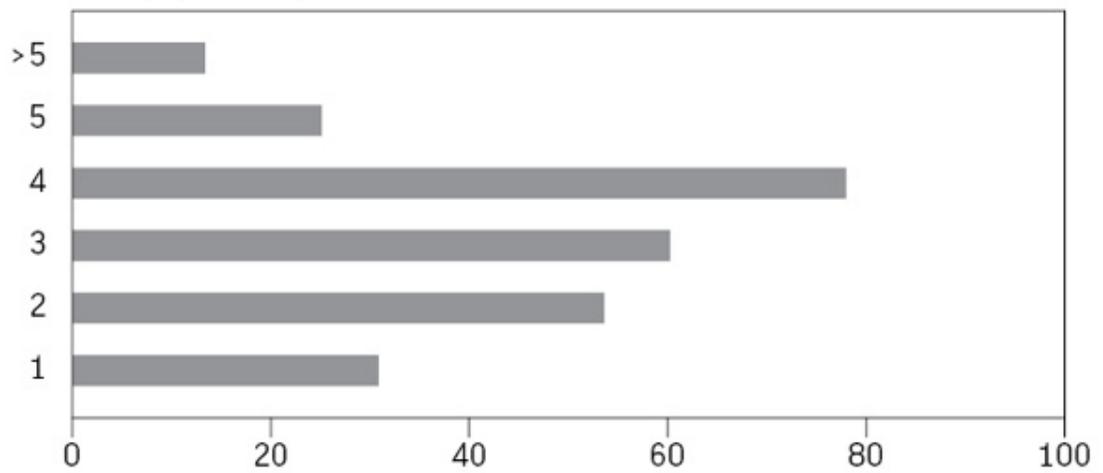
### Класс



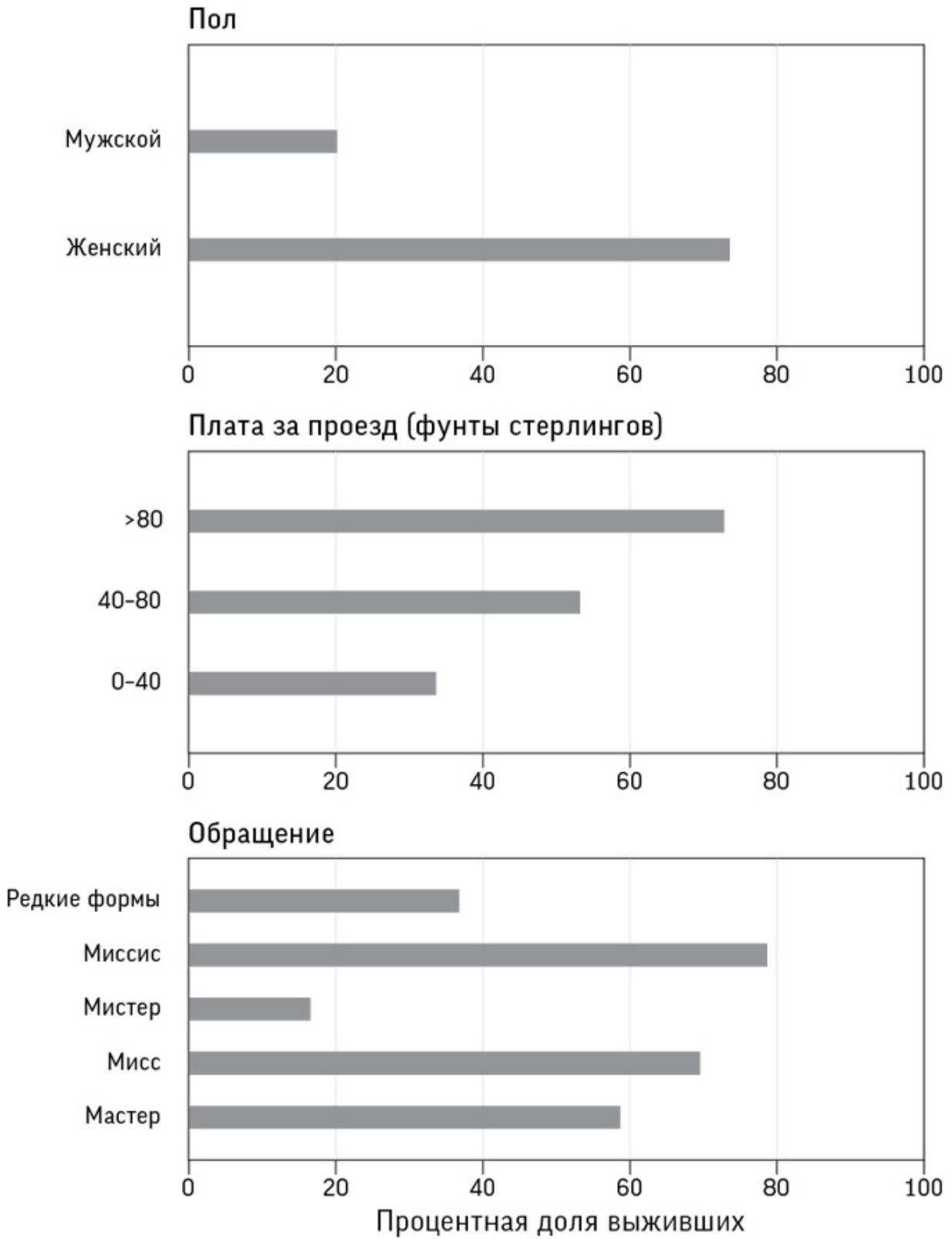
### Возраст



### Размер семьи



Процентная доля выживших



**Рис. 6.2**

Сводные данные о выживании для тренировочного набора из 897 пассажиров «Титаника», показывающие процентную долю выживших для различных категорий людей

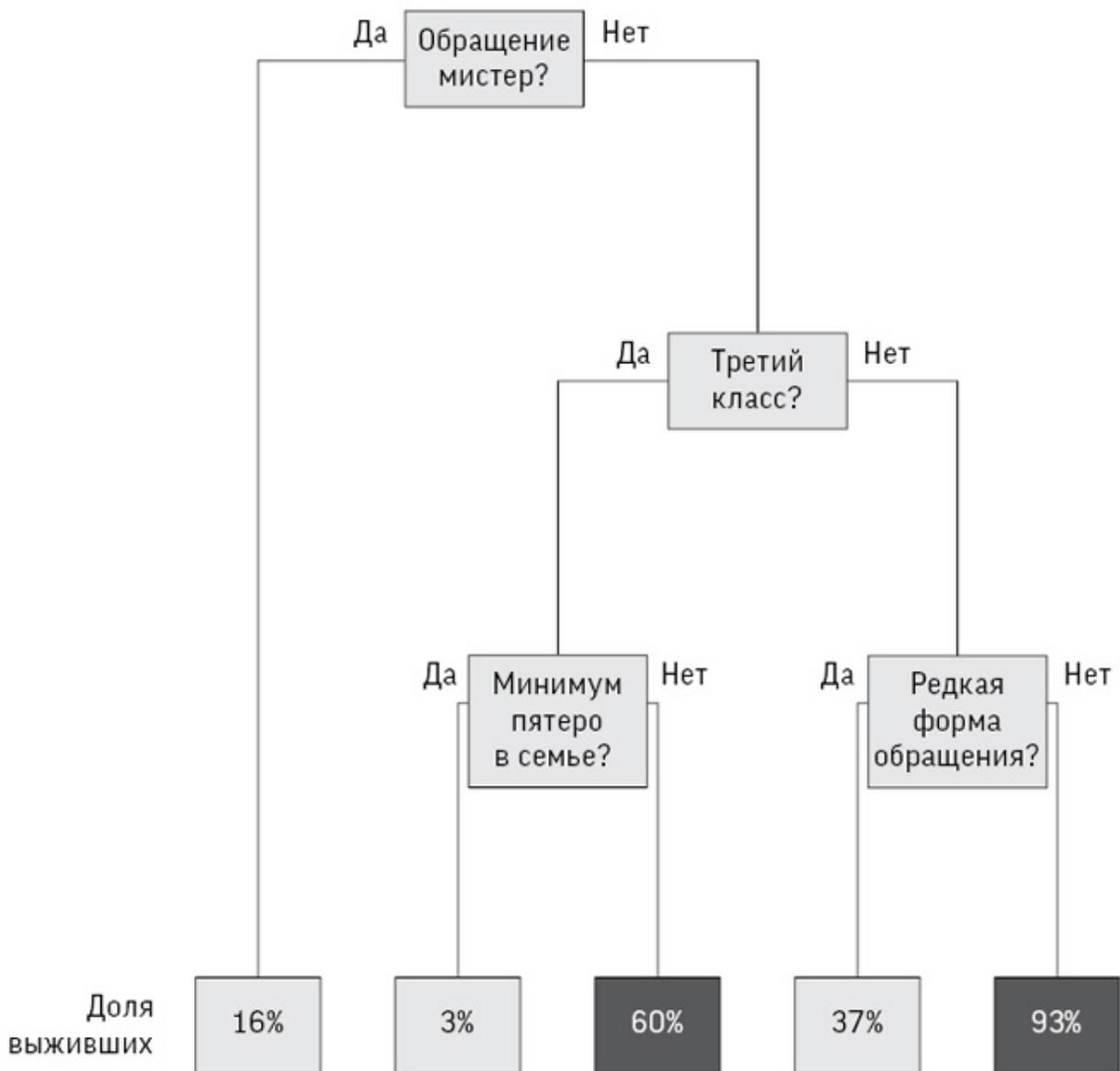
Однако эти параметры нельзя назвать независимыми. Пассажиры более высокого класса предположительно больше заплатили за билеты; можно также ожидать, что у них меньше детей, чем у бедных эмигрантов. Многие мужчины плыли в одиночку. Важным может быть и способ кодирования информации: следует ли рассматривать возраст как качественную переменную с делением на категории (см. рис. 6.2) или как непрерывную переменную? Участники конкурса потратили много времени на подробное рассмотрение таких деталей и кодирование с извлечением максимума информации, но мы перейдем прямо к прогнозированию.

Предположим, мы сделали (заведомо неверный) прогноз: «Никто не выжил». Тогда, учитывая, что зафиксирована смерть 61 % пассажиров, показатель правильности нашего прогноза для тренировочного набора данных составил бы 61 %. Если бы мы строили прогноз на более сложном правиле «Все женщины выживают, а все мужчины погибают», то мы бы верно классифицировали 78 % данных в тренировочном наборе. Эти примитивные правила – хорошие ориентиры, по которым можно измерять все улучшения, обеспечиваемые более изощренными алгоритмами.

### *Деревья классификации*

**Дерево классификации** – пожалуй, самая простая форма алгоритма, поскольку состоит из серии вопросов типа «да/нет», где ответ на каждый вопрос определяет формулировку следующего вопроса, и так до тех пор, пока не будет получено заключение. На рис. 6.3 показано дерево классификации для данных по «Титанику», в котором в конце каждой ветки указана доля выживших для соответствующей категории. Легко видеть и выбранные факторы, и окончательный вывод. Например, Фрэнсис Сомертон в базе данных отмечен как «мистер», а потому попадает на левую ветвь. Окончание этой ветки включает 58 % данных тренировочного набора, из которых 16 % выжило. Поэтому мы можем оценить на основании ограниченной информации, что шансы на выживание у Сомертона составляли 16 %.

Наш простой алгоритм определяет две группы с более чем 50-процентными шансами на выживание. Во-первых, это женщины и дети в первом и втором классе (если у них нет редкой формы обращения), 93 % которых спаслись. Во-вторых, женщины и дети в третьем классе, при условии, что они не из многодетных семей, – из них выжило 60 %.



**Рис. 6.3**

Дерево классификации для данных по «Титанику», в котором последовательность вопросов приводит пассажиров к концу ветви, где указаны доли выживших для групп из тренировочного набора. Согласно прогнозу, конкретный человек выживет, если в аналогичной

группе доля выживших превышает 50 %. Такой прогноз предлагается только для двух категорий пассажиров: женщин и детей из третьего класса из небольших семей, а также всех женщин и детей из первого и второго класса – при условии, что у них нет редких форм обращений

Прежде чем смотреть, как реально конструируется такое дерево, нам нужно решить, какие показатели эффективности следует использовать в нашем конкурсе.

### ***Оценивание эффективности алгоритма***

Если алгоритмы будут сравниваться по точности, нужно решить, что означает «точный». В конкурсной задаче о «Титанике» на платформе Kaggle это просто процентная доля пассажиров в тестовом наборе, которых алгоритм правильно классифицировал. Поэтому, после того как участники конкурса сконструируют алгоритмы, они дают свой прогноз на переменную отклика для тестового набора, а Kaggle измеряет точность прогнозов. Мы представим результаты сразу для всего тестового набора (подчеркнем, что это не то же самое, что тестовый набор Kaggle<sup>[128]</sup>).

Если применить дерево классификации, приведенное на рис. 6.3, к тренировочным данным, для которых оно разработано, то оно даст точность 82 %. Если этот алгоритм использовать на тестовом наборе, точность слегка упадет – до 81 %. В табл. 6.1 приведено число разных типов ошибок, допущенных алгоритмом; эта таблица называется **матрицей ошибок**. Когда мы пытаемся определить выживших, процент верно предсказанных из числа реально выживших именуется **чувствительностью** алгоритма, а процент верно предсказанных из числа реально погибших – **специфичностью**. Эти термины взяты из медицинских диагностических исследований<sup>[129]</sup>.

### **Таблица 6.1**

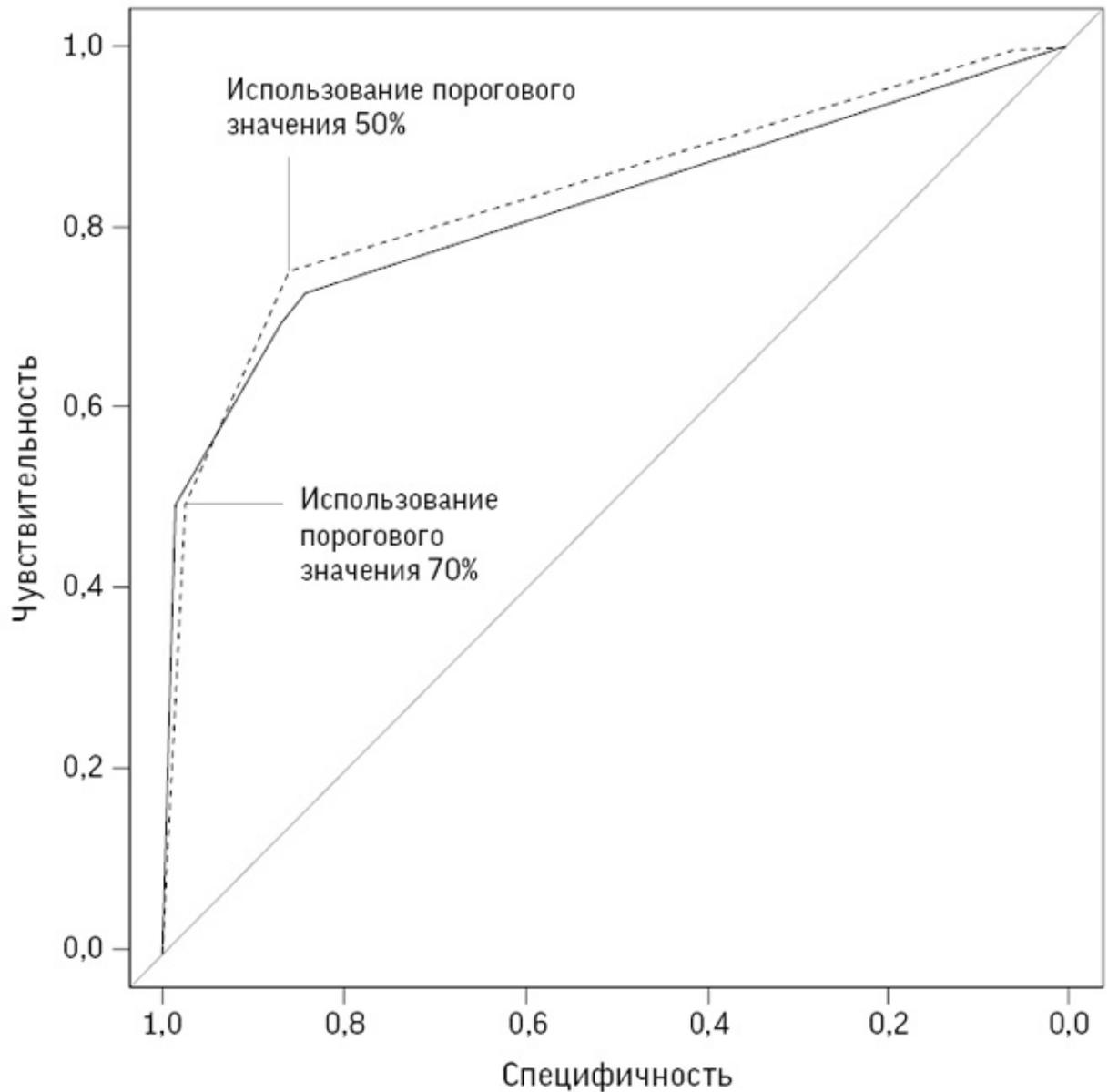
Матрица ошибок дерева классификации для тренировочных и тестовых данных, где отображается точность (% правильно классифицированных), чувствительность (% правильно классифицированных выживших) и специфичность (% правильно классифицированных погибших)

	ТРЕНИРОВОЧНЫЙ НАБОР			ТЕСТОВЫЙ НАБОР		
	Прогноз гибели	Прогноз выжива- ния		Прогноз гибели	Прогноз выжива- ния	
<b>Погибли</b>	475	93	568	228	45	273
<b>Выжили</b>	71	258	320	35	104	139
	546	351	897	263	149	412
<b>Точность</b>	= (475 + 258) / 897 = 82%			<b>Точность</b> = (228 + 104) / 412 = 81%		
<b>Чувствительность</b>	= 258 / 329 = 78%			<b>Чувствительность</b> = 104 / 139 = 75%		
<b>Специфичность</b>	= 475 / 568 = 84%			<b>Специфичность</b> = 228 / 273 = 84%		

Хотя общую точность выразить достаточно просто, это очень грубая мера эффективности, не учитывающая, с какой надежностью делается прогноз. Если мы посмотрим на кончики ветвей дерева классификации, то увидим, что разделение тренировочных данных не идеально: на всех ветвях кто-то выжил, а кто-то – нет. При грубом правиле распределения мы просто выбираем результат большинства, но можем поступить и иначе, присвоив каждому новому случаю *вероятность* выживания, соответствующую доле выживших в тренировочном наборе. Например, человеку с формой обращения «мистер» мы могли бы дать вероятность выживания 16 %, а не просто однозначно предсказать, что он погибнет.

Алгоритмы, которые дают не простую классификацию, а вероятность (или какое-то другое число), часто сравниваются с помощью **ROC-кривых**<sup>[130]</sup>, которые изначально были разработаны во время Второй мировой войны для анализа радиолокационных сигналов. Ключевая идея – возможность варьировать пороговое значение, при котором дается прогноз выживания. В табл. 6.1 показан эффект использования порогового значения 50 % для прогноза выживания, при этом значения специфичности и чувствительности в тренировочном наборе соответственно равны 0,84 и 0,78. Однако мы могли бы потребовать более высокую вероятность для предсказания, что кто-то выживет, например 70 %; в этом случае специфичность и чувствительность составили бы 0,98 и 0,50 соответственно. При таком более строгом пороговом значении мы правильно определим только половину реально выживших, но зато сделаем о них очень мало ложных утверждений. Если взять все пороговые значения для предсказания выживания, то все возможные значения для специфичности и чувствительности образуют кривую. Обратите внимание, что при изображении ROC-кривой значения специфичности традиционно идут по горизонтальной оси, уменьшаясь от 1 до 0.

На рис. 6.4 изображены ROC-кривые для тренировочного и тестового набора. Если алгоритм распределяет числа случайным образом (то есть абсолютно бесполезен), то его ROC-кривая будет диагональной линией. У самых лучших алгоритмов ROC-кривые подходят близко к левому верхнему углу. Стандартный способ сравнения разных ROC-кривых – измерить площадь под ними. Для бесполезного алгоритма она равна 0,5, а для идеального – 1. Для нашего тестового набора для данных о «Титанике» площадь под кривой составляет 0,82. Оказывается, для этой площади есть изящная интерпретация: если мы выбираем истинно выжившего и истинно погибшего случайным образом, то с вероятностью 82 % алгоритм дает истинно выжившему большую вероятность выживания, чем истинно погибшему. Области свыше 0,80 представляют весьма хорошую эффективность разделения. Площадь под ROC-кривой – это способ измерить, насколько точно алгоритм отделяет выживших от погибших, но она не отражает сами вероятности. Категория специалистов, которые лучше всего знакомы с вероятностными прогнозами, – это синоптики.



**Рис. 6.4**

ROC-кривые для дерева классификации, приведенного на [рис. 6.3](#), примененные к тренировочному (пунктирная линия) и тестовому (сплошная линия) набору данных. «Чувствительность» – доля правильно предсказанных выживших. «Специфичность» – доля правильно предсказанных погибших. Площади областей под кривыми равны 0,84 и 0,82 для тренировочного и тестового набора соответственно

**Как узнать, насколько точны прогнозы «вероятности осадков»?**

Предположим, мы хотим предсказать, будет ли завтра дождь в конкретном месте в конкретное время. Простейшие алгоритмы могут просто давать ответ в виде «да/нет», и он может оказаться правильным или неправильным. Более сложные модели могут выдавать вероятность дождя, что позволяет принимать более точные решения, ведь ваши действия при вероятности дождя в 50 % могут сильно отличаться от действий, если алгоритм выдаст 5-процентную вероятность.

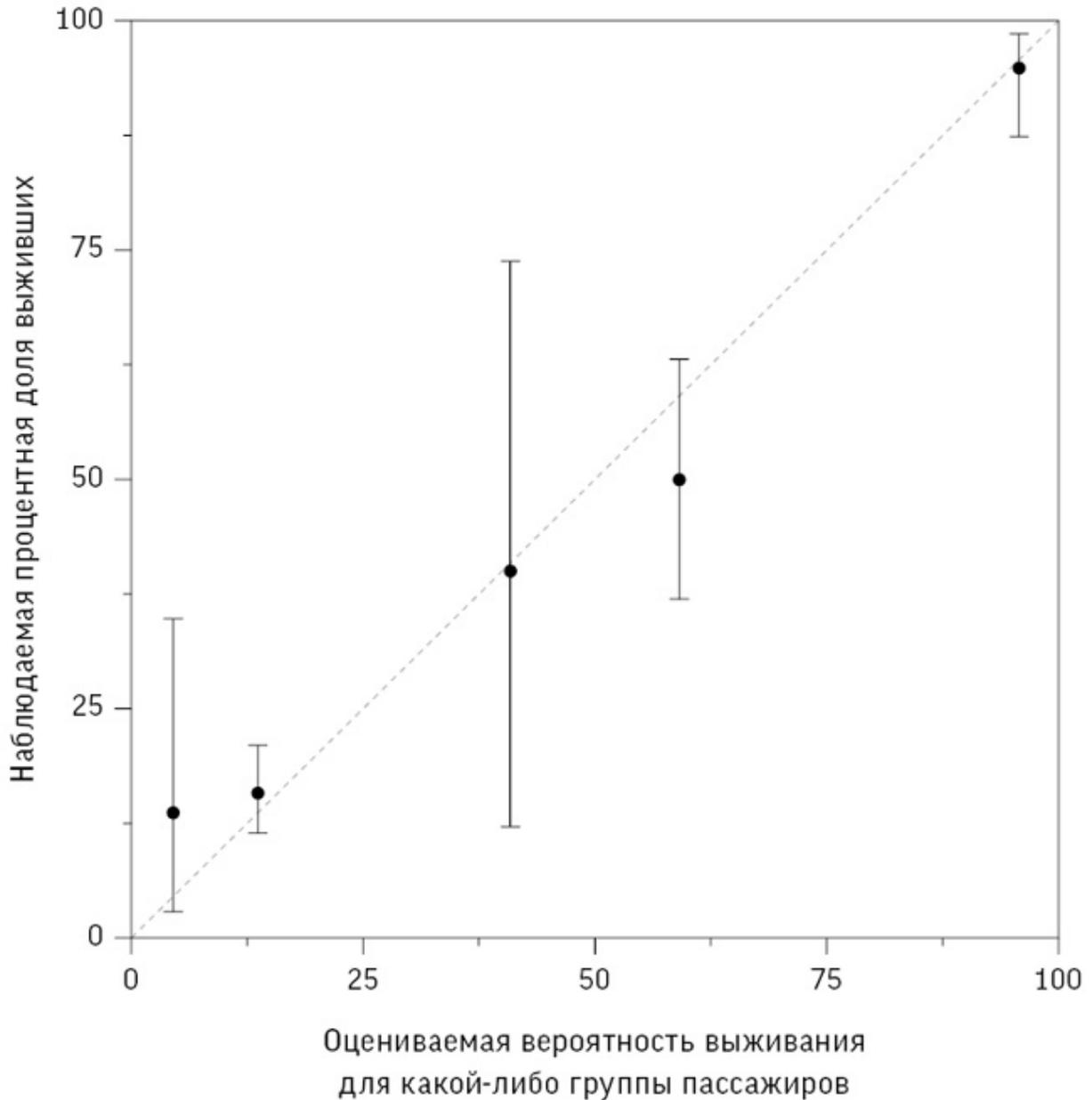
На практике прогнозы погоды основываются на крайне сложных компьютерных моделях, которые включают подробные математические формулы, отображающие развитие погоды из текущего состояния, и каждый прогон модели дает детерминистский прогноз «да/нет» для дождя в конкретном месте на конкретное время. Поэтому для получения **вероятностного прогноза** модель нужно прогнать много раз при слегка различных начальных условиях, что создаст в итоге список различных «возможных вариантов будущего», где в некоторых вариантах пойдет дождь, а в некоторых – нет. Синоптики запускают ансамбль, скажем, из 50 моделей, и если дождь будет в 5 из них, то они говорят о вероятности осадков в 10 %.

Но как проверить, насколько хороши такие вероятности? Мы не можем создать простую матрицу ошибок, как в случае дерева классификации, потому что алгоритм никогда не утверждает категорически, пойдет дождь или нет. Мы можем начертить ROC-кривые, но они только показывают, получают ли дни с дождем более частые прогнозы, чем дни без дождя. Ключевая идея – необходимость в проверке-калибровке, то есть если мы возьмем все дни, когда синоптики говорили о вероятности дождя в 70 %, то дождь действительно должен идти примерно в 70 % таких дней. Синоптики относятся к этому очень серьезно – вероятности должны соответствовать фактам.

Так называемые калибровочные графики позволяют увидеть, насколько достоверны заявленные вероятности; на них показывается общее количество событий с конкретной вероятностью появления, и доля случаев, когда событие действительно произошло.

На рис. 6.5 представлен калибровочный график для простого дерева классификации, использованного для тестового набора. Мы хотим, чтобы точки лежали ближе к диагонали (то есть заявленные вероятности соответствовали реальной доле наблюдений).

Вертикальные отрезки означают области, в которых будет находиться реальная доля в 95 % случаев для данной предсказанной вероятности. Если они пересекают диагонали, как на рис. 6.5, мы можем считать, что наш алгоритм хорошо откалиброван.



**Рис. 6.5**

Калибровочный график для простого дерева классификации, дающего вероятности выживания при катастрофе «Титаника». По горизонтальной оси отложены прогнозы, по вертикальной –

наблюдаемая доля выживших. Мы хотим, чтобы точки лежали на диагонали – это показывает, что вероятности надежны и означают именно то, о чем говорят

### ***Комбинированное измерение «точности» для вероятностей***

Хотя ROC-кривая оценивает, насколько хорошо алгоритм делит людей на группы, а калибровочный график показывает, означают ли вероятности то, о чем они говорят, было бы лучше найти какую-то простую комбинированную меру, объединяющую обе характеристики в одно число, которое мы могли бы использовать для сравнения алгоритмов. К счастью, синоптики еще в 1950-е годы придумали, как это сделать.

Если мы прогнозируем какую-нибудь числовую величину (например, температуру завтра в полдень в определенном месте), то точность прогноза обычно характеризуется ошибкой – разностью между предсказанной и реальной температурой. В отношении нескольких дней, как правило, вычисляют **среднеквадратичную ошибку (MSE)** – среднее значение квадратов отдельных ошибок; это аналог критерия наименьших квадратов, используемый в регрессионном анализе.

Особенность применения данного метода для вероятностей состоит в использовании критерия наименьших квадратов как при прогнозировании количества, но с учетом того, что будущее наблюдение дождя имеет значение 1, а его отсутствие – 0. Табл. 6.2 показывает, как это будет работать для некоей вымышленной синоптической системы. Для понедельника вероятность дождя в прогнозе равнялась 0,1, но дождя не было (истинный отклик 0), поэтому ошибка составляет  $0 - 0,1 = -0,1$ . При возведении в квадрат получим 0,01. Повторим это для всей недели. Тогда среднее арифметическое из всех ошибок  $B$  – мера точности прогнозов этого синоптика. В нашем случае  $B \approx 0,11$ <sup>[131]</sup>. Такая среднеквадратичная ошибка известна как **показатель Бриера** (названа в честь метеоролога Гленна Бриера, который описал этот метод в 1950 году).

### **Таблица 6.2**

Вымышленные прогнозы «вероятности осадков»: будет дождь завтра в полдень в определенном месте или нет. Наблюдаемые

результаты: 1 = был дождь, 0 = дождя не было. «Ошибка» – это разность между прогнозом и наблюдением, а показатель Бриера В – это среднеквадратичная ошибка. Показатель Бриера ВС для климатических данных основан на использовании средних долговременных данных для дождя в это время года, и в нашем случае предполагается, что вероятность дождя составит 20 % для всех дней

	Пн	Вт	Ср	Чт	Пт	Средне-квадратичная ошибка (показатель Бриера)
Вероятность осадков	0,1	0,2	0,5	0,6	0,3	
Был ли дождь на самом деле?	Нет	Нет	Да	Да	Нет	
Истинный отклик	0	0	1	1	0	
Ошибка	-0,1	-0,2	0,5	0,4	-0,3	
Квадратичная ошибка	0,01	0,04	0,25	0,16	0,09	$V = 0,54 / 5 \approx 0,11$
Климатическая вероятность	0,2	0,2	0,2	0,2	0,2	
Климатическая ошибка	-0,2	-0,2	0,8	0,8	0,2	
Квадратичная климатическая ошибка	0,04	0,04	0,64	0,64	0,04	$BC = 1,4 / 5 = 0,28$

К сожалению, сам по себе показатель Бриера не так легко истолковать, а потому трудно определить, насколько квалифицированно

работает тот или иной синоптик. Лучше всего сравнивать его с контрольным показателем, основанным на исторических записях о климате. Такие климатические прогнозы не учитывают текущих условий и просто оценивают вероятность осадков как долю тех случаев, когда в этот день шел дождь. Подобный прогноз может делать кто угодно без каких-либо навыков – в табл. 6.2 мы условно считаем, что для каждого дня на этой неделе вероятность дождя составит 20 %. Это даст нам показатель Бриера, рассчитанный по климатическим историческим данным (мы назвали его ВС), равный 0,28.

Любой приличный алгоритм прогнозирования должен работать лучше того, который основан только на исторических данных для этого дня, и наша система действительно улучшила показатель Бриера:  $BC - B = 0,28 - 0,11 = 0,17$ . Затем синоптик получает оценку мастерства, которая отражает пропорциональное уменьшение контрольного показателя – в нашем случае 0,61<sup>[132]</sup>. Иными словами, наш алгоритм на 61 % лучше, чем примитивный метод, использующий только исторические данные о климате.

Конечно, идеальная цель – оценка 100 %, однако такое возможно только в случае, когда показатель Бриера равен 0, то есть мы абсолютно точно предсказываем, будет дождь или нет. Это требует от синоптика особого мастерства, и реальные оценки качества работы при прогнозировании дождя сейчас составляют около 0,4 для прогнозов на следующий день и 0,2 для недельного прогноза<sup>[133]</sup>. Конечно, самый ленивый прогноз – это просто сказать: все, что происходит сегодня, будет происходить и завтра. Это обеспечивает идеальное соответствие с историческими данными (сегодняшним днем), но для предсказания будущего не слишком эффективно.

Если вернуться к примеру с «Титаником», то наивный алгоритм просто приписывает каждому пассажиру вероятность выживания 39 %, то есть долю выживших во всем тренировочном наборе. Такой алгоритм не использует никаких данных об отдельных людях и, по сути, эквивалентен прогнозированию погоды по историческим климатическим данным без учета текущих обстоятельств. Показатель Бриера для этого «не требующего умений» правила равен 0,232.

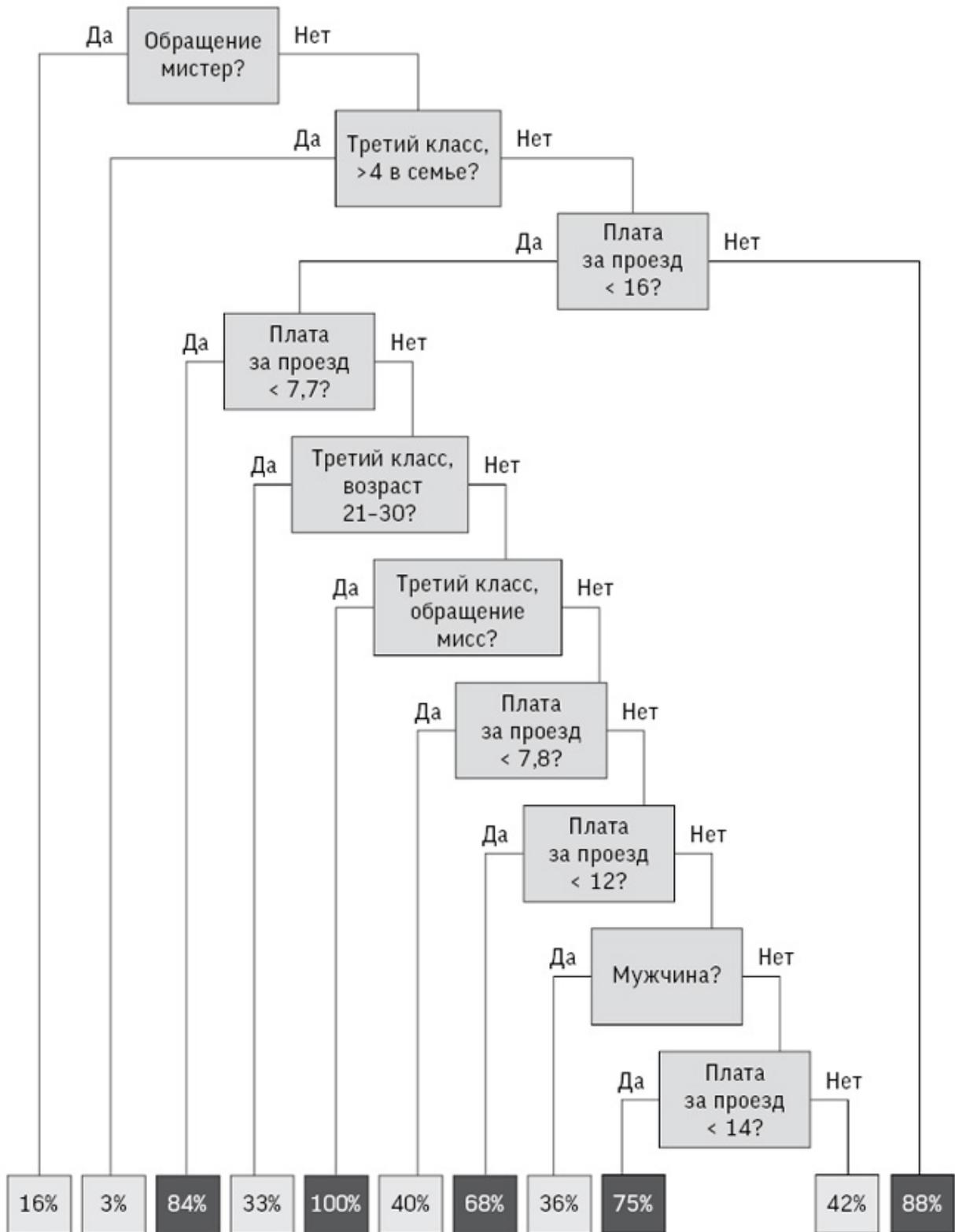
Напротив, показатель Бриера для простого дерева классификации составляет 0,139, то есть дает 40-процентное уменьшение по сравнению с наивным алгоритмом и поэтому демонстрирует

значительные умения. Другой способ интерпретации числа 0,139: именно его вы бы получили, если бы приписали всем выжившим 63-процентные шансы на выживание, а всем погибшим – 63-процентные шансы на гибель. Посмотрим, можно ли улучшить показатель с помощью более сложных моделей, но сначала предупреждаю: они не должны быть *слишком* сложными.

### ***Переобучение***

Нам необязательно останавливаться на простом дереве классификации, показанном на [рис. 6.3](#). Мы могли бы его усложнить, добавив больше ветвей, что позволило бы правильно классифицировать больше случаев из тренировочного набора, поскольку мы идентифицировали бы все больше и больше его особенностей.

На рис. 6.6 показано дерево, построенное с учетом многих уточняющих факторов. Его точность на тренировочном наборе составляет 83 %, то есть больше, чем у первоначального маленького дерева. Но при использовании алгоритма на тестовом наборе точность упадет до 81 %, то есть будет такой же, как и у маленького дерева, а показатель Бриера составит 0,150 – явно хуже, чем 0,139 для исходного дерева. Мы слишком хорошо подогнали наш алгоритм под тренировочный набор данных, и в результате его прогностическая способность начала снижаться.



**Рис. 6.6**

Чрезмерно подогнанное к данным дерево классификации для данных о «Титанике». Как на рис. 6.3, процентная доля в конце каждой ветви – это доля выживших из тренировочного набора; каждый новый пассажир, по прогнозу, выживет, если эта доля больше 50 %. Довольно странный перечень вопросов подтверждает мысль, что дерево слишком приспособлено к конкретным случаям в тренировочном наборе

Такое явление называется **переобучением**, или **переподгонкой**<sup>[134]</sup>, и считается одним из самых важных вопросов при конструировании алгоритмов. Слишком усложняя алгоритм, мы фактически начинаем подгонять его под шум, а не под сигнал. Рэндел Манро (карикатурист, автор веб-комикса *xkcd*) придумал прекрасную иллюстрацию к переобучению, предложив правдоподобные «правила», которых до определенного момента придерживались американские президенты, но в итоге каждое все равно нарушалось<sup>[135]</sup>. Например:

- Никто из республиканцев не становился президентом, не победив в Палате представителей или Сенате – до Эйзенхауэра в 1952 году.
- Католики не могут победить – до Кеннеди в 1960 году.
- Никто не становился президентом после развода – до Рейгана в 1980 году.

И так далее, включая некоторые совершенно вычурные правила, скажем, такое:

- Никто из действующих президентов-демократов без боевого опыта не выигрывал у человека, чье первое имя дает больше очков в «Скрэббл»<sup>[136]</sup> – до тех пор, пока Билл Клинтон не победил Боба Доула в 1996 году (имя Билл дает шесть очков в «Скрэббл», а имя Боб – семь).

Переобучение происходит, когда мы заходим слишком далеко в стремлении приспособиться к локальным обстоятельствам, в благородном, но ложном порыве устранить смещение и учесть всю имеющуюся информацию. В обычных условиях мы бы аплодировали такой цели, однако эта очистка означает, что у нас становится меньше данных для работы и поэтому падает надежность. Стало быть, переобучение приводит к уменьшению смещения, но за счет большей

неопределенности или разброса в оценках, поэтому защиту от него иногда называют **дилеммой смещения – дисперсии**.

Мы можем проиллюстрировать эту тонкую идею на следующем примере. Представьте огромную базу данных о продолжительности человеческой жизни, которая используется для предсказания вашего будущего здоровья, – например, ваши шансы дожить до 80 лет. Мы могли бы взять ваших сверстников с таким же, как у вас, социально-экономическим положением и посмотреть, что с ними случилось. Например, если их 10 тысяч и из них 80-летнего возраста достигли 8 тысяч, то можно оценить, что шансы дожить до 80 лет у вас и подобных вам людей составляют 80 %, причем можно быть вполне уверенным в этом числе, ведь оно основано на очень большой выборке.

Но такая оценка использует всего пару признаков, по которым вы сопоставляете с другими случаями из базы данных, и игнорирует множество индивидуальных характеристик, способных повлиять на наш прогноз, например недостаток внимания к своему здоровью или вредные привычки. Поэтому можно применить другую стратегию – найти более похожих на вас людей, скажем с тем же весом, ростом, артериальным давлением, уровнем холестерина, сходной физической активностью, которые курят, пьют, едят столько же, как и вы, и так далее. Предположим, что сопоставляя все больше и больше личных характеристик, мы сузили поиск до двух человек в базе данных, которые почти идеально совпадают с вами. Допустим, один дожил до 80, а второй – нет. Значит ли это, что ваши шансы достичь этого возраста равны 50 %? Эта величина в каком-то смысле имеет меньшее смещение, поскольку выборка вам максимально соответствует, но поскольку в ней всего два человека, оценка менее надежна, то есть у нее больше дисперсия.

Интуитивно понятно, что где-то между этими двумя крайними случаями есть золотая середина, и найти этот баланс трудно, но важно. Методы для устранения переобучения включают регуляризацию, когда поощряются сложные модели, но при этом влияние переменных приближается к нулю. Но, пожалуй, чаще всего используется простая, но мощная идея **перекрестной проверки** при конструировании алгоритма.

Важно проверять любые прогнозы на независимом тестовом наборе, который не использовался при обучении алгоритма, но мы это

можем сделать только после окончания процесса разработки. Это укажет нам на наличие переобучения, но не поможет строить алгоритм. Однако мы можем симитировать этот независимый тестовый набор, выделив на него, скажем, 10 % из тренировочных данных. Тогда мы отработаем алгоритм на оставшихся 90 %, а протестируем на выделенных 10 %. При этом процедуру можно провести не один раз – в нашем случае десять, ведь мы можем каждый раз брать в качестве тестового набора разные 10 процентов данных, и тогда у нас будет десять повторов – иными словами, десятикратная перекрестная проверка<sup>[137]</sup>.

Все алгоритмы, описанные в данной главе, имеют какие-то настраиваемые параметры, которые предназначены главным образом для контроля сложности итогового алгоритма. Например, стандартная процедура построения дерева классификации – сначала сконструировать очень сложное дерево со множеством ветвей, намеренно сделав его переобученным, а затем обрезать дерево до чего-то более простого и надежного. Такая обрезка контролируется параметром сложности, который можно выбирать с помощью процедуры перекрестной проверки. Для каждой из создающихся при этом десяти выборок строится дерево для каждого из ряда параметров сложности. Затем для каждого значения параметра вычисляется средняя предсказательная эффективность по всем десяти перекрестным проверкам. До определенного момента эта средняя эффективность улучшается, а потом падает – когда деревья становятся слишком сложными. Оптимальное значение для параметра сложности – то, которое дает оптимум при перекрестной проверке, и далее оно используется для построения дерева на всем тренировочном наборе, которое и будет итоговым вариантом.

Десятикратная перекрестная проверка применялась как для выбора параметра сложности для дерева на [рис. 6.3](#), так и для выбора параметров настройки во всех моделях, которые мы рассмотрим далее.

### ***Регрессионные модели***

Из главы 5 мы узнали, что суть регрессионной модели – построить простую формулу для предсказания результата. Переменной отклика для ситуации с «Титаником» будет результат типа «да/нет», который

указывает, выжил человек или нет, поэтому логистическая регрессия уместна (как и в случае с операциями на сердце у детей, см. рис. 5.2).

В табл. 6.3 приведены результаты подбора для логистической регрессии. При обучении использовался *бустинг*<sup>[138]</sup> – итеративная процедура, призванная уделять повышенное внимание сложным случаям: неверно классифицированные люди из тренировочного набора получали при следующей итерации повышенный вес. Последовательность итераций создавалась с помощью десятикратной перекрестной проверки.

### Таблица 6.3

Коэффициенты для признаков в логистической регрессии для данных о выживании пассажиров «Титаника»: отрицательные коэффициенты понижают шансы на выживание, а положительные – повышают

Характеристика	Оценка
Начальная оценка	3,20
Третий класс	-2,30
Мистер	-3,86
Мужчина в третьем классе	+1,43
Редкая форма обращения	-2,73
Возраст 51–60 во втором классе	-3,62
Каждый член семьи	-0,38

Для общей оценки выживаемости можно добавлять коэффициенты для признаков конкретного пассажира. Например, Фрэнсис Сомертон начинает с параметра 3,20. Затем вычитается 2,30 за то, что он плыл в третьем классе, и еще 3,86 за обращение «мистер», но потом добавляется 1,43, так как он был мужчиной в третьем классе. Еще 0,38 он теряет, поскольку его семья состоит из одного человека. В итоге его общая оценка составляет  $-1,91$ , что переводится в вероятность выживания 13 %, то есть чуть меньше, чем те 16 %, которые давало простое дерево классификации<sup>[139]</sup>.

Это «линейная» система, но обратите внимание, что в нее входят некоторые сложные **комбинированные признаки**, например положительный балл за сочетание мужчины и третьего класса несколько компенсирует значительные отрицательные баллы за третий класс и обращение «мистер», которые уже были учтены. Хотя мы сосредоточены на предсказательной эффективности, эти коэффициенты действительно дают определенное понимание важности различных признаков.

Существует множество более сложных регрессионных подходов для работы с масштабными сложными задачами – например, нелинейные модели или процесс LASSO, который одновременно оценивает коэффициенты и выбирает подходящие предикторные переменные, фактически считая их коэффициенты нулевыми<sup>[140]</sup>.

### ***Более сложные методы***

Деревья классификации и регрессионные модели возникают из нескольких различных философий моделирования: деревья пытаются построить простые правила, которые определяют группы случаев со сходными ожидаемыми результатами, в то время как регрессионные модели сосредоточены на весах, которые придаются тому или иному признаку, безотносительно к тому, что еще наблюдалось для этого случая.

Специалисты по машинному обучению используют не только деревья классификации и регрессии, они разработали множество альтернативных, более сложных методов для создания алгоритмов. Например:

- *Случайные леса* – состоят из большого количества деревьев, каждое из которых производит какую-то классификацию. Итоговая классификация определяется большинством голосов – процесс известен как бэггинг<sup>[141]</sup>.

- *Методы опорных векторов* – пытаются найти линейные комбинации признаков, которые лучше всего разделяют различные результаты.

- *Нейронные сети* – включают слои узлов, каждый узел зависит от предыдущего слоя с какими-то весами, то есть получается нечто вроде ряда логистических регрессий, наложенных друг на друга. Веса определяются с помощью процедур оптимизации; подобно случайным лесам, можно построить и усреднить несколько нейронных сетей. Нейронные сети с большим количеством слоев известны как модели глубокого обучения: говорят, что Inception – система распознавания образов Google – имеет больше двадцати слоев и свыше 300 тысяч параметров для оценки.

- *Метод k-ближайших соседей* – классифицирует объект в соответствии с большинством случаев среди ближайших соседей в тренировочном наборе, то есть присваивает объекту тот класс, который чаще всего встречается среди его k соседей.

В табл. 6.4 приведены результаты применения некоторых из этих методов к данным о «Титанике»; параметры настройки выбирались с помощью десятикратной перекрестной проверки, а в качестве критерия оптимизации использовались ROC-кривые.

#### **Таблица 6.4**

Эффективность различных алгоритмов для тестового набора данных о «Титанике»: полужирным шрифтом выделены наилучшие результаты. Сложные алгоритмы оптимизировались по максимизации площади под ROC-кривой

Метод	Точность (чем больше, тем лучше)	Площадь под ROC- кривой (чем больше, тем лучше)	Показатель Бриера (чем меньше, тем лучше)
Каждый имеет шансы на выживание 39%	0,639	0,500	0,232
Все женщины выживают, все мужчины погибают	0,786	0,578	0,214
Простое дерево классификации	<b>0,806</b>	0,819	<b>0,139</b>
Дерево классификации с переобучением	<b>0,806</b>	0,810	0,150
Логистическая регрессия	0,789	0,824	0,146
Случайный лес	0,799	<b>0,850</b>	0,148
Метод опорных векторов	0,782	0,825	0,153
Нейронная сеть	0,794	0,828	0,146
Усредненная нейронная сеть	0,794	0,837	0,142
Метод k-ближайших соседей	0,774	0,812	0,180

Высокая точность простейшего правила «все женщины выживают, все мужчины погибают», которая либо превосходит точность более сложных алгоритмов, либо лишь незначительно им уступает, демонстрирует неадекватность грубой «точности» как меры эффективности. Метод случайного леса дает наилучшее различение, отражаемое площадью под ROC-кривой, а (что, вероятно, удивительно) наилучший показатель Бриера дают вероятности из простого дерева классификации. Поэтому явного алгоритма-победителя нет. Позже, в главе 10, мы проверим, можно ли уверенно заявлять о победителе по каждому из этих критериев, поскольку преимущество лидеров настолько невелико, что его можно объяснить случайными отклонениями, например распределением, кто из пассажиров оказался в тренировочном наборе, а кто – в тестовом.

Это отражает общую озабоченность тем, что алгоритмы, выигрывающие соревнования на платформе Kaggle, имеют тенденцию к колоссальной сложности – и все ради крохотного итогового преимущества, необходимого для победы. Основная проблема в том, что эти алгоритмы напоминают непостижимые черные ящики – они выдают прогноз, но почти невозможно понять, что творится у них внутри. Здесь можно выделить три негативных аспекта. Во-первых, чрезвычайная сложность сильно затрудняет реализацию и модернизацию: когда Netflix предложила приз в миллион долларов за рекомендательные системы, победитель оказался настолько сложным, что Netflix его не использовала. Во-вторых, мы не знаем, откуда взялось такое заключение и насколько мы можем быть в нем уверены: нам приходится либо принимать, либо отказываться от него. Более простые алгоритмы легче для понимания. В-третьих, не понимая, как алгоритм выдает ответ, мы не можем исследовать его на наличие неявных, но систематических ошибок в отношении определенных участников сообщества – об этом мы подробно поговорим далее.

Все это указывает на то, что количественная эффективность может быть не единственным критерием алгоритма: как только она становится «достаточно хорошей», порой куда разумнее отказаться от попыток дальнейших небольших увеличений ради сохранения простоты.

**Так кто оказался самым везучим на «Титанике»?**

Таким выжившим можно считать человека с самым высоким показателем Бриера при усреднении по всем алгоритмам. Им стал Карл Даль, 45-летний норвежско-австралийский столяр<sup>[142]</sup>, путешествовавший в одиночку в третьем классе и заплативший за билет столько же, сколько и Фрэнсис Сомертон; два алгоритма даже дали ему 0 % шансов на выживание. Попав в ледяную воду, он забрался на спасательную шлюпку № 15, несмотря на то что некоторые на шлюпке пытались столкнуть его обратно. Возможно, он просто применил силу.

Это резко контрастирует с судьбой Фрэнсиса Сомертона из Илфракомба, чья смерть, как мы видели, вполне вписывается в общую закономерность. Его жене Ханне Сомертон досталось всего 5 фунтов (меньше, чем Фрэнсис потратил на билет), а не успешный муж в Америке.

### ***Проблемы алгоритмов***

Алгоритмы способны демонстрировать замечательную эффективность, однако по мере увеличения их роли в обществе актуализируются и их потенциальные проблемы. На данный момент можно выделить четыре основные.

- *Недостаток робастности (устойчивости).* Алгоритмы создаются по связям, и в случае непонимания лежащих в их основе процессов они могут оказаться слишком чувствительны к изменениям. Даже если нас волнует исключительно точность, а не научная истина, нам по-прежнему нужно помнить базовые принципы цикла PPDAC и этапы перехода от данных, полученных из какой-то выборки, к утверждениям, касающимся всей целевой совокупности. Для предсказательной аналитики эта целевая совокупность включает будущие случаи, и если все остается по-прежнему, то алгоритмы, сконструированные по прошлым данным, должны работать хорошо. Но порой мир меняется. Мы уже отмечали провал алгоритмов при изменениях в финансовом мире 2007–2008 годов. Еще один яркий пример – попытка компании Google предсказать тенденции распространения вируса гриппа на основании закономерностей в поисковых запросах пользователей. Сначала все работало хорошо, но в 2013 году алгоритм начал резко завышать прогнозы для гриппа. Одно

из объяснений – изменения, внесенные Google в поисковую систему, могли привести к большему количеству ключевых слов, указывавших на грипп.

- *Отсутствие учета статистического разброса.* Автоматическое ранжирование на основе ограниченного объема данных будет ненадежным. В США учителей оценивали и наказывали в соответствии с коэффициентом роста знаний их учеников за год, что проявлялось в невероятно резких изменениях в годовой оценке учителей: в Вирджинии, например, у четверти учителей фиксируется разница более чем в 40 баллов (по шкале 1–100 баллов) от года к году<sup>[143]</sup>. Но как такое может быть, ведь хорошие учителя обычно хороши и в этом году, и в следующем.

- *Неявное смещение.* Повторюсь, алгоритмы основаны на связях, а это может означать, что в итоге они используют признаки, которые мы, как правило, считаем не имеющими отношения к рассматриваемой задаче. Например, когда один алгоритм машинного зрения обучали отличать изображения хаски от немецких овчарок, он был эффективен, пока его не применили к хаски, которых содержали в квартирах в качестве домашних питомцев, – оказалось, что его эффективность основывалась на идентификации снега на заднем плане<sup>[144]</sup>. Менее тривиальные примеры включают алгоритм для определения красоты, которому не нравится темная кожа, и еще один алгоритм, идентифицирующий чернокожих людей как горилл. Некоторые алгоритмы способны серьезно повлиять на жизнь человека – например, присваивающие кредитный рейтинг или решающие вопросы страхования. Можно запретить использование расы в качестве одной из предикторных переменных, но применение почтовых индексов для указания местожительства может быть не менее мощным индикатором расы.

- *Недостаток прозрачности.* Некоторые алгоритмы могут быть непрозрачными из-за своей явной сложности. Но даже простые алгоритмы на базе регрессии становятся совершенно непостижимыми в случае закрытости (например, если речь идет о частном коммерческом продукте). Это одна из главных жалоб на алгоритмы, определяющие вероятность рецидива, такие как COMPAS компании Northpointe или LSI-R у MMR<sup>[145]</sup>. Они дают оценки или категории риска, которые можно использовать для принятия решений об условном осуждении

или вынесении приговора, но при этом способ взвешивания различных факторов неизвестен. Более того, поскольку собирается информация и о воспитании, и о прошлых соучастниках, решения принимаются на основании не только личной истории, но и с учетом фоновых факторов, которые (как было показано) должны быть связаны с будущими преступными действиями, даже если они обусловлены таким фактором, как бедность и лишения. Конечно, если бы все, что имеет значение, давало точный прогноз, то можно было бы использовать любой признак, даже расовую принадлежность. Однако многие утверждают, что справедливость требует, чтобы такие алгоритмы были контролируемы, прозрачны и подлежали обжалованию.

В какой-то степени можно объяснить даже собственные (проприетарные) алгоритмы – при условии, что мы можем экспериментировать с различными входными сигналами. При покупке онлайн-страховки предлагаемая сумма рассчитывается по некой неизвестной формуле, подчиняющейся только определенным юридическим ограничениям: например, в Великобритании расценки при страховании автомобиля не могут опираться на информацию о расе или на генетические данные (за исключением болезни Хантингтона<sup>[146]</sup>) и так далее. Однако мы все же можем получить представление о влиянии различных факторов, раз за разом давая ложные ответы и наблюдая, как меняются расценки: это предоставляет некоторую возможность обратного инжиниринга<sup>[147]</sup> для алгоритма – увидеть, какие факторы определяют предлагаемую сумму.

Необходимость в контролируемости алгоритмов, влияющих на жизнь людей, растет, и требования, чтобы выводы имели понятное объяснение, включаются в законодательство. Такие требования препятствуют использованию сложных черных ящиков и могут приводить к предпочтению (довольно старомодных) регрессионных алгоритмов, в которых влияние каждого фактора предельно ясно.

Однако, рассмотрев темную сторону алгоритмов, уместно закончить весьма полезным и обнадеживающим примером.

### **Какова ожидаемая польза от адьювантной терапии после операции при раке молочной железы?**

Почти всем женщинам, у которых диагностирован рак молочной железы, делают операцию, хотя степень хирургического вмешательства

может быть ограниченной. Критический вопрос – выбор адъювантной терапии<sup>[148]</sup> после операции, чтобы уменьшить вероятность рецидива и последующей смерти от рака. Возможны различные варианты – радиотерапия, гормональная терапия, химиотерапия и другие лекарственные средства. В терминологии цикла PPDAC это *проблема*.

План британских исследователей состоял в разработке алгоритма, который помогал бы принять такое решение с помощью *данных* о 5700 прошлых случаях рака молочной железы, имеющих в канцер-регистре страны. Анализ включал построение алгоритма, использующего подробную информацию о конкретной женщине и ее опухоли для вычисления ее шансов на выживание в течение 10 лет после операции и их изменения в зависимости от различных методов лечения. Однако при анализе прошлых результатов требуется осторожность, поскольку причины выбора методов лечения неизвестны и мы не можем использовать видимые результаты из базы данных. Вместо этого создается регрессионная модель с выживаемостью в качестве результата, но эффект лечения оценивается по обзорам крупномасштабных клинических испытаний. Получившийся алгоритм общедоступен, а его разделительная эффективность и калибровка проверены на независимых наборах данных, включавших 27 тысяч женщин<sup>[149]</sup>.

Созданное программное обеспечение называется Predict 2.1, и результаты его работы выдаются в виде доли схожих по анамнезу женщин, которые, как ожидается, проживут 5 и 10 лет при различных видах адъювантной терапии. В табл. 6.5 приведены некоторые результаты для воображаемой женщины, а на рис. 6.7 – кривые выживаемости из Predict 2.1 для периода до 15 лет после операции.

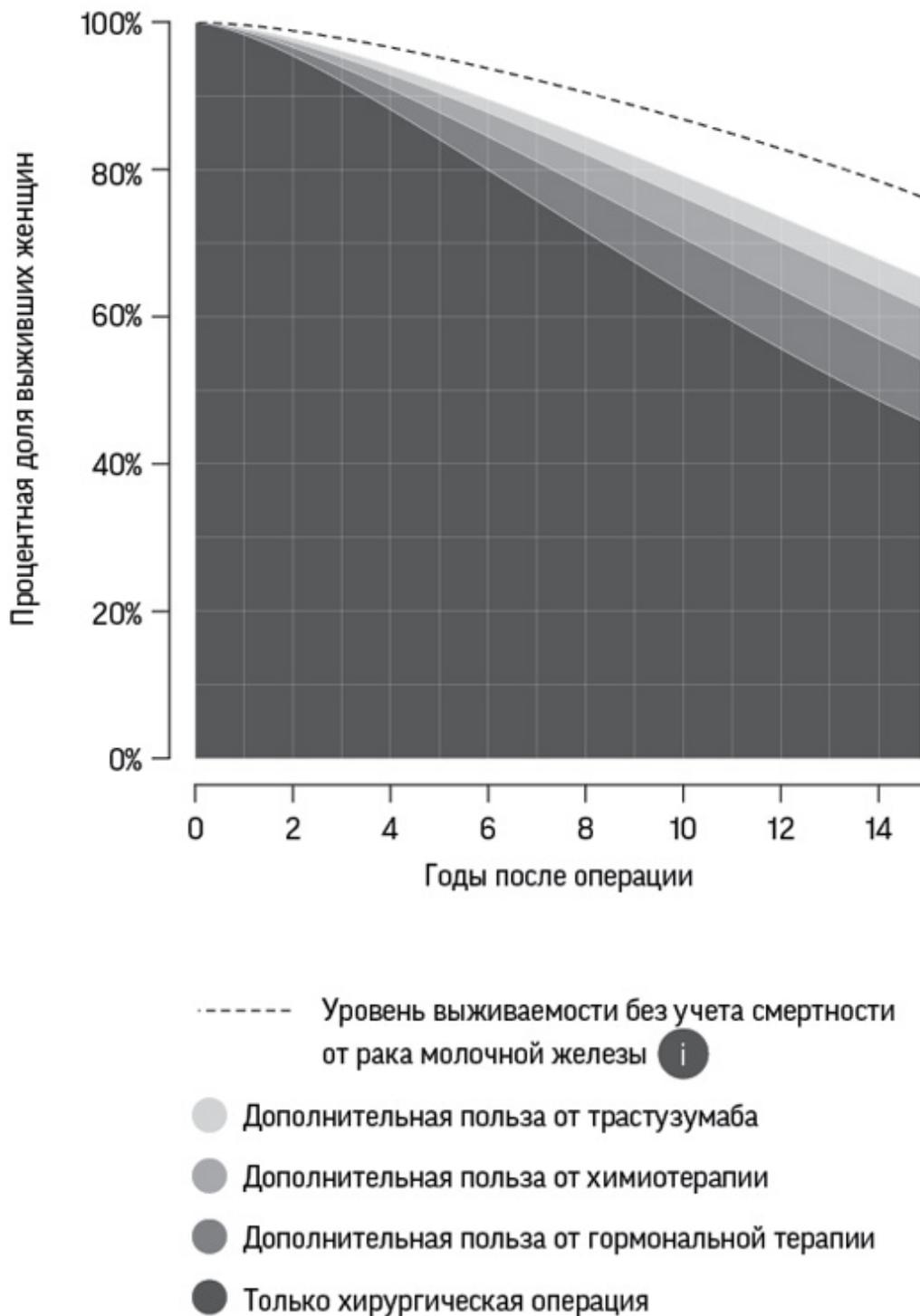
### **Таблица 6.5**

С помощью алгоритма Predict 2.1 определялась ожидаемая доля 65-летних женщин, которые проживут 10 лет после операции при раке молочной железы, когда при обследовании была обнаружена опухоль 2-й стадии размером 2 см, с двумя метастазами узлов и положительными индексами ER, HER2 и Ki-67. Показана кумулятивная ожидаемая польза для различных методов адъювантной терапии, хотя они могут иметь побочные эффекты. Доля выживаемости для «женщин без рака»

отражает наилучшую возможную выживаемость с учетом возраста женщины

<b>Метод лечения</b>	<b>Дополнительная польза сверх предыдущего лечения, %</b>	<b>Общая выживаемость, %</b>
Только хирургическая операция	—	64
+ гормональная терапия	7	70
+ химиотерапия	6	76
+ трастузумаб (герцептин)*	3	79
Для женщин без рака		87

\* Трастузумаб – противоопухолевый препарат. Герцептин – его торговое название. *Прим. пер.*



**Рис. 6.7**

Кривые выживаемости из Predict 2.1 для периода до 15 лет после операции – для женщин с признаками, перечисленными в подписи к табл. 6.5. Показано совокупное увеличение выживаемости в зависимости от дополнительных видов терапии. Область над

пунктирной линией отображает процент женщин с раком молочной железы, умерших от других причин

Программа Predict 2.1 не совершенна, и данные в табл. 6.5 могут использоваться только в качестве примерных ориентиров: они отображают, что (как можно ожидать) случится с женщинами, обладающими признаками, включенными в алгоритм, но для какой-то конкретной женщины нужно учитывать дополнительные факторы. Тем не менее Predict 2.1 регулярно используется для десятков тысяч случаев в месяц как на консилиумах, где обсуждаются варианты лечения, так и для передачи этой информации пациентке. Для женщин, желающих активно участвовать в своем лечении, используется процесс, известный как «совместный уход», когда пациентке предоставляется информация, обычно доступная только врачам, что может помочь ей лучше контролировать свою жизнь. Этот алгоритм не запатентован и представляет собой программное обеспечение с открытым исходным кодом, поэтому система регулярно обновляется, чтобы предоставлять дополнительную информацию, в том числе о и негативных последствиях лечения.

### ***Искусственный интеллект***

С момента появления в 1950-е годы идея искусственного интеллекта (ИИ) периодически переживает всплески ажиотажа и энтузиазма и последующие волны критики. Я работал над диагностикой с помощью компьютера и неопределенностью в ИИ в 1980-е, когда в целом эта тема обсуждалась в рамках конкуренции между различными подходами: как основанными на вероятности и статистике или на экспертных «правилах» для суждений, так и теми, которые пытались эмулировать когнитивные способности с помощью нейронных сетей. Сейчас эта область стала более зрелой, подход к ее основополагающей философии – более прагматичным и универсальным, хотя ажиотаж не исчез.

Демонстрируемый машинами интеллект – весьма широкая идея. Она намного масштабнее, чем ограниченная проблема алгоритмов, обсуждаемая в этой главе, и статистический анализ – всего лишь один компонент для построения систем ИИ. Однако, как показывают последние выдающиеся достижения в компьютерном зрении, речи,

играх и так далее, главную роль в успехах в «узком» ИИ играет статистическое обучение. Такие системы, как Predict, которые ранее считались системами принятия решений на базе статистики, теперь можно с полным основанием именовать ИИ<sup>[150]</sup>.

Многие из вышеописанных проблем сводятся к алгоритмам, моделирующим только связи и не имеющим понятия о процессах, лежащих в основе. Джуда Перл, благодаря которому повысилось внимание к причинным связям в ИИ, утверждает, что такие модели позволяют всего лишь отвечать на вопросы типа «Мы наблюдаем X, чего мы можем ожидать от следующего наблюдения?» В то же время общий ИИ нуждается в причинно-следственной модели того, как реально функционирует мир, что позволило бы ему отвечать на вопросы человеческого уровня, касающиеся результатов вмешательства («Что будет, если мы сделаем X?») и контрфактуального мышления («Что было бы, если бы мы не сделали X?»). Пока мы весьма далеки от искусственного интеллекта, обладающего такой способностью.

Эта книга подчеркивает классические статистические проблемы малых выборок, систематические ошибки (в статистическом смысле) и невозможность обобщения на новые ситуации. Список задач для алгоритмов показывает, что хотя беспокоиться о величине выборки можно меньше ввиду наличия колоссальных объемов данных, другие проблемы имеют тенденцию усугубляться и мы сталкиваемся с дополнительной проблемой – объяснением для обоснования алгоритма. Огромные массивы данных только увеличивают проблемы при получении надежных выводов. Смирение при построении алгоритмов крайне важно.

### **Выводы**

- Алгоритмы, построенные на основе данных, можно использовать в различных технологических приложениях для классификации и прогнозирования.
- Важно остерегаться переобучения алгоритма на тренировочных данных, когда, по сути, происходит подгонка не к сигналу, а к шуму.
- Алгоритмы можно оценивать по точности классификации, способности различать отдельные группы, а также общей точности прогнозирования.

- Сложным алгоритмам может не хватать прозрачности, поэтому, возможно, разумнее потерять немного в точности ради понимания.

- Использование алгоритмов и искусственного интеллекта сопряжено со многими трудностями, поэтому важно осознавать как мощь, так и ограничения методов машинного обучения.

## Глава 7. Насколько мы можем быть уверены в происходящем? Оценки и интервалы

### Сколько в Великобритании безработных?

В январе 2018 года новостной сайт «Би-би-си» объявил, что за три месяца до прошедшего ноября «уровень безработицы в Соединенном Королевстве снизился на 3 тысяч и составил 1,44 миллиона человек». О причинах такого сокращения много спорили, но, как ни странно, никто не усомнился в точности этой цифры. Однако при тщательной проверке Бюро национальной статистики Великобритании обнаружило, что **погрешность** этой величины составляет  $\pm 77\ 000$ . Иными словами, истинное изменение могло колебаться от снижения на 80 тысяч до увеличения на 74 тысячи. Таким образом, хотя журналисты и политики считали, что заявленное сокращение касается всей страны, фактически это была неточная оценка, основанная на опросе примерно 100 тысяч человек<sup>[151]</sup>. Аналогично, когда Бюро статистики труда США сообщило о росте безработицы среди гражданского населения на 108 тысяч человек между декабрем 2017 и январем 2018 года, эта оценка опиралась на выборку примерно из 60 тысяч домохозяйств, а погрешность (которую опять же трудно определить) составляла  $\pm 300\ 000$ <sup>[152]</sup>,<sup>[153]</sup>.

Осознавать неопределенность крайне важно. Сделать какую-нибудь оценку способен кто угодно, но умение реалистично определить ее возможную погрешность – важнейший компонент статистики. Даже притом, что это затрагивает некоторые сложные понятия.

Предположим, мы собрали какие-то точные данные, возможно, с помощью хорошо спланированного опроса, и хотим обобщить результаты на изучаемую совокупность. Если мы проявляли осторожность и избегали внутренних смещений (скажем, обеспечив случайную выборку), то можем ожидать, что характеристики выборки

будут близки к соответствующим характеристикам изучаемой совокупности.

Этот важный момент стоит уточнить. В хорошем исследовании мы ожидаем, что выборочное среднее будет близко к среднему всей совокупности, интерквартильный размах в выборке будет близок к интерквартильному размаху всей совокупности и так далее. В [главе 3](#) мы рассматривали идею характеристик всей совокупности на примере данных о весе новорожденных, где назвали выборочное среднее *статистикой*, а среднее всей совокупности – *параметром*. В более строгих статистических текстах эти две величины обычно обозначают римскими и греческими буквами соответственно – скорее всего, в обреченной (вероятно) попытке избежать путаницы. Например, латинской буквой  $m$  часто обозначают выборочное среднее, а греческой буквой  $\mu$  (мю) – среднее всей совокупности, буквой  $s$  – выборочное среднеквадратичное отклонение, а буквой  $\sigma$  (сигма) – среднеквадратичное отклонение всей совокупности.

Часто сообщают только итоговую статистику, и во многих случаях этого может быть достаточно. Например, мы видели, что большинство людей не знают, что показатели безработицы в США и Соединенном Королевстве основаны не на полном подсчете всех официально зарегистрированных безработных, а на масштабных опросах. Если такой опрос установил, что 7 % людей в выборке безработные, то национальные агентства и СМИ обычно преподносят это как факт, что 7 % всего населения страны безработные, вместо того чтобы признать, что 7 % – это всего лишь оценка. Выражаясь научно более точно, они просто путают выборочное среднее и среднее во всей совокупности.

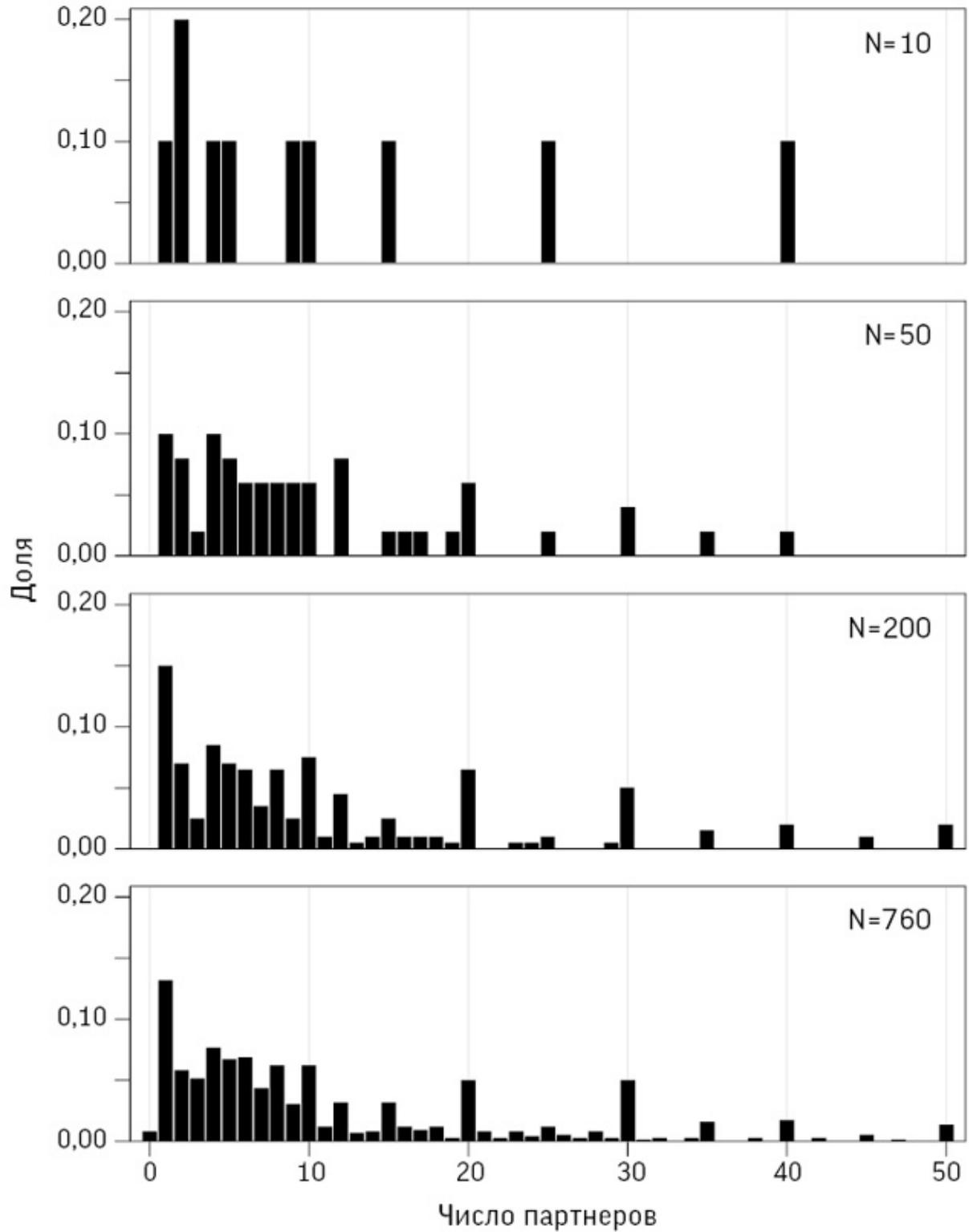
Это может оказаться неважным при намерении просто представить широкую картину происходящего в стране, когда опрос масштабен и надежен. Но давайте возьмем такой пример: вы услышали, что опрошены только 100 человек, из которых семь сказали, что не имеют работы. Оценка составляет 7 %, но, вероятно, вряд ли вы сочли бы ее надежной и были бы счастливы, если бы она описывала всю совокупность. А если бы в опросе участвовала 1000 человек? А 100 тысяч? При достаточном масштабе опроса вы, возможно, увереннее согласитесь с тем, что выборочная оценка – достаточно хорошая характеристика всей совокупности. Размер выборки должен влиять на вашу уверенность в оценке, а чтобы делать статистические

выводы, необходимо знать, насколько выборочная характеристика может отличаться от настоящей.

### *Количество сексуальных партнеров*

Давайте вернемся к опросу Natsal, описанному в [главе 2](#), в котором участников спрашивали, сколько сексуальных партнеров у них было в течение жизни. В качестве респондентов было привлечено 1125 женщин и 806 мужчин в возрасте 35–44 лет, так что это был солидный опрос. В [табл. 2.2](#) представлены вычисленные выборочные характеристики, например медиана – 8 для мужчин и 5 для женщин. Поскольку мы знаем, что этот опрос базировался на правильной случайной выборке, вполне разумно предположить, что изучаемая совокупность соответствует целевой совокупности, то есть взрослому населению Великобритании. Главный вопрос здесь таков: насколько близки найденные статистики к тому, что мы обнаружили бы, опросив всех жителей страны?

В качестве иллюстрации того, как точность статистики зависит от размера выборки, представим, что мужчины в нашем опросе фактически представляют собой всю генеральную совокупность, которая нас интересует. Их ответы приведены на нижней диаграмме рис. 7.1. Для иллюстрации извлечем последовательные случайные выборки из общей совокупности из 760 участников: сначала 10, затем 50, а потом 200 человек. Распределение данных для трех выборок показано на рис. 7.1. Ясно видно, что маленькие выборки «ухабистее», поскольку они чувствительны к отдельным точкам. Сводные характеристики этих постепенно увеличивающихся выборок представлены в табл. 7.1. В первой выборке из 10 человек наблюдается большое количество партнеров (среднее 8,4), но по мере роста выборки эта величина постепенно уменьшается, приближаясь к характеристике всей группы из 760 человек.



**Рис. 7.1**

Нижняя диаграмма отображает распределение ответов для всех 760 мужчин в опросе. Из этой группы случайным образом

последовательно выбираются 10, 50 и 200 человек. Соответствующие распределения построены на первых трех диаграммах. У меньших выборок видны значительные разбросы, но постепенно форма распределения приближается к распределению всей группы из 760 мужчин. Не показаны значения свыше 50 партнеров

**Таблица 7.1**

Сводные статистические данные о количестве сексуальных партнеров за всю жизнь у мужчин в возрасте 35–44 лет, которое они указывали в исследовании Natsal 3 (случайные выборки и характеристики всей группы из 760 мужчин)

<b>Размер выборки</b>	<b>Среднее число партнеров</b>	<b>Медианное число партнеров</b>
10	8,3	9
50	10,5	7,5
200	12,2	8
760	11,4	7

А теперь вернемся к фактической задаче: что мы можем сказать о среднем и медианном числе партнеров во всей изучаемой совокупности мужчин в возрасте 35–44 лет, основываясь на реальных выборках мужчин, показанных на рис. 7.1? Мы могли бы оценить эти параметры всей популяции по выборочной статистике каждой группы, представленной в табл. 7.1, предполагая, что статистики на основе больших выборок в каком-то смысле «лучше»: например, оценки среднего количества партнеров сходятся к 11,4, и при достаточно большой выборке мы, скорее всего, приблизились бы к истинному ответу с желаемой точностью.

Вот здесь мы подошли к критическому шагу. Чтобы понять, насколько точными могут быть такие характеристики, нам нужно подумать, как эти статистики могут измениться, если мы (в

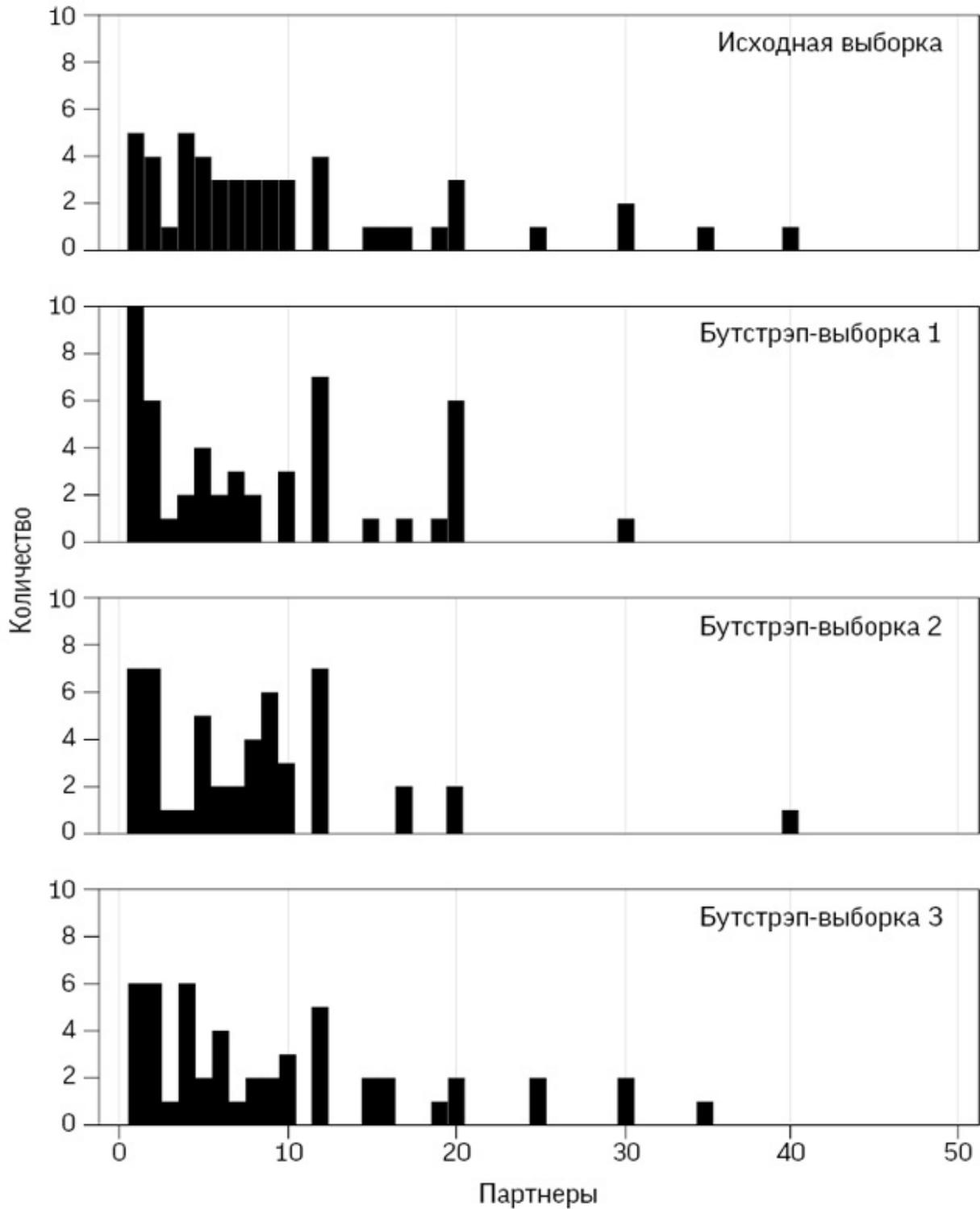
воображении) неоднократно повторим процесс составления выборки. Иначе говоря, если бы мы раз за разом формировали выборки из 760 британцев, насколько сильно менялись бы их статистики?

Если бы мы знали, как сильно они будут варьироваться, это помогло бы нам понять, насколько точна наша фактическая оценка. К сожалению, определить точный разброс оценок мы могли бы только в случае, если бы точно знали информацию о всей генеральной совокупности. Но как раз этого мы и не знаем.

Есть два способа выбраться из этого круга. Первый – сделать какие-то математические предположения о форме исходного распределения в генеральной совокупности, а затем с помощью методов теории вероятностей определить ожидаемый разброс для нашей оценки, а потом и то, чего можно ожидать для разницы между средним в выборке и средним во всей совокупности. Это традиционный способ, который включают в учебники по статистике; мы рассмотрим в главе 9, как он работает.

Но есть и альтернативный подход, основанный на правдоподобном предположении, что вся популяция должна быть примерно схожа с выборкой. Поскольку мы не можем извлечь еще несколько выборок из общей популяции, возьмем несколько раз новые выборки из нашей выборки!

Мы можем проиллюстрировать эту идею на примере нашей предыдущей выборки размером 50, показанной на верхней диаграмме на рис. 7.2; ее среднее значение равно 10,5. Предположим, что мы берем еще 50 точек, каждый раз с возвратом уже взятого наблюдения, и получаем распределение, показанное на второй диаграмме, где среднее значение равно 8,4. Обратите внимание, что это распределение может содержать только те величины, которые есть в исходном распределении, но количество таких наблюдений будет другим, поэтому форма распределения будет слегка отличаться, а вместе с ней будет немного отличаться и среднее. Процесс можно повторять; на рис. 7.2 отображены три повторные выборки, средние значения которых равны 8,4, 9,7 и 9,8.



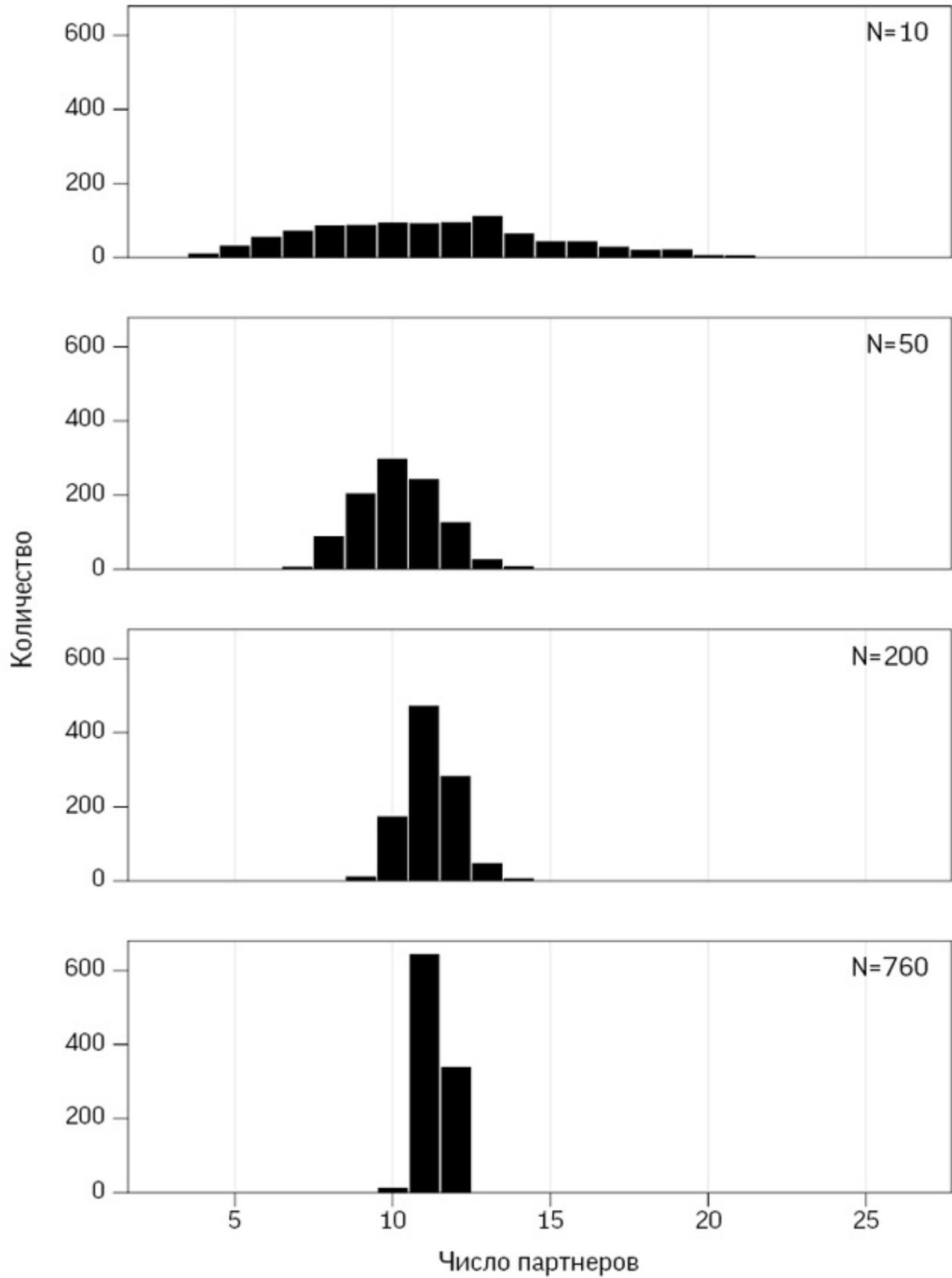
**Рис. 7.2**

Исходная выборка из 50 наблюдений и три «бутстрэп-выборки»<sup>[154]</sup>, каждая из которых состоит из 50 наблюдений, извлеченных случайным образом из исходного набора, каждый раз с

возвратом. Например, наблюдение в 25 партнеров в первоначальной выборке встречается один раз (справа). В первой и второй бутстрэп-выборках его не оказалось вовсе, а в третьей встретилось дважды

В результате мы получаем представление, как при перевыборках изменяется наша оценка. Процесс известен под названием **бутстрэппинг** – волшебная идея вытягивания себя за ремешки на обуви сопоставляется со способностью извлекать информацию из самой выборки без предположения о форме распределения всей генеральной совокупности<sup>[155]</sup>.

Если мы повторим эту процедуру, скажем, 1000 раз, то получим 1000 возможных оценок среднего. Они представлены в виде гистограммы на второй панели на рис. 7.3. Остальные гистограммы отражают бутстрэппинг для других выборок на [рис. 7.1](#), при этом каждая гистограмма показывает разброс бутстрэп-оценок вокруг среднего в исходной выборке. Это **выборочные распределения** оценок, поскольку они отражают разброс оценок, появляющийся вследствие повторных составлений выборок.



**Рис. 7.3**

Распределение средних значений для 1000 бутстрэп-выборок, построенных для размеров 10, 50, 200 и 760, отображенных на рис. 7.1. Разброс значений для среднего уменьшается по мере роста размера выборки

Рис. 7.3 отражает некоторые очевидные особенности. Первая и, возможно, самая примечательная – исчезновение практически всех следов асимметрии исходных выборок: распределения для оценок, основанных на данных из повторных выборок, почти симметричны относительно среднего в исходных данных. Это следствие центральной предельной теоремы, которая гласит, что распределение выборочных средних по мере увеличения размера выборки сходится к нормальному распределению – *практически вне зависимости от формы исходного распределения данных*. Этот важнейший результат мы рассмотрим в главе 9.

Важно отметить, что эти бутстрэп-распределения позволяют количественно выразить нашу неопределенность в оценках, показанных в [табл. 7.1](#). Например, мы можем найти диапазон, который будет содержать 95 % средних в бутстрэп-выборках, и назвать его 95-процентным интервалом неопределенности для исходных характеристик, или погрешностью. Соответствующие интервалы показаны в табл. 7.2 – симметрия бутстрэп-распределений означает, что интервалы неопределенности расположены примерно симметрично вокруг исходной оценки.

### **Таблица 7.2**

Выборочные средние для числа сексуальных партнеров за всю жизнь, указанного мужчинами в возрасте 35–44 лет в исследовании Natsal 3, для вложенных выборок размера 10, 50, 200 и полных данных о 760 мужчинах, с 95-процентными интервалами неопределенности, также называемыми погрешностями

Размер выборки	Среднее число партнеров	95-процентный интервал неопределенности
10	8,3	от 5,3 до 11,5
50	10,5	от 7,7 до 13,8
200	12,2	от 10,5 до 13,8
796	11,4	от 10,5 до 12,2

Вторая важная особенность [рис. 7.3](#) – сужение бутстрэп-распределений по мере роста выборки, что отражено в постепенном уменьшении размера 95-процентных интервалов неопределенности.

В этом разделе вы познакомились с некоторыми сложными, но важными идеями:

- разброс в статистиках, основанных на выборках;
- бутстрэппинг данных, когда мы не хотим делать предположения о форме распределения в генеральной совокупности;
- тот факт, что форма распределения статистики не зависит от формы исходного распределения, из которого взяты наблюдения.

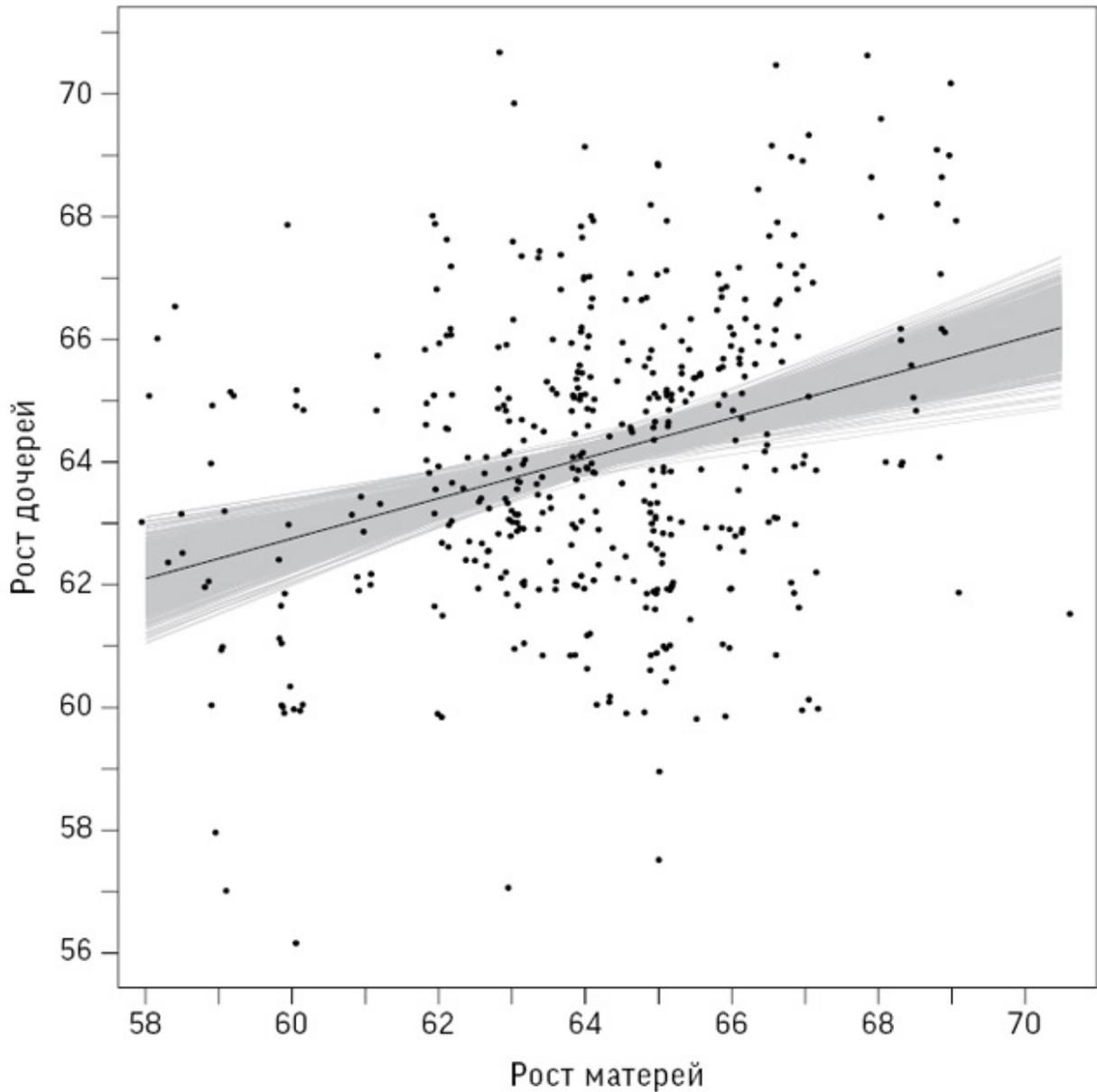
Весьма примечательно, что всего это мы достигли без помощи математики, за исключением идеи брать наблюдения случайным образом.

Теперь я покажу, что бутстрэппинг можно применять и в более сложных ситуациях.

В главе 5 мы проводили линии регрессии для данных Гальтона о росте, что позволяло предсказывать, например, рост дочерей на основе роста их матерей с помощью регрессионной прямой с угловым коэффициентом 0,33 (см. [табл. 5.2](#)). Но насколько мы можем быть

уверены в положении такой прямой? Бутстрэппинг предоставляет интуитивно понятный способ ответить на этот вопрос, не делая никаких предположений о генеральной совокупности, из которой взяты наблюдения.

Составим из 433 пар дочь/мать (рис. 7.4) повторную выборку из 433 элементов (с возвратом) и построим для нее прямую наилучшего соответствия по методу наименьших квадратов. Повторим процедуру столько раз, сколько считаем нужным: рис. 7.4 показывает построенные всего по 20 таким перевыборкам линии наилучшего соответствия, чтобы продемонстрировать их разброс. Поскольку исходный набор данных велик, разброс у этих прямых относительно небольшой – при 1000 бутстрэп-выборках угловой коэффициент с вероятностью 95 % лежит в интервале от 0,22 до 0,44.



**Рис. 7.4**

Регрессионные прямые для 20 перевыборок из данных Гальтона о росте матерей и дочерей, наложенные на исходные данные. Из-за большого размера выборки угловой коэффициент прямых изменяется относительно слабо

Бутстрэппинг обеспечивает интуитивно понятный, удобный для использования компьютера способ выразить неопределенность в оценках, не делая сильных предположений и не используя теорию вероятностей. Однако этот метод неэффективен, когда нужно найти,

например, погрешность в опросе 100 тысяч человек о безработице. Хотя бутстрэппинг – простая, блестящая и крайне эффективная идея, перерабатывать с его помощью такие огромные объемы данных неудобно, особенно при наличии теории, которая может предоставить готовые формулы для величины интервалов неопределенности. Но прежде чем мы ее рассмотрим в главе 9, познакомимся с восхитительной, хотя и непростой теорией вероятностей.

### **Выводы**

- Интервалы неопределенности – важная часть информации о характеристиках выборки.
- Бутстрэппинг – это метод создания из первоначальной выборки новых наборов данных одинакового размера посредством перевыборок с возвратом.
- Выборочные характеристики, вычисленные с помощью бутстрэп-выборок, для больших наборов данных близки к нормальному распределению – независимо от формы исходного распределения данных.
- Интервалы неопределенности, построенные с помощью бутстрэппинга, используют вычислительные мощности современных компьютеров, не требуют предположений о математическом виде генеральной совокупности и сложной теории вероятностей.

## Глава 8. Вероятность – язык неопределенности и случайности

В 1650-х годах самозванный шевалье<sup>[156]</sup> де Мере столкнулся во время игры с дилеммой. Не то чтобы он был уж слишком азартным игроком (хотя играл довольно увлеченно), но тем не менее хотел знать, в какой из двух игр у него больше шансов на победу.

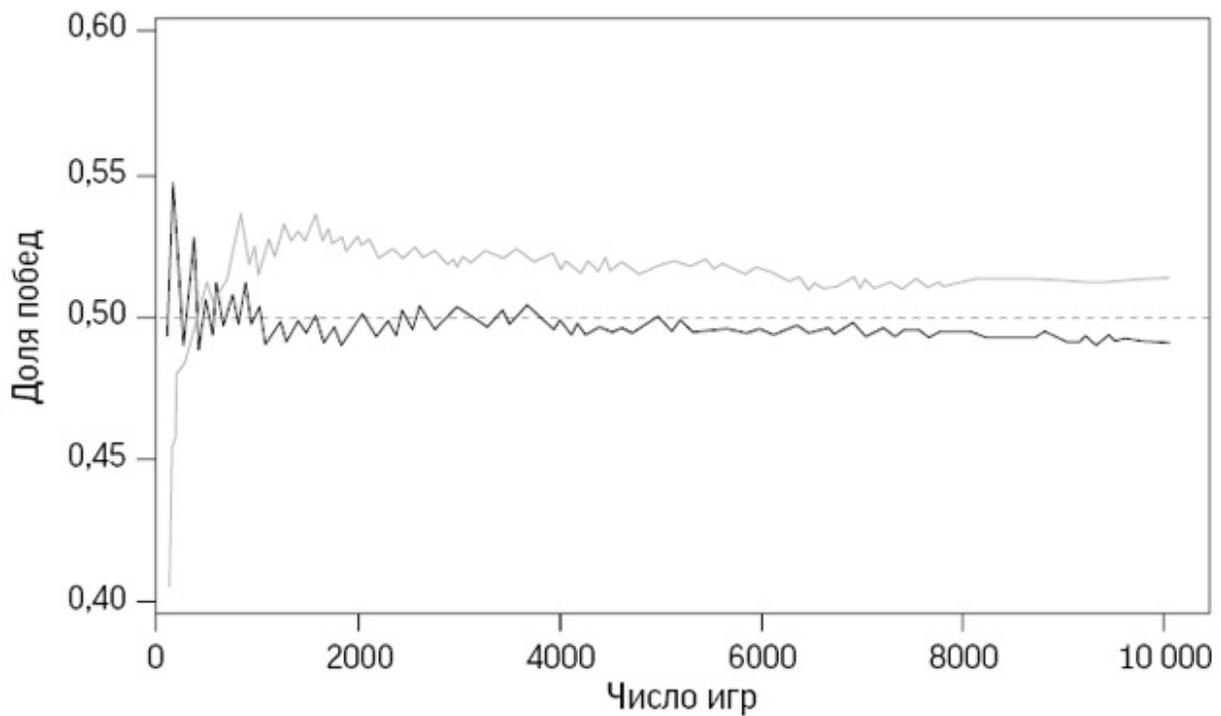
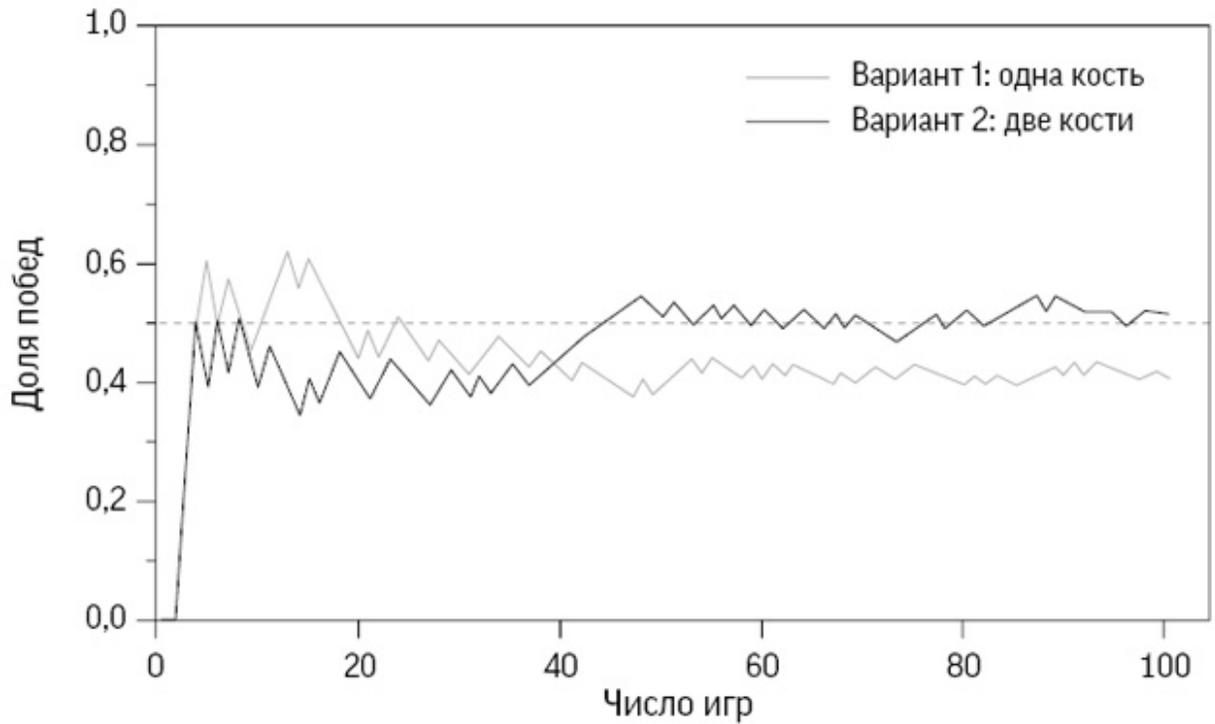
Вариант 1. Правильная игральная кость бросается четыре раза, игрок побеждает, если хотя бы раз выпадает шестерка.

Вариант 2. Пара правильных игровых костей бросается 24 раза, игрок побеждает, если хотя бы раз выпадает пара шестерок.

На что выгоднее поставить?

В соответствии с эмпирическими статистическими принципами шевалье де Мере решил сыграть в обе игры много раз и посмотреть, насколько часто он выигрывает. Это потребовало немало времени и усилий, но в причудливой параллельной вселенной, где были компьютеры, но не было теории вероятностей, шевалье не потратил бы столько времени на сбор данных, а просто смоделировал бы тысячи игр.

На рис. 8.1 представлены результаты такого моделирования – доля побед по мере увеличения количества прохождений игр. Хотя какое-то время Вариант 2 кажется выгоднее, примерно после 400 игр становится ясно, что Вариант 1 лучше и что в (очень) долгосрочной перспективе шевалье может рассчитывать на победу примерно в 52 % игр для Варианта 1 и только 49 % игр для Варианта 2.



**Рис. 8.1**

Компьютерное моделирование 10 тысяч повторений двух вариантов игр. В Варианте 1 вы выигрываете, если шестерка выпадает хотя бы раз при четырех бросаниях кости, а в Варианте 2 – если пара шестерок выпадет хотя бы раз при 24 бросаниях пары костей. После

первых 100 подбрасываний в каждом из вариантов (верхняя диаграмма) вроде бы выгоднее кажется Вариант 2, однако после тысяч игр (нижняя диаграмма) становится ясно, что Вариант 1 несколько лучше

Примечательно, что де Мере играл достаточно часто, чтобы прийти к аналогичному выводу: Вариант 1 немного лучше. Это шло вразрез с его (ошибочными) попытками вычислить шансы на победу<sup>[157]</sup>, поэтому он обратился за помощью в модный парижский салон Мерсенна<sup>[158]</sup>. К счастью, его частым посетителем был философ Блез Паскаль, который, познакомившись с задачей, написал о ней своему другу Пьеру де Ферма (да-да, автору той самой Великой теоремы!). Вместе в последующей переписке они сделали первые шаги на пути к созданию теории вероятностей.

Несмотря на то что люди тысячелетиями играли в азартные игры и делали ставки на то, какой стороной упадут игральные кости, формальная теория вероятностей – сравнительно недавняя идея. В течение следующих пятидесяти лет после работ Паскаля и Ферма в 1650-х годах были заложены математические основы, и сегодня вероятность используется в физике, страховании, пенсионных расчетах, торговле на финансовых рынках, прогнозировании и, конечно же, в азартных играх. Но почему нужно использовать теорию вероятностей при статистических расчетах?

Мы уже встречались с концепцией «случайного выбора» из общего распределения в совокупности – ваша подруга из [главы 3](#), родившая ребенка с низким весом, была нашим первым примером знакомства с вероятностью. Мы должны предположить, что любой элемент генеральной совокупности с равными шансами может попасть в нашу выборку: вспомните [аналогию Гэллапа о перемешивании супа](#) перед тем, как его попробовать. И мы видели, что при намерении делать какие-то статистические заключения о неизвестных аспектах мира, включая прогнозы, наши выводы неизбежно будут иметь некоторую неопределенность.

В предыдущей главе мы обсудили, как использовать бутстрэппинг, чтобы узнать, какого разброса в характеристиках выборки можно ожидать, делая раз за разом перевыборку, а затем применить эти данные для указания степени неопределенности в отношении истинной, но

неизвестной характеристики всей генеральной совокупности. Опять же для этого нужна концепция «случайного выбора» – идея, которую легко улавливают даже маленькие дети как выразители справедливого выбора.

Традиционно курс статистики начинается с вероятности – именно так я всегда делал, когда преподавал в Кембридже, – однако такое математическое вступление может быть препятствием в понимании важных идей, изложенных в предыдущих главах, где теория вероятности не требуется. Напротив, эта книга – часть того, что можно назвать новой волной в преподавании статистики, в которой формальная теория вероятностей как основа для статистических выводов появляется гораздо позже<sup>[159]</sup>. Мы уже видели, что компьютерное моделирование – очень мощный инструмент как для изучения возможных будущих событий, так и для бутстрэппинга с помощью прошлых данных, однако это довольно неуклюжий и грубый способ проведения статистического анализа. Поэтому, несмотря на то что мы долгое время избегали формальной теории вероятностей, настало время познакомиться с ее жизненно важной ролью в обеспечении «языка неопределенности».

Но почему за последние 350 лет развилось нежелание использовать эту блестящую теорию? Меня часто спрашивают, почему люди склонны считать вероятность сложной и интуитивно неясной идеей, и я отвечаю, что после 40 лет исследований и преподавания пришел к выводу, что вероятность действительно сложная и интуитивно неясная идея. Я сочувствую любому, кто считает вероятность трудной и запутанной. Даже после десятилетий работы статистиком, когда мне задают школьный вопрос на вероятность, я предпочитаю уединиться, чтобы молча посидеть в тишине с ручкой и бумагой, попробовать несколько разных способов и наконец озвучить (как я надеюсь) правильный ответ.

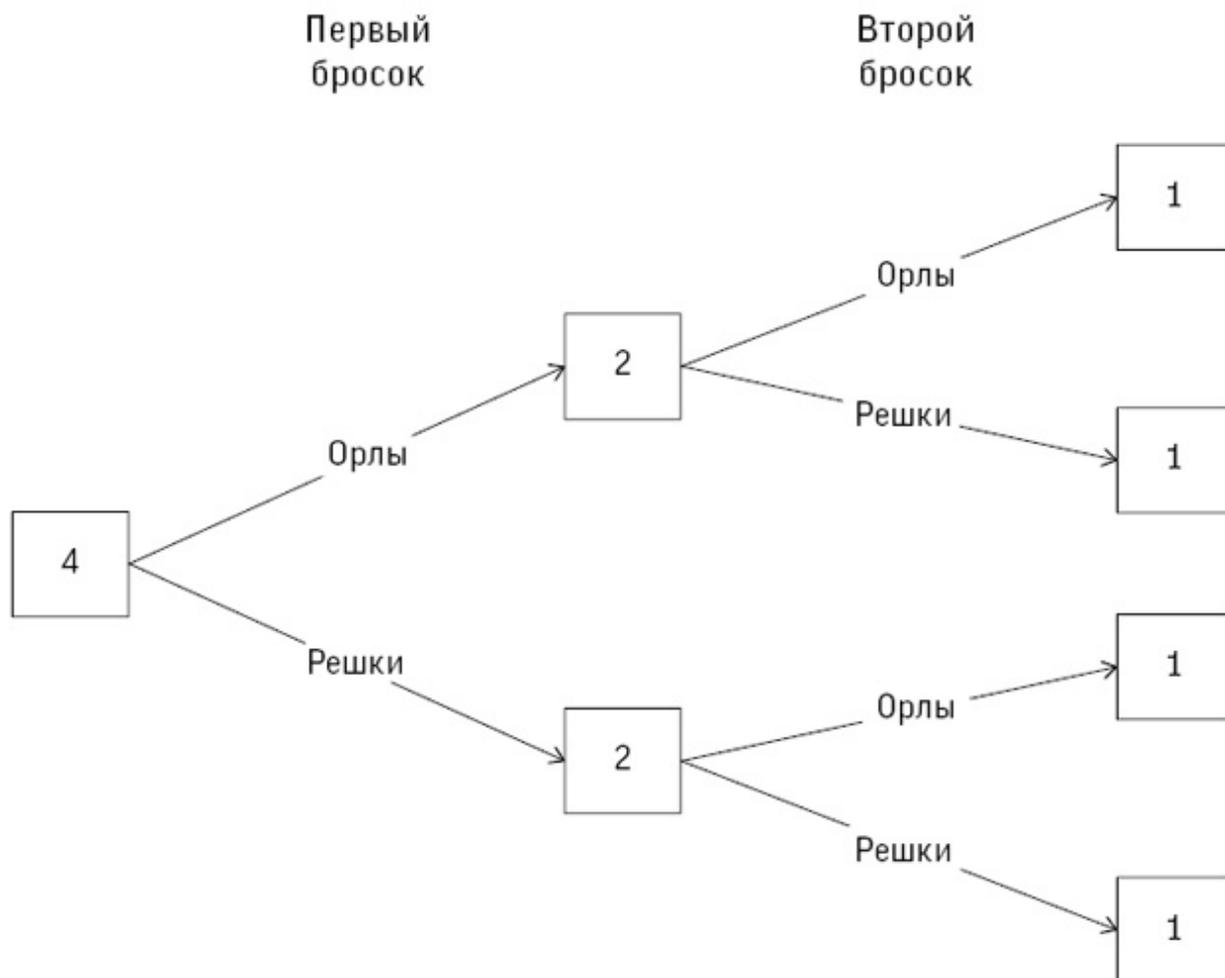
Давайте начнем с моего любимого метода решения задач, который мог бы избавить от смущения некоторых политиков.

***Правила для вероятностей, изложенные, возможно, чуть проще***

В 2012 году 97 парламентариев спросили: «Если вы подбросите монетку дважды, какова вероятность выпадения двух орлов?» Большинство – 60 из 97 – не смогли дать правильный ответ<sup>[160]</sup>. Как политики могли бы улучшить результаты?

Возможно, им стоило бы знать правила работы с вероятностями, но большинство людей их не знают. Однако в качестве альтернативы можно использовать более интуитивную идею, которая (как показали многочисленные психологические эксперименты) позволяет людям лучше понять суть вероятностей.

Это идея «ожидаемого количества». Столкнувшись с задачей о двух монетах, вы спрашиваете себя: «Что будет, если я проведу такой эксперимент несколько раз?» Например, вы подбрасываете одну монету, потом вторую – всего делаете так четыре раза. Подозреваю, что даже политик мог бы, слегка подумав, прийти к выводу, что можно ожидать результатов, показанных на рис. 8.2.



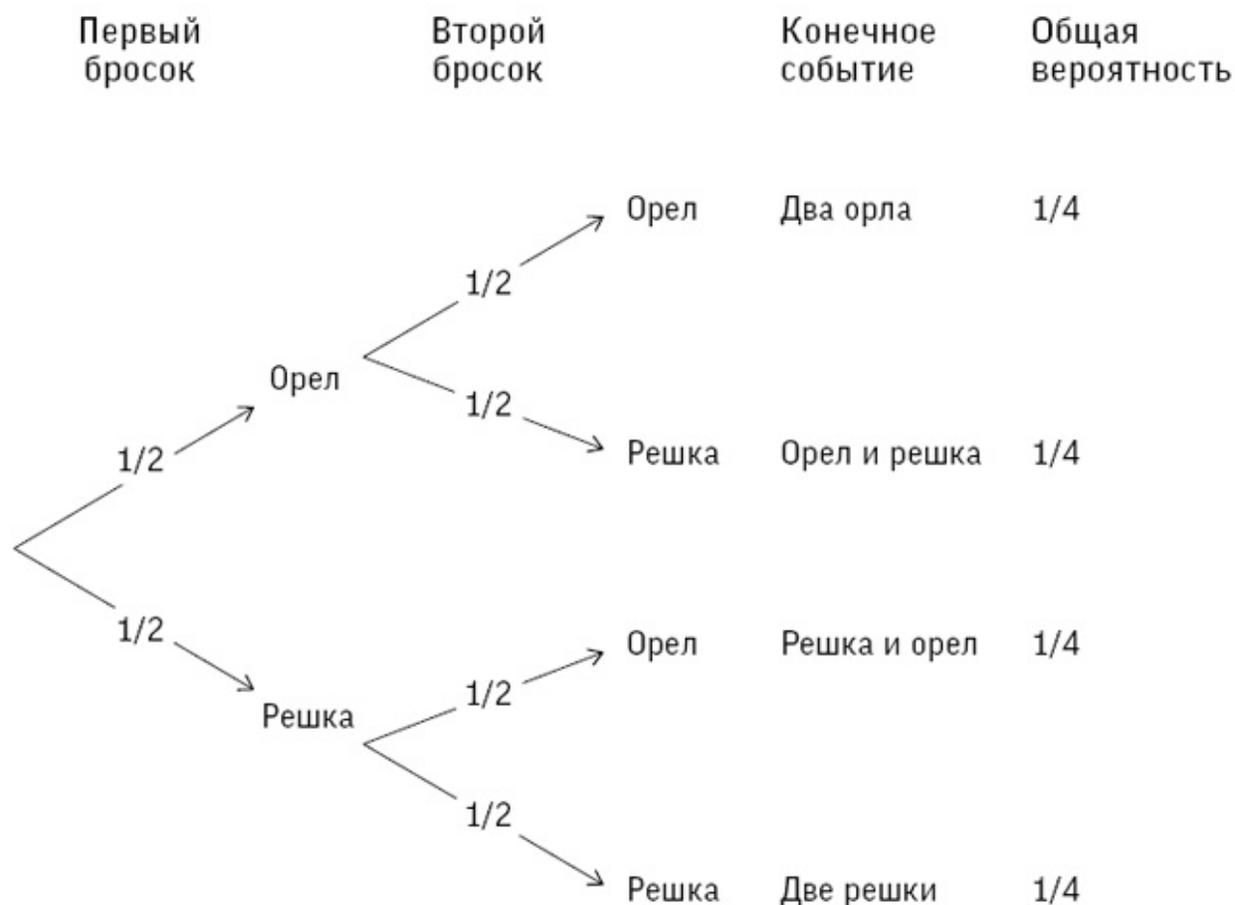
**Рис. 8.2**

Дерево ожидаемых частот для подбрасывания двух монет, повторенного четыре раза. Например, вы ожидаете, что среди первых четырех подбрасываний будут два орла, а на втором подбрасывании в одном случае выпадет орел, а во втором – решка

Таким образом, один раз из четырех вы могли бы ожидать выпадения двух орлов. Поэтому вероятность, что оба орла выпадут в единственной попытке, составляет  $1 / 4$ . К счастью, это и есть правильный ответ.

Дерево ожидаемых частот можно преобразовать в «дерево вероятностей», если для каждой «развилки» указать долю соответствующих случаев (см. рис. 8.3). Тогда становится ясно, что общая вероятность всей ветви дерева (например, выпадения орла после

орла) получается путем умножения дробей, стоящих на частях ветви, то есть  $1/2 \times 1/2 = 1/4$ .



**Рис. 8.3**

Дерево вероятностей для подбрасывания двух монет. На каждой «развилке» указана доля событий. Вероятность целой ветви дерева определяется путем умножения дробей на всех ее частях

Деревья вероятностей – весьма распространенный и крайне эффективный способ изучения вероятностей в школе. В самом деле, мы можем использовать этот простой пример с двумя монетами для ознакомления со всеми правилами вероятностей. Дерево показывает следующее:

1. *Вероятность события* – это число от 0 до 1, где 0 – вероятность невозможных событий (например, не выпали ни орлы, ни

решки), а 1 – вероятность достоверных событий (выпала какая-то из четырех возможных комбинаций).

2. *Правило дополнения.* Дополнением к событию А называется событие, которое произойдет в случае, если А не произошло. Вероятность его наступления равна единице минус вероятность события А. Например, вероятность события «выпала хотя бы одна решка» равна единице минус вероятность события «выпало два орла»:  $1 - 1/4 = 3/4$ .

3. *Правило сложения (правило «ИЛИ»):* если события несовместны (то есть не могут произойти одновременно), то вероятность того, что произойдет хотя бы какое-то одно из них, равна сумме вероятностей отдельных событий. Например, вероятность «выпадения хотя бы одного орла» составляет  $3/4$ , так как включает три несовместных события: «выпало два орла», ИЛИ «выпал сначала орел, а потом решка», ИЛИ «сначала выпала решка, а потом орел» – каждое с вероятностью  $1/4$ .

4. *Правило умножения (правило «И»):* при наличии последовательности **независимых событий** (то есть одно не влияет на другое) вероятность наступления всех событий в последовательности равна произведению вероятностей отдельных событий. Например, вероятность выпадения двух орлов равна  $1/2 \times 1/2 = 1/4$ .

Эти основные правила позволяют решить задачу шевалье де Мере, показывая, что на самом деле в варианте 1 его шансы на победу составляли 52 %, а в варианте 2 – 49 %<sup>[161]</sup>.

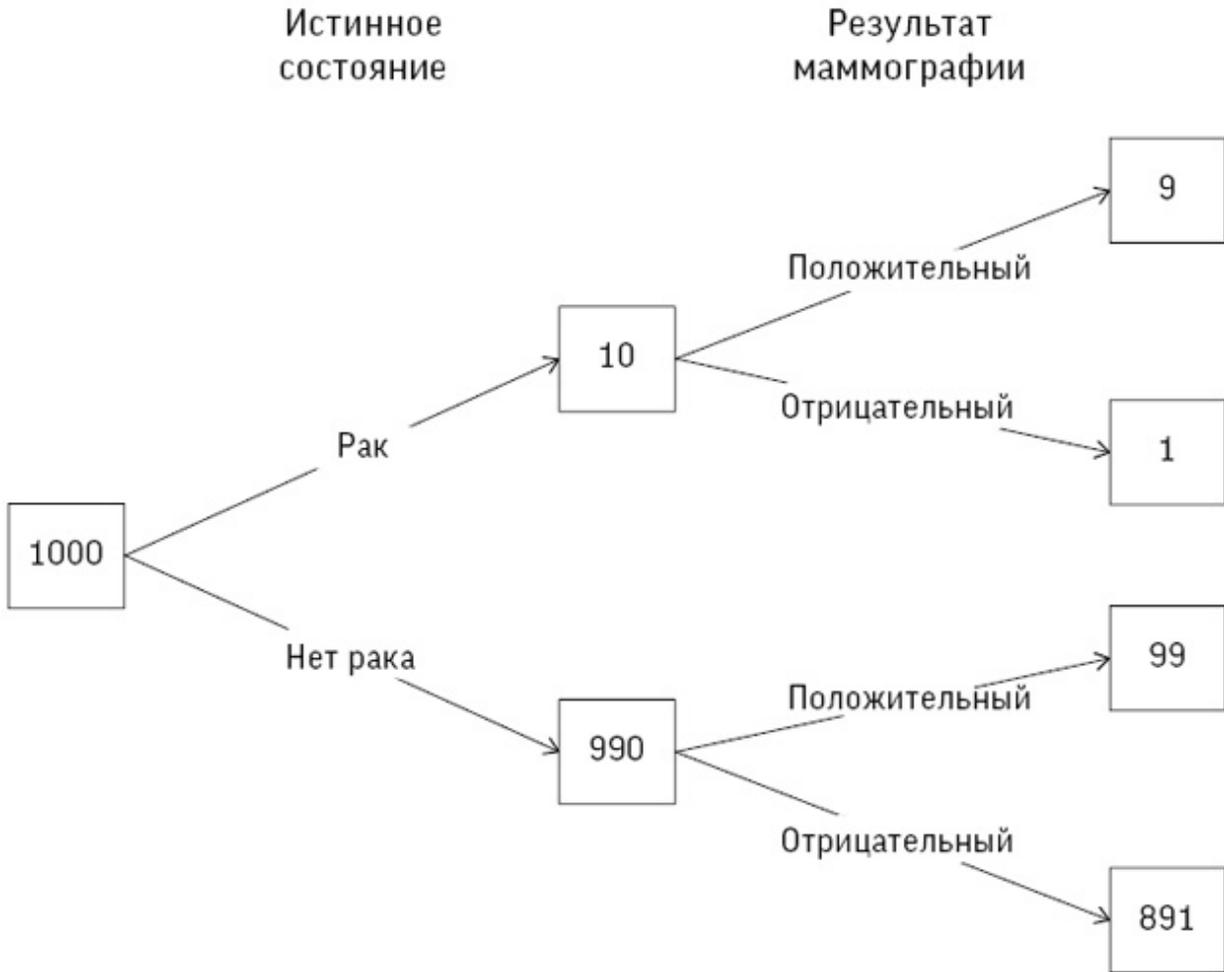
Мы по-прежнему делаем сильные предположения – даже в простейшем примере с подбрасыванием монет. Мы полагаем, что монета симметрична, что результат при ее подбрасывании не будет предсказуем, что она не упадет на ребро, что после первого броска в Землю не врежется астероид и так далее. Задача всех этих серьезных (за исключением, пожалуй, падения астероида) соображений – подчеркнуть, что все используемые нами вероятности *условны*: не существует безусловной вероятности события; всегда есть какие-то предположения и иные факторы, которые могут на нее влиять. И, как мы сейчас увидим, нам нужно проявлять осторожность в отношении того, на чем мы основываемся.

*Условная вероятность – когда вероятности зависят от других событий*

**При диагностике рака молочной железы точность маммографии – примерно 90 %, то есть она правильно определяет 90 % женщин с раком и 90 % женщин без рака. Предположим, что 1 % обследуемых женщин действительно больны. Какова вероятность, что у случайно выбранной женщины окажется положительная маммограмма, и если так, то какова вероятность, что у женщины на самом деле рак?**

В случае с двумя монетами события независимы, поскольку вероятность выпадения орла на второй монете не зависит от результата подбрасывания первой монеты. В школе мы обычно узнаем о **зависимых событиях**, когда нам начинают задавать несколько утомительные вопросы, скажем, о разноцветных носках, которые вытаскивают из ящика. Пример выше гораздо ближе к реальной жизни.

Подобные задачи – классические в тестах оценки интеллекта, и их не так легко решать. Однако идея ожидаемого количества существенно упрощает проблему. Ее суть – подумать, чего можно ожидать для большой группы женщин (скажем, 1000), как показано на рис. 8.4.



**Рис. 8.4**

Дерево ожидаемых частот, отображающее наши ожидания для 1000 женщин, проходящих скрининг рака молочной железы. Мы предполагаем наличие рака у 1 % женщин, а маммография верно классифицирует 90 % женщин с раком молочной железы и 90 % женщин без рака. Всего мы можем ожидать  $9 + 99 = 108$  положительных маммограмм, из которых девять окажутся истинно правильными

Из 1000 женщин у 10 (1 %) действительно выявляют рак молочной железы. Из этих 10 у девяти (90 %) обследование даст положительный результат. Однако из 990 здоровых женщин (без рака) у 99 (10 %) маммография будет ложноположительной. В общей сложности мы получим  $9 + 99 = 108$  положительных маммограмм, а значит, вероятность того, что у случайно выбранной женщины будет положительный результат, равна  $108 / 1000 \approx 11\%$ . Но среди этих 108

реально больны раком только 9, поэтому вероятность, что у женщины на самом деле рак, равна  $9 / 108 \approx 8 \%$ .

Это упражнение на условную вероятность помогает понять весьма парадоксальный результат: несмотря на «90-процентную точность» маммографии, подавляющее большинство женщин с положительной маммограммой на самом деле не больны. Легко перепутать «вероятность положительного теста при условии наличия рака» с «вероятностью рака при условии положительного теста».

Такая путаница известна как «**заблуждение прокурора**», поскольку часто встречается в судебных разбирательствах, связанных с анализом ДНК. Например, судебно-медицинский эксперт может утверждать, что «если обвиняемый невиновен, то вероятность того, что его ДНК совпадет с ДНК, найденной на месте преступления, только один шанс на миллиард». Но это неверно интерпретируется как «учитывая данные анализа ДНК, есть только один шанс на миллиард, что обвиняемый невиновен»<sup>[162]</sup>.

Подобная ошибка не редкость, но логика здесь так же неправильна, как и в переходе от утверждения «если вы папа римский, то вы католик» к утверждению «если вы католик, то вы папа римский», где абсурдность выражения сразу бросается в глаза.

### ***Так что же такое вероятность?***

В школе нас учат математике расстояний, масс и времени, которые мы можем измерить с помощью рулетки, весов или часов. Но как измерить вероятность? Не существует никакого вероятностемера. Словно вероятность – это некая «виртуальная» величина, которой мы можем присвоить какое-то число, но не измерить напрямую.

Еще больше настораживает вполне закономерный вопрос: а что вообще означает вероятность? Есть какое-то доходчивое определение этого понятия? Это может выглядеть как схоластика, но философия вероятности не только захватывающая тема сама по себе, но и играет огромную роль в практическом применении статистики.

Не ждите консенсуса от всевозможных «экспертов». Они могут соглашаться с математикой вероятностей, но философы и статистики выдвигают разные идеи о том, что на самом деле означают эти неуловимые числа, и активно их обсуждают. Вот некоторые популярные предложения.

- *Классическое определение вероятности.* Это то, чему нас учат в школе. Оно основано на симметрии монет, костей, перетасованных колод карт и так далее и может быть сформулировано как «отношение числа благоприятных исходов к числу всех исходов, если все исходы равновозможны». Например, вероятность выпадения единицы на правильной кости равна  $1/6$ , потому что возможны 6 исходов, а нас устраивает один. Однако это определение в какой-то степени носит круговой характер, поскольку прежде мы должны уяснить, что значит *равновозможны*.

- *«Перечислительная» вероятность*<sup>[163]</sup>. Предположим, в ящике лежат три белых и четыре черных носка. Если вытаскивать носок случайным образом, то чему равна вероятность, что он белый? Ответ  $3/7$  можно получить путем простого перечисления всех возможностей. Многие из нас страдали от таких вопросов в школе, и здесь мы фактически имеем дело с расширением рассмотренной выше классической идеи, где требуется случайный выбор из группы физических объектов. Мы уже использовали эту идею при описании случайного выбора элемента данных из общей генеральной совокупности.

- *Вероятность как частота.* Такое определение говорит о вероятности как о доле случаев, когда интересующее нас событие наступает в бесконечной последовательности идентичных экспериментов – в точности так как при моделировании двух вариантов игры шевалье де Мере. Для бесконечно повторяющихся событий это может быть разумно (хотя бы теоретически), но как насчет уникальных одноразовых событий, например скачек или завтрашней погоды? На деле практически любая реальная ситуация даже в принципе не может быть бесконечно воспроизводимой.

- *Пропенситивная интерпретация вероятности.* Основная идея состоит в том, что у каждой ситуации есть объективная склонность порождать какое-то событие<sup>[164]</sup>. Внешне идея выглядит привлекательно: если бы вы были прозорливым существом, то могли бы сказать, что существует вероятность того, что ваш автобус скоро придет или что вас сегодня собьет машина. Однако у нас, простых смертных, похоже, нет возможности оценивать такие скорее метафизические «истинные шансы».

- *Субъективная, или «личная», вероятность.* Это степень веры конкретного человека в какое-либо событие, основанная на его нынешних знаниях. Обычно субъективные вероятности интерпретируются в терминах пари. Допустим, мне предлагают 1 фунт, если я смогу пять минут жонглировать тремя шариками, а я готов сделать на это безвозвратную ставку в 60 пенсов. Тогда моя личная вероятность события оценивается в 0,6.

У различных «экспертов» собственные предпочтения относительно этих альтернатив, но лично я предпочитаю последний вариант – субъективную вероятность. Это означает, что я придерживаюсь мнения, что любая численная вероятность фактически *строится* в соответствии с тем, что известно в нынешней ситуации, – и на самом деле вероятность вообще не «существует» (за исключением, возможно, субатомного уровня). Такой подход лежит в основе **байесовской** школы статистики, о чем мы подробно поговорим в главе 11.

К счастью, вы не обязаны соглашаться с моим (довольно спорным) тезисом, что численные вероятности объективно не существуют. Можно предположить, что монеты и другие устройства для рандомизации объективно случайны – в том смысле, что генерируют настолько непредсказуемые данные, что они могут быть неотличимы от тех, которые мы ожидаем получить от «объективных» вероятностей. Поэтому в целом мы действуем так, *будто* наблюдения случайны, даже если знаем, что это не совсем верно. Наиболее яркие примеры – генераторы псевдослучайных чисел, по сути, основанные на полностью предсказуемых, детерминированных вычислениях. В них вообще нет никакой случайности, но их механизм настолько сложен, что на практике они неотличимы от настоящих случайных последовательностей, скажем, полученных из источника субатомных частиц [\[165\]](#).

Такая отчасти странная способность действовать, как будто что-то истинно, хотя вы знаете, что это не так, обычно считается опасно иррациональной. Однако это полезно, когда дело доходит до использования вероятности в качестве основы для статистического анализа данных.

Сейчас мы подошли к крайне важной, хотя и сложной стадии изложения общей взаимосвязи между теорией вероятностей, данными и изучением любой интересующей нас целевой совокупности.

Теория вероятностей естественным образом вступает в игру, когда мы имеем дело с ситуацией 1 (назовем ее так):

1. Когда можно считать, что данные *сгенерированы* каким-то рандомизирующим устройством, например, при подбрасывании монет, костей или путем случайного распределения пациентов по методам лечения с помощью генератора псевдослучайных чисел с последующей регистрацией результатов лечения.

Однако на практике мы можем столкнуться с ситуацией 2:

2. Когда рандомизирующее устройство *выбирает* уже существующий элемент данных, скажем, отбирает людей для участия в опросе.

И большую часть времени наши данные появляются из ситуации 3:

3. Когда случайности нет вообще, но мы действуем так, как если бы данные были сгенерированы каким-то случайным процессом, например при интерпретации веса новорожденного ребенка вашей подруги.

В большинстве описаний эти различия четко не разграничиваются: вероятность в целом объясняют с помощью рандомизирующих устройств (ситуация 1), статистике учат с помощью идеи «случайной выборки» (ситуация 2), но на самом деле большинство статистических приложений вообще не задействуют никаких рандомизирующих устройств или случайных выборок (ситуация 3).

Однако сначала рассмотрим ситуации 1 и 2. Непосредственно перед тем, как запустить рандомизирующее устройство, мы предполагаем, что у нас есть набор возможных результатов, которые можно наблюдать, а также их соответствующие вероятности – например, монета может выпасть орлом или решкой с вероятностью каждого исхода  $1/2$ . Связав все возможные исходы с вероятностями их появления, мы можем сказать, что у нас есть случайная величина с каким-то вероятностным распределением. В ситуации 1

рандомизирующее устройство гарантирует, что наши наблюдения случайным образом извлекаются из этого распределения, но когда наблюдение сделано, вся случайность пропадает и все потенциально возможные пути развития будущего события сводятся к одному фактическому варианту. Аналогично, в ситуации 2, если мы случайным образом выбираем человека и, например, измеряем его доход, то мы фактически извлекаем случайное наблюдение из распределения доходов в генеральной совокупности.

Таким образом, вероятность явно важна при работе с рандомизирующим устройством. Но большую часть времени мы просто рассматриваем все доступные на какой-то момент измерения, которые могли быть собраны без соблюдения формальностей или (как мы видели в [главе 3](#)) даже могут представлять все возможные наблюдения: вспомните об уровне выживаемости после операций на сердце у детей в различных больницах или результатах экзаменов у британских детей – оба включают все имеющиеся данные и никакой случайной выборки здесь просто нет.

В [главе 3](#) мы обсуждали идею *метафорической* генеральной совокупности, включающей все возможные случайности, которые могли бы произойти, но не произошли. Сейчас нам надо подготовиться к явно иррациональному шагу – действовать так, как будто данные получены каким-то случайным механизмом из общей совокупности, хотя мы прекрасно знаем, что это не так.

***Если мы все наблюдаем, то откуда появляется вероятность?***

**Как часто мы ожидаем семь или более отдельных случаев убийства в Англии и Уэльсе за один день?**

Когда несколько экстремальных событий происходят в тесной последовательности (например, череда крушений самолетов или природных катастроф), появляется естественное подозрение, что между ними существует какая-то связь. В этом случае важно выяснить, насколько необычны такие события, в чем нам и поможет следующий пример.

Чтобы оценить, насколько редок «кластер» из как минимум семи убийств в день, давайте изучим данные за три года (1095 дней) между апрелем 2014-го и мартом 2016-го. За этот период в Англии и Уэльсе

было совершено 1545 убийств, то есть в среднем  $1545/1095 = 1,41$  в день. Ни одного дня с семью и более случаями убийства<sup>[166]</sup> за это время не наблюдалось, однако было бы весьма наивно полагать, что такое событие невозможно. Если мы сумеем построить разумное вероятностное распределение для количества убийств в день, то сможем ответить на поставленный вопрос.

Но каковы обоснования для построения такого вероятностного распределения? Число убийств, регистрируемых в стране, – это просто факт, тут нет никакой случайной выборки и явного случайного элемента, генерирующего каждое преступление. Просто невообразимо сложный и непредсказуемый мир. Но какова бы ни была наша личная философия по отношению к удачам и неудачам, оказывается, полезно действовать так, *словно* все эти события были порождены каким-то случайным процессом, основанным на вероятности.

Давайте представим, что в начале каждого дня у нас есть огромная популяция людей, в которой у каждого ее члена есть очень малая вероятность стать жертвой убийства. Такого рода данные можно считать наблюдениями из **распределения Пуассона**, предложенного французским математиком Симеоном Пуассоном в 1837 году для описания вероятности вынесения неправомерных обвинительных приговоров за год. С тех пор оно использовалось для моделирования всего – от количества голов, забитых футбольной командой в матче, и еженедельного числа выигранных лотерейных билетов до ежегодного числа прусских офицеров, убитых ударом копыта их лошадей. Во всех этих ситуациях для наступления события есть очень большое число предпосылок, но каждая с ничтожно малым шансом на реализацию, что и приводит к необычайно универсальному распределению Пуассона.

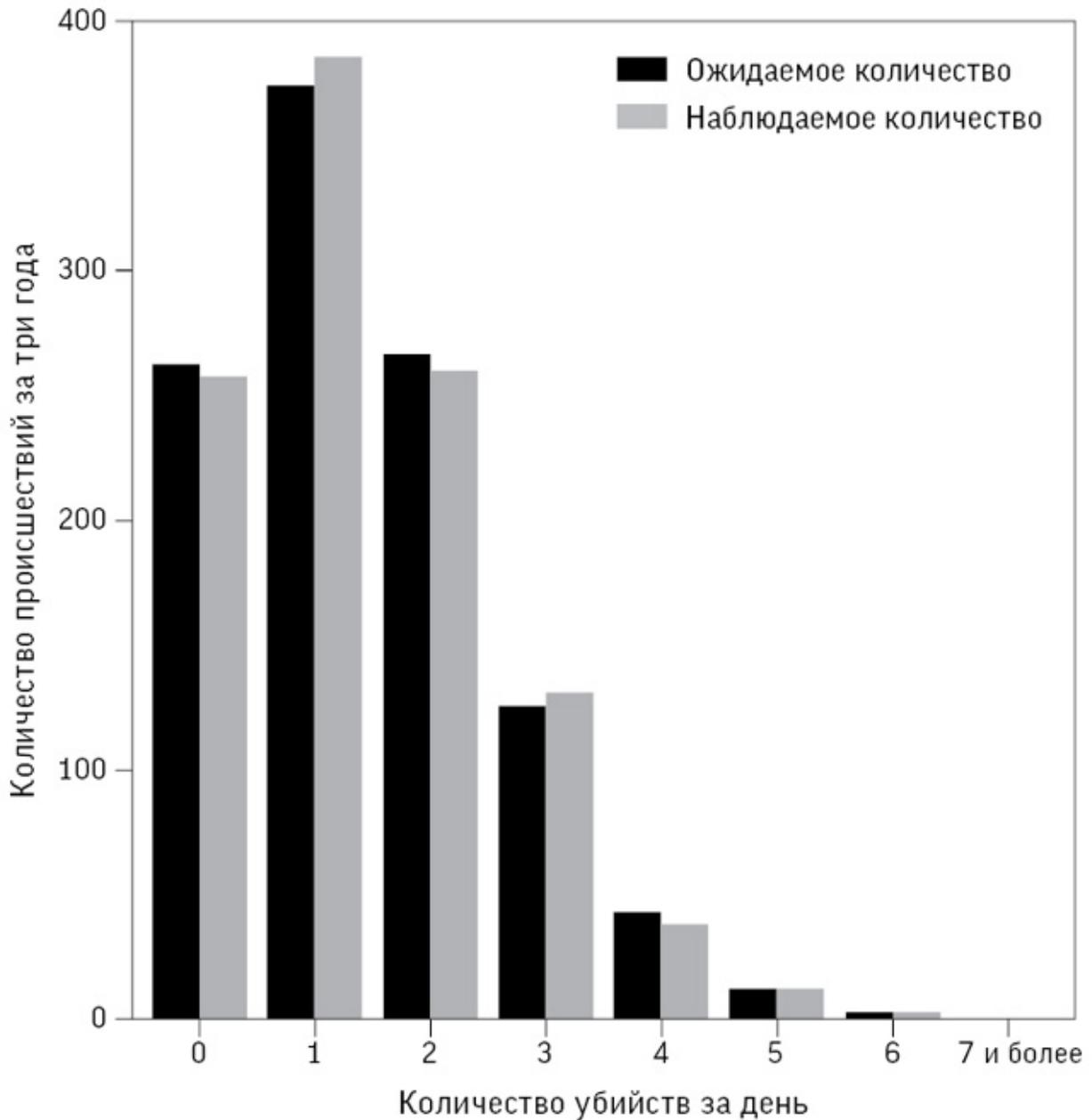
Тогда как нормальное (гауссовское) распределение, описанное в [главе 3](#), требует двух параметров (среднее значение и среднеквадратичное отклонение), у распределения Пуассона только один параметр (он имеет смысл среднего). В нашем конкретном примере это ожидаемое ежедневное число случаев убийства, которое мы принимаем равным 1,41, поскольку таково среднее значение за трехлетний период. Однако нам нужно тщательно проверить, насколько разумно предположение о распределении Пуассона, чтобы мы могли обращаться с количеством убийств так, словно это случайное

наблюдение, взятое из пуассоновского распределения с параметром 1,41.

Например, зная это среднее, мы можем использовать формулу для распределения Пуассона или стандартное программное обеспечение, чтобы вычислить, что вероятность совершения пяти убийств в день равна 0,001134. А значит, за 1095 дней можно ожидать  $1095 \times 0,001134 = 12,4$  дней, когда будут наблюдаться ровно пять случаев убийства.

Удивительно, но реальное число дней с пятью убийствами за трехлетний период... 13.

На рис. 8.5 приведено сравнение ожидаемого распределения для ежедневного числа убийств на основании распределения Пуассона и фактического эмпирического распределения для 1095 дней. Соответствие очень хорошее, и в главе 10 я покажу, как формально проверить, оправдано ли предположение о пуассоновском распределении данных.



**Рис. 8.5**

Наблюдаемое и ожидаемое (при условии распределения Пуассона) ежедневное количество зарегистрированных убийств за 2014–2016 годы в Англии и Уэльсе<sup>[167]</sup>

Чтобы ответить на вопрос, поставленный в начале этого раздела, мы можем вычислить вероятность семи и более убийств в день, исходя из распределения Пуассона. Она равна 0,07 %, а значит, такое событие можно ожидать в среднем раз в 1535 дней, то есть примерно раз в

четыре года. Напрашивается вывод, что при нормальном ходе вещей оно маловероятно, но не невозможно.

Соответствие между этим математическим распределением и эмпирическими данными подозрительно хорошее. Несмотря на то что за каждой трагедией стоит какая-то личная история, и практически любая из них непредсказуема, данные ведут себя так, словно их сгенерировал какой-то известный случайный механизм. Благодаря способности представлять, что могли бы быть (но не были) убиты другие люди, мы наблюдаем один из множества возможных миров, которые могли реализоваться; точно так же как, подбрасывая монету, наблюдаем одну из возможных последовательностей.

Адольф Кетле – бельгийский статистик, социолог и астроном XIX века – одним из первых привлек внимание к потрясающей предсказуемости общей картины, составленной из отдельных непредсказуемых событий. Он был заинтригован появлением нормального распределения при различных явлениях (например, распределении веса новорожденного, как описывалось в [главе 3](#)) и предложил идею «среднего человека» (*l'homme moyen*), который вобрал в себя среднее значение всех характеристик. Кетле развил идею «социальной физики», поскольку регулярные закономерности социальной статистики, казалось, отражали какой-то почти механический процесс, лежащий в ее основе. Так же как случайные молекулы газа, соединяясь, обеспечивают предсказуемые физические свойства, непредсказуемые действия миллионов отдельных людей в совокупности генерируют национальный уровень самоубийств, который из года в год практически не меняется.

К счастью, нам незачем верить, что реальные события обусловлены чистой случайностью (что бы это ни было). Просто предположение о «случайности» включает в себе всю неизбежную непредсказуемость мира или то, что иногда называют *естественной изменчивостью*. Поэтому мы установили, что вероятность образует надлежащий математический фундамент как для «чистой» случайности, проистекающей из субатомных процессов, монет, костей и так далее, так и для «естественной» неизбежной изменчивости, проявляющейся в весе новорожденных, уровне выживаемости после операций, результатах экзаменов, количестве убийств и других явлениях, которые нельзя точно предсказать.

В следующей главе мы обратимся к поистине замечательной теме: как объединить эти два аспекта вероятности, чтобы получить строгую основу для формальных статистических выводов.

### **Выводы**

- Теория вероятностей предоставляет формальный язык и математические инструменты для работы со случайными явлениями.
- Вероятностные выводы не бывают интуитивно понятными, однако понимание можно улучшить с помощью идеи ожидаемого количества.
- Вероятности полезны даже тогда, когда нет явного применения механизма рандомизации.
- Многие социальные явления в целом демонстрируют удивительную закономерность, в то время как отдельные события совершенно непредсказуемы.

## Глава 9. Объединяем вероятность и статистику

*Предупреждение. Это, пожалуй, самая сложная глава в книге, но, проявив настойчивость и изучив ее, вы обретете ценное понимание статистических выводов.*

**Мы обнаружили, что в случайной выборке из 100 человек 20 – левши. Что можно сказать о доле левшей во всей генеральной совокупности?**

В предыдущей главе мы обсуждали идею случайной величины – одного элемента данных, извлеченного из какого-то вероятностного распределения, описываемого определенными параметрами. Но нас редко интересует только один элемент – обычно у нас большой массив данных, для которого мы вычисляем среднее, медиану и другие статистики. Фундаментальный шаг, который мы сделаем в этой главе, – рассмотрим эти статистики как случайные величины, извлеченные из их собственных распределений.

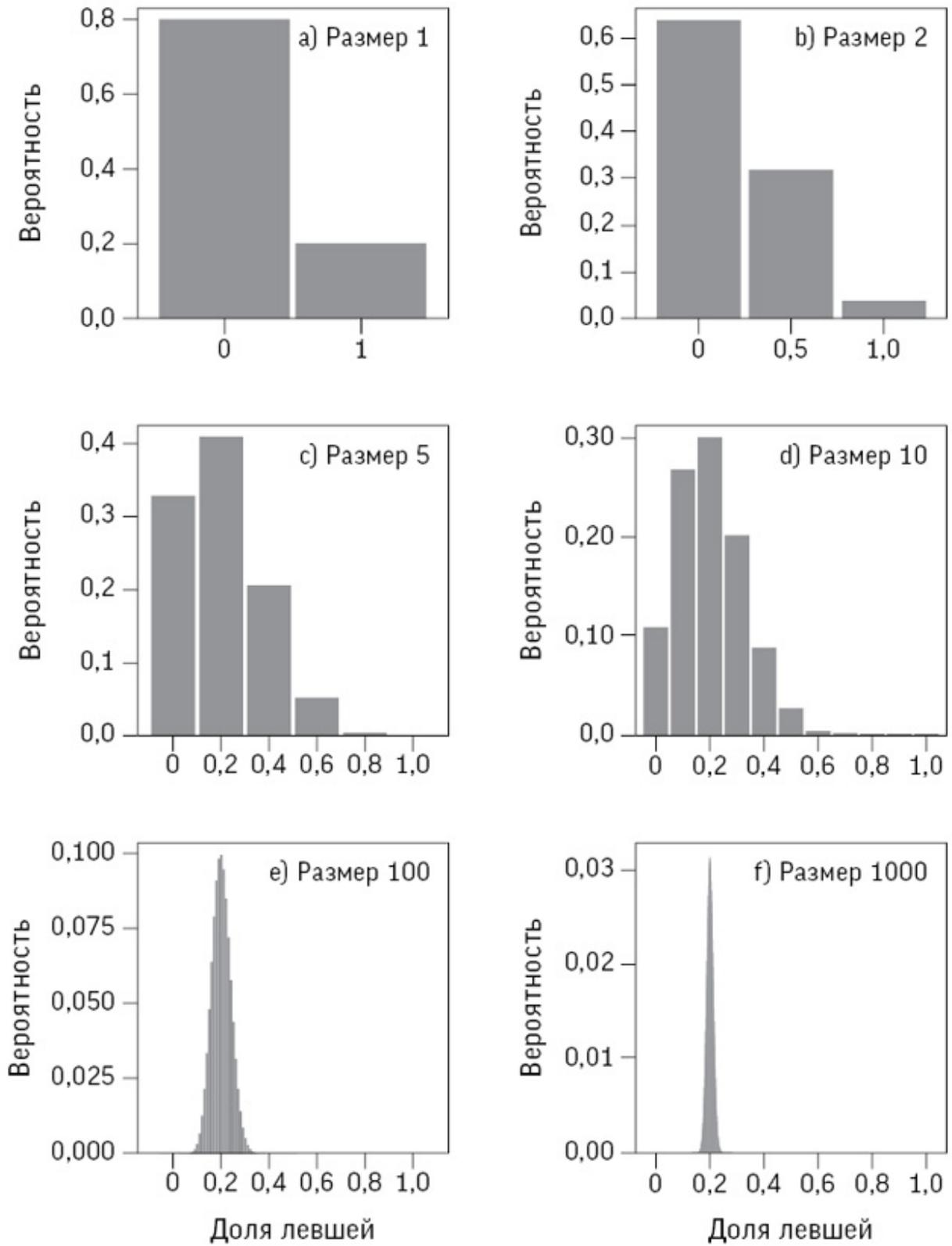
Это существенный шаг, создавший проблемы не только поколениям статистиков, но и математикам, которые пытались выяснить, из каких распределений мы извлекаем эти статистики. С учетом обсуждения бутстрэппинга в [главе 7](#) разумно задаться вопросом, зачем нам вообще нужна вся эта математика, когда мы можем узнать интервалы неопределенности и прочее, используя моделирование методом бутстрэппинга. Например, на вопрос, поставленный в начале главы, можно было ответить, взяв наблюдаемую выборку из 20 левшей и 80 правшей и многократные повторные выборки с возвратом по 100 наблюдений из этого набора, посмотреть на распределение наблюдаемой доли левшей.

Но такое моделирование неуклюже и затратно по времени, особенно для больших объемов данных, да и в более сложных ситуациях не так просто решить, что нужно моделировать. Напротив, формулы, предлагаемые теорией вероятностей, обеспечивают и понимание, и удобство и (в отличие от моделирования) всегда дают

один и тот же ответ. Обратная сторона в том, что эта теория опирается на предположения, и мы должны быть очень осторожны, чтобы впечатляющие выкладки не ввели нас в заблуждение и не привели к необоснованным выводам. Позже мы поговорим об этом подробнее, а пока, уже оценив полезность нормального и пуассоновского распределений, введем еще одно важное вероятностное распределение.

Предположим, что мы составляем выборки разного размера из совокупности, содержащей ровно 20 % левшей и 80 % правшей, и вычисляем вероятность получения различных возможных долей левшей. Конечно, здесь все наоборот – мы хотим по известной выборке узнать о неизвестной генеральной совокупности. Однако для этого нужно сначала исследовать, как известная совокупность порождает различные выборки.

Простейший случай – выборка из одного человека. Тогда доля леворуких будет 0 или 1 (в зависимости от того, выберем мы правшу или левшу) и вероятность этого события составит 0,8 и 0,2 соответственно. Полученное распределение вероятностей представлено на рис. 9.1(a).



**Рис. 9.1**

Вероятностное распределение наблюдаемой доли левшей в случайных выборках по 1, 2, 5, 10 и 1000 человек, где истинная доля левшей в генеральной совокупности равна 0,2. Вероятность получения не менее 30 % левшей в выборке вычисляется путем сложения вероятностей для всех значений справа от 0,3

Если мы выберем случайным образом двух человек, то доля левшей может быть 0 (оба правши), 0,5 (один левша и один правша) или 1 (оба левши). Вероятность таких событий равна 0,64, 0,32 и 0,04 соответственно<sup>[168]</sup>, и это распределение показано на рис. 9.1(b). Аналогично с помощью теории вероятностей мы можем найти распределение для наблюдаемых долей левшей в выборках по 5, 10, 100 и 1000 человек (рис. 9.1). Такое распределение известно как **биномиальное**, а часть диаграммы, лежащая правее какого-либо значения, называется его хвостом.

Среднее значение случайной величины также известно как **математическое ожидание**, и в наших выборках мы можем ожидать долю левшей 0,2, или 20 %: все распределения, представленные на рис. 9.1, имеют среднее 0,2. Среднеквадратичное отклонение для каждого из них зависит от параметров распределения (в нашем случае 0,2) и размера выборки. Обратите внимание, что стандартное отклонение какой-то статистики обычно называют **стандартной ошибкой**, чтобы отличить от стандартного (среднеквадратичного) отклонения в распределении, из которого взяты данные.

Рис. 9.1 демонстрирует некоторые отличительные особенности. Во-первых, по мере увеличения выборки форма распределения становится более правильной и симметричной (так же как мы наблюдали при использовании бутстрэппинга), во-вторых, распределения сужаются. В следующем примере показано, как простое применение этих идей позволяет быстро определить, насколько статистическое утверждение обоснованно.

### **Действительно ли в некоторых регионах Великобритании смертность от колоректального рака в три раза выше?**

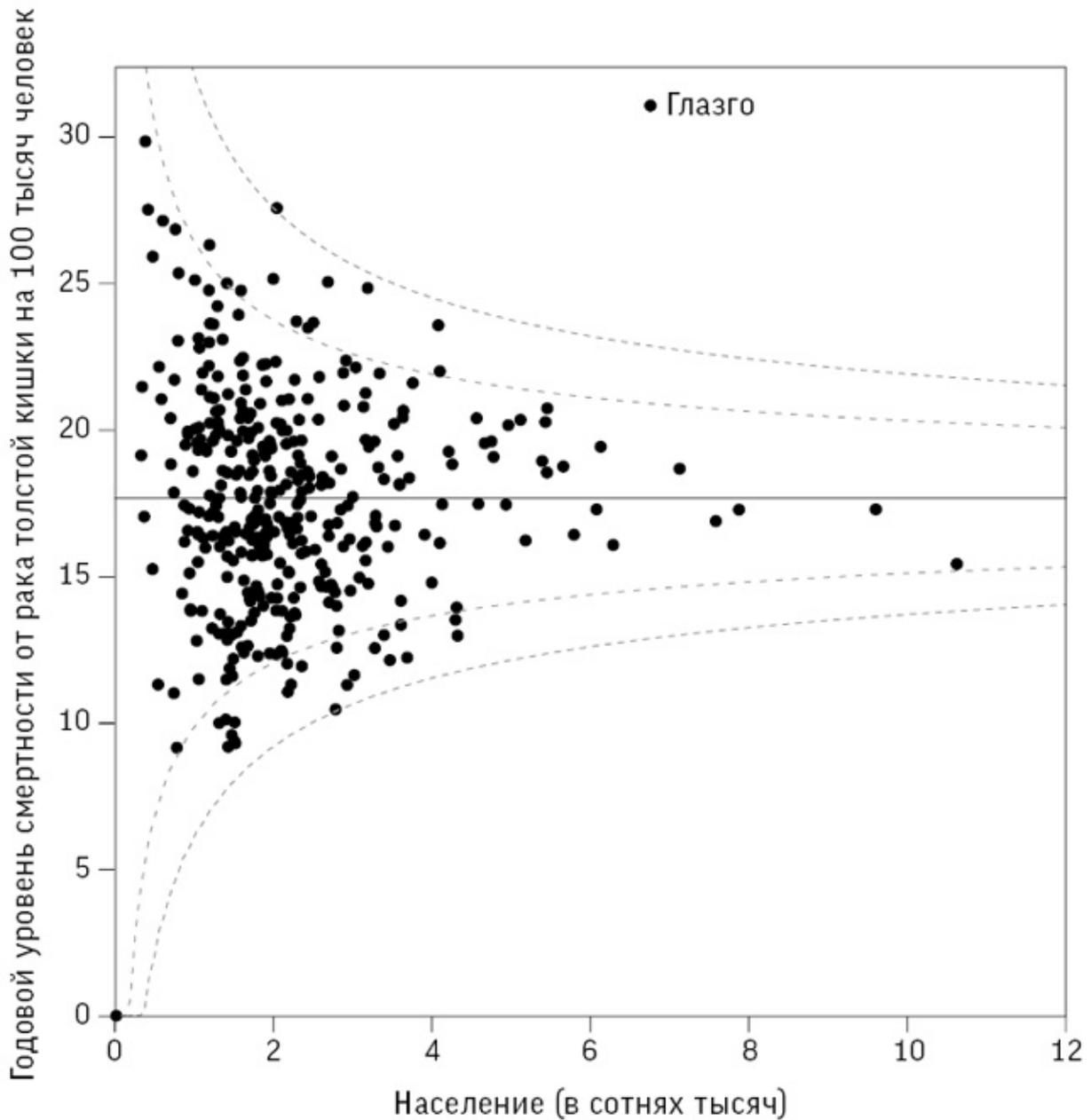
Заголовок на уважаемом новостном сайте «Би-би-си» в сентябре 2011 года настораживал: «Трехкратное различие в уровне смертности

от колоректального рака в Великобритании». Далее в статье объяснялось, что в различных округах страны показатели смертности от рака толстой кишки значительно разнятся, а комментатор добавлял, что «местным органам здравоохранения крайне важно изучить эту информацию и использовать ее для оповещения о потенциальных изменениях в оказании услуг».

«Трехкратное различие» звучит необычайно драматично. Но когда блогер Пол Барден наткнулся на эту статью, он задался вопросом: «Неужели люди в разных частях страны действительно сталкиваются со столь значительной разницей рисков умереть от рака? Чем объяснить такое расхождение?» Он счел это настолько неправдоподобным, что решил заняться этой темой. К счастью, все данные были в открытом доступе в интернете, и Барден обнаружил, что они подтверждают заявление «Би-би-си»: ежегодные показатели смертности от этого вида рака действительно отличались в три раза между разными регионами страны – от 9 случаев на 100 тысяч человек в районе Россендейл (Ланкашир) до 31 на 100 тысяч в округе Глазго-Сити<sup>[169]</sup>.

Однако расследование на этом не закончилось. Барден построил диаграмму смертности населения в каждом округе, что дало картину, представленную на рис. 9.2. Видно, что точки (за исключением экстремального случая с Глазго-Сити) расположены в форме воронки, причем чем население округов меньше, тем разброс больше. Затем Пол добавил **контрольные граничные значения**, которые показывают, куда могли бы попасть точки, если бы разница между наблюдаемыми уровнями определялась исключительно естественной неизбежной изменчивостью числа людей, ежегодно умирающих от рака толстой кишки, а не какими-то систематическими отклонениями в рисках для различных округов. Эти предельные значения получены из предположения, что число смертей – это наблюдение, взятое из выборки с биномиальным распределением, размер которой равен количеству взрослого населения округа: вероятность того, что любой конкретный человек умрет от рака в течение года, составляет 0,000176 (это средний риск смерти по всей стране). Граничные значения включают 95 % и 99,8 % всех наблюдений соответственно. График такого типа называется **воронкообразным** и широко используется при работе с несколькими медицинскими организациями или

учреждениями, поскольку позволяет отобразить выбросы, не создавая упорядоченных таблиц.



**Рис. 9.2**

Ежегодные показатели смертности от колоректального рака на 100 тысяч человек в 380 округах Великобритании в зависимости от численности населения округа. Две пары пунктирных линий, полученные исходя из предположения о биномиальном распределении, обозначают области, куда должны были бы попасть 95 % и 99,8 %

округов, если бы между ними не было никакой разницы в рисках. Только Глазго демонстрирует риск, отличный от среднего. Такой способ представления данных называется воронкообразным графиком

Данные достаточно хорошо укладываются в указанные пределы, а значит, различия между округами как раз такие, как мы бы ожидали в результате случайной изменчивости. В маленьких округах меньше случаев заболевания, поэтому они более уязвимы к случайным отклонениям и поэтому их показатели рассеяны сильнее: в Россендейле зафиксировано всего семь смертей, поэтому один лишний случай сильно изменяет уровень смертности. Следовательно, несмотря на драматический заголовок «Би-би-си», никаких сверхоткрытий здесь нет – трехкратное различие в уровне смертности мы могли бы ожидать даже в случае, если бы вероятность заболеть была бы в точности одинаковой во всех округах.

Этот простой пример преподает нам важный урок. Даже в эпоху открытых данных, науки о данных и журналистики данных нам по-прежнему нужны базовые статистические принципы, чтобы нас не ввели в заблуждение видимые закономерности в числах.

Наша диаграмма показывает, что единственное наблюдение, требующее внимания, – это точка, соответствующая Глазго. Неужели колоректальный рак – это, некий шотландский феномен? Действительно ли верно это наблюдение? Более поздние данные за 2009–2011 годы показывают, что уровень смертности от колоректального рака в Большом Глазго<sup>[170]</sup> составлял 20,5 на 100 тысяч человек, в Шотландии в целом – 19,6, а в Англии – 16,4: эти результаты ставят под сомнение вышеуказанное наблюдение для Глазго, но демонстрируют, что в Шотландии уровень смертности выше, чем в Англии. Как правило, заключения, сделанные после одного цикла решения задачи, поднимают новые вопросы и цикл начинается заново.

### *Центральная предельная теорема*

Отдельные наблюдения могут быть взяты из самых разных распределений, которые порой бывают сильно асимметричными или имеют длинные хвосты (как в случае дохода или числа сексуальных

партнеров). Однако мы сделали решительный шаг в сторону изучения распределения статистик, а не отдельных наблюдений, и эти статистики в каком-то смысле обычно более усреднены. Мы уже видели в [главе 7](#), что распределение выборочных средних у бутстрэп-выборок сходится к симметричной форме независимо от вида исходного распределения данных, и теперь можем пойти дальше, к более глубокой и замечательной идее, которая появилась около 300 лет назад.

Пример с левшами показывает, что по мере увеличения размера выборки отклонения для наблюдаемой доли уменьшаются – вот почему воронка на [рис. 9.2](#) сужается вокруг среднего значения. Это классический **закон больших чисел**, который в начале XVIII века вывел швейцарский математик Якоб Бернулли. Испытанием Бернулли называется эксперимент с двумя исходами – «успехом» и «неудачей», которые обычно обозначаются 1 и 0. Соответствующая случайная величина, принимающая значение 1 с вероятностью  $p$  и 0 с вероятностью  $1 - p$  имеет **распределение Бернулли**. Например, если вы один раз подбрасываете симметричную монету, то число выпавших орлов – это случайная величина, имеющая распределение Бернулли с  $p = 0,5$ . Предположим, что вы с помощью монеты будете производить последовательность испытаний Бернулли. Тогда доля орлов будет постепенно приближаться к 0,5, и мы скажем, что наблюдаемая доля орлов сходится к реальной вероятности их выпадения. Конечно, поначалу эта доля может отличаться от 0,5, и после нескольких выпавших подряд орлов появляется искушение поверить, что решки теперь как-то «обязаны» появляться чаще, чтобы восстановить баланс. Это заблуждение известно как *ошибка игрока*, и такое психологическое препятствие преодолеть довольно сложно (могу судить по личному опыту). Однако у монеты нет памяти – ключевая идея в том, что монета не может *компенсировать* прошлый дисбаланс и просто выдает все новые и новые результаты очередных подбрасываний.

В [главе 3](#) мы представили классическую колоколообразную кривую, также известную как нормальное (гауссовское) распределение, когда показывали, что оно хорошо описывает распределение веса новорожденных в США, и объяснили, что вес детей зависит от огромного количества факторов, каждый из которых оказывает небольшое влияние; складывая все эти маленькие воздействия, в итоге мы получаем нормальное распределение.

Именно это лежит в основе так называемой центральной предельной теоремы, впервые доказанной в 1733 году французским математиком Абрахамом де Муавром<sup>[171]</sup> для частного случая биномиального распределения. Однако к нормальному распределению сходится среднее не только для биномиальных случайных величин – примечательно то, что какое бы распределение для наших наблюдений мы ни взяли, можно считать, что при больших размерах выборки среднее значение наблюдений имеет нормальное распределение<sup>[172]</sup>. При этом его среднее совпадает со средним исходного распределения, а среднеквадратичное отклонение (как уже упоминалось, его часто называют стандартной ошибкой) имеет простую связь со среднеквадратичным отклонением для исходного распределения<sup>[173]</sup>.

Фрэнсис Гальтон не только написал работы о мудрости толпы, корреляции, регрессии и на многие другие темы, но и считал настоящим чудом то, что нормальное распределение (называемое в то время законом распределения ошибок) каким-то упорядоченным образом возникает из видимого хаоса:

Я едва ли знаю что-либо, способное воздействовать на воображение так, как чудесная форма космического порядка, выраженная «Законом Распределения Ошибок». Если бы древние греки знали этот закон, они бы персонифицировали и обожествили его. Он безмятежно царит среди самой дикой сумятицы. Чем больше толпа, чем больше видимая анархия, тем совершеннее его владычество. Это высший закон среди неразумности. Всякий раз, когда мы берем множество хаотичных элементов и расставляем их по величине, появляется неожиданная и доселе скрытая прекраснейшая закономерность.

Он был прав – это действительно выдающийся закон природы.

### ***Как теоретические рассуждения помогают определить точность наших оценок***

Вся эта теория хорошо помогает при попытке что-то узнать о распределении статистик, основанных на данных, взятых из известных совокупностей, но не это нас больше всего интересует. Мы должны

найти способ развернуть данный процесс: то есть вместо того чтобы по известным исходным распределениям говорить что-то о возможных выборках, попробовать по одной выборке что-то сказать о возможном распределении. Это процесс индуктивного вывода, описанный в главе 3.

Предположим, у меня есть монета, и я спрашиваю вас, с какой вероятностью выпадет орел. Вы радостно отвечаете «50 процентов» или нечто подобное. Затем я подбрасываю ее и накрываю, пока никто не увидел результат, и снова спрашиваю, с какой вероятностью будет орел. Если вы типичный человек, то, как показывает мой опыт, после паузы, скорее всего, довольно неохотно скажете: «50 процентов». Потом я смотрю на монету, не показывая вам, и повторяю вопрос еще раз. И снова, если вы относитесь к большинству, вы бормочете: «50 процентов».

Это простое упражнение показывает главное различие между двумя типами неопределенности: **стохастической неопределенностью**<sup>[174]</sup> до подбрасывания монеты (когда мы имеем дело с будущим непредсказуемым событием) и **эпистемической неопределенностью**<sup>[175]</sup> после подбрасывания монеты (выражением недостатка наших знаний об уже произошедшем событии). Это как разница между лотерейным билетом (где результат зависит от случая) и билетом мгновенной лотереи (где результат уже предопределен, просто вы его еще не знаете).

Статистика используется при наличии эпистемической неопределенности в отношении какой-то величины. Например, мы проводим опрос, когда не знаем истинной доли людей в популяции, считающих себя религиозными, или фармакологическое испытание, когда не знаем истинного среднего эффекта какого-то препарата. Как мы уже говорили, эти фиксированные, но неизвестные величины называются параметрами и часто обозначаются греческими буквами<sup>[176]</sup>. Как и в примере с подбрасыванием монеты, до проведения экспериментов у нас есть стохастическая неопределенность в отношении их результатов из-за случайного составления выборок или случайного назначения пациентам препарата или плацебо. После проведения исследования и получения данных мы используем эту вероятностную модель, чтобы справиться с текущей эпистемической неопределенностью – точно так же, как вы говорили «50 процентов»

о накрытой монете. Таким образом, теория вероятностей, которая говорит нам, чего ожидать в будущем, используется, чтобы сказать, что можно узнать из наших наблюдений в прошлом. Это и есть (довольно примечательная) основа для статистических выводов.

На этой фундаментальной идее построена процедура получения интервала неопределенности вокруг нашей оценки или погрешности, включающая три этапа.

1. Мы используем теорию вероятностей, чтобы для конкретных параметров генеральной совокупности получить интервал, в котором наблюдаемая статистика будет лежать с вероятностью 95 %. На рис. 9.2 такие 95-процентные интервалы прогнозирования изображены в виде внутренней воронки.

2. Затем мы наблюдаем конкретную статистику.

3. И наконец (и это самое трудное) определяем диапазон возможных параметров генеральной совокупности, для которых наша статистика попадает в 95-процентные интервалы прогнозирования. Этот диапазон мы называем «95-процентным **доверительным интервалом**». Он включает величину 95 %, поскольку при большом числе повторений 95 % таких интервалов будут содержать истинное значение параметра<sup>[177]</sup>.

Все ясно? Если нет, не расстраивайтесь: вы просто присоединились ко многим поколениям озадаченных студентов. Конкретные формулы приведены в глоссарии, но детали не так важны, как сам фундаментальный принцип: доверительный интервал – это тот диапазон параметров генеральной совокупности, при котором наша наблюдаемая статистика будет правдоподобным следствием.

### ***Вычисление доверительных интервалов***

Понятие доверительных интервалов было формализовано в 1930-е годы в Университетском колледже Лондона Ежи Нейманом, блестящим польским математиком и статистиком, и Эгоном Пирсоном, сыном Карла Пирсона<sup>[178]</sup>. До этого работа по определению необходимых вероятностных распределений для коэффициентов корреляции и коэффициентов регрессии велась десятилетиями; математические детали таких распределений входят в стандартные академические

курсы статистики. К счастью, результаты всех этих трудов теперь содержатся в статистическом программном обеспечении, так что практики могут сосредоточиться на важных вопросах и не отвлекаться на сложные формулы.

В [главе 7](#) мы узнали, как с помощью бутстрэппинга получить 95-процентные интервалы для углового коэффициента регрессионной прямой, связывающей рост матерей и дочерей. Гораздо проще получить точные интервалы, основанные на теории вероятностей и включенные в стандартные программы. Табл. 9.1 показывает, что они дают весьма сходные результаты. «Точные» интервалы, основанные на теории вероятностей, требуют больше предположений, чем метод бутстрэппинга, и, строго говоря, будут точными только в случае нормального распределения. Но центральная предельная теорема говорит, что при настолько большом объеме выборки разумно считать, что наши оценки имеют нормальное распределение, поэтому такие интервалы приемлемы.

**Таблица 9.1**

Оценки коэффициента регрессионной прямой, демонстрирующей связь между ростом дочерей и матерей. Стандартные ошибки и 95-процентные интервалы точные и для бутстрэппинга, основанного на 1000 перевыборок

**Угловой коэффициент регрессионной прямой  
для родителей и потомства**

	<b>Оценка</b>	<b>Стандартная ошибка</b>	<b>95-процентный интервал</b>
<b>Точный метод</b>	0,33	0,05	от 0,23 до 0,42
<b>Бутстрэппинг</b>	0,33	0,06	от 0,22 до 0,44

Традиционно используются 95-процентные интервалы, которые обычно отклоняются от среднего на две стандартные ошибки в обе стороны<sup>[179]</sup>; однако иногда интервалы берутся уже (например, 80 %) или шире (99 %). Статистическое управление США использует для

определения уровня безработицы 90-процентные интервалы, в то время как Национальное статистическое управление Великобритании – 95 %. Важно уточнять, какой именно интервал используется.

### ***Погрешности опросов***

Когда какое-то заявление базируется на опросе (например, опросе общественного мнения), стандартная практика – указать статистическую погрешность. У статистики безработицы, приведенной в [главе 7](#), на удивление большая погрешность (оценка в 3000 имеет погрешность  $\pm 77\ 000$ ). Это значительно влияет на интерпретацию исходного числа – в нашем случае такая погрешность показывает, что мы даже не знаем, выросла безработица или сократилась.

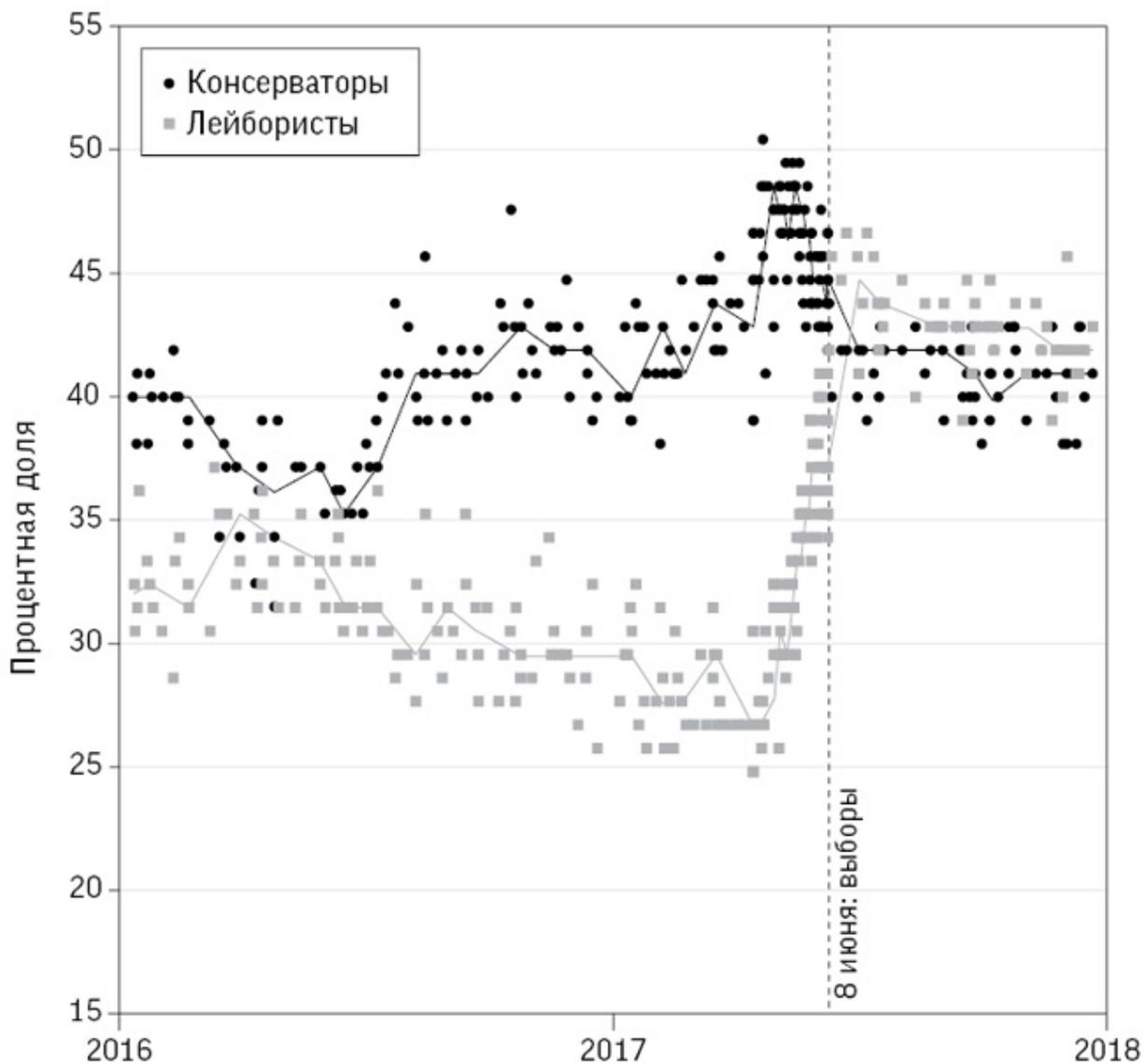
Существует простое эмпирическое правило: если вы оцениваете процент людей, предпочитающих, скажем, на завтрак чай, а не кофе, и рассматриваете случайную выборку из генеральной совокупности, то ваша погрешность (в процентах) будет максимум плюс-минус 100, деленное на квадратный корень из размера выборки<sup>[180]</sup>. Поэтому при выборке в 1000 человек (стандартный объем в таких опросах) погрешность обычно указывается как  $\pm 3\%$ <sup>[181]</sup>. Если 400 человек предпочитают кофе, а 600 – чай, то вы можете примерно оценить реальную долю любителей утреннего кофе в популяции следующим образом:  $40 \pm 3\%$ , то есть от 37 до 43 %.

Конечно, это верно только в случае, если устроители опроса действительно взяли случайную выборку, а все респонденты ответили, причем правду. Таким образом, хотя мы и можем вычислить погрешность, мы должны помнить, что вычисления верны, если примерно верны и наши предположения. Но можем ли мы на них опираться?

### ***Можно ли доверять погрешностям?***

Перед всеобщими выборами в Соединенном Королевстве в июне 2017 года публиковались многочисленные опросы общественного мнения с участием в каждом примерно 1000 респондентов. Если бы это были идеально случайные опросы, где участники давали бы правдивые ответы, то максимальная погрешность составила бы  $\pm 3\%$  и разброс результатов опросов относительно их среднего значения находился бы в этом диапазоне, поскольку предполагалось, что выборка каждый раз

берется из одной и той же генеральной совокупности. Однако рис. 9.3, основанный на диаграмме, использованной «Би-би-си», показывает, что рассеяние было намного больше. А значит, погрешности не могли быть верными.



**Рис. 9.3**

Способ визуализации данных социологических опросов, проведенных «Би-би-си» перед всеобщими выборами в Великобритании 2017 года<sup>[182]</sup>. Линия тренда – это медиана предыдущих семи опросов. В каждом опросе, как правило, участвовали 1000 человек, поэтому максимальная погрешность предполагалась  $\pm 3\%$ . Однако разбросы у разных опросов значительно превосходят эту

величину. Данные приведены только для двух партий – Консервативной и Лейбористской

Мы уже знаем много причин, почему опросы бывают неточными, не считая неизбежной (поддающейся количественному определению) погрешности из-за случайного разброса. В этом случае вину за излишнее рассеяние можно возложить на методы составления выборки, в частности на телефонные (причем в основном с использованием стационарных телефонов) опросы с очень низким коэффициентом ответов, вероятно, от 10 до 20 %. Я лично придерживаюсь эвристического правила, что для учета допущенных в опросе систематических ошибок заявленную погрешность нужно удвоить.

Мы не можем ожидать полной точности от предвыборных опросов, но могли бы ожидать большего от ученых, занимающихся измерением физических констант, например скорости света. Однако долгая история заявляемых погрешностей в таких экспериментах впоследствии оказалась безнадежно подпорченной: в первой половине XX века интервалы неопределенности вокруг оценок скорости света не включали значение, принятое сейчас.

В результате организациям, занимающимся *метрологией* (наукой об измерениях), пришлось указать, что погрешности всегда должны базироваться на двух компонентах:

- Тип А: стандартные статистические показатели, обсуждаемые в этой главе, которые при увеличении числа измерений предположительно станут снижаться.
- Тип В: систематические ошибки, которые, как ожидается, не уменьшатся при увеличении числа наблюдений и должны обрабатываться с использованием нестатистических средств, таких как экспертные суждения или внешние свидетельства.

Эти идеи должны пробудить в нас некоторое смирение в отношении статистических методов, которые мы можем применить к отдельному источнику данных. При наличии фундаментальных проблем со способом сбора данных никакие умные методы не помогут устранить такие ошибки, и нам нужно использовать знания и опыт, чтобы скорректировать свои заключения.

### *Что происходит, когда у нас есть все возможные данные?*

Вполне естественно использовать теорию вероятностей для определения погрешностей в результатах опроса, поскольку его участники случайно выбираются из более крупной совокупности, поэтому понятно, как в генерирование данных проникает случайность. Но давайте снова зададимся вопросом: а если наши статистические данные полные, то есть учитывают все, что произошло? Например, ежегодно некая страна учитывает все убийства. Если предположить, что в подсчетах нет ошибок (и согласовать определение термина «убийство»), то это будет просто описательная статистика без погрешностей.

Но, допустим, мы хотим сделать заявление о каких-то существующих тенденциях, скажем «количество убийств в Соединенном Королевстве растет». Например, Национальная статистическая служба Великобритании сообщила, что с апреля 2014 года по март 2015-го совершено 497 убийств и 557 в следующем таком же периоде. Конечно, число убийств возросло, но мы знаем, что оно меняется из года в год без видимых причин. Так есть ли здесь реальное изменение годового уровня убийств? Мы хотим сделать заключение об этом неизвестном количестве, поэтому нам нужна вероятностная модель для наблюдаемых величин.

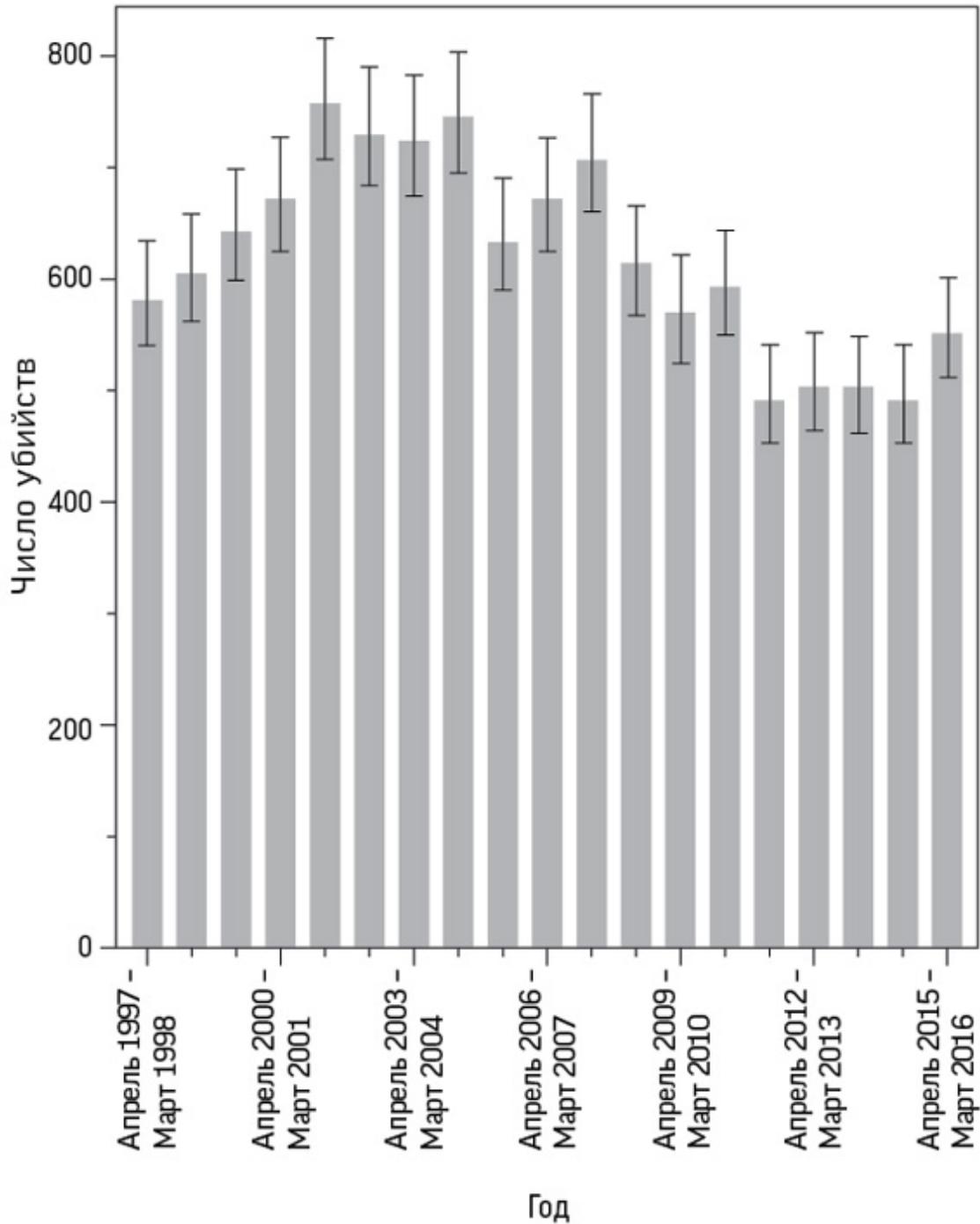
К счастью, в предыдущей главе мы видели, что ежедневные количества убийств ведут себя как случайные наблюдения с распределением Пуассона – словно взятые из какой-то метафорической совокупности альтернативных возможных историй. В свою очередь, это означает, что общее число убийств за год можно рассматривать как одно наблюдение с пуассоновским распределением со средним значением  $m$ , равным (гипотетическому) «истинному» годовому уровню. Нас интересует, меняется ли это  $m$  от года к году.

Среднеквадратичное (стандартное) отклонение у распределения Пуассона – это корень из  $m$ , то есть  $\sqrt{m}$ ; такова же стандартная ошибка нашей оценки. Это позволяет нам определить доверительный интервал, если мы будем знать  $m$ . Но мы его не знаем (в этом-то и суть проблемы). У нас есть период 2014–2015 годы, когда было совершено 497 убийств; это наша оценка для за этот год. С ее помощью можно

найти стандартное отклонение: оно равно  $\sqrt{m} = \sqrt{497} = 22,3$ . Это

дает погрешность  $\pm 1,96 \times 22,3 \pm \pm 43,7$ . В итоге мы получаем приблизительный доверительный интервал для:  $4 \pm \pm 43,7$ , то есть от 453,3 до 540,7. Мы можем быть уверены на 95 %, что «истинный» уровень убийств за это время находится между 453 и 541.

На рис. 9.4 отображено наблюдаемое число убийств в Англии и Уэльсе с 1998 по 2016 год, а также 95-процентные доверительные интервалы для «истинного» уровня. Ясно, что, несмотря на неизбежные разбросы между ежегодными числами, доверительные интервалы показывают, что нужно весьма осторожно делать заключения о временных трендах. Например, 95-процентный интервал за 2015–2017 годы для числа 557 простирается от 511 до 603, то есть с существенным перекрытием с доверительным интервалом для предыдущего года.



**Рис. 9.4**  
 Число ежегодных убийств в Англии и Уэльсе между 1998 и 2016 годами, а также 95-процентные доверительные интервалы для «истинного» уровня убийств<sup>[183]</sup>

Итак, как же нам решить, произошло реальное изменение риска стать жертвой убийства или наблюдаемые изменения можно просто отнести к неизбежным случайным отклонениям? Если бы доверительные интервалы не перекрывались, то мы могли бы быть уверены, по крайней мере на 95 %, что изменение реально. Однако это довольно строгий критерий, и нам действительно следует построить 95-процентный интервал для изменения уровня убийств. Если такой интервал будет включать в себя 0, то мы не можем быть уверены в реальности изменения.

Между числом убийств за 2014–2015 и 2015–2016 годы произошло увеличение на  $557 - 477 = 60$ . Оказывается, 95-процентный доверительный интервал для этого наблюдаемого изменения простирается от  $-4$  до  $+124$ . Это включает 0 (правда, едва-едва). Формально это означает, что мы не можем с 95-процентной уверенностью заключить, что истинный уровень изменился, но, поскольку нулевое значение находится на самом краю доверительного интервала, было бы неразумно утверждать, что изменений вовсе нет.

У доверительных интервалов вокруг числа убийств на [рис. 9.4](#) совершенно иная природа по сравнению с погрешностями, скажем, для безработицы. Последние выражают нашу эпистемическую неопределенность в отношении фактического числа безработных, в то время как интервалы вокруг числа убийств не выражают неопределенности для их фактического количества (мы полагаем, что они подсчитаны верно), а относятся к истинным рискам убийств в обществе. Эти два вида интервалов могут похоже выглядеть и даже использовать одинаковую математику, однако их интерпретации принципиально разнятся.

В этой главе содержался довольно сложный материал, что неудивительно: фактически в ней заложен весь формальный фундамент для статистических выводов, основанных на вероятностном моделировании. Но усилия того стоят, поскольку теперь мы можем использовать эту конструкцию для выхода за рамки простых описаний и оценок характеристик мира и понимания того, как статистическое моделирование может нам помочь ответить на важные вопросы о реальном мироустройстве и таким образом обеспечить прочную основу для научных открытий.

## **Выводы**

- Теорию вероятностей можно использовать для получения распределения для выборочных статистик, из которых могут быть выведены формулы для доверительных интервалов.

- 95-процентный доверительный интервал определяется так: если мы проведем большое количество независимых экспериментов, для которых верны определенные предположения, то в 95 % этих испытаний построенный доверительный интервал будет содержать истинное значение параметра. Нельзя утверждать, что какой-то интервал с вероятностью 95 % содержит истинное значение.

- Из центральной предельной теоремы следует, что для больших выборок выборочное среднее и некоторые другие статистики имеют приблизительно нормальное распределение.

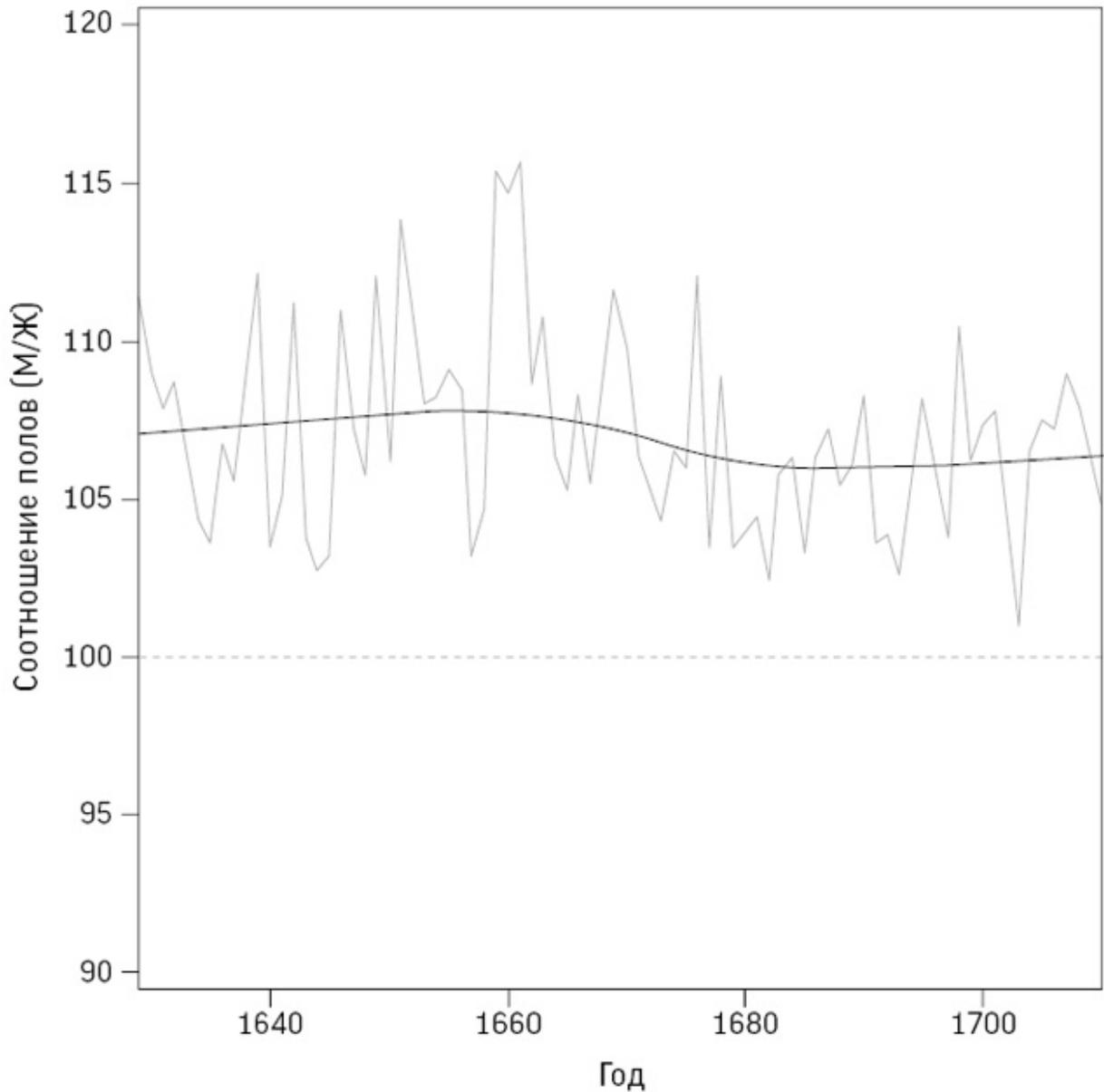
- Погрешности обычно не включают систематическую ошибку, вызванную не стохастическими причинами, – для ее оценивания нужны внешние знания и рассуждения.

- Доверительные интервалы можно вычислять, даже если мы наблюдаем все данные. Они отражают неопределенность параметров базовой метафорической совокупности.

## **Глава 10. Отвечаем на вопросы и заявляем об открытиях**

### **Рождается ли мальчиков больше, чем девочек?**

Врач Джон Арбетнот, ставший в 1705 году придворным лекарем королевы Анны, задался целью ответить на этот вопрос и проанализировал данные об обрядах крещения, проведенных в Лондоне за 82 года – с 1629 по 1710 год. Результаты его исследования приведены на рис. 10.1 в виде соотношения полов, то есть числа родившихся мальчиков на 100 родившихся девочек.



**Рис. 10.1**

Данные о соотношении полов (число мальчиков на 100 девочек) при обряде крещения в Лондоне между 1629 и 1710 годами, опубликованные Джоном Арбетнотом в 1710 году. Сплошная линия отображает равное число мальчиков и девочек; эта кривая построена по эмпирическим данным. Ежегодно мальчиков было окрещено больше, чем девочек

Арбетнот обнаружил, что ежегодно было окрещено больше мальчиков, чем девочек, причем соотношение колебалось от 101 до 116

и в целом составляло 107. Но он хотел вывести более общий закон, поэтому предположил, что если бы на самом деле никакой разницы между истинной долей мальчиков и девочек не было, то каждый год вероятность того, что мальчиков рождалось бы больше, чем девочек, а девочек рождалось бы больше, чем мальчиков, составила бы 50 на 50, то есть так же, как при подбрасывании монеты.

Но если предположить, что мальчики рождаются так же часто, как и девочки, и 82 года подряд наблюдается их избыток, то это все равно что 82 раза подбросить симметричную монету и каждый раз получить орла. Вероятность этого события составляет  $1/2^{82}$ . Это очень маленькое число, с 24 нулями после запятой. Если бы мы наблюдали 82 выпадения орла в реальном эксперименте, то уверенно бы заявили, что монета нечестная. Точно так же и Арбетнот заключил, что некая сила заставляет рождаться больше мальчиков в целях компенсации повышенной смертности мужского пола: «Чтобы восстановить эти Потери, предусмотрительная Природа по промыслу мудрого Творца рождает больше Мужчин, чем Женщин; и это почти постоянное соотношение»<sup>[184]</sup>.

Впоследствии данные Арбетнота не раз перепроверялись. И хотя в них могут быть ошибки подсчета и учтены только англиканские<sup>[185]</sup> обряды крещения, тем не менее его основной вывод по-прежнему верен: считается, что «естественное» соотношение полов – около 105, то есть на каждые 20 девочек рождается 21 мальчик. Название опубликованного им труда служит прямым статистическим подтверждением сверхъестественного вмешательства: «Аргумент в пользу Божественного провидения, извлеченный из постоянной регулярности, наблюдаемой при рождении обоих полов». И хотя Арбетнот тогда об этом не знал, он вошел в историю как человек, который выполнил первую проверку статистической значимости.

Пожалуй, мы подошли к самой важной части цикла решения проблем, где мы ищем ответы на конкретные вопросы о том, как устроен мир. Например:

1. Соответствует ли ежедневное число убийств в Соединенном Королевстве распределению Пуассона?

2. Изменился ли уровень безработицы в Соединенном Королевстве за последний квартал?

3. Снижает ли употребление статинов риск инфарктов и инсультов у людей вроде меня?

4. Связан ли рост матерей с ростом их сыновей, если учитывать рост отцов?

5. Существует ли бозон Хиггса?

Этот список показывает, что можно задавать самые разные вопросы – от преходящих до вечных.

1. Убийства и распределение Пуассона: общее правило, которое не представляет особого интереса для общества, но помогает ответить, произошли ли изменения в реальном уровне преступности.

2. Изменения в уровне безработицы: специфический вопрос, относящийся к конкретному времени и конкретному месту.

3. Статины: научное утверждение, но относящееся к конкретной группе.

4. Рост матерей: возможно, представляет научный интерес.

5. Бозон Хиггса: может изменить основные представления о физических законах Вселенной.

У нас есть данные, которые могут помочь ответить на некоторые из этих вопросов. Мы уже строили графики и делали какие-то неформальные заключения для подходящих статистических моделей. Однако сейчас мы подошли к составляющей этапа *анализа* цикла PPDAC, известной как **проверка гипотез**.

### ***Что такое «гипотеза»?***

Гипотезу можно определить как предлагаемое объяснение явления. Это не абсолютная истина, а временное рабочее предположение, которое, возможно, лучше всего представлять как подозреваемого в уголовном деле.

При обсуждении регрессии в главе 5 мы столкнулись с утверждением, что

наблюдение = детерминистская модель + остаточная ошибка.

Оно отражает идею, что статистические модели – это математические представления наших наблюдений, где сочетаются детерминистский и стохастический компоненты. Стохастический компонент отражает непредсказуемость, или случайную «ошибку», обычно выраженную в терминах какого-нибудь вероятностного распределения. В рамках статистики гипотезой считается какое-то конкретное предположение об одном из компонентов статистической модели, носящее смысловой оттенок «рабочей версии», а не «истины».

### ***Зачем нужно формально тестировать нулевые гипотезы?***

Открытия ценятся не только учеными – восторг от обнаружения чего-то нового универсален и настолько соблазнителен, что у нас есть врожденная склонность ощущать, будто мы нашли что-то новое, даже когда на самом деле этого нет. Ранее для описания способности видеть закономерности и связи там, где их не существует, мы использовали термин *апофения*; даже было высказано предположение, что такая склонность обеспечивает определенное эволюционное преимущество – те наши предки, которые, заслышав шорох в кустах, тут же убегали, не дожидаясь момента, чтобы выяснить, действительно ли там тигр, выживали с большей вероятностью.

Однако такое отношение может быть приемлемым для охотников-собирателей, но не для науки, ведь когда наши утверждения остаются всего лишь плодом нашего воображения, подрывается сама суть научного процесса. Должен существовать способ защитить нас от ложных открытий, и проверка гипотез претендует на эту роль.

Центральной здесь становится идея **нулевой гипотезы**: это упрощенная форма статистической модели, с которой мы будем работать, пока не получим достаточное количество аргументов против нее. Для вышеуказанных вопросов нулевыми гипотезами могут быть:

1. Ежедневное количество убийств в Соединенном Королевстве *имеет* распределение Пуассона.
2. Уровень безработицы в Соединенном Королевстве за последний квартал *не изменился*.
3. Статины *не уменьшают* риск инфарктов и инсультов у людей вроде меня.
4. Рост матерей *не влияет* на рост сыновей, если учесть рост отцов.

## 5. Бозона Хиггса *не существует*.

Нулевая гипотеза – это то, что мы готовы принять, пока не докажем обратное. Она безжалостно негативна в своем отрицании прогресса и перемен. Но это не означает, что на самом деле мы верим в ее безусловную правильность: ясно, что ни одна из вышеперечисленных гипотез не может быть в точности верной (за исключением, возможно, несуществования бозона Хиггса). Поэтому мы никогда не можем заявить, что нулевая гипотеза фактически доказана. Как говорил великий британский статистик Рональд Фишер, «нулевая гипотеза в ходе экспериментов никогда не доказывается, но, возможно, опровергается. Можно сказать, что любой эксперимент существует только для того, чтобы дать фактам шанс опровергнуть нулевую гипотезу»<sup>[186]</sup>.

Весьма хороша аналогия с уголовным судопроизводством в английской правовой системе: подсудимого можно признать виновным, но никого нельзя признать невиновным просто потому, что не доказана его вина. Точно так же мы можем что-то найти, что способно опровергнуть нулевую гипотезу, но если у нас недостаточно доказательств, это вовсе не значит, что мы можем считать ее истинной. Это всего лишь рабочее предположение, пока не найдется что-нибудь получше.

**Скрестите руки на груди. У вас сверху левая или правая рука? Согласно исследованиям, примерно у половины людей сверху правая рука и у половины – левая. Но связано ли это с тем, мужчина вы или женщина?**

Хотя это не самый насущный научный вопрос, который я исследовал, когда преподавал в Африканском институте математических наук<sup>[187]</sup> в 2013 году, он был прекрасным упражнением для аудитории, а мне действительно хотелось узнать ответ<sup>[188]</sup>. Я получил данные по 54 аспирантам со всей Африки. Табл. 10.1 показывает общее распределение ответов по полу и положению правой или левой руки сверху. Такой тип таблицы в статистике называется таблицей сопряженности, или факторной таблицей.

**Таблица 10.1**

Таблица сопряженности полов и положения рук при скрещивании для 54 аспирантов

	Женщины	Мужчины	Всего
Левая рука сверху	5	17	22
Правая рука сверху	9	23	32
Всего	14	40	54

В целом большинство кладут сверху правую руку ( $32/54 = 59\%$ ). Однако у женщин доля таких «праворуких» ( $9/14 = 64\%$ ) выше, чем у мужчин ( $23/40 = 57\%$ ): наблюдаемая разница между долями составляет  $64 - 57 = 7\%$ . В этом случае нулевая гипотеза состоит в том, что между скрещиванием рук и полом нет никакой связи, а потому наблюдаемая разница в долях между полами должна равняться  $0\%$ . Ключевой вопрос: может ли наблюдаемое отклонение в  $7\%$  считаться достаточно большим, чтобы противостоять нулевой гипотезе?

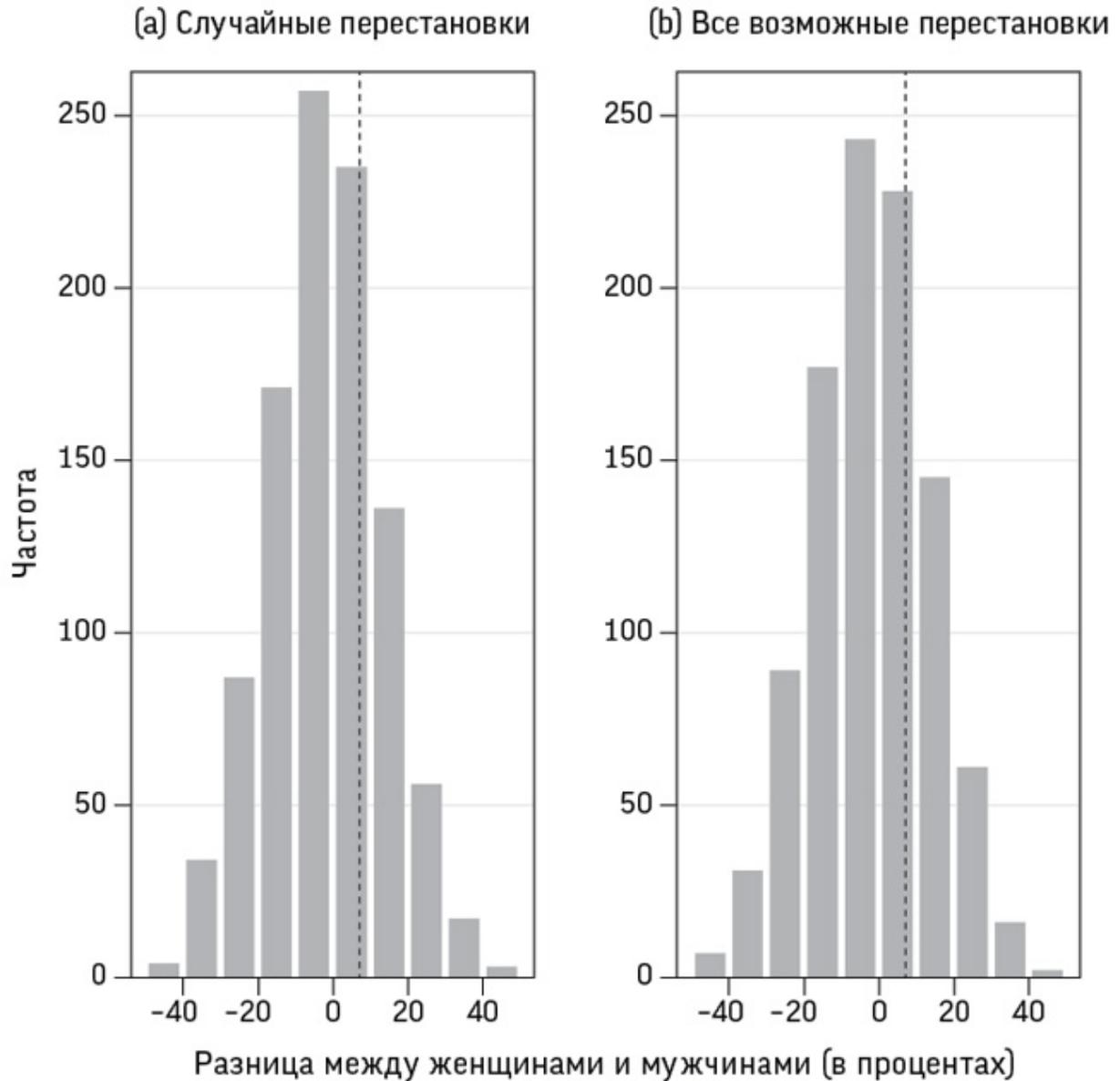
Чтобы ответить на этот вопрос, нужно знать, какой разницы в долях мы можем ожидать просто вследствие случайного разброса при справедливости нулевой гипотезы, то есть независимости скрещивания рук и пола. Более формально: совместима ли наблюдаемая разница  $7\%$  с нулевой гипотезой?<sup>[189]</sup>

Это сложная, но важная идея. Когда Арбетнот проверял свою нулевую гипотезу, согласно которой мальчики и девочки рождаются равновероятно, он смог легко выяснить, что наблюдаемые данные ни в малейшей степени не совместимы с нулевой гипотезой – шансы, что по чистой случайности мальчики будут численно превосходить девочек 82 года подряд, ничтожно малы. В более сложных ситуациях выяснить, совместимы ли данные с нулевой гипотезой, не так просто. Однако приведенный ниже **тест перестановки** отображает мощную процедуру, позволяющую избежать сложной математики.

Представьте, что все 54 человека выстроились в ряд, сначала 14 женщин, а затем 40 мужчин, и каждому присвоен номер от 1 до 54.

Допустим, у каждого есть билет, указывающий, какая рука у него при скрещивании сверху – левая или правая. А теперь вообразите, что все эти билеты смешали в шляпе и раздали присутствующим наугад. Это пример того, каких результатов можно ожидать, если бы нулевая гипотеза была верна, ведь при случайной раздаче скрещивание рук и пол никак не связаны.

Но даже при случайном распределении доля «державших сверху правую руку» не будет в точности совпадать для мужчин и женщин (просто из-за чистой случайности), и мы можем вычислить наблюдаемую разницу в долях для этой случайной раздачи билетов. Затем мы могли бы повторить процесс, скажем 1000 раз, и посмотреть, какое распределение будет у этой разницы. Результаты приведены на рис. 10.2(a): показан разброс наблюдаемых разниц – некоторые в пользу мужчин, некоторые в пользу женщин – с центром в нуле. Фактически наблюдаемая разница находится недалеко от центра распределения.



**Рис. 10.2**

Эмпирическое распределение разницы между долями женщин и мужчин, которые при скрещивании рук кладут сверху правую руку: (а) для 1000 случайных перестановок, (б) для всех равновероятно возможных перестановок по отношению к скрещиванию рук. Наблюдаемое различие в пропорциях (7 %) обозначено вертикальной пунктирной линией

В качестве альтернативы (при наличии времени) можно взять все возможные перестановки билетов, не ограничиваясь моделированием 1000 симуляций. Каждая перестановка даст какую-то наблюдаемую

разницу в долях «праворуких» у мужчин и женщин, и, нанеся на график все результаты, мы получим более гладкое распределение, чем построенное по 1000 симуляциям.

К несчастью, таких перестановок масса, и даже если вычислять их со скоростью миллион в секунду, на это уйдет число лет с 57 нулями<sup>[190]</sup>. К счастью, нам незачем производить эти вычисления, поскольку распределение для наблюдаемой разницы в таких долях при нулевой гипотезе можно найти теоретически: оно представлено на рис. 10.2(b) и основано на так называемом **гипергеометрическом распределении**.

Рис. 10.2 показывает, что реально наблюдаемая разница в долях «праворуких» мужчин и женщин (7 % в пользу женщин) лежит достаточно близко к центру распределения для разниц, которых можно было бы ожидать, если бы никакой связи вообще не было. Нам нужна мера, характеризующая, насколько близко к центру лежит наблюдаемое значение, и одна из таких характеристик – это площадь хвоста распределения. Например, площадь части фигуры, расположенной справа от вертикальной пунктирной линии, составляет 0,45, или 45 %.

Это число именуется **P-значением**<sup>[191]</sup> и считается одним из самых полезных понятий в статистике, а потому заслуживает строгого определения: *P-значение – это вероятность получить результат, по крайней мере такой же или более экстремальный, чем наблюдаемый, если нулевая гипотеза (и все другие предположения моделирования) на самом деле верна.*

Но тут есть важный нюанс, что мы подразумеваем под «экстремальным» результатом? Наше P-значение 0,45 **одностороннее**, так как указывает, насколько вероятно получить в эксперименте не меньшую разницу в пользу только женщин, если нулевая гипотеза верна. Это P-значение используется при так называемых **односторонних критериях**. Но ведь большая разница в пользу мужчин тоже заставила бы нас сомневаться в справедливости нулевой гипотезы. Поэтому мы должны также вычислить вероятность получить отклонение не меньше 7 % в *обоих* направлениях. Так появляются **двусторонние P-значения**, соответствующие **двусторонним критериям**. Общая площадь двух частей фигуры, отдаленных от центра-нуля больше чем на 7 %, равна примерно 0,89, а поскольку это значение близко к единице, следовательно, наблюдаемое значение

находится близко к центру нулевого распределения. Конечно, на [рис. 10.2](#) это видно и так, но, учитывая, что подобные гистограммы доступны не всегда, нам нужно число, формально выражающее «экстремальность» наших данных.

Арбетнот предоставил первый зафиксированный пример такой процедуры: при нулевой гипотезе (когда девочки и мальчики рождаются с равной вероятностью) вероятность того, что 82 года подряд мальчики будут рождаться чаще девочек, равна  $1/2^{82}$ . Но так определяется «экстремальность» только в терминах превосходства мальчиков. А поскольку мы можем сомневаться и в нулевой гипотезе, что 82 года подряд девочки будут рождаться чаще мальчиков, то должны удвоить это число, чтобы получить экстремальный результат в обоих направлениях. Поэтому число  $1/2^{82}$  можно считать первым установленным двусторонним Р-значением, хотя этот термин появился только через 250 лет.

Кстати, моя небольшая выборка не выявила никакой связи между полом и скрещиванием рук, да и другие, более научные исследования не обнаружили взаимосвязи между поведением при скрещивании рук, полом, леворукостью и остальными признаками.

### ***Статистическая значимость***

Идея **статистической значимости** проста: когда Р-значение достаточно мало, мы говорим, что результаты статистически значимы. Этот термин был популяризирован Рональдом Фишером в 1920-х годах и, несмотря на критику, которую мы рассмотрим позже, продолжает играть в статистике важную роль.

Рональд Фишер был незаурядным, но трудным человеком. Незаурядным потому, что его считают пионером в двух областях – генетике и статистике. А трудным, поскольку, имея весьма скверный характер, мог крайне негативно отзываться о тех, кто (по его мнению) оспаривал его идеи; к тому же его репутации сильно повредила поддержка евгеники и критика доказательств связи между курением и раком легких. И хотя его личная репутация пострадала в результате обнаружения его финансовых связей с табачной промышленностью, на научной репутации ученого это никак не сказалось, так как его идеи неизменно находят новое применение при анализе больших массивов данных.

Как упоминалось в [главе 4](#), Фишер развил идею рандомизации для сельскохозяйственных испытаний во время работы на опытной сельскохозяйственной станции в Ротамстеде. Потом он продемонстрировал идеи рандомизации в своем знаменитом тесте с дегустацией чая, в ходе которого некая женщина (по имени Мюриэль Бристоль) заявила, что может по вкусу определить, добавляли в чашку молоко *до* или *после* чая.

В четыре чашки налили сначала чай, а затем молоко, а в четыре – сначала молоко, а потом чай. Все восемь чашек в случайном порядке выставили в ряд и сообщили Мюриэль, что здесь по четыре чашки каждого вида наливания. Говорят, она правильно определила все чашки. Если считать нулевой гипотезой то, что Мюриэль просто угадывала, то с помощью гипергеометрического распределения нетрудно показать, что вероятность этого равна  $1/70 \approx 1,4\%$ <sup>[192]</sup>. Такое Р-значение считается маленьким<sup>[193]</sup>, а потому результат можно объявить статистически значимым подтверждением того, что Мюриэль не угадывала, а действительно умела различать, в какой последовательности доливали молоко.

Подводя итог, мы действуем следующим образом.

1. Ставим вопрос в терминах нулевой гипотезы, которую хотим проверить. Обычно она обозначается  $H_0$ .

2. Выбираем какую-нибудь статистику критерия, которая, если ее величина будет достаточно экстремальной, позволит нам поставить под сомнение нулевую гипотезу (часто большие значения такой статистики указывают на несовместимость с нулевой гипотезой).

3. Создаем выборочное распределение этой статистики при условии, что нулевая гипотеза верна.

4. Проверяем, находится ли наблюдаемая величина в хвостах этого распределения, что определяем с помощью Р-значения: какова вероятность наблюдаемого экстремального распределения в случае, если верна нулевая гипотеза. Численно эта вероятность представляет собой площадь части распределения, лежащей правее наблюдаемой величины.

5. Аккуратно подходим к определению, что такое «экстремальная» величина, – например, если с нулевой гипотезой несовместимы и

большие положительные, и большие отрицательные значения статистики критерия, то P-значение должно это учитывать.

6. Объявляем результат статистически значимым, если P-значение меньше некоторой критической пороговой величины.

Рональд Фишер использовал в качестве удобных порогов значимости  $P < 0,05$  и  $P < 0,01$  и составил таблицы критических значений статистики критерия, которые нужно превзойти, чтобы получить такие уровни значимости. Ввиду популярности этих таблиц числа 0,05 и 0,01 стали общепринятыми, хотя сейчас рекомендуется указывать точные P-значения. Важно подчеркнуть, что точное P-значение зависит не только от истинности нулевой гипотезы, но и от всех других допущений, лежащих в основе статистической модели, например отсутствия систематической ошибки, независимости наблюдений и так далее.

Весь этот процесс известен как проверка значимости нулевой гипотезы (NHST – Null Hypothesis Significance Testing), и, как мы увидим далее, он стал источником серьезных разногласий. Но сначала посмотрим, как идеи Фишера работают на практике.

### ***Использование теории вероятностей***

Пожалуй, самый сложный в проверке значимости нулевой гипотезы третий шаг – определение распределения выбранной статистики при нулевой гипотезе. Мы всегда можем вернуться к методам компьютерного моделирования (как с тестом перестановки для данных о скрещивании рук на груди), однако намного удобнее работать с хвостами статистического критерия непосредственно с помощью теории вероятностей, как это делали Арбетнот (в простейшем случае) и Фишер (применивший гипергеометрическое распределение в эксперименте с чашками).

Часто мы используем приближения (аппроксимации), разработанные пионерами статистики. Например, около 1900 года Карл Пирсон разработал несколько критериев для проверки зависимости для таблиц сопряженности (таких как [табл. 10.1](#)). Из этого вырос классический критерий согласия  $\chi^2$  (хи-квадрат).

Эти проверки включают вычисление ожидаемого числа событий, попадающих в каждую ячейку таблицы при условии справедливости

нулевой гипотезы (отсутствие зависимости), после чего статистика хи-квадрат измеряет общее расхождение между наблюдаемыми и ожидаемыми значениями. В табл. 10.2 приведены ожидаемые значения в ячейках таблицы при условии нулевой гипотезы: например, ожидаемое количество женщин, кладущих сверху левую руку, равно общему числу женщин (14), умноженному на долю всех «леворуких» ( $22/54$ ), и составляет 5,7.

**Таблица 10.2**

Наблюдаемое и ожидаемое (в скобках) число людей, кладущих сверху правую или левую руку, в зависимости от пола. Ожидаемые количества вычислены при нулевой гипотезе, согласно которой скрещивание рук не зависит от пола

	Женщины	Мужчины	Всего
Левая рука сверху	5 (5,7)	17 (16,3)	22
Правая рука сверху	9 (8,3)	23 (23,7)	32
Всего	14	40	54

Из табл. 10.2 видно, что наблюдаемое и ожидаемое число довольно близки, то есть реальные данные соответствуют тому, что мы могли бы ожидать при нулевой гипотезе. Статистика хи-квадрат – это общая мера расхождения между наблюдаемыми и ожидаемыми значениями (ее формула приводится в глоссарии), в данном случае она равна 0,02. Соответствующее Р-значение (есть в таблицах или программах) составляет 0,90, что не противоречит нулевой гипотезе. Обнадуживает то, что оно фактически то же, что и «точный» критерий, основанный на гипергеометрическом распределении.

Разработка и использование статистических критериев и Р-значений традиционно составляют значительную часть стандартного курса статистики и, к сожалению, обеспечивают этой области репутацию места, где в основном следует брать нужную формулу и использовать нужную таблицу. И хотя цель этой книги – сформировать более широкий взгляд на предмет, тем не менее полезно рассмотреть

примеры, которые мы обсуждали, с точки зрения статистической значимости.

### **1. Соответствует ли ежедневное число убийств в Англии и Уэльсе распределению Пуассона?**

На [рис. 8.5](#) отображено наблюдаемое количество дней с различным числом убийств в Англии и Уэльсе за 2014–2016 годы. Всего за 1095 дней зафиксировано 1545 случаев убийства, в среднем – 1,41 в день. Если в качестве нулевой гипотезы принять, что убийства имеют распределение Пуассона со средним 1,41, то можно ожидать чисел, указанных в последнем столбце табл. 10.3. Используя тот же подход, что и для [табл. 10.2](#), для расхождения между наблюдаемыми и ожидаемыми данными можно применить **критерий согласия хи-квадрат** (см. подробности в [гlossарии](#)).

#### **Таблица 10.3**

Наблюдаемое и ожидаемое количество дней с определенным числом случаев убийства в Англии и Уэльсе с апреля 2014 по март 2016 года. Критерий согласия хи-квадрат дает Р-значение 0,96, что указывает на отсутствие расхождений с нулевой гипотезой о распределении Пуассона

Количество случаев убийства в день	Наблюдаемое количество дней	Ожидаемое (при нулевой гипотезе) количество дней
0	259	267,1
1	387	376,8
2	261	265,9
3	131	125,0
4	40	44,1
5	13	12,4
6 или более	3	3,6
Всего	1095	1095

Наблюдаемое Р-значение 0,96 не значимо, поэтому нет оснований отклонять нулевую гипотезу (на самом деле согласие настолько хорошее, что это почти подозрительно). Конечно, нам не стоит предполагать, что нулевая гипотеза однозначно истинна, но было бы разумно использовать ее в качестве исходного предположения, например, при оценке изменения уровня убийств, описанного в [главе 9](#).

## 2. Изменился ли уровень безработицы в Великобритании в недавнем прошлом?

В [главе 7](#) мы показали, что квартальное изменение уровня безработицы на 3000 имело погрешность  $\pm 77\ 000$  (то есть  $\pm 2$  стандартные ошибки). Это означает, что 95-процентный доверительный

интервал простирается от  $-80\ 000$  до  $+74\ 000$  и явно содержит  $0$ , соответствующий отсутствию изменения уровня безработицы. Но то, что 95-процентный доверительный интервал включает  $0$ , логически эквивалентно тому, что оценка  $-3000$  отклоняется от  $0$  меньше чем на 2 стандартные ошибки, а значит, такое изменение не отличается значимо от  $0$ .

Это обнаруживает принципиальное сходство между проверкой гипотез и доверительными интервалами:

- двустороннее Р-значение меньше  $0,05$ , если 95-процентный доверительный интервал не включает нулевую гипотезу (обычно  $0$ );
- 95-процентный доверительный интервал – это набор нулевых гипотез, которые не отвергаются при  $P < 0,05$ .

Эта тесная связь между проверкой гипотез и доверительными интервалами должна помешать людям неправильно интерпретировать результаты, которые статистически значимо не отличаются от  $0$ , – это означает не то, что нулевая гипотеза действительно верна, а то, что доверительный интервал для истинного значения содержит  $0$ . К сожалению, как мы увидим далее, этот урок часто игнорируется.

### **3. Снижает ли употребление статинов риск инфарктов и инсультов у людей вроде меня?**

Табл. 10.4 воспроизводит результаты исследования по защите сердца (HPS), ранее представленные в [табл. 4.1](#), но с добавлением столбцов, демонстрирующих степень уверенности в улучшении показателей. Между стандартными ошибками, доверительными интервалами и Р-значениями существует тесная связь. Доверительные интервалы для снижения риска – это, грубо говоря, оценка  $\pm 2$  стандартные ошибки (обратите внимание, что в исследовании по защите сердца относительные уменьшения округляются до целых чисел). Доверительные интервалы легко исключают нулевую гипотезу  $0\%$ , соответствующую отсутствию воздействия статина, а Р-значения ничтожно малы – фактически Р-значение для  $27\%$  снижения риска инфаркта составляет 1 на 3 миллиона. Это следствие масштабности исследования.

**Таблица 10.4**

Результаты исследования по защите сердца, показывающие оцениваемые относительные эффекты, их стандартные ошибки, доверительные интервалы и Р-значения при проверке нулевой гипотезы «эффект приема отсутствует»

Событие	Процентная доля в группе из 10 269 человек, которым назначен статин	Процентная доля в группе из 10 267 человек, которым назначено плацебо	Процентное (относительное) снижение риска у людей, которым назначен статин	Стандартная ошибка для снижения риска	Доверительный интервал для процентного снижения	Р-значение
Инфаркт	8,7	11,8	27%	4%	от 21 до 33%	P < 0,0001
Инсульт	4,3	5,7	25%	5%	от 15 до 34%	P < 0,0001
Смерть по любой причине	12,9	14,7	13%	4%	от 6 до 19%	P = 0,0003

Можно использовать и другие статистики, например разницу в абсолютных рисках, но все они должны давать близкие Р-значения. Специалисты, проводившие HPS, сосредоточились на пропорциональном снижении, поскольку оно почти постоянно в отдельных подгруппах и поэтому обеспечивает хорошую единую меру. Существует несколько способов расчета доверительных интервалов, но они дают лишь небольшие расхождения.

#### **4. Связан ли рост матерей с ростом их сыновей, если учитывать рост отцов?**

В [главе 5](#) мы продемонстрировали множественную линейную регрессию, с ростом сыновей в качестве зависимой переменной (переменной отклика) и ростом отца и матери в качестве независимых (объясняющих) переменных. Коэффициенты регрессии приведены в [табл. 5.3](#), но без указания, можно ли их считать значимо отличными от 0. Чтобы проиллюстрировать, как эти результаты появляются в статистических программах, табл. 10.5 воспроизводит результаты работы популярной (бесплатной) программы R.

**Таблица 10.5**

Выдаваемый программой R результат для множественной линейной регрессии по данным Гальтона. Переменная отклика – рост сыновей, объясняющие переменные – рост матерей и отцов;  $t$ -значение – это оценка, деленная на стандартную ошибку. Столбец  $\text{Pr}( > |t| )$  представляет двустороннее P-значение; вероятность получения какого-то большего  $t$ -значения (положительного или отрицательного) при нулевой гипотезе, что истинное значение 0. Обозначение « $2 \text{e} - 16$ » означает, что P-значение меньше 0,0000000000000002 (то есть 15 нулей). Под таблицей дана расшифровка звездочек в терминах P-значений

	Оценка	Стандартная ошибка	$t$ -значение	$\text{Pr}( >  t  )$
(Отсекаемый отрезок)	69,22882	0,10664	649,168	$< 2 \text{e} - 16^{***}$
Рост матери	0,33355	0,04600	7,252	$1,74 \text{e} - 12^{***}$
Рост отца	0,41175	0,04668	8,820	$< 2 \text{e} - 16^{***}$

Обозначения уровня значимости:  $*** = 0,001$   $** = 0,01$   $* = 0,05$

Как и в [табл. 5.3](#), отсекаемый отрезок – это средний рост сыновей, а коэффициенты (в столбце оценок) – ожидаемое изменение роста при отклонении роста матери и отца на один дюйм от среднего роста матерей и отцов. Стандартная ошибка рассчитывается по известной формуле и явно мала по сравнению с величиной коэффициентов.

Основное внимание сосредоточено на  $t$ -значении, также известном как  **$t$ -статистика**, поскольку именно оно указывает на то, можно ли считать связь между объясняющей переменной и переменной отклика статистически значимой. Это частный случай так называемого  $t$ -критерия Стьюдента. Стьюдент – псевдоним Уильяма Госсета, разработавшего этот метод в 1908 году для оценки качества пива пивоваренной компании «Гиннесс», которая хотела сохранить анонимность своих сотрудников. Для получения  $t$ -значения оценка делится на стандартную ошибку (это можно проверить для чисел в табл. 10.5), поэтому его можно интерпретировать как отклонение оценки от нуля, выраженное в стандартных ошибках. Учитывая  $t$ -значение и размер выборки, программа может выдать точное  $P$ -значение; для больших выборок  $t$ -значения больше 2 или меньше  $-2$  соответствуют  $P < 0,05$ , хотя для меньших размеров выборок эти пороговые значения будут больше. Программа R использует простую систему звездочек для  $P$ -значений – от одной, означающей  $P < 0,05$ , до трех, означающих  $P < 0,001$ . В табл. 10.5  $t$ -значения настолько велики, что  $P$ -значения исчезающе малы.

В [главе 6](#) мы показали, что алгоритм может выиграть конкурс прогнозов с очень незначительным преимуществом. Например, при прогнозе выживания для тестового набора данных о «Титанике» простое дерево классификации дало наилучший показатель Бриера (среднеквадратичная ошибка прогноза) 0,139, что лишь чуть-чуть отличается от величины 0,142 у усредненной нейронной сети (см. [табл. 6.4](#)). Вполне резонно спросить, действительно ли эта крохотная разница  $-0,003$  статистически значима или все можно объяснить случайными отклонениями?

Это несложно проверить,  $t$ -статистика составляет  $-0,54$ , а двустороннее  $P$ -значение равно 0,59<sup>[194]</sup>. Поэтому достаточно веских оснований для утверждений, что дерево классификации – наилучший алгоритм, нет! Для конкурсов вроде устраиваемых Kaggle подобный анализ не считается тривиальным, но важно помнить, что статус победителя зависит от выбора тестового набора.

Исследователи тратят свои жизни на тщательное изучение результатов работы компьютерных программ наподобие представленных в табл. 10.5 в надежде увидеть мерцающие звезды, указывающие на существенный результат, который они могут получить

и затем включить в следующую научную статью. Но, как мы видим, такой навязчивый поиск статистической значимости довольно легко приводит к заблуждениям.

### ***Опасность выполнения нескольких проверок на значимость***

Стандартные пороговые значения для «значимости»  $P < 0,05$  и  $P < 0,01$  Рональд Фишер выбрал для своих таблиц весьма произвольно, поскольку в те времена вычислять точные  $P$ -значения без механических и электрических калькуляторов было невозможно. Но что произойдет, если провести много проверок на значимость, каждый раз наблюдая, не превышает ли наше  $P$ -значение величину 0,05?

Предположим, что лекарство на самом деле не помогает, тогда нулевая гипотеза истинна. Проведя одно клиническое испытание, мы назовем результат статистически значимым, если  $P$ -значение меньше 0,05. Поскольку препарат неэффективен, такая вероятность составляет 0,05, или 5 %, что, собственно, и есть определением  $P$ -значения. Это будет считаться **ложноположительным** результатом, так как мы (неправильно) решим, что лекарство помогает. Если мы проведем два испытания и посмотрим на результаты, то вероятность получить хотя бы один значимый, то есть ложноположительный, результат близка к 0,10, или 10 %<sup>[195]</sup>. При увеличении количества испытаний шансы на получение хотя бы одного ложноположительного результата быстро растут: если провести десять испытаний бесполезных препаратов, вероятность получить хотя бы один значимый результат при  $P < 0,05$  достигает 40 %. Такая ситуация известна как проблема **множественной проверки гипотез**, она возникает всякий раз, когда проверок выполняется много, а сообщается о самом значимом результате.

Еще одна проблема возникает, когда исследователи делят данные на много подклассов, проверяют гипотезу на каждом из них, а затем рассматривают самые значимые результаты. Классический пример – эксперимент, проведенный авторитетными исследователями в 2009 году, в котором испытуемому показывали серию фотографий людей с различными эмоциями на лице и проводили сканирование мозга (функциональную магнитно-резонансную томографию, фМРТ), чтобы посмотреть, какая его зона даст значимый отклик, приняв  $P < 0,001$ .

Изюминка заключалась в том, что «испытуемым» был двухкилограммовый атлантический лосось, который «не был жив на момент сканирования». Из 8064 участков мозга этой крупной мертвой рыбыны 16 продемонстрировали статистически значимый отклик на фотографии. Ученые не стали утверждать, что мертвый лосось обладает уникальными умениями, а сделали верный вывод<sup>[196]</sup>, что проблема в многократном тестировании – более 8 тысяч проверок обязательно приведут к ложноположительному результату<sup>[197]</sup>. Даже при строгом критерии  $P < 0,001$  мы бы ожидали 8 значимых результатов по чистой случайности.

Один из способов обойти эту проблему – потребовать очень маленькое  $P$ -значение для уровня значимости, и здесь проще всего применить **поправку Бонферрони**<sup>[198]</sup>, то есть использовать пороговое значение  $0,05/n$ , где  $n$  – число проведенных тестов. Таким образом, проверки для каждого участка мозга лосося можно выполнять, требуя  $P$ -значение, равное  $0,05/8000 = 0,00000625$ , или 1 на 160 000. Этот метод стал стандартным при поиске в геноме человека участков, связанных с болезнями: поскольку существует примерно 1 000 000 участков для генов, прежде чем заявлять об открытии, положено потребовать значение  $P$  меньше  $0,05/1000000 = 1$  на 20 миллионов.

Таким образом, при одновременной проверке большого количества гипотез (например, в области исследований мозга или геномике) метод Бонферрони позволяет решать, значимы ли наиболее экстремальные результаты. Также разработаны несложные методы, слегка смягчающие критерий Бонферрони для второго по экстремальности результата, третьего и так далее. Так контролируется общая доля «открытий», которые оказываются ложными, – так называемый **уровень ложноположительных результатов**.

Еще один способ избежать ложноположительных результатов – потребовать воспроизведения первоначального исследования, с проведением повторного эксперимента в других условиях, но с тем же протоколом. Чтобы американское Управление по санитарному надзору за качеством пищевых продуктов и медикаментов одобрило новый препарат, необходимо провести два независимых клинических испытания, причем в каждом должна быть показана клиническая польза с уровнем значимости  $P < 0,05$ . Это означает, что вероятность

одобрить неэффективный препарат составляет всего  $0,05 \times 0,05 = 0,0025$ , или 1 на 400.

## 5. Существует ли бозон Хиггса?

На протяжении XX века физики разрабатывали стандартную модель, предназначенную для объяснения сил, действующих на субатомном уровне. Но одна часть модели оставалась недоказанной теорией – «поле Хиггса», которое объясняет наличие масс у частиц-переносчиков слабого взаимодействия. Квантом такого поля должна была стать гипотетическая частица – так называемый бозон Хиггса. В 2012 году исследователи из ЦЕРН<sup>[199]</sup> заявили о его открытии, как о результате «пять сигма»<sup>[200]</sup>. Однако мало кто понимал, что это показывало уровень статистической значимости.

Когда ученые построили график появления определенных событий для различных уровней энергии, оказалось, что кривая имеет четко выраженный «горб» именно в том месте, где его и следовало ожидать, если бы бозон Хиггса существовал. Важно то, что критерий согласия хи-квадрат дает Р-значение меньше 1 на 3,5 миллиона при нулевой гипотезе, что бозона Хиггса не существует и горб был просто результатом случайного отклонения. Но почему об открытии сообщалось как о «пяти сигма»?

В теоретической физике стандартом считается объявление об открытии в терминах «сигм», где результат «два сигма» означает наблюдение, которое на две стандартные ошибки отклонилось от нулевой гипотезы (вспомните, что мы использовали греческую букву *сигма* ( $\sigma$ ) для обозначения среднеквадратичного отклонения в генеральной совокупности): «сигмы» в теоретической физике точно соответствуют  $t$ -значению в результатах работы компьютерной программы для множественной линейной регрессии, показанных в [табл. 10.5](#). Наблюдение, которое в соответствии с критерием хи-квадрат давало двустороннее Р-значение 1 на 3,5 миллиона, отличалось бы от нулевой гипотезы на пять стандартных ошибок. Поэтому об открытии бозона Хиггса объявили как о результате уровня пять сигма.

Команда из ЦЕРН не хотела объявлять о своем «открытии» до тех пор, пока Р-значение не стало крайне малым. Во-первых, им нужно было учесть тот факт, что проверки значимости выполнялись для всех уровней энергии, а не только однократно в итоговой проверке по

методу хи-квадрат – в физике такой эффект при многократном тестировании известен как Look-elsewhere effect<sup>[201]</sup>. Во-вторых, они хотели быть уверены, что любая попытка воспроизвести результат даст тот же вывод. Было бы слишком неловко делать ложные заявления о законах физики.

Что касается ответа на вопрос, поставленный в начале этого раздела, то сейчас разумнее предположить, что бозон Хигса существует. Это утверждение становится новой нулевой гипотезой, пока, возможно, не появится более глубокая теория.

### *Теория Неймана – Пирсона*

#### **Почему для исследования по защите сердца понадобилось больше 20 тысяч участников?**

Исследование по защите сердца было масштабным, но его размер не определялся произвольным образом. При планировании испытания исследователи должны были указать, сколько людей нужно включить в группу, употребляющую статины или плацебо, причем требовалось серьезное статистическое обоснование, чтобы оправдать стоимость такого эксперимента. План опирался на идеи, развитые Ежи Нейманом и Эгоном Пирсоном, которых мы уже упоминали как разработчиков понятия доверительных интервалов.

Идею Р-значений и проверок значимости Рональд Фишер внедрил в 1920-х годах как способ проверки разумности какой-то конкретной гипотезы. Если наблюдается маленькое Р-значение, то либо случилось нечто удивительное, либо нулевая гипотеза ложна: чем меньше Р-значение, тем обоснованнее утверждение, что она неверна. Эта методика задумывалась как довольно неформальная процедура, но разработанная Нейманом и Пирсоном в 1930-х теория **индуктивного поведения** попыталась поставить ее на более строгий математический фундамент.

Конструкция ученых требовала указывать не только нулевую, но и альтернативную гипотезу, которая предоставляла более сложное объяснение данных. Затем они рассматривали возможные решения после проверки – либо отвергнуть нулевую гипотезу в пользу альтернативы, либо не отвергать<sup>[202]</sup>. Это приводило к вероятности появления двух видов ошибок – **ошибки первого рода**, возникающей,

когда мы отклоняем истинную нулевую гипотезу, и **ошибки второго рода** – когда принимаем неверную нулевую гипотезу. Строгая юридическая аналогия проиллюстрирована в табл. 10.6: ошибка первого рода – это осудить невиновного человека, а ошибка второго рода – признать кого-то невиновным, тогда как на самом деле он совершил преступление.

**Таблица 10.6**

Возможные результаты проверки гипотезы, по аналогии с судебным разбирательством

Истина	Результат проверки гипотезы	
	Не отвергать нулевую гипотезу (признать подозреваемого невиновным)	Отвергнуть нулевую гипотезу в пользу альтернативы (признать подозреваемого виновным)
Нулевая гипотеза (подозреваемый невиновен)	Правильное решение: не отклонять нулевую гипотезу Правильно признать невиновного человека невиновным	Ошибка первого рода: неправильно отклонить нулевую гипотезу Неправильно осудить невиновного
Альтернативная гипотеза (подозреваемый виновен)	Ошибка второго рода: неправильно отклонить нулевую гипотезу Неправильно не осудить виновного	Правильное решение: отвергнуть нулевую гипотезу Правильно осудить виновного

Нейман и Пирсон предложили при планировании эксперимента выбирать две величины, которые вместе определяют, насколько масштабным он должен быть. Во-первых, следует заранее зафиксировать значение для вероятности ошибки первого рода (при условии, что нулевая гипотеза верна), скажем 0,05; это называется **размером критерия** и, как правило, обозначается буквой  $\alpha$  (альфа). Во-

вторых, нужно заранее определить вероятность ошибки второго рода (при условии, что альтернативная гипотеза верна); она традиционно обозначается  $\beta$  (бета). На самом деле исследователи обычно работают с величиной  $1 - \beta$ , которая именуется **мощностью критерия** и представляет собой вероятность отклонить нулевую гипотезу в пользу альтернативной, когда последняя верна. Другими словами, мощность в каком-нибудь эксперименте – это вероятность, что будет правильно обнаружен реальный эффект.

Налицо тесная связь между размером  $\alpha$  и Р-значением Фишера. Если в качестве пороговой величины, при которой результаты считаются значимыми, взять число  $\alpha$ , то результаты, которые заставят нас отказаться от нулевой гипотезы, будут в точности теми же, для которых  $P < \alpha$ . Поэтому  $\alpha$  можно рассматривать как пороговый уровень значимости:  $\alpha = 0,05$  означает, что мы отвергнем нулевую гипотезу для всех Р-значений меньше 0,05.

Существуют формулы для размера и мощности при различных видах экспериментов, и каждая зависит от размера выборки. Но если у выборки фиксированный размер, то компромисс неизбежен: чтобы увеличить мощность, мы можем ослабить порог для «значимости» и тем самым с большей вероятностью идентифицировать истинный эффект, однако это означает увеличение вероятности ошибки первого рода (размера). В юридической аналогии мы можем ослабить критерии для осуждения (например, снизив требования для стандарта доказывания «вне разумных сомнений»), что в результате приведет к большему количеству правильно осужденных преступников, но, увы, и к большему количеству невинно осужденных.

Теория Неймана – Пирсона берет начало в процессах контроля качества на производстве, но в настоящее время широко используется при проверке новых методов лечения. Перед началом рандомизированного клинического исследования в протоколе указывается нулевая (лечение неэффективно) и альтернативная (как правило, эффект одновременно правдоподобен и важен) гипотезы. Затем исследователи устанавливают размер и мощность критерия, нередко выбирая  $\alpha = 0,05$  и  $1 - \beta = 0,80$ . Это означает, что для объявления результата значимым организаторы эксперимента требуют, чтобы Р-значение было меньше 0,05, и с 80-процентной вероятностью достигнут этого, если лечение на самом деле эффективно. Вместе эти

два числа позволяют оценить необходимое количество участников эксперимента.

Если исследователи хотят провести какое-то определяющее клиническое испытание, им нужно быть строже. Например, исследование по защите сердца пришло к выводу, что

если холестеринснижающая терапия за 5 лет сокращает смертность от сердечно-сосудистых заболеваний примерно на 25 %, а смертность от всех причин – на 15 %, то исследование именно такого размера имеет прекрасные шансы для демонстрации подобных эффектов на убедительных уровнях статистической значимости (то есть мощность  $> 90 \%$ ,  $p < 0,01$ ).

Другими словами, если истинный эффект лечения составляет 25-процентное снижение смертности от сердечно-сосудистых заболеваний и 15-процентное – от всех причин (альтернативные гипотезы), то исследование имеет мощность примерно 90 %, а размер  $\alpha = 1 \%$ . Такие требования диктуют размер выборки свыше 20 000. Фактически, как показывает [табл. 10.4](#), итоговые результаты дают 13-процентное уменьшение смертности от всех причин, что весьма близко к тому, что планировалось.

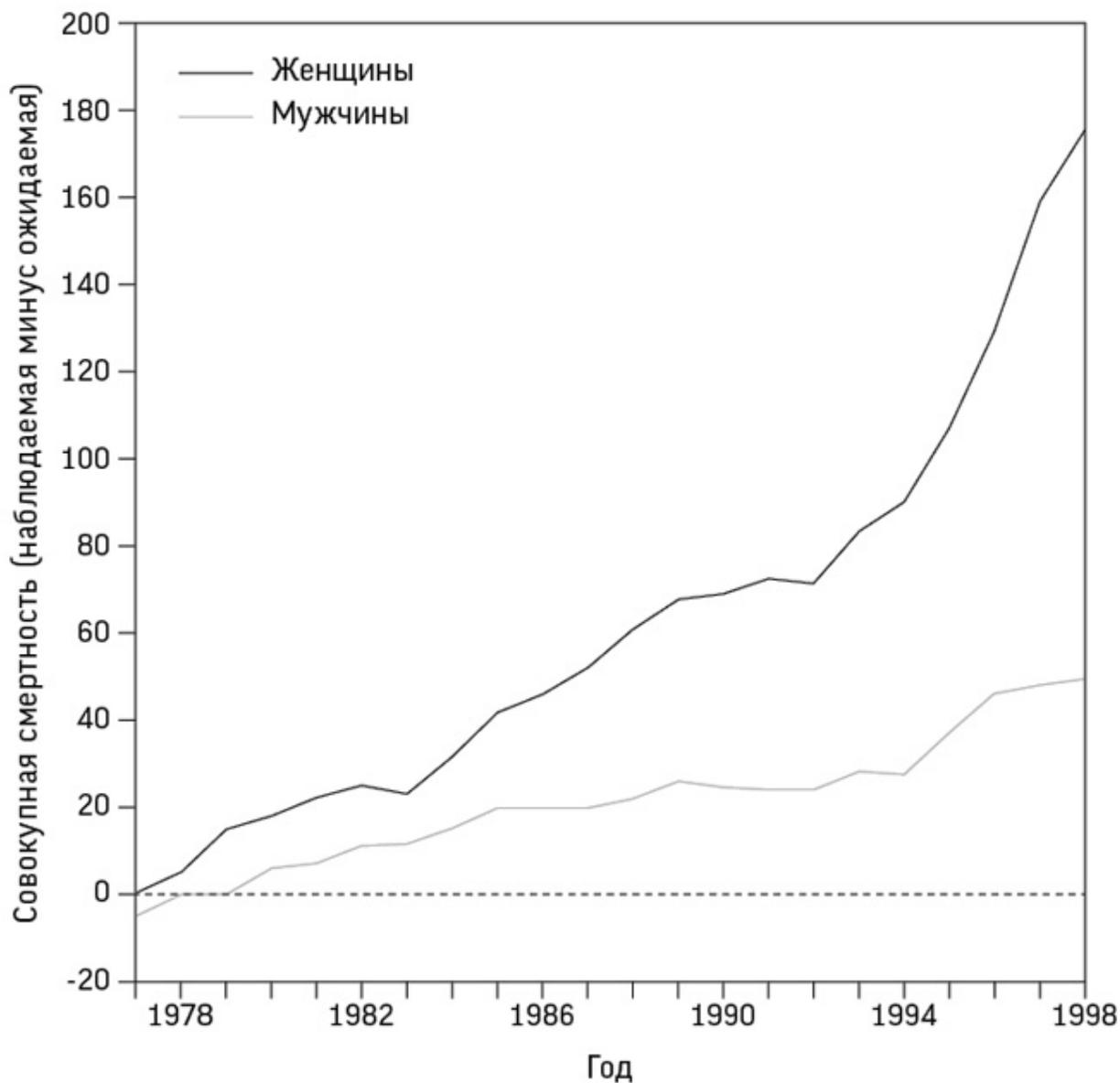
Идея о необходимости достаточно большой выборки для выявления правдоподобной альтернативной гипотезы прочно укоренилась при планировании медицинских испытаний. Однако исследования в психологии и нейробиологии часто используют размер выборки, определяемый удобством или традицией, и он может ограничиваться всего 20 объектами для каждого исследуемого состояния. В слишком маленьких исследованиях верные и интересные альтернативные гипотезы можно просто пропустить, поэтому сейчас наконец признали, что и в других областях исследований нужно задуматься о мощности проводимых экспериментов.

Как мы увидим в следующей главе, Нейман и Пирсон вели яростные, порой оскорбительные споры с Фишером по поводу правильного метода проверки гипотез, и этот конфликт так и не разрешился принятием какого-то единого «правильного» подхода. Как показывает исследование по защите сердца, клинические испытания, как правило, разрабатываются по теории Неймана – Пирсона, однако,

строго говоря, размер и мощность не имеют значения, когда эксперимент уже фактически проведен. В этот момент испытания анализируются с помощью доверительных интервалов, демонстрирующих правдоподобные значения для эффектов лечения, а фишеровские Р-значения показывают степень свидетельств против нулевой гипотезы. Таким образом, некая странная смесь идей Фишера и Неймана – Пирсона оказалась на удивление эффективной.

### **Можно ли было поймать Гарольда Шипмана раньше?**

Из [введения](#) мы узнали, что доктор Гарольд Шипман за двадцать лет убил более двухсот пациентов, прежде чем был разоблачен. Семьи его жертв, естественно, очень взволновало то, что ему удавалось так долго совершать преступления, не вызывая подозрений, поэтому последовавшее общественное расследование должно было установить, существовал ли шанс заподозрить его раньше. До начала расследования подсчитали количество свидетельств о смерти, подписанных Шипманом для его пациентов с 1977 года, а затем сравнили это число с тем, которого можно было бы ожидать, исходя из возраста всех пациентов Шипмана и уровней смертности у других врачей, практикующих в данном районе. При проведении таких сравнений учитываются местные условия, например изменения температуры воздуха или вспышки гриппа. На рис. 10.3 представлены результаты, полученные путем вычитания ожидаемого количества из наблюдаемого числа свидетельств о смерти, выданных Шипманом с 1977 года до своего ареста в 1998 году. Эту разницу можно называть его «избыточной» смертностью.



**Рис. 10.3**

Совокупное количество свидетельств о смерти, подписанных Шипманом для пациентов 65 лет и старше, с вычетом числа ожидаемых смертей (с учетом возраста пациентов)

К 1998 году его избыточная смертность для людей в возрасте 65 лет и старше составила 174 женщины и 49 мужчин. Это почти точное количество пожилых людей, которые в ходе расследования были признаны его жертвами, что показывает поразительную точность этого чисто статистического анализа, куда не входили сведения о конкретных случаях<sup>[203]</sup>.

Предположим, в какой-то вымышленной истории некто год за годом отслеживал смерти пациентов Шипмана и производил вычисления, необходимые для составления рис. 10.3. В какой момент ему следовало бить тревогу? Например, такой человек мог бы проводить проверку значимости в конце каждого года. У большого количества людей малая вероятность такого события, как смерть, поэтому можно считать, что количество смертей, подобно количеству убийств, имеет распределение Пуассона, а значит, нулевая гипотеза будет состоять в том, что совокупное число наблюдаемых смертей соответствует распределению Пуассона со средним значением, которое определяется числом ожидаемых смертей.

Если бы это было сделано с общим числом смертей для мужчин и женщин, которые показаны на [рис. 10.3](#), то уже в 1979 году, то есть всего через три года наблюдений, появилось бы одностороннее Р-значение 0,004, отражающее разницу между наблюдаемыми 40 смертями и ожидаемыми 25,3<sup>[204]</sup>. Результаты могли бы быть объявлены статистически значимыми, и Шипман был бы разоблачен.

Однако существуют две причины, по которым такая статистическая процедура была бы крайне неуместна для отслеживания уровня смертности пациентов у врачей общей практики. Во-первых, если у нас нет веской причины подозревать именно Шипмана и наблюдать только за ним, то нам придется вычислять Р-значения для всех семейных врачей в Соединенном Королевстве, а на тот момент их было около 25 тысяч. Но по примеру с мертвым лососем мы знаем, что при проведении достаточно большого количества проверок мы обязательно получим ложные сигналы. Если при пороге 0,05 протестировать 25 тысяч врачей, то каждый двадцатый из совершенно невинных докторов (то есть 1300 человек) покажет «статистически высокий уровень» при каждой проведенной проверке, а значит, в отношении него нужно вести расследование, что абсолютно абсурдно. А вот Шипман, наоборот, имел шанс потеряться среди всех этих ложноположительных случаев.

Альтернативой может считаться метод Бонферрони, то есть требование, чтобы Р-значение равнялось  $0,05/25\ 000$ , или 1 на 500 000. В этом случае Шипман был бы пойман в 1984 году, когда при ожидаемом количестве 59,2 у него было 105 смертей, то есть на 46 больше.

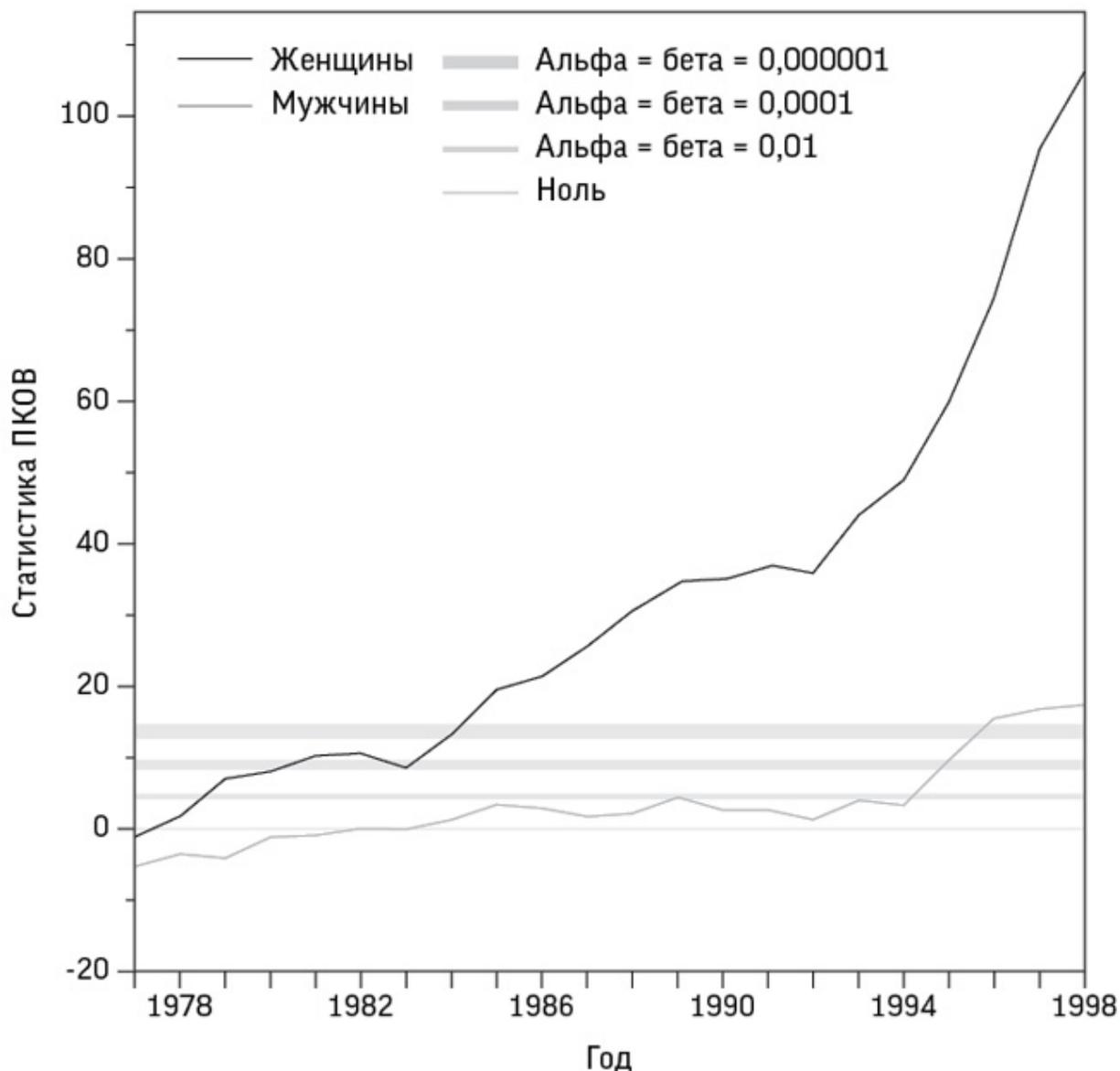
Но даже это не будет надежной процедурой для всех врачей в стране. Вторая проблема заключается в проведении повторных проверок на значимость, поскольку ежегодно добавляются новые данные и производится очередная проверка. Существует один замечательный, но сложный теоретический результат, именуемый очаровательным термином «закон повторного логарифма», который показывает, что, выполняя такое повторное тестирование, даже при справедливости нулевой гипотезы, мы *определенно* отвергнем ее при любом выбранном уровне значимости.

Это настораживает, потому что означает, что при долгосрочной проверке какого-нибудь доктора в итоге мы гарантированно посчитаем, что нашли доказательства избыточной смертности, хотя в реальности его пациенты не подвергаются никакому излишнему риску. К счастью, существуют статистические методы для решения проблемы **последовательного тестирования**, изначально разработанные во время Второй мировой войны группой статистиков, которая не имела ничего общего со здравоохранением, а трудилась над задачами контроля качества при производстве вооружения и других военных материалов.

Изделия, сходящие с производственной линии, проверяли на соответствие стандарту, а весь процесс контролировался посредством постепенно накапливавшегося общего количества отклонений от стандарта – ровно так же, как при отслеживании избыточной смертности. Ученые поняли, что из закона повторного логарифма следует, что повторное тестирование всегда приводит к предупреждению, что промышленный процесс вышел из-под контроля, даже если на самом деле все функционирует нормально. Независимо друг от друга статистики из США и Соединенного Королевства разработали метод, известный как последовательный критерий отношения вероятностей (ПКОВ), представляющий собой статистику, которая отслеживает накапливающуюся информацию об отклонениях и может быть в любой момент сопоставлена с простыми пороговыми значениями. Как только один из этих порогов преодолевается, срабатывает сигнал тревоги и производственная линия исследуется<sup>[205]</sup>. Эти методы позволили создать более эффективные промышленные процессы, а позже были адаптированы к использованию в так называемых последовательных клинических испытаниях, при которых

накапливаемые результаты регулярно отслеживаются, чтобы понять, не пересечен ли порог, указывающий на полезное лечение.

Я был одним из команды разработчиков варианта ПКОВ, который можно применить к данным о Шипмане. Рис. 10.4 отображает картину для мужчин и женщин в случае, когда в качестве альтернативной гипотезы выступает предположение, что смертность у Шипмана вдвое больше, чем у его коллег. У этого критерия есть пороговые значения, которые контролируют вероятность ошибки первого (альфа) и второго (бета) рода для значений 1 из 100, 1 из 10 000 и 1 из 1 000 000: ошибка первого рода – это общая вероятность того, что статистика пересечет пороговое значение в какой-нибудь точке, если бы у Шипмана был ожидаемый уровень смертности, а ошибка второго рода – общая вероятность статистики *не* пересечь пороговое значение в какой-нибудь точке, если бы у Шипмана был удвоенный ожидаемый уровень смертности<sup>[206]</sup>.



**Рис. 10.4**

Статистика для последовательного критерия отношения вероятностей (ПКОВ) для обнаружения удвоения риска смертности: пациенты в возрасте >64 лет, умершие дома / на приеме. Прямые линии отображают пороговые значения «сигнала тревоги» для показанных общих величин ошибок первого (альфа) и второго (бета) рода – предполагается, что они совпадают. Если смотреть на линию для женщин, видно, что Шипман пересек бы внешний порог в 1985 году

Поскольку насчитывается около 25 тысяч семейных врачей, разумно взять пороговое Р-значение 0,05/25 000, или 1 на 500 000. Для

одних только женщин Шипман преодолел бы более строгий порог  $\alpha = 0,000001$ , или 1 на миллион, в 1985 году, а суммарно для женщин и мужчин – в 1984-м. Поэтому последовательный критерий забил бы тревогу в тот же момент, что и примитивный повторяемый тест значимости.

Мы пришли к выводу, что если бы кто-то занимался таким отслеживанием и на Шипмана в 1984 году завели бы уголовное дело и осудили, то тем самым спасли бы примерно 175 жизней. И все исключительно с помощью применения простой процедуры статистического мониторинга.

После этого была запущена система наблюдения для врачей, которая немедленно выявила врача с еще более высоким уровнем смертности, чем у Шипмана! Расследование показало, что он работал в городке на южном побережье, где было много домов престарелых и множество стариков, и сознательно помогал многим пациентам оставаться дома до смерти, не настаивая на госпитализации в последние дни жизни. Было бы несправедливо осуждать этого доктора за выдачу большого числа свидетельств о смерти. Урок заключается в том, что, хотя статистические системы способны обнаружить выбросы, они не могут предложить их объяснения, поэтому нужно тщательно разбираться в каждом из них, чтобы избежать ложных обвинений. Еще одна причина быть осторожными с алгоритмами.

### **Что может быть не так с P-значениями?**

Рональд Фишер развил идею P-значения как меры совместимости данных с какой-то предварительно сформулированной гипотезой. Таким образом, если вы вычислите P-значение и оно будет маленьким, это означает, что если ваша гипотеза верна, а значение статистики получилось крайне большим или малым, то это маловероятно; стало быть, либо произошло нечто удивительное, либо ваша исходная гипотеза неверна. Такая логика довольно запутанна, но мы видели, насколько полезной может быть эта базовая идея. Так что же может пойти не так?

Оказывается, многое. Фишер описывал ситуации, как в первых примерах этой главы, – с одним набором данных, одной характеристикой результата и одной проверкой совместимости. Но за последние несколько десятилетий P-значения существенно

распространились в научной литературе – одно исследование насчитало 30 тысяч  $t$ -статистик и соответствующих  $P$ -значений всего лишь за три года публикации в восемнадцати журналах по психологии и нейробиологии<sup>[207]</sup>.

Итак, давайте посмотрим, что можно ожидать при, скажем, 1000 исследований, каждое с размером 5 % ( $\alpha$ ) и мощностью 80 % ( $1 - \beta$ ), хотя заметим, что на практике у большинства исследований мощность значительно ниже 80 %. Да, в реальном мире эксперименты проводятся в надежде сделать открытие, тем не менее нужно признать, что большинство нулевых гипотез верны (хотя бы приблизительно). Итак, предположим, что только 10 % проверенных нулевых гипотез на самом деле ложны: при испытаниях новых препаратов даже это число, вероятно, завышено – процент успехов здесь весьма низкий. Тогда, аналогично описанной в [главе 8](#) схеме, рис. 10.5 показывает, чего мы можем ожидать при 1000 исследований.



## Рис. 10.5

Ожидаемые количества для результатов 1000 проверок гипотез с размером 5 % (вероятность ошибки первого рода,  $\alpha$ ) и мощностью 80 % ( $1 - \beta$ , при вероятности ошибки второго рода  $\beta$ ). Только 10 % (100) нулевых гипотез ложны, и мы правильно обнаружим 80 % из них (80). Из 900 нулевых гипотез, которые истинны, мы неправильно отвергнем 5 % (45). В целом из 125 «открытий» ложными окажутся 36 % (45)

Получается, что можно ожидать заявления о 125 «открытиях», из которых 45 ложноположительные: иными словами, 36 % (больше трети) отклоненных нулевых гипотез («открытий») – это ложные утверждения. Столь мрачная картина усугубляется еще сильнее, если учесть, что на самом деле происходит в научной литературе, ориентированной на публикацию положительных результатов. После проведения аналогичного анализа Джон Иоаннидис, профессор школы медицины Стэнфордского университета, сделал в 2005 году свое знаменитое заявление, что «большинство публикуемых результатов исследований ложны»<sup>[208]</sup>. Мы вернемся к причинам его столь печального заключения в [главе 12](#).

Поскольку все эти ложные открытия основаны на Р-значениях, указывающих на «значимый» результат, в потоке неверных научных выводов все чаще стали винить именно их. В 2015 году один авторитетный журнал по психологии даже объявил, что запретит проверку значимости нулевой гипотезы. Наконец, в 2016 году Американской статистической ассоциации (ASA) удалось согласовать с группой статистиков шесть принципов, касающихся Р-значений.

Первый принцип просто описывает, что могут делать Р-значения.

1. Р-значения могут указывать на то, насколько несовместимы данные с конкретной статистической моделью.

Как мы не раз видели, Р-значения делают это, по сути, измеряя, насколько удивительны имеющиеся данные, при условии нулевой гипотезы, что чего-то не существует. Например, мы спрашиваем, насколько несовместимы данные с утверждением, что лекарство не работает? Такая логика может быть изощренной, но полезной.

Второй принцип помогает исправить ошибки в интерпретации P-значений.

2. P-значения не измеряют вероятность того, что изучаемая гипотеза верна или что данные получены исключительно по случайности.

В [главе 8](#) мы очень внимательно различали соответствующие утверждения об условных вероятностях, такие как «только 10 % женщин без рака молочной железы имеют положительную маммограмму» и (ложное) «только у 10 % женщин с положительной маммограммой нет рака молочной железы». Эта ошибка известна как «ошибка прокурора», и мы видели, что есть аккуратные способы ее исправить, представив, чего можно ожидать для 1000 тестируемых женщин.

Аналогичные проблемы могут возникать с P-значениями, которые измеряют вероятность *появления* таких экстремальных данных при условии, что нулевая гипотеза верна, но не измеряют вероятность того, что нулевая гипотеза верна, при *наличии* таких данных. Это тонкое, но существенное различие.

Когда команда ЦЕРН сообщила о результате «пять сигма» для бозона Хиггса, что соответствует P-значению примерно 1 на 3,5 миллиона, «Би-би-си» правильно это интерпретировала, сказав, что это означает «вероятность примерно 1 на 3,5 миллиона, что такой сигнал появился бы при отсутствии частицы Хиггса». Однако почти во всех остальных источниках это P-значение истолковали неверно. Например, журнал *Forbes* писал: «Шансы на то, что это не бозон Хиггса, составляют меньше одной миллионной» – яркий пример ошибки прокурора. Типичной реакцией был текст в газете *The Independent*: «Вероятность, что их результат – статистическая случайность, составляет меньше одной миллионной». Это, возможно, не так явно вводит в заблуждение, как у *Forbes*, но все равно приписывает малую вероятность тому, что «их результат – статистическая случайность», что, по логике, то же самое, что сказать о вероятности проверяемой нулевой гипотезы. Вот почему ASA пытается подчеркнуть, что P-значение – это *не* «вероятность того, что данные получены исключительно случайно».

Третий принцип ASA пытается противостоять одержимости статистической значимостью.

3. Научные заключения и процесс принятия решений не должны основываться только на том, переходит ли P-значение определенный порог.

Когда Рональд Фишер начал публиковать таблицы со значениями статистик, которые соответствовали результатам  $P < 0,05$  или  $P < 0,01$ , он вряд ли представлял, что такие довольно произвольные значения станут доминировать в научных публикациях, причем все результаты будут стремиться поделить на «значимые» и «незначимые». Отсюда уже недалеко и до того, чтобы расценивать «значимые» результаты как доказанные открытия, что создает крайне упрощенный и опасный прецедент перехода от данных прямо к заключениям – без паузы на размышление.

Губительное следствие такой дихотомии – неправильное толкование «незначимого». Незначимое P-значение подразумевает, что данные совместимы с нулевой гипотезой, но это не говорит о том, что нулевая гипотеза определенно верна. В конце концов, отсутствие прямых доказательств пребывания преступника на месте преступления еще не означает, что он невиновен. Но эта ошибка на удивление распространена.

Рассмотрим крупный научный спор о пользе алкоголя, скажем одной порции<sup>[209]</sup> в день. Как показало одно исследование, умеренное потребление алкоголя полезно только пожилым женщинам, однако тщательная проверка выявила и другие группы населения, извлекающие из него пользу, но она не была статистически значимой, поскольку доверительные интервалы вокруг оценки предполагаемой выгоды в этих группах были очень широкими. Хотя доверительные интервалы включали 0 (и потому эффект не был статистически значим), данные полностью согласовывались с высказанным ранее предположением о 10–20-процентном снижении риска смертности. Между тем The Times провозгласила, что «алкоголь не приносит никакой пользы здоровью»<sup>[210]</sup>.

Подытоживая, можно сказать, что крайне неправильно интерпретировать выражение «незначимо отличается от 0» как

означающее, что реальный эффект действительно *равен 0*, особенно в небольших исследованиях с низкой мощностью и широкими доверительными интервалами.

Четвертый принцип ASA звучит вполне безобидно.

4. Правильный вывод требует полной отчетности и прозрачности.

Самое очевидное – необходимо четко указать, сколько проверок фактически проведено. Если подчеркивается самый значимый результат, то можно сделать какую-то поправку (например, методом Бонферрони). Но проблемы с выборочной отчетностью могут быть намного тоньше, как мы увидим в следующей главе. Только зная план исследования и то, что было на самом деле сделано, можно избежать проблем с Р-значениями.

Вы запланировали исследование, собрали данные, провели анализ и получили «значимый» результат. Обязательно ли это должно быть важным открытием? Пятый принцип ASA просит вас быть не слишком самонадеянным.

5. Р-значение или статистическая значимость не измеряет величину эффекта или важность результата.

Наш следующий пример показывает, что (особенно в случае больших выборок) мы можем быть достаточно уверены в наличии связи, но при этом не сильно впечатляться ее важностью.

### **Почему поступление в университет повышает риск развития опухоли мозга?**

Мы рассматривали этот вопрос в [главе 4](#). Сделав поправку в регрессионном анализе на семейное положение и уровень дохода, шведские ученые обнаружили относительное повышение риска на 19 % между низким (начальная школа) и более высоким (университетский диплом) уровнем образования, с 95-процентным доверительным интервалом от 7 до 33 %. Интересно, что в работе не указывалось никаких Р-значений, однако в силу того что 95-процентный интервал для относительного риска не включает 1, можно заключить, что  $P < 0,05$ .

К этому моменту читатель уже должен иметь наготове список потенциальных вопросов к такому выводу, однако авторы упредили их, обнародовав одновременно с полученными результатами следующее:

- вывод о причинно-следственной связи невозможен;
- никакие поправки относительно факторов, потенциально влияющих на образ жизни (например, потребление алкоголя), не вносились;
- люди с более высоким экономическим статусом обычно чаще обращаются за медицинской помощью, поэтому может возникнуть так называемая предвзятость отчетности.

Но одна важная характеристика все же не была упомянута: небольшая величина видимой связи. Повышение риска на 19 % между низким и высоким уровнем образования гораздо ниже, чем для многих видов рака. В статье сообщалось, что в группе из более чем 2 миллионов человек старше 18 лет было диагностировано 3715 опухолей головного мозга (примерно 1 на 600). Следовательно, если мы, как в [главе 1](#), переведем относительные риски в абсолютные, то получим такие расчеты:

- мы можем ожидать, что среди примерно 3 тысяч человек с минимальным уровнем образования будет диагностировано пять опухолей (если базовый риск 1 на 600);
- а среди примерно 3 тысяч человек с максимальным уровнем образования – шесть опухолей (относительное увеличение на 19 %).

Такой расчет формирует несколько иное восприятие результатов и весьма обнадеживает. Столь небольшой повышенный риск развития редкого вида рака может оказаться статистически значимым только при изучении огромного количества людей: в нашем случае – свыше двух миллионов.

Поэтому из этого научного исследования можно извлечь два важных урока:

- «большие данные» способны легко привести к статистически значимым, но не имеющим **практической значимости** результатам;

- не следует беспокоиться, что учеба в вузе приведет к развитию опухоли головного мозга.

Последний принцип ASA довольно тонкий.

6. Само по себе Р-значение не дает надежного подтверждения модели или гипотезы. Например, Р-значение, близкое к 0,05, взятое само по себе, предлагает лишь слабое свидетельство против нулевой гипотезы.

Это утверждение, частично основанное на «байесовской» аргументации, описанной в следующей главе, побудило группу известных статистиков выдвинуть предложение изменить стандартное пороговое значение для «открытия» нового эффекта на  $P < 0,005$ <sup>[211]</sup>.

Какой эффект это может дать? Если на [рис. 10.5](#) мы изменим уровень значимости с 0,05 (1 на 20) на 0,005 (1 на 200), то вместо 45 ложноположительных «открытий» получим только 4,5. Это уменьшит общее количество открытий до 84,5, и всего лишь 4,5 из них (5 %) будут ложными. Выглядит как значительное улучшение по сравнению с 36 %, не так ли?

Исходная идея Фишера для проверки гипотез оказалось очень полезной для практического использования и предотвращения необоснованных научных заявлений. Однако статистики часто жаловались на готовность некоторых исследователей перейти от Р-значений, полученных в плохо спланированных экспериментах, к уверенным обобщающим выводам: своего рода алхимия для превращения неопределенности в определенность, механически применяющая статистические критерии к разделению результатов на «значимые» и «незначимые». В главе 12 мы рассмотрим некоторые из печальных последствий такого поведения, но сначала обратимся к альтернативному подходу к статистическим выводам, который полностью отвергает саму идею проверки значимости нулевой гипотезы.

Итак, еще одно расширяющее кругозор требование статистической науки: будет полезно, если вы сможете (временно) забыть все, что узнали из этой и предыдущих глав.

## **Выводы**

- Проверки нулевых гипотез – предположений о статистических моделях – составляют основную часть статистической практики.

- Р-значение – это мера несовместимости между наблюдаемыми данными и нулевой гипотезой: формально это вероятность наблюдения в эксперименте настолько же экстремального результата, если нулевая гипотеза верна.

- Традиционно для утверждений о статистической значимости используются пороговые значения 0,05 и 0,01.

- При проведении множественных проверок (например, по различным подмножествам данных или различным характеристикам) такие пороги требуют корректировки.

- Существует точное соответствие между доверительными интервалами и Р-значениями: например, если 95-процентный интервал не включает 0, то мы можем отвергнуть нулевую гипотезу о 0 при  $P < 0,05$ .

- Теория Неймана – Пирсона определяет альтернативную гипотезу и фиксирует вероятности ошибок первого и второго рода для двух возможных типов ошибок при проверке гипотезы.

- Для последовательного анализа разработаны отдельные формы проверки гипотез.

- Р-значения часто интерпретируются неправильно, в частности они не выражают вероятность того, что нулевая гипотеза верна, равно как и незначимый результат не означает, что нулевая гипотеза верна.

## Глава 11. Учимся на опыте – байесовский путь

*Я совсем не уверен, что «доверие» – это не злоупотребление доверием.*

*Артур Боули[137], 1934 год*

[\[212\]](#)

Сначала я должен сделать признание от имени всего статистического сообщества. Формальная основа для обучения на данных несколько запутанна. Несмотря на многочисленные попытки создать единую теорию статистических выводов, ни одна версия так и не была полностью принята. Неудивительно, что математики не любят преподавать статистику.

Мы уже познакомились с конкурирующими идеями Фишера и Неймана – Пирсона. Пришло время исследовать третий, байесовский подход к работе. Хотя он получил известность только в последнее пятидесятилетие, его базовые принципы восходят к далекому прошлому, фактически к преподобному Томасу Байесу, пресвитерианскому священнику и математику из Танбридж-Уэллса, занимавшемуся философией и теорией вероятностей<sup>[213]</sup>.

Хорошая новость состоит в том, что байесовский подход открывает новые возможности для создания сложных данных. Плохая – он означает, что вам придется отложить в сторону почти все, что вы узнали из этой и других книг об оценивании, доверительных интервалах, Р-значениях, проверке гипотез и так далее.

### *В чем суть байесовского подхода?*

Первым крупным вкладом Томаса Байеса в науку было использование вероятности как выражения недостатка наших знаний о мире или, что одно и то же, нашего незнания о происходящем в данный момент. Он показал, что вероятность может использоваться не только

для будущих событий, подверженных случайности, – стохастической неопределенности, если пользоваться термином, введенным в главе 8, но и для реальных событий, хорошо известных некоторым людям, просто мы этого пока не знаем, то есть для эпистемической неопределенности.

Если задуматься, то мы окружены эпистемической неопределенностью в отношении вещей, которые определены, но нам пока неизвестны. Игроки ставят на следующую карту, мы покупаем билеты мгновенной лотереи, обсуждаем пол будущего ребенка, ломаем голову над детективом, спорим о количестве тигров, оставшихся в дикой природе, и получаем оценки возможного числа мигрантов или безработных. Все это объективно существующие факты или числа, просто мы их не знаем. Снова подчеркну, что с байесовской точки зрения для представления нашего личного незнания этих фактов и чисел удобно использовать вероятности. Мы можем даже подумать о присвоении вероятностей альтернативным научным теориям, но этот вопрос более спорный.

Конечно, эти вероятности будут зависеть от наших нынешних знаний: вспомните пример из главы 8, где вероятность выпадения орла или решки зависит от того, посмотрели мы на монету или нет. Байесовские вероятности с необходимостью субъективны – они зависят от наших отношений с окружающим миром, а не являются свойствами самого мира. Такие вероятности должны меняться по мере получения нами новой информации.

Это приводит нас ко второму крупному вкладу Байеса – результату, который позволяет постоянно пересматривать текущие вероятности в свете новых доказательств. Он известен как **теорема Байеса** и фактически предоставляет формальный механизм обучения на опыте – блестящее достижение для малоизвестного священника из маленького английского курортного городка<sup>[214]</sup>.

Наследие Байеса обеспечивает фундаментальное понимание того, что данные не говорят сами за себя – центральную роль здесь играет наше внешнее знание и наши суждения. Это может показаться несовместимым с научным процессом, тем не менее наши фоновые знания и понимание всегда были частью извлечения информации из данных, разница лишь в том, что в байесовском подходе они обрабатываются формальным математическим образом.

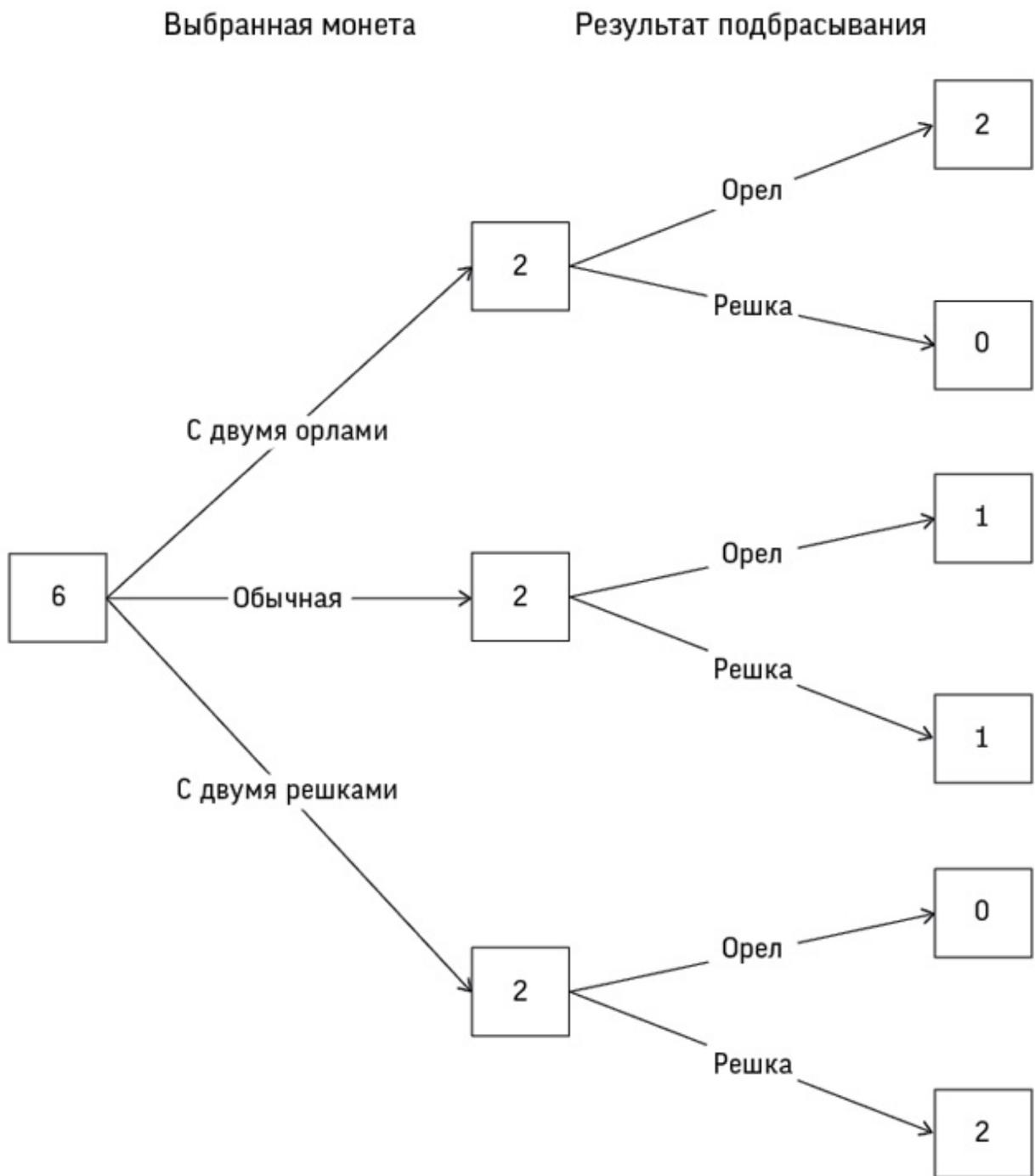
О выводах из работы Байеса рьяно спорили многие статистики и философы, возражающие против идеи, что субъективное суждение играет в статистике какую-либо роль. Поэтому будет справедливо, если я проясню собственную позицию: меня познакомили с субъективистской байесовской школой статистических рассуждений в начале моей карьеры<sup>[215]</sup>, и она до сих пор кажется мне наиболее удовлетворительным подходом.

**У вас в кармане три монеты: на одной два орла, на другой две решки, третья обычная. Вы наугад вытаскиваете монету, подбрасываете ее, и выпадает орел. Какова вероятность, что на другой стороне монеты тоже орел?**

Это классическая задача с эпистемической неопределенностью: как только монета падает после подбрасывания, никакой случайности не остается и любое высказывание о вероятности – всего лишь выражение вашего нынешнего личного незнания о другой стороне монеты.

Многие бы решили, что ответ –  $1/2$ , поскольку монета либо обычная, либо с двумя орлами, и вероятность выбрать одну из них одинакова. Существует много способов это проверить, но проще всего использовать идею с ожидаемыми количествами, описанную в [главе 8](#).

На рис. 11.1 показано, чего можно ожидать, если проделать такой эксперимент шесть раз. В среднем каждая монета будет выбрана дважды, и каждая из сторон выпадет по разу. Орел выпадает в трех случаях, причем в двух на второй стороне также будет орел. Поэтому вероятность того, что на второй стороне монеты тоже орел, равна  $2/3$ , а не  $1/2$ . По сути, выпадение орла повышает вероятность выбора монеты с двумя орлами, ведь у такой монеты есть два варианта упасть орлом вверх, а у симметричной – только один.



**Рис. 11.1**

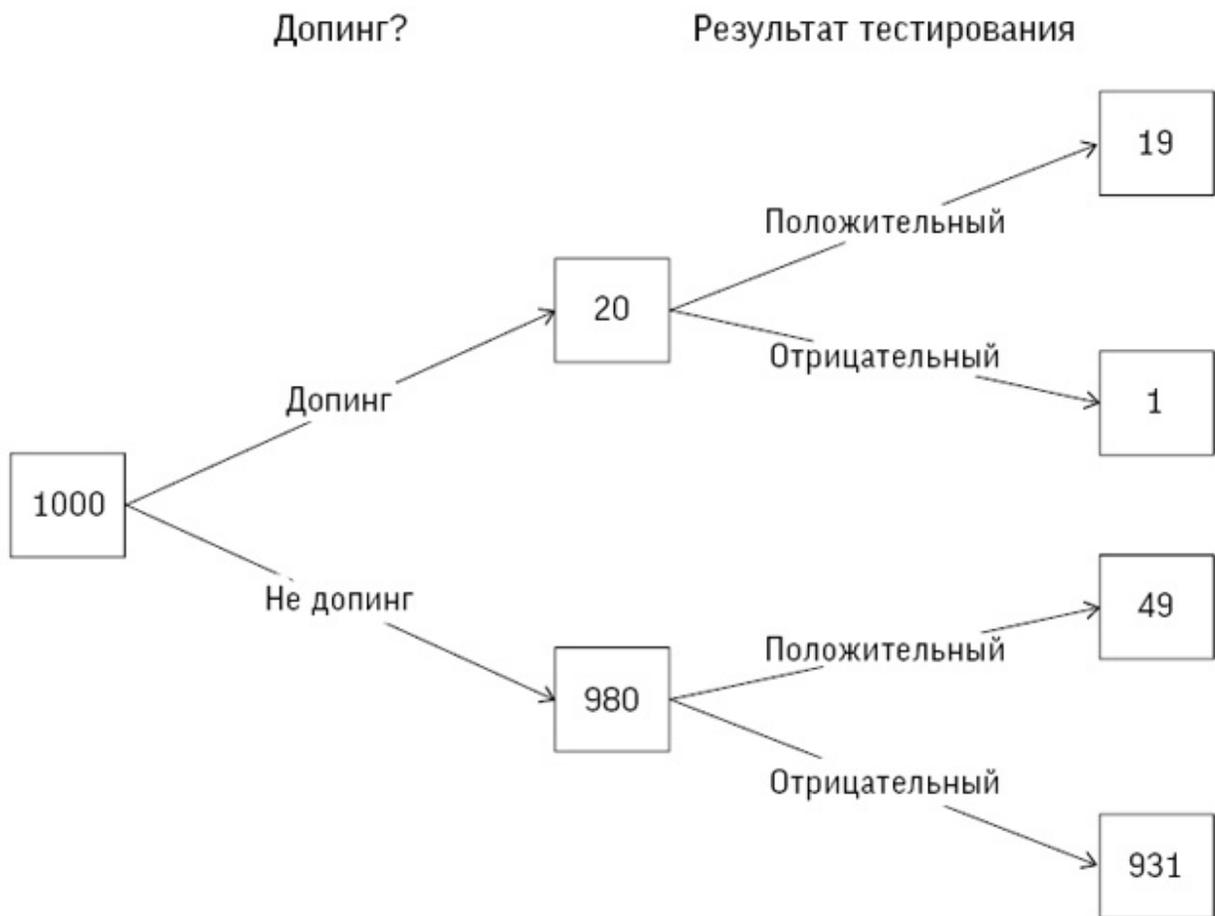
Дерево ожидаемых количеств для задачи с тремя монетами, показывающее, чего можно ожидать в случае шести экспериментов

Если этот результат не кажется вам интуитивно понятным, то следующий пример удивит вас еще больше.

**Предположим, что точность некой проверки на допинг в спорте – 95 %, то есть правильно будут определены 95 % тех, кто принимает допинг, и 95 % тех, кто не принимает. Допустим, что 1 из 50 атлетов действительно принимает допинг. Если тест спортсмена показал положительный результат, то какова вероятность, что он точно допингист?**

Этот тип потенциально сложной задачи опять же лучше всего решать с помощью ожидаемых количеств, аналогично проверке женщин на рак молочной железы из главы 8 и ситуации с высокой долей неверных результатов в научных публикациях из главы 10.

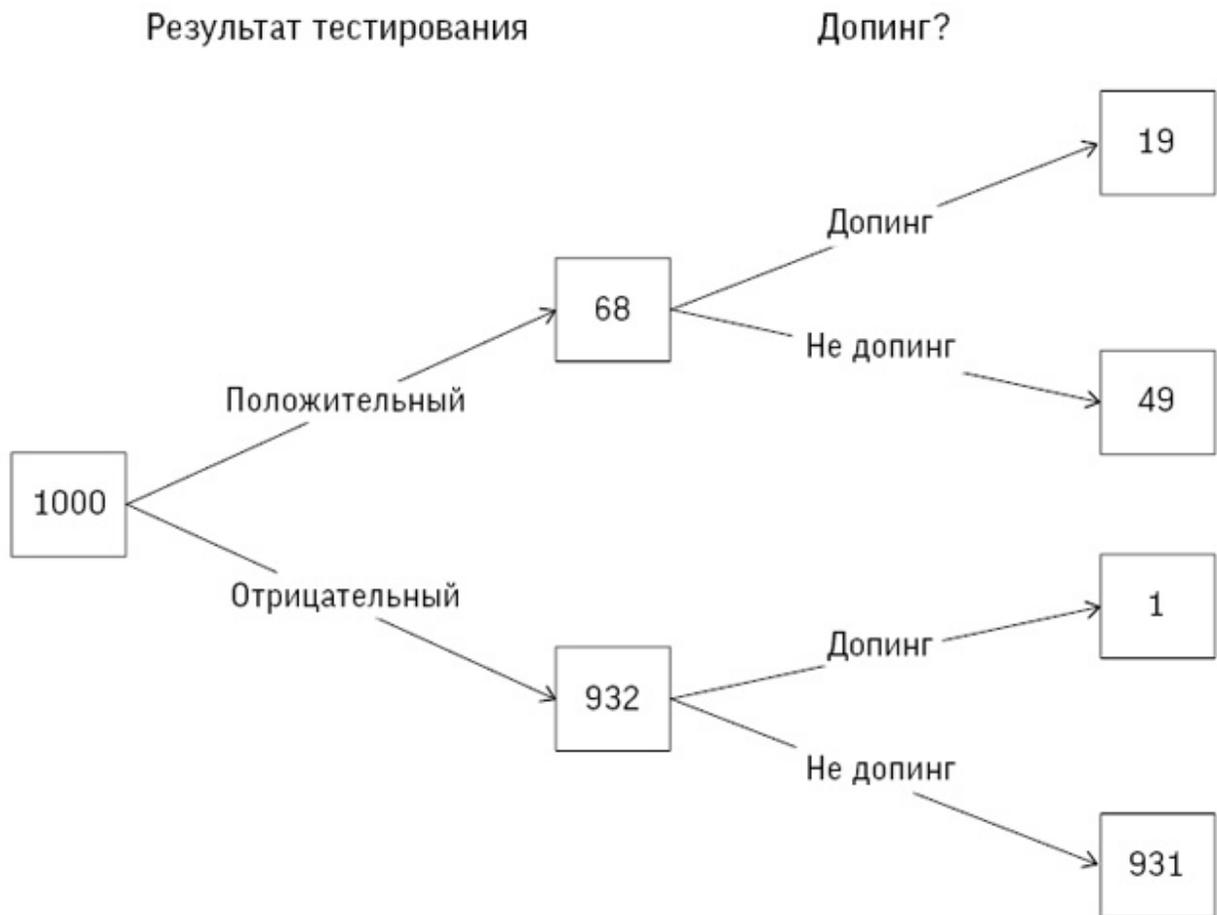
Дерево на рис. 11.2 начинается с 1000 спортсменов, из которых 20 употребляли допинг, а 980 нет. Все допингисты, кроме одного, выявлены (95 % от 20 = 19), однако положительные тесты также оказались у 49 атлетов, не употреблявших допинг (95 % от 980 = 931). Следовательно, в общей сложности мы можем ожидать  $19 + 49 = 68$  положительных тестов, из которых только 19 действительно отражают допинг. Поэтому вероятность, что атлет с положительным допинг-тестом истинный допингист, составляет всего  $19/68 = 28 \%$ , а оставшиеся 72 % будут ложными обвинениями. Итак, хотя объявлено, что точность тестирования на допинг 95 %, большинство людей с положительным допинг-тестом на самом деле будут невиновными. Нетрудно представить все проблемы, которые этот парадокс может вызвать в реальной жизни, когда спортсменов незаслуженно клеймят за проваленный допинг-тест.



**Рис. 11.2**

Дерево ожидаемых количеств для задачи о допинге, показывающее, чего можно ожидать при проверке 1000 спортсменов, когда допинг принимает 1 из 50, а «точность» тестирования составляет 95 %

Один из способов осмыслить этот процесс – «поменять порядок» в дереве, сначала поставив тестирование, а затем раскрыв истину. Это показано на рис. 11.3.



**Рис. 11.3**

«Обращенное» дерево ожидаемых количеств для задачи о допинге, перестроенное так, чтобы сначала шли результаты тестов, а затем истинное положение вещей

Это «обращенное» дерево дает в точности те же числа, но учитывает временной порядок, в котором мы получаем информацию (тестирование → допинг), а не порядок по фактической временной шкале (допинг → тестирование). Это «обращение» как раз и есть тем, что делает теорема Байеса; на самом деле байесовское мышление до 1950-х именовалось «обратной вероятностью».

Пример со спортивным допингом показывает, насколько легко спутать вероятность наличия допинга при условии положительного теста (28 %) с вероятностью положительного теста при условии наличия допинга (95 %). Мы уже сталкивались со случаями, когда вероятность события А при условии, что произошло событие В, путали с вероятностью события В при условии, что произошло событие А:

- неправильная интерпретация Р-значений, когда вероятность какого-то факта при условии нулевой гипотезы смешивается с вероятностью нулевой гипотезы при условии этого факта;
- ошибка прокурора в судебных разбирательствах, когда вероятность факта при условии невиновности путается с вероятностью невиновности при условии такого факта.

Разумный наблюдатель может подумать, что формальное байесовское мышление внесло бы ясность и строгость в работу с доказательствами в судебных разбирательствах, а потому точно удивится, узнав, что британские суды фактически запрещают теорему Байеса. Прежде чем объяснить, почему, нам нужно рассмотреть статистическую величину, которая в суде разрешена, – **отношение правдоподобия**.

### ***Отношение шансов и отношение правдоподобия***

Пример с допингом демонстрирует логические шаги, позволяющие добраться до той величины, которая действительно важна при принятии решения: *среди спортсменов с положительным тестом доля реальных допингистов 19/68*. Дерево ожидаемых количеств показывает, что эта величина зависит от трех ключевых чисел: доли атлетов, принимающих допинг (1/50, или 20 человек из 1000 в нашем дереве), доли допингистов, которые тест определяет правильно (95 %, или 19/20 в дереве), и доли честных атлетов с ложноположительным результатом теста (5 %, или 49/980 в дереве).

С помощью дерева ожидаемых количеств анализ становится вполне интуитивно понятным, хотя теорема Байеса также предоставляет удобную формулу для выражения в вероятностях. Но сначала мы должны вернуться к идее шансов, введенной в [главе 1](#), хотя опытные игроки, по крайней мере в Британии, прекрасно с ней знакомы. Шансы на какое-то событие – это вероятность того, что оно произойдет, деленная на вероятность того, что оно не произойдет. Например, если мы бросаем игральную кость, то шансы на выпадение шестерки – 1 к 5. На самом деле вероятность выпадения шестерки равна 1/6, а вероятность выпадения нешестерки – 5/6; поэтому шансы на выпадение шестерки равны  $1/6 : 5/6 = 1/5$ <sup>[216]</sup> (обычно именуется

«один к пяти» или «пять против одного», если вы используете британский метод выражения шансов в азартных играх).

Теперь нам нужно ввести идею отношения правдоподобия – понятия, ставшего критически важным при выражении силы судебных доказательств в уголовном судопроизводстве. Судьи и адвокаты постепенно учатся понимать отношения правдоподобия, которые фактически сравнивают относительную поддержку, предоставляемую неким фактом (свидетельством, доказательством) для двух конкурирующих гипотез, назовем их А и В, часто означающих виновность и невиновность. С технической точки зрения отношение правдоподобия – это вероятность факта при условии гипотезы А, деленная на вероятность факта при условии гипотезы В.

Давайте посмотрим, как это работает в случае с пробой на допинг, когда судебный факт – это положительный результат теста, гипотеза А – спортсмен принимал допинг, а гипотеза В – не принимал. Мы приняли, что 95 % допингистов сдают положительный тест, значит, вероятность такого факта при условии гипотезы А равна 0,95. Мы знаем, что 5 % атлетов, не принимающих допинг, сдают положительный тест, поэтому вероятность такого факта при условии гипотезы В составляет 0,05. Следовательно, отношение правдоподобия равно  $0,95/0,05 = 19$ , то есть положительный результат теста в 19 раз более вероятен, когда спортсмен виновен, чем когда невиновен. На первый взгляд это может показаться довольно веским доказательством, но вскоре мы увидим отношения правдоподобия, составляющие миллионы и миллиарды.

Давайте соединим все это в теореме Байеса, которая просто говорит, что

начальные шансы какой-то гипотезы × отношение правдоподобия = конечные шансы для этой гипотезы<sup>[217]</sup>.

В примере с допингом начальные шансы гипотезы «спортсмен принимал допинг» составляют 1 к 49, а отношение правдоподобия равно 19, поэтому теорема Байеса гласит, что конечные шансы равны

$$1/49 \times 19 = 19/49.$$

Шансы 19 к 49 можно преобразовать в вероятность  $19/(19 + 49) = 28 \%$ . Таким образом, вероятность, которую мы получили более простым путем из дерева ожидаемых количеств, можно вывести и из теоремы Байеса.

На более формальном языке начальные шансы именуются *априорными*, а конечные – *апостериорными*<sup>[218]</sup>. Формулу можно применить еще раз, и тогда апостериорные шансы после первого факта станут априорными перед учетом второго, независимого, факта. При объединении всех этапов процесс эквивалентен умножению независимых отношений правдоподобия и формированию общего составного отношения правдоподобия.

Теорема Байеса выглядит обманчиво простой, но на самом деле в ней заключен чрезвычайно мощный способ получения информации из данных.

### ***Отношения правдоподобия и судебная экспертиза***

**В субботу 25 августа 2012 года археологи начали раскопки под автостоянкой в Лестере в поисках останков Ричарда III. Через несколько часов был найден первый скелет. Какова вероятность, что он принадлежит Ричарду III?**

Согласно популярному мнению, распространению которого активно способствовал сторонник Тюдоров Уильям Шекспир, Ричард III (последний король из династии Йорков) был злобным горбуном. И хотя это довольно спорная (впоследствии опровергнутая) точка зрения, доподлинно известно, что он был убит в битве при Босворте 22 августа 1485 года в возрасте 32 лет, что фактически положило конец войне Алой и Белой розы. Тело его было после смерти изувечено и захоронено в монастыре Грейфрайерс в Лестере, который впоследствии был разрушен, а через какое-то время на его месте построили автостоянку.

С учетом только предоставленной информации мы можем предположить, что скелет принадлежит Ричарду III, если истинны все нижеперечисленные условия:

- он действительно похоронен в монастыре Грейфрайерс;

- его тело не было выкопано, перемещено или уничтожено за следующие 527 лет;
- первый обнаруженный скелет действительно принадлежит Ричарду.

Предположим с изрядной долей пессимизма, что история о его похоронах правдива с вероятностью 50 % и что вероятность того, что скелет по-прежнему находится в месте захоронения в Грейфрайерсе, тоже 50 %. Представьте, что в указанном месте похоронено еще 100 тел (археологи хорошо знали, где копать, поскольку сообщалось, что Ричард похоронен в хоре монастыря). Тогда вероятность того, что все вышеуказанные события подлинны, равна  $1/2 \times 1/2 \times 1/100 = 1/400$ . Это довольно низкая вероятность того, что найденный скелет – Ричард III; исследователи, которые первоначально проводили анализ, приняли «скептическую» априорную вероятность равной 1/40, так что мы гораздо скептичнее<sup>[219]</sup>.

Однако детально исследовав скелет, археологи получили несколько примечательных результатов криминалистических экспертиз: 1) данные радиоуглеродного анализа костей (который дал 95-процентную вероятность, что они датируются периодом с 1456 по 1536 год); 2) подтверждение, что это был мужчина в возрасте около 30 лет с признаками сколиоза (искривления позвоночника); 3) доказательства того, что тело было изувечено после смерти. Генетический анализ с участием известных потомков его близких родственников (сам Ричард детей не имел) показал общую митохондриальную ДНК (через его мать). Связь по мужской Y-хромосоме не подтвердилась, но это можно легко объяснить разрывом в мужской линии из-за неправильно определенного отцовства.

Ценность каждого из фактов-доказательств можно выразить через их отношения правдоподобия, которые в данном случае определяются как

отношение правдоподобия = вероятность факта при  
условии, что это скелет Ричарда III / вероятность факта при  
условии, что это скелет НЕ Ричарда III.

В табл. 11.1 показаны отдельные отношения правдоподобия для каждого из фактов-доказательств, при этом исследователи были

осторожны и намеренно занижали оценки в сторону наименьших отношений правдоподобия, то есть не в пользу того, что это скелет Ричарда III. Но если мы предположим независимость всех результатов, это даст нам право перемножить все эти отношения и получить общую оценку силы всех фактов-доказательств: значение достигнет 6,5 миллиона, что означает «крайне сильное подтверждение». Словесные формулировки, приведенные в табл. 11.1, взяты из шкалы, рекомендованной для использования в суде (см. табл. 11.2)<sup>[220]</sup>.

### **Таблица 11.1**

Отношения правдоподобия для отдельных фактов-доказательств в отношении скелета, найденного в Лестере. Сравняются гипотезы, это скелет Ричарда III или нет. Объединенное отношение правдоподобия получается путем перемножения отдельных отношений правдоподобия и достигает 6,5 миллиона

<b>Доказательство</b>	<b>Отношение правдоподобия (консервативная оценка)</b>	<b>Словесная формулировка</b>
Радиоуглеродное датирование (1456–1530)	1,8	Слабое подтверждение
Возраст и пол	5,3	Слабое подтверждение
Сколиоз	212	Умеренно сильное подтверждение
Раны после смерти	42	Умеренное подтверждение
Совпадение митохондриальной ДНК	478	Умеренно сильное подтверждение
Несовпадение Y-хромосомы	0,16	Слабое свидетельство против
Объединенное подтверждение	6,5 миллиона	Более чем чрезвычайно сильное подтверждение

**Таблица 11.2**

Рекомендуемые словесные интерпретации для отношений правдоподобия при предоставлении результатов криминалистической

## экспертизы в суде

<b>Значение отношения правдоподобия</b>	<b>Словесная формулировка</b>
1–10	Слабое подтверждение
10–100	Умеренное подтверждение
100–1000	Умеренно сильное подтверждение
1000–10 000	Сильное подтверждение
10 000–100 000	Очень сильное подтверждение
100 000–1 000 000	Чрезвычайно сильное подтверждение

Насколько убедительны эти доказательства? Вспомните, что, прежде чем перейти к вычислениям отношений правдоподобия, мы сделали консервативную оценку  $1/400$ , что это скелет Ричарда III. Это соответствует примерным начальным шансам 1 к 400. Тогда по теореме Байеса мы получаем для апостериорных шансов число  $6,7$  миллиона /  $400 = 16\,750$ . Таким образом, даже будучи предельно осторожными с оценкой априорных шансов и отношения правдоподобия, мы можем сказать, что шансы на то, что это скелет короля Ричарда, составляют примерно 16 750 против 1.

Поскольку исследователи брали число 40, а не 400, то полученные ими шансы составили примерно 167 000 против 1, то есть они нашли Ричарда III с вероятностью 0,999994. Это было сочтено достаточным доказательством для торжественного перезахоронения скелета в соборе Лестера.

В судебных делах отношения правдоподобия обычно прилагаются к данным ДНК при обнаружении какой-то степени «совпадения» между ДНК подозреваемого и следами, найденными на месте преступления. Две конкурирующие гипотезы в этом случае таковы: следы ДНК оставил подозреваемый либо это сделал кто-то другой. Следовательно, отношение правдоподобия можно записать так:

$$\text{отношение правдоподобия} = \frac{\text{вероятность совпадения ДНК при условии, что следы оставил подозреваемый}}{\text{вероятность совпадения ДНК при условии, что следы оставил кто-то другой}}$$

Число в числителе обычно принимается равным 1, а в знаменателе считается вероятностью того, что случайно выбранный из совокупности человек обеспечит случайное совпадение ДНК, – это называется **вероятностью случайного совпадения**. Типичные отношения правдоподобия для подтверждений по ДНК могут составлять миллионы и миллиарды, хотя точные величины можно оспаривать, например в случае затруднений из-за наличия в следах ДНК нескольких разных людей.

В британских судах разрешены отдельные отношения правдоподобия, но их нельзя перемножать, как в случае с Ричардом III, поскольку считается, что процедура объединения отдельных доказательств возложена на жюри присяжных<sup>[221]</sup>. Юридическая система, по-видимому, еще не готова принять научную логику.

### **Жульничает ли архиепископ Кентерберийский при игре в покер?**

Мало кто знает, что известный экономист Джон Кейнс, изучая теорию вероятностей, придумал мысленный эксперимент, демонстрирующий важность учета начальных шансов при оценке последствий. В этом упражнении он просил представить, что вы

играете в покер с архиепископом Кентерберийским, который в первом круге сдает себе роял-флеш<sup>[222]</sup>. Следует ли нам подозревать его в жульничестве?

Отношение правдоподобия для этого события равно:

отношение правдоподобия = вероятность комбинации роял-флеш при условии, что архиепископ жульничает / вероятность комбинации роял-флеш при условии, что архиепископу просто повезло.

Будем считать, что числитель равен единице, а вероятность в знаменателе можно вычислить как  $1 / 72\,000$ <sup>[223]</sup>. Тогда отношение правдоподобия составит 72 000, что, согласно стандартам из [табл. 11.2](#), означает «очень сильное подтверждение», что архиепископ жульничает. Но должны ли мы делать этот вывод? Как говорит теорема Байеса, апостериорные шансы равны произведению отношения правдоподобия на априорные шансы. Кажется разумным предположить, что (по крайней мере, пока мы не начали играть) шансы на то, что архиепископ не жульничает, крайне высоки, возможно, миллион против 1, учитывая его высокий духовный сан<sup>[224]</sup>. Поэтому произведение таких шансов и отношения правдоподобия даст нам  $72\,000 / 1\,000\,000$ , то есть примерно 7 к 100, что соответствует вероятности 7/107, или 7 %, что он жульничает. Таким образом, на этом этапе мы можем себе позволить дать ему кредит доверия (чего не сделали бы по отношению к человеку, с которым, скажем, только что столкнулись в пабе). И, возможно, нам надо держать ухо востро во время игры с архиепископом!

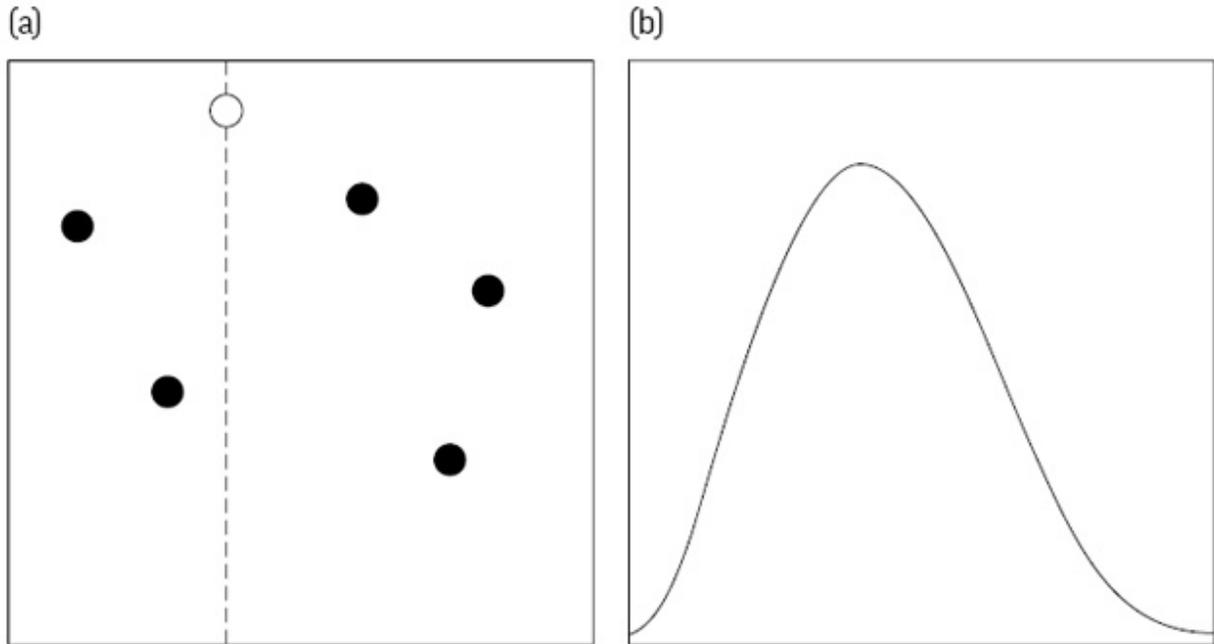
### ***Байесовские статистические выводы***

Теорема Байеса, даже если она и не разрешена в британских судах, – это научно корректный способ менять наше мнение на основании новых фактов. Ожидаемые количества делают байесовский анализ достаточно простым для несложных ситуаций, где есть всего две гипотезы, например, заболел человек или не заболел, совершил преступление или не совершил. Однако все усложняется, когда мы хотим применить эти же идеи к выводам относительно неизвестных величин, которые могут принимать целый диапазон значений, таких как параметры в статистических моделях.

Оригинальная работа преподобного Томаса Байеса, опубликованная в 1763 году, давала ответ на один очень простой вопрос: если известно, что нечто произошло или не произошло определенное количество раз, то какова вероятность, что это произойдет в следующий раз?<sup>[225]</sup> Например, если канцелярскую кнопку подбросили 20 раз и она 15 раз упала острием вверх, а 5 раз – острием вниз, то чему равна вероятность ее падения острием вверх в следующий раз? Возможно, вы подумаете, что ответ очевиден:  $15 / 20 = 75 \%$ . Однако ответ преподобного был бы другим –  $16 / 22$  (73 %). Как бы он к нему пришел?

Байес использовал метафору бильярдного стола<sup>[226]</sup>, который от вас скрыт. Предположим, на стол случайно брошен белый шар; его положение на столе отмечается линией, после чего белый шар убирают. Затем на стол случайным образом бросают несколько красных шаров, но вам сообщают только их число слева и справа от линии. Как думаете, где может проходить линия и чему, по-вашему, равна вероятность того, что следующий красный шар будет слева от линии?

Допустим, после того как было брошено пять красных шаров, вам сказали, что три шара приземлились слева от линии, где лежал белый шар, а три – справа, как на рис. 11.4(a). Байес показал, что наше представление о положении линии должно описываться вероятностным распределением, представленным на рис. 11.4(b), – математические рассуждения тут довольно сложные и приведены в примечании<sup>[227]</sup>. Оценка положения пунктирной линии, указывающей, куда упал белый шар, –  $3/7$  длины стола, что является средним (математическим ожиданием) для этого распределения.



**Рис. 11.4**

«Бильярдный» стол Байеса. (а) На стол бросают белый шар и его конечное положение отмечают пунктирной линией. Затем на стол бросают пять красных шаров – их положение обозначено темными точками. (б) Наблюдатель не видит стола, но ему говорят, что два красных шара приземлились слева от линии, а три – справа. Кривая отображает вероятностное распределение положения пунктирной линии (белого шара) для наблюдателя, наложенное на стол. Среднее значение кривой равно  $3/7$ , и это также текущая вероятность для наблюдателя, что следующий красный шар окажется слева от линии

Значение  $3/7$  может показаться странным, поскольку интуитивная оценка –  $2/5$  (доля красных шаров, оказавшихся слева от линии)<sup>[228]</sup>. Однако Байес показал, что в такой ситуации следует оценивать положение по формуле

количество красных шаров, лежащих слева, +1 / общее количество красных шаров +2.

Это, в частности, означает, что, перед тем как бросать красные шары, мы можем оценить положение белого шара как  $(0 + 1) / (0 + 2) = 1/2$ , в то время как интуитивный подход подсказывает, что нельзя дать никакого ответа, так как пока нет никаких данных. В

сущности, Байес использует информацию о том, как изначально была проведена линия, ведь мы знаем, что она определялась случайным броском белого шара. Эта первоначальная информация играет ту же роль, что и известная частотность случаев, используемая при маммографии или проверке на допинг, – она называется априорной информацией и влияет на наши окончательные выводы. Фактически, учитывая, что вышеприведенная формула добавляет один шар к числу красных шаров слева от линии и два шара к общему числу красных шаров, мы можем считать это эквивалентным тому, что вы уже бросили два «воображаемых» красных шара – по одному с каждой стороны от пунктирной линии.

Обратите внимание, что если ни один из пяти шаров не попадает слева от пунктирной линии, то мы оцениваем его положение не как  $0/5$ , а как  $1/7$ , что выглядит более осмысленно. Байесовская оценка не может быть 0 или 1, она всегда ближе к  $1/2$ , чем простая доля: при таком «сжатии» оценки всегда стягиваются к центру исходного распределения, в нашем случае к  $1/2$ .

Байесовский анализ берет знание о положении пунктирной линии, чтобы определить его **априорное распределение**, добавляет новые факты, используя понятие **правдоподобия**, и делает заключение об **апостериорном распределении**, выражающем наши текущие знания об этой неизвестной величине. Например, с помощью компьютера можно вычислить, что промежуток от 0,12 до 0,78 содержит 95 % вероятности на [рис. 11.4\(b\)](#), поэтому мы можем с 95-процентной уверенностью сказать, что линия, отмечающая положение белого шара, лежит между этими граничными значениями. Чем больше красных шаров будут бросать на стол и сообщать об их положении относительно пунктирной линии, тем уже будет такой доверительный интервал, постепенно сходясь к правильному ответу.

Основное расхождение в отношении байесовского анализа – источник априорного распределения. В примере со столом белый шар бросается наугад, поэтому любой согласится, что априорное распределение – это равномерное распределение от 0 до 1. Когда знание такого рода недоступно, предположения об априорном распределении приходится делать с помощью субъективных суждений, исторических данных или определения **объективного априорного распределения**,

чтобы данные могли говорить сами за себя без добавления субъективных суждений.

Пожалуй, в этом заключена самая важная идея – что не существует никакого «истинного» априорного распределения и любой анализ должен включать анализ чувствительности к ряду альтернативных гипотез, охватывающих целый ряд возможных мнений.

### **Как лучше анализировать предвыборные опросы?**

Мы видели, как байесовский анализ обеспечивает формальный механизм использования имеющихся знаний для более реалистичных выводов о конкретной, стоящей перед нами задаче. Эти идеи можно (буквально) перенести на другой уровень, поскольку многоуровневое, или **иерархическое, моделирование** одновременно анализирует различные отдельные величины: мощность таких моделей отражена в успехах предвыборных опросов.

Мы знаем, что в идеале опросы должны основываться на больших случайных репрезентативных выборках, однако их формирование обходится все дороже, а люди все чаще отказываются участвовать в опросах. Поэтому сегодня компании, занимающиеся опросами, по большей части полагаются на онлайн-панели<sup>[229]</sup>. Поскольку, как известно, они не являются репрезентативными группами, впоследствии используется сложное статистическое моделирование, которое выясняет, какими могли бы быть ответы, если бы компании обеспечили надлежащую случайную выборку. Здесь на ум может прийти старое предупреждение о невозможности сделать шелковый кошелек из свиного уха<sup>[230]</sup>.

Ситуация усугубляется еще больше, когда дело доходит до предвыборных опросов, поскольку политические взгляды по стране распределяются неравномерно и заявления об общей картине на национальном уровне нужно делать на основе объединения результатов по многим различным штатам или избирательным округам. В идеале выводы следует делать на местном уровне, однако люди в онлайн-панели сильно неслучайным образом разбросаны по этим локальным областям, а значит, для такого локального анализа имеется весьма ограниченный объем данных.

Байесовский ответ на эту проблему – **многоуровневая регрессия и постстратификация (MRP)**. Основная идея – разбить всех

потенциальных избирателей на маленькие «ячейки», состоящие из однородной группы людей, например жителей одной области, людей одного возраста, пола, сходных политических взглядов и прочих измеримых характеристик. Для оценки числа людей в каждой ячейке можно использовать имеющиеся демографические данные; предполагается, что все ее члены голосуют за определенную партию с равной вероятностью. Проблема в том, чтобы выяснить, какова эта вероятность, когда наши неслучайные данные могут означать, что у нас в конкретной ячейке всего несколько человек, а возможно, и ни одного.

Первый шаг – построение регрессионной модели для вероятности голосования определенным образом при данных характеристиках ячейки, поэтому наша задача сводится к оцениванию коэффициентов уравнения регрессии. Но их по-прежнему слишком много для надежной оценки с помощью стандартных методов, вот тут и приходят на помощь байесовские идеи. Коэффициенты для различных областей предполагаются *сходными* – своего рода промежуточная точка между предположением, что они в точности одинаковы, и предположением, что они совершенно не связаны.

Можно показать, что это предположение эквивалентно тому, что все эти неизвестные величины извлечены из одного и того же априорного распределения, и это позволяет нам смещать многие отдельные, довольно неточные оценки ближе друг к другу, что в итоге приводит к более уверенным выводам, на которые не так сильно влияет несколько странных наблюдений. Сделав такие более надежные оценки поведения при голосовании внутри каждой из тысяч ячеек, можно объединить все результаты и спрогнозировать, как проголосует вся страна.

На президентских выборах в США в 2016 году опросы, основанные на многоуровневой регрессии и постстратификации, правильно определили победителя в 50 случаях из 51 (50 штатов и округ Колумбия), исходя из ответов всего 9485 человек за несколько недель до выборов, и ошиблись только для Мичигана. Аналогичные хорошие прогнозы были сделаны и для выборов 2017 года в Соединенном Королевстве, где компания YouGov опросила 50 тысяч человек, не заботясь о репрезентативности выборки, а затем с помощью метода MRP предсказала подвешенный парламент<sup>[231]</sup>, где консерваторы получают 42 % голосов, что в действительности и

произошло. А вот опросы, использовавшие более традиционные методы, с треском провалились [\[232\]](#).

Так можем ли мы сделать пресловутый шелковый кошелек из подходящего неслучайного свиного уха? MRP не панацея – если большое количество респондентов систематически дают недостоверные ответы и тем самым не представляют свою «ячейку», то никакой сложный статистический анализ не компенсирует этой ошибки. Однако, по-видимому, байесовское моделирование полезно использовать для каждого отдельного участка голосования и, как мы увидим позже, это на удивление эффективно в экзитполах, проводимых в день голосования.

Байесовское «сглаживание» может добавить точность очень скудным данным, и такие методы все чаще применяются, например, для моделирования распространения болезней во времени и пространстве. Байесовское обучение сейчас рассматривается как фундаментальный процесс осознания человеком окружающей обстановки, когда у нас есть априорные ожидания того, что мы увидим в каком-то контексте, а далее нужно обращать внимание только на неожиданные изменения в нашем видении, которые затем используются для обновления наших текущих представлений. Эта идея лежит в основе так называемого байесовского мозга [\[233\]](#). Те же самые процедуры обучения были реализованы в самоуправляемых автомобилях, которые имеют вероятностную «ментальную карту» окружающей местности, постоянно обновляющуюся по мере распознавания светофоров, людей, других машин и так далее. «По сути, робот-автомобиль “думает” о себе как о вероятностном пузырьке, путешествующем по байесовской дороге» [\[234\]](#).

Эти проблемы касаются оценки величин, описывающих мир, однако использование байесовских методов для оценки научных гипотез более спорно. Как и при проверке гипотез методом Неймана – Пирсона, нам сначала нужно сформулировать две конкурирующие гипотезы. Нулевая гипотеза  $H_0$  обычно означает отсутствие чего-либо, например отсутствие бозона Хиггса или эффекта от какого-то метода лечения. Альтернативная гипотеза  $H_1$  утверждает, что нечто важное существует.

Идеи, лежащие в основе проверки байесовских гипотез, по сути, те же, что и в судебных разбирательствах, когда нулевая гипотеза обычно

означает невиновность, а альтернативная – вину, и мы каждым фактом-доказательством выражаем поддержку той или иной гипотезы в виде отношения правдоподобия. Для проверки научных гипотез точным эквивалентом отношения правдоподобия служит **коэффициент Байеса**, с той лишь разницей, что научные гипотезы обычно содержат неизвестные параметры, например реальный эффект при альтернативной гипотезе. Коэффициент Байеса можно получить только посредством усреднения по отношению к априорному распределению неизвестных параметров, что делает именно априорное распределение критически важным. Это самая спорная часть байесовского анализа. Поэтому попытки заменить стандартные проверки значимости байесовскими коэффициентами (в частности, в психологии) стали источником серьезных споров; при этом критики указывают, что за любым байесовским коэффициентом скрываются предположительные априорные распределения для любых неизвестных параметров в обеих – нулевой и альтернативной – гипотезах.

Роберт Касс и Адриан Рафтери – два известных байесовских статистика – предложили широко используемую шкалу для байесовских коэффициентов (табл. 11.3). Обратите внимание на ее контраст со шкалой из [табл. 11.2](#) для словесных интерпретаций отношения правдоподобия, применяемых в юриспруденции, где для объявления какого-то факта «очень сильным подтверждением» отношение правдоподобия должно быть 10 000, в отличие от научных гипотез, для которых нужен байесовский коэффициент больше 150. Возможно, это отражает необходимость установить вину на уровне «вне разумных сомнений», в то время как о научных открытиях заявляют на основании более слабых доказательств, многие из которых опровергаются в ходе дальнейших исследований.

### **Таблица 11.3**

Шкала Касса и Рафтери для интерпретации коэффициентов Байеса в пользу какой-либо гипотезы<sup>[235]</sup>

Коэффициент Байеса	Весомость доказательств
1–3	Заслуживает всего лишь упоминания
3–20	Положительная
20–150	Сильная
> 150	Очень сильная

В главе о проверке гипотез говорилось, что Р-значение 0,05 эквивалентно только «слабому подтверждению». Частично такое утверждение основано на коэффициентах Байеса: можно показать, что  $P = 0,05$  соответствует (при некоторых разумных априорных условиях при альтернативной гипотезе) коэффициентам Байеса, лежащим между 2,4 и 3,4, что, согласно табл. 11.3, будет «слабым подтверждением». Как мы узнали из главы 10, это привело к предложению понизить пороговый уровень Р-значения для объявления об «открытии» до 0,005.

В отличие от проверки значимости нулевой гипотезы, коэффициенты Байеса обращаются с обеими гипотезами симметрично и поэтому могут активно поддерживать нулевую гипотезу. И при готовности поставить в гипотезы априорные вероятности мы могли бы даже вычислить апостериорные вероятности для альтернативных теорий об устройстве мира.

Предположим, что мы, основываясь исключительно на теоретических соображениях, оценили бы вероятность существования бозона Хиггса в 50 %, то есть шансы на его существование – 1:1. Данные, рассмотренные в предыдущей главе, дали Р-значение, примерно равное  $1 / 3\,500\,000$ . Можно вычислить, что это приводит к коэффициенту Байеса 80 000 в пользу существования бозона Хиггса,

что считается очень сильным подтверждением даже по юридической шкале.

Соединив это значение коэффициента и априорные шансы 1:1, мы получим апостериорные шансы 80 000:1, или вероятность 0,99999 существования бозона Хиггса. Однако ни юридическое, ни научное сообщество не одобряют подобный анализ, даже если его использовать для идентификации останков Ричарда III.

### *Идеологическая битва*

В этой книге мы перешли от неформального изучения данных путем знакомства с характеристиками выборки (статистиками) к использованию вероятностных моделей для получения доверительных интервалов, Р-значений и так далее. Эти стандартные математические инструменты, с которыми сражались поколения учащихся, известны как «классические» или «частотные» методы, поскольку они основаны на свойствах больших выборок.

Альтернативный байесовский подход базируется на совершенно иных принципах. Как мы видели, внешние факты о неизвестных величинах, выраженные в виде априорного распределения в сочетании с вероятностной моделью для данных (правдоподобие) дают итоговое апостериорное распределение, которое становится основой для всех заключений.

Если мы всерьез принимаем такую статистическую философию, выборочные свойства становятся неактуальными. И, потратив годы на изучение того, что 95-процентный доверительный интервал не означает, что истинное значение лежит в нем с вероятностью 95 %<sup>[236]</sup>, бедный студент теперь должен все это забыть: байесовский 95-процентный интервал неопределенности означает в точности последнее.

Однако дискуссии о «правильном» способе статистических выводов еще сложнее, чем простые споры между «частотниками» и «байесовцами». Как и политические движения, каждая школа делится на несколько фракций, которые нередко конфликтуют друг с другом.

В 1930-е годы в научных кругах вспыхнула трехсторонняя схватка. Площадкой для спора стало Королевское статистическое общество, которое тогда (как и сейчас) тщательно протоколировало и публиковало дискуссии о работах, представленных на его заседаниях. Когда в 1934 году Ежи Нейман предложил теорию доверительных

интервалов, Артур Боули, ярый сторонник байесовского подхода, тогда известного как обратная вероятность, отмечал: «Я совсем не уверен, что “доверие” – это не “злоупотребление доверием”». А затем предложил байесовский подход: «Действительно ли это продвигает нас дальше?.. Действительно ли ведет нас к тому, что нам необходимо, – к шансам, что во Вселенной, где мы берем выборки, эта доля находится в... определенных границах? Я думаю, что нет». Издевательское связывание доверительных интервалов со злоупотреблением доверием в последующие десятилетия продолжилось.

В следующем, 1935 году началась открытая война между двумя небайесовскими лагерями – Рональдом Фишером с одной стороны и Ежи Нейманом и Эгоном Пирсоном – с другой. Подход Фишера базировался на оценивании с использованием функции правдоподобия, выражающей относительное подтверждение для различных значений параметра, которое давали данные, а проверка гипотез основывалась на  $P$ -значениях. Напротив, подход Неймана – Пирсона, известный как «индуктивное поведение», в значительной степени фокусировался на принятии решений: если вы решаете, что истинный ответ находится в 95-процентном доверительном интервале, то будете правы 95 % времени и должны контролировать ошибки первого и второго рода при проверке гипотез. Ученые даже предлагали «принимать» нулевую гипотезу, если она включала 95-процентный доверительный интервал, – концепция, которая Фишеру казалась кощунством (и впоследствии была отвергнута статистическим сообществом).

Сначала Фишер обвинил Неймана «в ряде заблуждений, выявленных в его статье». Тогда на защиту Неймана встал Пирсон, сказав, что, «хотя он знает о распространенной вере в непогрешимость профессора Фишера, он должен в первую очередь просить позволения усомниться в мудрости обвинений какого-нибудь коллеги в некомпетентности, если при этом не продемонстрированы успехи в овладении предметом спора». Желчные дискуссии между Фишером и Нейманом длились десятилетиями.

Борьба за идеологическое лидерство в статистике продолжилась и после Второй мировой войны, но со временем более классические небайесовские школы стали применять прагматичное сочетание подходов: эксперименты в целом разрабатывались с использованием теории ошибок первого и второго рода по Нейману – Пирсону, а их

анализ проводился с фишеровской точки зрения – с  $P$ -значениями в качестве меры подтверждения. Как мы видели в контексте клинических испытаний, этот странный сплав, похоже, неплохо себя проявил, и в итоге выдающийся (байесовский) статистик Джером Корнфилд заметил: «Парадокс состоит в том, что, несмотря ни на что, возникла прочная конструкция непреходящей ценности, которой не хватает всего лишь надежного логического фундамента, на котором она, как изначально предполагалось, должна быть построена»<sup>[237]</sup>.

Предполагаемые преимущества традиционных статистических методов перед байесовским подходом включают явное отделение фактов в данных от субъективных факторов, общую простоту вычислений; распространенность и установившиеся критерии «значимости»; доступность программного обеспечения; существование робастных методов, при которых нет нужды делать сильные предположения о форме распределения. В то же время сторонники байесовской теории утверждают, что сама возможность использовать внешние и даже явно субъективные элементы – это то, что позволяет делать более мощные выводы и прогнозы.

Статистическое сообщество долгое время было втянуто в злобные споры об основах предмета, но сейчас объявлено перемирие и нормой стал более универсальный подход, когда методы выбираются в соответствии с практическими потребностями, а не идеологическими сертификатами, выданными школами Фишера, Неймана – Пирсона или Байеса. Это выглядит разумным и прагматичным компромиссом в дискуссии, которая нестатистикам может показаться довольно запутанной. Я думаю, что разумные статистики в целом придут к сходным заключениям, несмотря на расхождения в отношении фундаментальных основ. Проблемы, возникающие в статистике, обычно появляются не из-за различной философии, лежащей в основе используемых методов. Чаще их причина – не лучший проект эксперимента, данные со смещением, неправильные предположения и – возможно, самое важное – отсутствие надлежащей научной практики. И в следующей главе мы рассмотрим эту темную сторону статистики<sup>[238]</sup>.

### **Выводы**

- Байесовские методы объединяют свидетельства, полученные из данных (выраженные в виде правдоподобия), с

первоначальными представлениями (априорным распределением) и выдают апостериорное вероятностное распределение для неизвестной величины.

- Теорема Байеса для двух конкурирующих гипотез может быть сформулирована так: апостериорные шансы = априорные шансы  $\times$  отношение правдоподобия.

- Отношение правдоподобия выражает относительную поддержку обеих гипотез, которую дает какой-либо факт-свидетельство, и иногда используется в качестве характеристики при результатах судебной экспертизы в уголовных разбирательствах.

- Когда априорное распределение появляется из какого-нибудь физического процесса создания выборки, байесовские методы не вызывают споров. Однако в целом необходима определенная степень суждения.

- Иерархические модели позволяют проводить несколько небольших анализов по отдельным группам, где, как предполагается, параметры будут общими.

- Коэффициенты Байеса эквивалентны отношениям правдоподобия для научных гипотез и представляют собой спорную замену проверки значимости нулевой гипотезы.

- У теории статистических выводов долгая история споров, но вопросы качества данных и научной надежности гораздо важнее.

## Глава 12. Когда дела идут не так

### Существует ли экстрасенсорное восприятие (ЭСВ)?

В 2011 году выдающийся американский социальный психолог Дэрил Бем опубликовал в известном психологическом журнале важную статью, описывающую следующий эксперимент. Перед экраном компьютера с двумя шторками усадили сто человек, которые выбирали, какая из них – левая или правая – скрывает какое-то изображение. Затем шторы «открывались», чтобы проверить правильность выбора, и все повторялось для серии из 36 изображений. Подвох был в том, что участники не знали главного: положение картинки определялось наугад *после* того, как испытуемый делал выбор, поэтому любое превышение числа правильных выборов над тем, что можно было бы ожидать при выборе наугад, приписывалось умению *предвидеть*, где появится картинка.

Бем сообщал, что вместо ожидаемой доли успехов 50 % (при нулевой гипотезе об отсутствии предвидения) участники правильно выбирали в 53 % случаев, когда показывали эротическое изображение ( $P = 0,01$ ). В статье описывались результаты еще восьми экспериментов по предвидению, проводившихся в течение 10 лет и включавших свыше 1000 участников. Автор наблюдал статистически значимые результаты в пользу предвидения в восьми из девяти исследований. Можно ли считать это убедительным доказательством существования экстрасенсорного восприятия?

Надеюсь, эта книга проиллюстрировала некоторые способы приложения статистики к решению реальных проблем, при этом практики пользуются этими методами умело и осторожно, помня об ограничениях и потенциальных ловушках. Однако реальный мир не всегда достоин восхищения. Пришло время посмотреть, что происходит, когда наука и искусство статистики не столь хороши. А затем я расскажу, как была воспринята и оценена статья Бема.

Существует причина, почему сегодня так много внимания уделяется ненадлежащей статистической практике: то, в чем ее обвиняют, известно как **кризис воспроизводимости** в науке.

В [главе 10](#) мы упоминали о сделанном в 2005 году печально известном заявлении Джона Иоаннидиса, что большинство опубликованных результатов исследований ложны. С тех пор многие ученые утверждают, что в опубликованной научной литературе наблюдается фундаментальная нехватка достоверности. Ученые не могут воспроизвести эксперименты, выполненные их коллегами, а это наводит на мысль, что оригинальные исследования не так надежны, как считалось ранее. Несмотря на то что изначально эти обвинения сосредоточились на медицине и биологии, впоследствии они распространились на психологию и другие социальные науки, хотя фактическая процентная доля преувеличенных или ложных утверждений оспаривается.

Исходное заявление Иоаннидиса основывалось на теоретической модели, но в качестве альтернативного подхода можно взять прошлые исследования и попробовать повторить их, то есть провести аналогичные эксперименты и посмотреть, дадут ли они сходные результаты. Был инициирован запуск крупного совместного проекта «Воспроизводимость»<sup>[239]</sup>, в рамках которого проверялись результаты 100 психологических исследований, но с большим размером выборок, чтобы точно обнаружить эффект, если он существует. Хотя в 97 из 100 исходных исследований сообщалось о статистически значимых результатах, в повторных экспериментах они подтвердились только в 36 % случаев<sup>[240]</sup>.

К сожалению, это почти везде преподносилось как то, что оставшиеся 64 % «значимых» исследований оказались ложными заявлениями. Однако здесь мы попадаем в ловушку строгого разделения исследований на значимые и незначимые. Выдающийся американский статистик и блогер Эндрю Гельман заявлял, что «различие между значимым и незначимым само по себе не может считаться статистически значимым»<sup>[241]</sup>. Фактически только у 23 % исходных и повторных исследований результаты значимо отличались друг от друга, и это, возможно, более удачная оценка для доли оригинальных экспериментов с преувеличенными или ложными заявлениями.

Вместо того чтобы определять «открытие» в терминах значимости или незначимости, лучше сосредоточиться на размерах оцениваемых эффектов. Проект «Воспроизводимость» установил, что эффект в повторных экспериментах в среднем имел ту же направленность, что и в исходных, но был примерно вдвое меньше по величине. Это указывает на важное смещение в научной литературе: исследование, обнаружившее нечто «большое», скорее приведет к серьезной публикации. По аналогии с регрессией к среднему это можно назвать «регрессией к нулю»: первоначальные преувеличенные оценки эффекта позднее уменьшаются в сторону нулевой гипотезы.

Кризис воспроизводимости – сложная проблема, которая коренится в чрезмерном давлении на исследователей: им нужно делать «открытия» и публиковаться в престижных научных журналах, а это зависит от получения статистически значимых результатов. Нельзя винить ни одно учреждение и ни одну профессию. При обсуждении проверки гипотез мы уже показали, что даже при идеальной статистической практике редкость истинных и существенных эффектов означает, что среди результатов, объявленных «значимыми», немалую долю неизбежно будут составлять ложноположительные (см. [рис. 10.5](#)). Впрочем, как мы видим, статистическая практика далека от совершенства.

На каждом этапе цикла PPDAC работа может быть сделана плохо.

Прежде всего мы можем взяться за *проблему*, которую просто нельзя решить с помощью имеющейся информации. Например, при попытке выяснить, почему уровень подростковой беременности в Соединенном Королевстве за последнее десятилетие так резко упал, никакие наблюдаемые данные не дадут объяснения<sup>[242]</sup>.

Далее могут возникнуть неувязки и с *планированием*.

- Использование удобной и недорогой, но не репрезентативной выборки (например, при телефонных опросах перед выборами).
- Наводящие вопросы при опросе или вводящие в заблуждение формулировки (например: «Как думаете, сколько вы можете сэкономить на покупках в интернете?»).
- Неспособность провести честное сравнение (скажем, оценивать эффект гомеопатии, наблюдая только принимающих ее добровольцев).

- Разработка исследования, которое слишком мало и обладает низкой мощностью, а значит, вы обнаружите меньше истинных альтернативных гипотез.

- Неспособность собрать данные о потенциальных возмущающих факторах, отсутствие слепых рандомизированных испытаний и так далее.

Как выразился Рональд Фишер, «чтобы проконсультироваться со статистиком после окончания эксперимента, часто достаточно попросить его провести посмертное вскрытие. Возможно, он скажет, от чего умер эксперимент»<sup>[243]</sup>,<sup>[244]</sup>.

Типичные проблемы на этапе сбора *данных* – чрезмерное количество тех, кто отказался отвечать на вопросы, выбывание участников из исследования, набор испытуемых медленнее ожидаемого, обеспечение эффективного кодирования данных. Все эти проблемы надо предусмотреть и устранить в режиме тестирования.

Простейший досадный промах на этапе *анализа* – обычная ошибка. Многие из нас ошибались при кодировании или создании электронных таблиц, но, вероятно, не с такими последствиями, как в следующих примерах:

- Выдающиеся экономисты Кармен Рейнхарт и Кеннет Рогофф в 2010 году опубликовали работу, которая сильно повлияла на меры жесткой экономии. Позже один аспирант обнаружил, что из основного анализа по недосмотру были исключены пять стран – из-за простой ошибки в электронной таблице<sup>[245]</sup>,<sup>[246]</sup>.

- Программист крупной инвестиционной компании AXA Rosenberg неправильно запрограммировал статистическую модель, из-за чего некоторые из вычисленных элементов рисков были уменьшены в десять тысяч раз, что привело к убыткам клиентов в 217 миллионов долларов. В 2011 году Комиссия по ценным бумагам и биржам США (SEC) оштрафовала AXA Rosenberg на эту сумму плюс дополнительные 25 миллионов долларов пени. Итоговый штраф компании за несообщение клиентам об ошибке в модели рисков составил 242 миллиона<sup>[247]</sup>.

Расчеты могут быть верными с точки зрения математики, но при этом использовать некорректные статистические методы. Вот некоторые популярные примеры неправильных методов.

- Провести кластерное рандомизированное испытание, при котором для какого-либо конкретного вмешательства целые группы людей распределить случайным образом, а потом анализировать результаты так, как будто случайно распределялись отдельные люди.

- Измерить две группы на исходном уровне и после вмешательства, а потом заявить, что группы различны, если одна значительно отличается от исходного уровня, а изменения во второй незначимы. Правильная процедура в этом случае – провести статистическую проверку того, отличаются ли группы одна от другой (проверка взаимодействия).

- Истолковать «незначимость» как «отсутствие эффекта». Например, в исследовании связи между употреблением алкоголя и смертностью, упомянутом в главе 10, мужчины в возрасте 50–64 лет, употреблявшие 15–20 стандартных доз в неделю, продемонстрировали значительное уменьшение риска смертности, в то время как снижение для мужчин, пьющих чуть меньше или чуть больше, незначимо отличалось от нуля. В работе это было заявлено как важное отличие, но доверительные интервалы показали, что разница между этими группами незначительна. Еще раз заметим: разница между значимым и незначимым не обязательно значима.

Что касается этапа *заклучений*, то здесь, пожалуй, самая вопиющая практика – проведение множества статистических проверок с последующим обнародованием только наиболее значимых результатов, которые выдаются за типичные. Мы видели, как сильно это повышает шансы найти значимость – вплоть до «оживления» мертвой рыбы. Это все равно что смотреть по телевизору только забитые командой голы и в упор не видеть тех, которые она пропускает: при такой избирательной отчетности невозможно получить истинное представление о матче.

Избирательная отчетность начинает переходить границы между простой некомпетентностью и нарушением научной этики, и есть тревожные подтверждения того, что это не редкость. В США даже был вынесен обвинительный приговор за избирательное сообщение о

значимых результатах в одном анализе для подмножеств. Скотт Харконен возглавлял компанию InterMune, занимавшуюся клиническими испытаниями нового препарата от идиопатического легочного фиброза. Испытание в целом не выявило никакой пользы, но у небольшой группы пациентов (с легкой и умеренной степенью заболевания) отмечалось значительное снижение смертности. Харконен выпустил для инвесторов пресс-релиз с указанием этого результата и добавил, что, по его мнению, такое исследование может привести к увеличению объемов продаж. Хотя это и не была заведомая ложь, жюри присяжных в 2009 году осудило его за мошенничество с использованием электронных средств коммуникации, а конкретно – за намерение обмануть инвесторов. Государство требовало 10-летнего заключения и штрафа в 20 тысяч долларов, однако Харконена приговорили к шести месяцам домашнего ареста и трем годам условно. Последующее клиническое испытание не выявило никакой пользы от лекарства для указанного подмножества больных<sup>[248]</sup>.

Нарушения в статистике могут быть сознательными или нет. Они даже намеренно использовались, чтобы показать недостатки научного рецензирования и публикации. Йоханнес Боханнон из немецкого института диеты и здоровья провел исследование, в котором людей разделили на три группы: 1) придерживающихся обычной диеты; 2) низкоуглеводной; 3) низкоуглеводной с добавлением шоколада. После ряда измерений, проводившихся в течение трех недель, было сделано заключение, что потеря веса в группе людей, диета которых включала шоколад, превышает потерю веса в группе с низкоуглеводной диетой на 10 % ( $P = 0,04$ ). Этот «значимый» результат предоставили в один журнал, который назвал его «выдающимся» и сообщил, что за 600 евро «он может быть опубликован в нашем основном журнале». После публикации пресс-релиза Институтом диеты и здоровья в СМИ появились многочисленные статьи под заголовками наподобие «Шоколад ускоряет потерю веса».

Но потом выяснилось, что все это было преднамеренным обманом. Йоханнес Боханнон оказался журналистом Джоном Боханноном, Института диеты и здоровья вообще не существовало; правда, данные исследования оказались несфабрированными. Однако в каждой группе было всего по пять испытуемых, тесты проводились многократно, но сообщили только о существенных различиях.

Авторы этой сфальсифицированной работы сразу же признались в обмане. Однако далеко не все статистические махинации осуществляются с благими намерениями указать таким способом на слабые места экспертной оценки.

### ***Преднамеренный обман***

Умышленная фабрикация данных действительно практикуется, но считается, что достаточно редко. Проверка анонимных самоотчетов показала, что 2 % ученых признались в фальсификации данных, в то время как Национальный научный фонд и Отдел по обеспечению добросовестности в исследованиях сталкиваются с довольно небольшим числом заведомо нечестных действий, хотя обнаруженное количество наверняка занижено<sup>[249]</sup>.

Кажется вполне логичным, чтобы мошенничество в статистике выявила сама статистика. Ури Симонсон, психолог из Пенсильванского университета, проверял статистические данные, описывающие предположительно рандомизированные испытания, которые должны демонстрировать типичный случайный разброс, но оказывались либо неправдоподобно похожими, либо неправдоподобно различными. Например, он заметил, что в одном отчете все три стандартных отклонения составили 25,11, хотя предполагалось, что их вычисляли для трех разных групп по 15 человек. Симонсон получил исходные данные и показал с помощью моделирования, что шансы получить такие одинаковые стандартные отклонения исчезающе малы, после чего исследователь, ответственный за этот отчет, был снят с должности<sup>[250]</sup>.

Британский психолог Сирил Берт, известный своим исследованием наследственности IQ, был посмертно обвинен в мошенничестве, когда выяснилось, что коэффициенты корреляции, которые он приводил для IQ разлученных близнецов, практически не менялись со временем, несмотря на постоянное увеличение группы близнецов: коэффициент был равен 0,770 в 1943 году, 0,771 в 1955-м и 0,771 в 1966-м. Психолога обвинили в подтасовке данных, но поскольку все его записи были после его смерти сожжены, вопрос до сих пор остается спорным. Некоторые утверждают, что тут просто закралась ошибка, ведь обман слишком очевиден, вряд ли ученый мог его совершить.

Все было бы гораздо проще, если бы единственными проблемами статистики – пусть даже серьезными – были только некомпетентность и нечестность. Мы могли бы обучать, проверять, воспроизводить, открывать данные для проверки и так далее, об этом мы поговорим в последней главе, посвященной правильным методам работы. Но, увы, существует более масштабная и тонкая проблема, и именно она, по мнению некоторых, и есть главный фактор кризиса воспроизводимости.

### *«Сомнительные исследовательские практики»*

Даже если данные подлинные, анализ произведен правильно, а статистика и соответствующее  $P$ -значение корректны, могут возникнуть затруднения с интерпретацией полученных результатов, если мы точно не знаем, исходя из чего исследователи сделали такие выводы.

Мы видели какие проблемы возникают, когда исследователи сообщают только о значимых результатах, но, возможно, более важен тот сознательный или неосознанный набор мелких решений, которые принимает исследователь в зависимости от того, что, как ему кажется, показывают данные. Такие поправки могут касаться изменения структуры эксперимента; решения о прекращении сбора данных; того, какие данные нужно исключить, какие коэффициенты подправить, какие группы выделить, на каких характеристиках сосредоточиться, на какие группы разделить непрерывные переменные, как обработать недостающие данные, и так далее. Симонсон называет такие решения «степенями свободы исследователя», в то время как Эндрю Гельман описывает их более поэтично – «сад расходящихся тропок». Все эти ухищрения увеличивают шансы на получение статистической значимости и все подпадают под общее название «сомнительной исследовательской практики».

Важно различать **поисковые** и **подтверждающие исследования**. Поисковые эксперименты – как раз то, о чем говорит их название: это гибкие исследования с целью рассмотреть многие возможности и выдвинуть гипотезы для последующей проверки с помощью более формальных подтверждающих экспериментов. В поисковых исследованиях можно применять самые разные настройки, но подтверждающие исследования нужно проводить в соответствии с заранее установленным и предпочтительно публичным протоколом. Любой может использовать  $P$ -значения, чтобы охарактеризовать силу

доказательств для своих выводов, но эти Р-значения надо четко различать и по-разному интерпретировать.

Действия, направленные на получение статистически значимых результатов, известны как Р-хакинг<sup>[251]</sup>, и хотя самый очевидный метод – выполнить несколько проверок, а опубликовать только наиболее значимый результат, есть гораздо более тонкие способы, позволяющие исследователям реализовать свои «степени свободы».

### **Делает ли вас прослушивание песни «Битлз» When I'm Sixty-Four моложе?**

Вы можете быть вполне уверены в правильном ответе на этот вопрос. Что делает его еще более впечатляющим, учитывая, что Симонсон с коллегами ухитрились (нужно признать, довольно мудреными средствами) получить существенный положительный результат<sup>[252]</sup>.

Студенты Пенсильванского университета, распределенные случайным образом, слушали композиции When I'm Sixty-Four группы «Битлз», или Kalimba, или Hot Potato группы The Wiggles. Затем испытуемых спрашивали, когда они родились, на сколько лет себя ощущают и еще задавали несколько абсолютно не относящихся к делу вопросов<sup>[253]</sup>.

Симонсон с коллегами постоянно анализировали эти данные всеми способами, до которых смогли додуматься, и продолжали набирать участников, пока не обнаружили некоторую значимую связь. Это случилось после 34 испытуемых, и хотя между их возрастом и записями, которые они слушали, не было выявлено никакой связи, при сравнении только When I'm Sixty-Four и Kalimba удалось получить  $P < 0,05$  в регрессии, учитывавшей возраст отца. Естественно, исследователи сообщили только о значимом результате, не упомянув о бесчисленном количестве манипуляций и избирательной отчетности – все это было раскрыто только в конце статьи, которая стала классической преднамеренной демонстрацией практики, получившей название «харкинг» – выдвижение гипотез после того, как известны результаты<sup>[254]</sup>.

### ***Сколько людей на самом деле участвуют в сомнительных исследовательских практиках?***

В опросе 2155 американских психологов, проведенном в 2012 году<sup>[255]</sup>, только 2 % признались в фальсификации данных. Но когда им задавали вопросы по списку, состоящему из десяти сомнительных исследовательских практик:

- 35 % сказали, что сообщили о неожиданном результате, как будто прогнозировали его изначально;
- 58 % признались, что продолжили собирать данные после проверки значимости полученных результатов;
- 67 % заявили, что не сообщили обо всех ответах в исследовании;
- 94 % признались, что использовали как минимум одну из вышеуказанных сомнительных исследовательских практик.

### ***Проблемы с коммуникацией***

Независимо от того, насколько хороша (или нет) статистическая работа, в какой-то момент ее нужно представить на суд аудитории – коллегам-профессионалам или широкой публике. Ученые – не единственные, кто делает заявления на основании статистических данных. Правительства, политики, благотворительные и другие общественные организации – все сражаются за наше внимание, используя цифры и науку, чтобы обеспечить «объективную» основу для своих утверждений. Технологии способствуют увеличению разнообразия источников, постоянно расширяется общение в социальных сетях, но средств контроля их надежности у нас немного.

На рис. 12.1 представлен сильно упрощенный процесс поступления к нам статистических доказательств<sup>[256]</sup>. Он начинается с первоисточника данных, затем они, пройдя через лиц, принимающих решения, их пресс-службы, информационные подразделения, поступают к журналистам, которые пишут тексты, и редакторам, которые сочиняют к ним заголовки, и наконец доходят до нас – отдельных членов общества. Ошибки и искажения могут возникать на протяжении всего процесса.



Рис. 12.1

Упрощенная схема движения информационных потоков от первоисточников статистической информации до широкой публики. На каждом этапе есть фильтры, генерируемые сомнительными методами исследования, интерпретации и коммуникации, например избирательная отчетность, отсутствие контекста, преувеличение важности и так далее

### ***Что происходит при печати?***

Первый фильтр появляется при публикации статистического исследования. Многие работы вообще не издаются – либо из-за недостаточно интересных результатов, либо из-за несоответствия целям исследовательской организации: в частности, фармацевтические компании в прошлом часто обвиняли в сокрытии результатов исследований, которые их не устраивали. В итоге ценные данные оседают в «ящике стола», и создается положительное смещение для тех, которые появляются в печати. А мы не получаем необходимой информации.

Это положительное смещение усугубляется «открытиями», которые с большей вероятностью примут к печати в серьезных журналах, нежеланием публиковать повторные результаты и, конечно же, всеми сомнительными исследовательскими практиками, которые, как мы видели, могут привести к преувеличенной статистической значимости.

### ***Пресс-служба***

Еще больше потенциальных проблем возникает на следующем этапе схемы, когда научные материалы попадают в пресс-службы для освещения в СМИ. Мы уже видели, с каким энтузиазмом было воспринято исследование о социально-экономическом положении и риске опухолей головного мозга, в результате чего появился классический заголовок: «Почему поступление в университет повышает вероятность развития опухоли мозга». Эта пресс-служба не одинока в своих преувеличениях: одно исследование установило, что из 462 пресс-релизов британских университетов 2011 года

- 40 % содержали преувеличенные заявления;
- 33 % – преувеличенные утверждения о причинности;

- 36 % – преувеличенные последствия для людей на основании исследований на животных.

Та же группа обнаружила, что большинство преувеличений, появляющихся в прессе, можно отследить до выхода пресс-релиза. Несколько более обнадеживающие результаты та же команда нашла в 534 пресс-релизах от крупных биомедицинских журналов: преувеличения были в 21 % соответствующих пресс-релизов, хотя они не получили более масштабного освещения в прессе<sup>[257]</sup>.

В [главе 1](#) мы говорили, что подача чисел может влиять на их интерпретацию: например, «обезжиренный на 90 %» звучит лучше, чем «10 % жирности». Прекрасный пример изобретательного сочинительства появился, когда одно приличное, но довольно скучное исследование установило, что 10 % людей имеют ген, защищающий их от повышения кровяного давления. Отдел коммуникаций переформулировал это так: «Девять из десяти человек несут ген, который увеличивает риск развития гипертонии», и такая негативная формулировка получила широкое освещение в международной прессе<sup>[258]</sup>.

### ***Средства массовой информации***

Журналистов часто винят за плохое освещение научных событий, хотя они во многом зависят от того, какую информацию им скамливают в пресс-релизах и научных статьях и как отражает их текст заголовков, позже вставленный редактором. Ведь мало кто из читателей газет понимает, что автор статьи, как правило, не имеет никакого отношения к созданию заголовков, задача которых – привлечь внимание читателей.

Основная проблема при освещении в СМИ не в откровенной неправде, а в манипуляции и преувеличении путем некорректной интерпретации «фактов»: они могут быть технически верными, но искажены тем, что мы называем «сомнительными методами интерпретации и коммуникации». Вот краткий список способов, как оживить подачу материалов по статистике в СМИ. И те, чья карьера зависит от привлечения читателей, слушателей и кликов, считают многие из этих сомнительных практик вполне оправданными.

1. Выбирать тексты, которые идут вразрез с текущим общим мнением.
2. Популяризировать тексты вне зависимости от качества исследований.
3. Не сообщать уровень неопределенности.
4. Не предоставлять контекст или сравнительные перспективы (например, долгосрочные тенденции).
5. Предполагать причинность, когда наблюдалась всего лишь связь.
6. Преувеличивать важность результатов.
7. Утверждать, что факты поддерживают какую-то установку или линию.
8. Использовать положительную или отрицательную подачу в зависимости от цели – успокоить или напугать.
9. Пренебрегать конфликтами интересов и альтернативными точками зрения.
10. Использовать яркую, но неинформативную графику.
11. Информировать только об относительных, но не об абсолютных рисках.

Последний метод практически универсален. В [главе 1](#) мы наблюдали, насколько впечатляюще может звучать история о повышении риска развития рака кишечника при употреблении бекона, если выражать данные в относительных, а не в абсолютных рисках. Журналисты знают, что относительные риски (часто именуемые в СМИ просто «повышенными рисками» вне зависимости от величины) – это эффективный способ сделать текст более захватывающим, хотя результаты большинства биомедицинских исследований выдаются в виде относительных рисков (**отношения шансов, отношения показателей или отношения рисков**).

Приковывающий внимание заголовок «Почему просмотр телевизора в пьяном виде может вас убить» появился в результате эпидемиологического исследования, которое дало скорректированный относительный риск 2,5 для фатальной легочной эмболии, связанной с просмотром телевизора более пяти часов в сутки по сравнению с просмотром менее двух с половиной часов. Однако внимательный анализ абсолютного показателя в группе высокого риска (13 на 158 тысяч человеко-лет) можно истолковать как означающий, что

для наступления такого события вы в среднем должны сидеть по 5 часов перед телевизором в течение 12 тысяч лет. Это несколько снижает воздействие факта<sup>[259]</sup>. Заголовок явно писался с целью привлечь внимание и в этом преуспел – я определенно считал его неотразимым.

В эпоху, когда все мы стремимся к новой информации и новым впечатлениям, неудивительно, что СМИ придают остроту текстам об исследованиях и провоцируют необычные (часто преувеличенные) заявления, выходящие за рамки серьезных статистических фактов<sup>[260]</sup>. В следующей главе мы поговорим о том, как можно улучшить ситуацию, а пока вернемся к примечательным утверждениям Дэрила Бема о предвидении.

Дэрил Бем знал, что публикует необычные утверждения, и, к его чести, активно поощрял попытки воспроизвести свои результаты и даже предоставлял для этого материалы. Однако когда другие исследователи приняли его вызов и попытались сделать то же самое, но потерпели неудачу, журнал, опубликовавший оригинальную статью Бема, отказался писать о провалившихся попытках.

Так как же Бем получал свои результаты? Иногда он корректировал проект в соответствии с данными и выделял определенные группы – например, сообщал о положительном предвидении при показе эротических картинок, а не о негативных результатах с неэротическими. Бем признавал: «Я начинаю один [эксперимент] и, если ничего не выходит, бросаю его и начинаю заново с изменениями». Некоторые из этих изменений описывались в статье, другие – нет<sup>[261]</sup>,<sup>[262]</sup>. Эндрю Гельман заметил, что выводы Бема

основаны на Р-значениях, которые являются утверждениями о том, как выглядели бы характеристики выборки, если бы данные были другими; однако Бем не предоставил никаких доказательств того, что его анализ был бы таким же, если бы данные были другими. В самом деле, девять исследований, описанных в его статье, основаны на разных методиках анализа данных<sup>[263]</sup>.

Этот случай – классический пример использования исследователем чрезмерного количества степеней свободы. Тем не менее Бем оказал большую услугу психологии и науке в целом: его статья 2011 года стала катализатором для самоанализа в среде ученых в отношении возможных причин отсутствия достоверности в научной литературе. Даже высказывалось предположение, что весь этот процесс, как и другие исследования, описанные в главе, был намеренно спланирован Бемом, чтобы показать слабые стороны в психологических исследованиях.

### **Выводы**

- Часть ответственности за кризис воспроизводимости в науке лежит на недобросовестных статистиках-практиках.
- Намеренная фабрикация данных – явление довольно редкое, в отличие от ошибок в статистических методах.
- Еще большую проблему представляют сомнительные исследовательские практики, которые, как правило, преувеличивают уровень статистической значимости.
- Вместо того чтобы представить статистические факты широкой публике, пресс-службы, журналисты и редакторы раздувают поток недостоверных сведений, неверно интерпретируя результаты и распространяя их.

## Глава 13. Как можно улучшить статистику

### В чем польза скрининговых исследований при раке яичников?

В 2015 году в Великобритании были опубликованы результаты масштабного скринингового исследования рака яичников. Оно стартовало еще в 2001 году, когда после тщательных расчетов необходимой мощности свыше 200 тысяч женщин случайным образом распределили на три группы – два вида скрининга и контрольная группа. Исследователи тщательно составили протокол, в котором в первичный анализ входило наблюдение за снижением смертности от рака яичников, оцениваемое с помощью статистического метода, предполагающего, что пропорциональное уменьшение риска будет одинаковым в течение всего периода наблюдения<sup>[264]</sup>.

Когда после среднего 11-летнего периода наблюдения данные были проанализированы, установленный первичный анализ не показал статистически значимой пользы и авторы должным образом сообщили об этом незначимом результате в качестве своего основного вывода. Но тогда почему в газете Independent появился заголовок «Прорыв в определении рака яичников по анализу крови: колоссальный успех нового метода тестирования может привести к национальному скрининговому обследованию в Британии»?<sup>[265]</sup>

Мы еще вернемся к тому, правильно ли интерпретировались результаты этого масштабного и очень дорогостоящего исследования.

В предыдущей главе мы говорили о том, как плохая практика может проявиться на любом этапе создания статистических текстов. А значит, если мы хотим использовать статистику более эффективно, следует задействовать три группы людей.

1. *Поставщики статистической информации*: ученые, статистики, исследовательские компании, промышленность. Они могут лучше генерировать данные.

2. *Коммуникаторы*: научные журналы, благотворительные организации, работники пресс-служб, журналисты и редакторы. Они могут лучше подавать статистические данные.

3. *Аудитория*: общественность, лица, принимающие решения, и эксперты. Они могут лучше проверять статистические данные.

Давайте поочередно рассмотрим, что может делать каждая группа.

### *Улучшение качества данных*

Как можно улучшить научный процесс? Широкое сотрудничество выдающихся исследователей привело к появлению «манифеста воспроизводимости», в котором говорится о совершенствовании методов исследования и обучения, содействии предварительной регистрации проектов и анализу исследований, улучшении качества отчетности о реально проделанной работе, стимулировании повторных экспериментов, развитии экспертных оценок и поощрении открытости и прозрачности<sup>[266]</sup>. Многие из этих идей отражены в Open Science Framework – программном проекте с открытым кодом, который, в частности, способствует обмену данными и предварительной регистрации исследований<sup>[267]</sup>.

С учетом примеров из предыдущей главы неудивительно, что многие предложения из этого манифеста касаются статистической практики, в частности обращение к предварительно зарегистрированным исследованиям призвано оградить от поведения, описанного в предыдущей главе, когда проект, гипотезы и анализ подстраиваются под уже полученные данные. Однако можно утверждать, что полная предварительная определенность нереальна, она не дает исследователю проявить воображение, а также лишает гибкости в процессе адаптации к новым данным. Опять же ответ, похоже, заключается в строгом различии между поисковыми и подтверждающими исследованиями, при этом нужно четко сообщать ту последовательность выбора, к которой прибегли исследователи.

Предварительная определенность анализа не лишена недостатков, поскольку может ограничить исследователей каким-то одним видом анализа, который по мере поступления данных они могут счесть

неподходящим. Например, группа, проводившая скрининговое исследование рака яичников, планировала включить в анализ всех рандомизированных пациентов, однако обнаружила, что если исключить из анализа «распространенные» случаи (когда рак яичников был выявлен до начала испытаний), что может показаться вполне разумным, то стратегия мультимодального скрининга *продемонстрирует* значимое 20-процентное снижение смертности от рака яичников ( $P = 0,02$ ). Кроме того, даже если были включены все случаи безотносительно того, был или не был выявлен рак на момент начала испытания, значимое 23-процентное снижение смертности в мультимодальной группе проявилось также в период между 7 и 14 годами после рандомизации. Поэтому проблемы, которые нельзя предусмотреть (например, рандомизация людей, у которых уже есть рак, или скрининг, требующий для большей точности определенного времени), мешают тому, чтобы предварительно запланированный общий результат оказался значимым.

Авторы педантично сообщали, что их первичный анализ не показал значимого результата, и с сожалением отмечали, что «основным ограничением в испытании была наша неспособность предвидеть в своем статистическом проекте отдаленный эффект скрининга». Это не помешало некоторым СМИ интерпретировать незначимый результат как подтверждение нулевой гипотезы и сообщить, что скрининговые исследования вообще не работают. Заголовок в *Independent*, провозглашающий, что скрининг может спасти тысячи жизней, хотя и слишком смел, тем не менее лучше отражает результаты исследования.

### ***Улучшение коммуникации***

Выше мы говорили о некорректном освещении СМИ содержания научных статей, основанных на статистических данные. Мы не знаем простых способов повлиять на журналистскую деятельность и СМИ – особенно сейчас, в условиях жесткой конкуренции с интернет-публикациями и социальными сетями, а также сокращения доходов от рекламы. Однако то, что статистики участвуют в составлении рекомендаций для СМИ и обучающих программ для журналистов и

сотрудников пресс-служб, вселяет надежду. Хорошая новость заключается в том, что журналистика данных процветает и сотрудничество с журналистами может обогатить тексты, основанные на данных, правильным и качественным контентом и визуализацией.

Однако преобразование сухих чисел в истории сопряжено с немалым риском. Традиционно хороший рассказ строится на эмоциях, увлекательном сюжете и эффектной развязке, а наука редко может все это предоставить, поэтому возникает соблазн все сильно упростить, а эффект преувеличить. Тексты должны опираться на факты, которые помогут автору разносторонне осветить поднятую проблему. В идеале в публикации может сообщаться, что какое-то лекарство или метод решения имеет как преимущества, так и побочные эффекты, которые люди могут оценивать по-разному и, соответственно, приходиться к разным выводам. Журналисты, похоже, избегают подобных текстов, однако настоящий профессионал должен уметь делать такие истории захватывающими (например, включив мнения людей с разными взглядами). Так, Кристи Ашванден<sup>[268]</sup>, работающая на портале FiveThirtyEight, обсуждала статистические данные, полученные в ходе скрининговых исследований молочной железы, после чего решила отказаться от этой практики, в то время как ее подруга, располагая теми же фактами-свидетельствами, приняла противоположное решение<sup>[269]</sup>. Это подтверждает, как важно иметь собственное мнение, но при этом с уважением относиться к статистическим данным.

Мы могли бы также детальнее исследовать вопрос о наиболее оптимальной передаче статистических данных. Например, как сообщать о неуверенности в отношении наблюдений и прогнозов, не ставя под угрозу надежность информации, авторитет статистики и доверие к ней, и как адаптировать наши методы к аудитории с различными взглядами и знаниями. Это важные, требующие углубленного изучения вопросы. Кроме того, удручающий уровень статистических дискуссий во время британской кампании по Брекситу говорит о необходимости исследовать новые способы передачи информации о том, как политические решения могут влиять на общество.

### ***Помощь в обнаружении плохой практики***

Разные люди и группы играют определенную роль в обнаружении плохой статистической практики: это рецензенты готовящихся к публикации статей; те, кто проводит систематические обзоры опубликованных доказательств; журналисты; организации, занимающиеся проверкой фактов (фактчекинг), и отдельные члены общества.

Ури Симонсон особенно настаивал на том, чтобы рецензенты строже проверяли соответствие работ требованиям журнала, побуждая авторов предоставлять убедительные доказательства надежности результатов их исследований, а в случае сомнений могли настаивать на повторении опыта и расчетов. Но при этом он предлагал рецензентам терпимее относиться к несовершенству результатов, что способствовало бы составлению правдивых отчетов<sup>[270]</sup>.

Однако как человек, который ссылаясь на сотни научных работ, хочу сказать, что определить наличие проблемы не всегда просто. Четкие требования, безусловно, полезны, но авторы всегда могут их проигнорировать, чтобы статья выглядела убедительнее. Должен признаться, что у меня развилось особое чутье на выявление неправдивых данных и недомолвок – например, если было выполнено большое число сравнений, а сообщено только об «интересных».

Мое чутье моментально реагирует, когда результат кажется уж больно хорошим, чтобы быть правдой, скажем, когда маленькая выборка дает слишком большой эффект. Классический пример – широко известное исследование 2007 года, утверждающее, что у привлекательных людей чаще рождаются дочери. В опросе американских подростков по пятибалльной шкале оценивалась их физическая привлекательность, а через пятнадцать лет у тех, кто в подростковом возрасте был оценен как «очень привлекательный», только 44 % первенцев были мальчиками, хотя стандартная величина для всех людей – 52 % (как показал еще Арбетнот, в среднем рождается чуть больше мальчиков, чем девочек). Этот результат статистически значим, но, как указал Эндрю Гельман, эффект слишком большой, чтобы быть правдоподобным, и наблюдается только в «самой привлекательной» группе. Информация, приведенная в статье, не позволяет понять, насколько маловероятно описываемое наблюдение, – здесь требуются специальные знания<sup>[271]</sup>.

### *Систематическая ошибка публикации*

При проведении систематических обзоров, чтобы свести воедино всю имеющуюся информацию и представить текущее понимание явления, ученые просматривают огромное количество статей. Однако это занятие оказывается абсолютно бесполезным, если опирается на разбор работ, искажающих факты. Например, из-за того, что отрицательные результаты даже не пытаются опубликовать и потому, что значимые результаты, полученные с применением сомнительных исследовательских практик, печатаются в избытке.

Для выявления такой систематической ошибки были разработаны специальные статистические методы. Предположим, у нас есть ряд исследований для проверки одной и той же нулевой гипотезы, скажем, что некоторое вмешательство неэффективно. Вне зависимости от реально проведенных экспериментов, если вмешательство действительно неэффективно, то можно доказать, что Р-значение для проверки нулевой гипотезы имеет равномерное распределение от 0 до 1, а потому Р-значения из множества исследований, проверявших гипотезу, должны распределяться равномерно. Тогда, если эффект действительно существует, Р-значения должны смещаться в сторону малых значений.

Идея такой «Р-кривой» – рассмотреть все указанные в исследованиях Р-значения для значимых результатов теста, то есть для  $P < 0,05$ . Подозрение вызывают две вещи. Во-первых, если кластер Р-значений чуть ниже 0,05, значит, какие-то результаты были искажены, для того чтобы передвинуть некоторые значения Р через эту границу. Во-вторых, предположим, что эти значимые Р-значения не смещены к нулю, а довольно равномерно распределены между 0 и 0,05. Тогда это в точности то, что могло возникнуть, если нулевая гипотеза верна, а нам сообщили как о значимых только о тех результатах, для которых  $P < 0,05$  и которые в одном случае из двадцати попадают в этот диапазон по чистой случайности. Симонсон и его коллеги просмотрели опубликованные работы по психологии, поддерживавшие популярную идею, согласно которой предоставление людям излишнего выбора ведет к негативным последствиям. Анализ Р-кривой указал на наличие ошибки в публикациях и отсутствие достаточно веских подтверждений этой идеи<sup>[272]</sup>.

### ***Оценивание статистических утверждений или текстов***

Кем бы мы ни были – журналистами, специалистами по фактчекингу, учеными, бизнесменами, политиками, работниками общественных организаций или просто представителями общественности, мы регулярно слышим какие-то заявления, основанные на статистических фактах. И оценивание их достоверности – жизненно важный навык для современного мира.

Давайте сделаем смелое предположение, что все, кто причастен к сбору, анализу и использованию статистических данных, придерживаются этических норм, для которых доверие имеет превалирующее значение. Онора О’Нил, специалист по философии Канта и авторитет в области доверия, подчеркивала, что люди не должны стремиться к тому, чтобы им доверяли, поскольку это выбор других людей, но должны демонстрировать *достоверность* своей работы. О’Нил сформулировала несколько простых принципов – например, доверие требует честности, компетентности и надежности. Но она также отмечает, что требуются подтверждения достоверности, а это подразумевает *прозрачность* – нужно не просто сбрасывать массу данных на аудиторию, а обеспечить «разумную прозрачность»<sup>[273]</sup>. Это означает, что утверждения, основанные на данных, должны быть:

- *Доступными*: аудитория должна иметь доступ к информации.
- *Доходчивыми*: аудитория должна быть способна понять информацию.
- *Поддающимися оценке*: при желании аудитория должна иметь возможность проверить достоверность утверждений.
- *Полезными*: аудитория должна иметь возможность использовать информацию для своих нужд.

Но оценка достоверности – сложная задача. Статистики и другие специалисты тратят десятилетия, чтобы научиться взвешивать утверждения и формулировать вопросы, которые помогут выявить недостатки. Это не какой-то очередной контрольный список, с которым нужно просто свериться, здесь нужны опыт и разумная доля скептицизма. С учетом этой оговорки предлагаю набор вопросов, вобравших в себя всю мудрость, содержащуюся в этой книге. Перечисленные термины и темы либо говорят сами за себя, либо

рассматривались ранее. Я нахожу этот перечень вопросов полезным, надеюсь, и вы тоже.

### *Десять вопросов, которые нужно задать, столкнувшись с утверждением, основанным на статистических фактах*

#### НАСКОЛЬКО НАДЕЖНЫ ЧИСЛА?

1. *Насколько тщательно проведено исследование?* Например, проверьте «внутреннюю валидность», правильность проекта и формулировки вопросов, предварительную регистрацию протокола, репрезентативность выборки и обеспечение случайности при ее составлении, корректное сравнение с контрольной группой.

2. *Какова статистическая неопределенность / доверительный уровень для результатов?* Проверьте погрешности, доверительные интервалы, статистическую значимость, размер выборки, множественные сравнения, систематические ошибки.

3. *Верна ли представленная характеристика?* Проверьте правильное использование средних, разбросы, относительные и абсолютные риски.

#### НАСКОЛЬКО НАДЕЖЕН ИСТОЧНИК?

4. *Насколько надежен источник текста?* Рассмотрите вероятность искажения из-за конфликта интересов и проверьте, рецензировали ли публикацию независимые эксперты. Спросите себя: «Почему автор хочет, чтобы я услышал эту историю?»

5. *Как преподносится история?* Помните о способах подачи (эффект фрейминга), апеллировании к эмоциям посредством упоминания экстремальных случаев, вводящих в заблуждение графиках, гипертрофированных заголовках, громко звучащих числах.

6. *О чем мне не сказали?* Пожалуй, это самый важный вопрос. Подумайте о тенденциозно отобранных результатах, о пропущенной информации, которая бы противоречила изложенному в тексте, и отсутствию независимого комментария.

#### НАСКОЛЬКО НАДЕЖНА ИНТЕРПРЕТАЦИЯ?

7. *Как это утверждение соотносится с тем, что уже известно?* Взгляните на контекст, подходящие факторы сравнения, включая

прошлые данные, и то, что показывали другие исследования, в идеале метаанализ.

8. *Какое объяснение дано тому, что было замечено?* Корреляция или причинно-следственная связь? Некорректно утверждение, что незначимый результат означает «отсутствие эффекта»? Важны регрессия к среднему, влияние возмущающих факторов, атрибуция, ошибка прокурора.

9. *Насколько эта публикация актуальна для аудитории?* Подумайте о возможности обобщения, являются ли испытуемые каким-то особым случаем, не проводили ли экстраполяцию с мышей на людей?

10. *Важен ли заявленный эффект?* Проверьте, значима ли практически величина эффекта, и особенно остерегайтесь утверждений о «повышенном риске».

### ***Этика работы с данными***

Растущая обеспокоенность потенциально неправильным использованием персональных данных (особенно при их сборе с аккаунтов в социальных сетях) сосредоточивает внимание на этических аспектах науки о данных и статистики. Хотя государственные статистики связаны официальным кодексом поведения, в целом этика при работе с данными находится на стадии разработки.

В этой книге говорилось о том, что алгоритмы, влияющие на жизнь людей, должны быть честными и прозрачными, о важности честности и воспроизводимости в науке, о требованиях к надежной коммуникации. Все это составляющие этики работы с данными, а нашумевшие истории показали, как пагубно влияет конфликт интересов и даже просто чрезмерный энтузиазм, искажая полученные данные. Можно было бы выделить многие другие важные темы: конфиденциальность и право собственности на данные, информированное согласие на их более широкое использование, юридические аспекты объяснения алгоритмов и тому подобные.

Хотя статистика может показаться сугубо технической наукой, ее всегда нужно рассматривать в контексте общества, и ее представители несут за это ответственность. В ближайшем будущем можно ожидать, что этика работы с данными станет неотъемлемой частью преподавания статистики.

### *Пример хорошей статистической практики*

**Перед всеобщими выборами 8 июня 2017 года в Великобритании большинство опросов общественного мнения предполагало, что консерваторы получат значительный перевес. Через несколько минут после окончания голосования, в 22:00, группа статистиков предсказала, что консерваторы потеряли много мест, а с ними и абсолютное большинство, поэтому парламент будет подвешенным. Это заявление было встречено с недоверием. Как они смогли сделать столь смелый прогноз и оказались ли правы?**

Завершить книгу, которая была написана не для того, чтобы разоблачить недобросовестных исследователей, а для того, чтобы показать, какую пользу способно принести владение искусством и наукой работы с данными, вполне уместно ярким примером применения статистики.

Вопрос, кто выиграл выборы, сразу же после того, как они закончились, может показаться странным: в конце концов, можно посидеть ночь и подождать итогов. Но это уже стало традицией: буквально через несколько минут после окончания опросов эксперты делают прогнозы относительно результатов. Обратите внимание, что результаты уже фиксированы, просто неизвестны, так что мы имеем дело с классическим примером эпистемической неопределенности, возникающей при рассмотрении уровня безработицы и прочих величин, которые «существуют», но неизвестны.

Рассмотрим цикл PPDAC. *Проблема* состояла в том, чтобы дать прогноз результатов выборов в стране в течение нескольких минут после окончания голосования. Команда, в которую входили статистики Дэвид Фёрт и Джуни Куха, а также психолог Джон Кертис, разработала *план* проведения экзитполов, согласно которому в опросах участвовали примерно 200 респондентов, вышедших из каждого из 144 участков (из общего количества в 40 тысяч участков), причем эти участки должны были быть теми же, что и в предыдущих экзитполах. Данные включали ответы избирателей не только о том, как они проголосовали, но и как они голосовали на предыдущих выборах.

Анализ использовал ряд методов, о которых мы говорили в [главе 3](#).

- *Переход от данных к выборке.* Поскольку данные собирали после ухода с участков и респонденты говорили о том, что уже сделали, а не что намереваются сделать, опыт подсказывает, что ответы будут достаточно точной характеристикой того, как люди голосовали на этих и предыдущих выборах.

- *Переход от выборки к изучаемой совокупности.* Репрезентативная выборка берется из числа тех, кто проголосовал на каждом участке, так что результаты, полученные от этой выборки, можно использовать для примерной оценки изменения в голосовании («качели») в этой небольшой области.

- *Переход от изучаемой к целевой совокупности.* Используя знания о демографии каждого избирательного участка, строится регрессионная модель, которая пытается объяснить, как доля людей, поменявших свое мнение между выборами, зависит от характеристик избирателей на этом участке. При этом такие «качели» (свинг) необязательно будут одинаковыми по всей стране, а могут меняться в разных районах – например, в зависимости от того, какое население там преобладает, сельское или городское. Затем на основании этой регрессионной модели, знания демографических характеристик населения в каждом из примерно 600 избирательных округов и количества голосов избирателей, отданных на предыдущих выборах, можно сделать прогноз голосования на этих выборах для каждого отдельного избирательного округа, хотя на большинстве избирательных участков вообще не проводился экзитпол. По сути, это процедура многоуровневой регрессии и постстратификации (MRP), описанная в [главе 11](#).

Ограниченная выборка означает наличие у коэффициентов регрессионной модели неопределенности, которая при масштабировании до всей голосовавшей совокупности дает вероятностное распределение того, как люди голосовали, а следовательно, и вероятность для каждого кандидата получить максимальное количество голосов.

Сложив все эти данные со всех избирательных участков, мы получаем ожидаемое количество мест в парламенте, причем в каждом

случае будет свой уровень неопределенности (хотя в ночь после выборов о погрешностях не сообщалось)<sup>[274]</sup>.

В табл. 13.1 приведены прогнозы и конечные результаты для июньских выборов 2017 года. Предсказанное количество мест удивительно близко к реальному, ошибка максимум в четыре места для всех партий. Таблица показывает, что для трех последних выборов в Великобритании эта сложная статистическая методология имела исключительную точность. В 2015 году она предсказала колоссальные потери у либерал-демократов, оценив снижение с 57 мест до 10, и известный представитель этой партии Пэдди Эшдаун заявил в прямом телеэфире, что готов «съесть свою шляпу», если прогноз окажется правильным. На самом деле либеральные демократы получили всего 8 мест<sup>[275]</sup>.

### **Таблица 13.1**

Прогнозы числа мест, полученных каждой партией на трех последних национальных выборах в Великобритании, сделанные на основе экзитполов сразу по окончании голосования, в сравнении с фактическими результатами выборов. Прогнозы представляют собой оценки с определенными погрешностями

Год	Результаты	Консерваторы	Лейбористы	Либерал-демократы	Шотландские националисты	Прочие
2010	Прогнозируемые	307	255	59		29
	Фактические	307	258	57		28
2015	Прогнозируемые	316	239	10	58	27
	Фактические	331	232	8	56	23
2017	Прогнозируемые	314	266	14	34	21
	Фактические	318	262	12	35	22

В ночь выборов все СМИ обнародовали только прогнозируемое количество мест для каждой партии, хотя погрешность составляла около 20 мест. В прошлом удавалось добиться несколько большей точности, возможно, просто благодаря везению статистиков. Нельзя, однако, сказать, что их удача была неза заслуженной, поскольку проявилась она после использования мощных научных инструментов, высокая эффективность которых способна удивить как профессионалов, так и неосведомленных. Люди слабо представляют сложность лежащих в основе расчетов методов, а также то, что этот превосходный результат обусловлен тщательным вниманием к деталям всего цикла решения задач.

### **Выводы**

- Поставщики статистической информации, коммуникаторы и аудитория – все играют определенную роль в улучшении способов применения статистики в обществе.

- Поставщикам данных необходимо обеспечить воспроизводимость результатов. Чтобы продемонстрировать надежность, информация должна быть доступной, доходчивой, поддающейся оценке и полезной.

- Коммуникаторы должны быть осторожны, пытаясь подогнать статистические тексты под стандартные правила повествования.

- Аудитории нужно выявлять недостоверные данные, задавая вопросы о методах подсчета, источниках информации и методах анализа и интерпретации.

- Когда вы сталкиваетесь с каким-то утверждением, основанным на статистических данных, сначала посмотрите, правдоподобно ли оно.

## Глава 14. В заключение

Честно говоря, статистика может быть сложной. Хотя я в этой книге старался познакомить вас с базовыми идеями и не углубляться в технические подробности, в ходе повествования неизбежно пришлось опираться на некоторые сложные концепции. Так что поздравляю тех, кто добрался до конца.

Вместо того чтобы сводить информацию, приведенную в предыдущих главах, к короткому перечню мудрых советов, я воспользуюсь следующими десятью правилами эффективной статистической практики. Они позаимствованы у группы статистиков, которые, как и я в этой книге, старались подчеркнуть нетехнические вопросы, обычно не изучаемые в курсе статистики<sup>[276]</sup>. Я добавил свои комментарии. Эти правила вполне самоочевидны и довольно точно подытоживают вопросы, рассмотренные в книге.

1. *Статистические методы должны позволять данным отвечать на научные вопросы.* Спрашивайте: «Почему я это делаю?», а не фокусируйтесь на используемом методе.

2. *Сигнал всегда сопровождается шумом.* Именно попытки отделить их друг от друга делают эту область интересной. Случайный разброс неизбежен, а вероятностные модели полезны в качестве абстракции.

3. *Планируйте, и делайте это как можно раньше.* Это включает идею предварительной подготовки для подтверждающих исследований – во избежание степеней свободы исследователя.

4. *Обеспечивайте подходящее качество данных.* Это фундамент вашей работы.

5. *Статистический анализ – это нечто большее, чем просто набор вычислений.* Не используйте формулы или программы, если не понимаете, почему вы это делаете.

6. *Будьте проще.* Основная коммуникация должна быть максимально простой – не демонстрируйте умение строить сложные модели, если они не нужны.

7. *Обеспечьте оценки для разброса.* С предупреждением, что погрешности, как правило, больше заявленных.

8. *Проверяйте свои исходные предположения.* Если это невозможно, обязательно разьясните ситуацию.

9. *При наличии возможности повторите!* Или побуждайте других воспроизводить ваш опыт.

10. *Обеспечьте воспроизводимость вашего анализа.* Другие должны иметь доступ к вашим данным и коду.

Статистика играет важную роль в нашей жизни и постоянно меняется в ответ на увеличение объема и глубины доступных данных. Но изучение этой науки влияет не только на общество в целом, но и на его отдельных членов. Что касается меня, то написание этой книги позволило мне понять, насколько обогатилась моя жизнь благодаря статистике. Я надеюсь, что и вы ощутите то же самое – если не сейчас, то в будущем.

## Глоссарий

**P-значение:** мера расхождения между данными и нулевой гипотезой. Пусть имеется нулевая гипотеза  $H_0$  и критерий  $T$ , большие значения которого указывают на расхождение с  $H_0$ . Предположим, что мы наблюдаем некоторое значение  $t$ . Тогда (одностороннее) P-значение – это вероятность наблюдения не меньшего экстремального значения при условии истинности  $H_0$ , то есть  $P(T \geq t | H_0)$ . Если о несовместимости с  $H_0$  говорят и большие, и малые значения  $T$ , то двустороннее P-значение – это вероятность наблюдения таких экстремальных значений в обоих направлениях. Часто двустороннее P-значение берут как удвоенное одностороннее P-значение, в то время как программное обеспечение R использует общую вероятность событий, где вероятность появления ниже, чем реально наблюдаемая;

**ROC-кривая:** для алгоритма, вырабатывающего какую-то оценку, можно выбрать конкретное пороговое значение, при превышении которого объект классифицируется как «положительный». По мере изменения порогового значения на графике формируется ROC-кривая: получающаяся чувствительность (истинно положительная доля) по оси  $y$ , а единица минус специфичность (ложноположительная доля) – по оси  $x$ ;

**t-статистика:** статистика, используемая для проверки нулевой гипотезы, что какой-то параметр равен нулю; это отношение оценки к ее стандартной ошибке. Для больших выборок значения больше 2 или меньше  $-2$  соответствуют двустороннему P-значению 0,05; точные P-значения можно получить из статистических программ;

**Z-оценка:** способ стандартизации наблюдения  $x_i$  в терминах расстояния от среднего выборочного значения  $m$ , выраженного в стандартных отклонениях  $s$ , так что  $z_i = (x_i - m)/s$ . Наблюдение с Z-оценкой 3 соответствует трем стандартным отклонениям от среднего, то есть представляет собой довольно серьезным выброс. Z-оценку можно также определять в терминах среднего всей популяции и стандартного отклонения  $\sigma$ , в этом случае  $z_i = (x_i - \mu)/\sigma$ ;

**абсолютный риск:** доля людей в определенной группе, с которыми за указанный период времени произошло интересующее нас

событие;

**алгоритм:** правило или формула, которые получают входные данные/переменные и дают на выходе некоторый результат, например прогноз, классификацию или вероятность;

**анализ по назначенному лечению:** принцип, согласно которому участники рандомизированных испытаний анализируются в соответствии с вмешательством, которое им назначено, вне зависимости от того, получили ли они его на самом деле;

**апостериорное распределение:** в байесовском анализе вероятностное распределение неизвестных параметров, определенное с учетом наблюдаемых данных по теореме Байеса;

**априорное распределение:** в байесовском анализе начальное вероятностное распределение для неизвестных параметров. После наблюдения каких-то данных его пересматривают, получая апостериорное распределение с помощью теоремы Байеса;

**асимметричное распределение:** распределение (выборки или генеральной популяции), которое несимметрично и имеет длинный левый или правый хвост. Распространено у величин со значительной неравномерностью, например доход или продажи книг. Для таких распределений величины выборочного среднего и стандартного отклонения могут вводить в заблуждение;

**Байеса коэффициент:** относительное подтверждение, которое дает какой-то набор данных двум альтернативным гипотезам. Для гипотез  $H_0$ ,  $H_1$  и данных  $x$  это отношение равно  $p(x|H_0)/p(x|H_1)$ ;

**Байеса теорема:** утверждение, которое показывает, как наступление события  $A$  изменяет наше априорное представление об утверждении  $B$  (априорную вероятность  $p(B)$ ) и дает апостериорное представление (апостериорную вероятность  $p(B|A)$ ) с помощью

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

формулы . Ее нетрудно доказать: поскольку  $p(BA) = p(AB)$ , то правило умножения для вероятностей означает, что  $p(B|A)p(A) = p(A|B)p(B)$ , и деление обеих частей на  $p(A)$  дает утверждение теоремы;

**байесовский подход:** подход к статистическим выводам, при котором вероятность используется не только для стохастической, но и для эпистемической неопределенности в отношении неизвестных

фактов. Затем с помощью теоремы Байеса можно пересмотреть представления в свете новых фактов;

**Бернулли распределение:** если  $X$  – случайная величина, которая принимает значение 1 с вероятностью  $p$  и значение 0 с вероятностью  $1-p$ , то  $X$  имеет распределение Бернулли. Математическое ожидание (среднее) такой величины равно  $p$ , а дисперсия составляет  $p(1-p)$ . Сам эксперимент с двумя исходами (успех и неудача) называется испытанием Бернулли;

**бинарные (двоичные) данные:** переменные, которые могут принимать два значения, часто это ответы типа «да»/«нет» на какой-нибудь вопрос. Математически их можно представить с помощью распределения Бернулли;

**биномиальное распределение:** если у нас есть  $n$  независимых испытаний Бернулли с одной и той же вероятностью успеха, то число успехов в  $n$  испытаниях имеет биномиальное распределение. Формально: пусть  $X_1, \dots, X_n$  – независимые случайные величины, имеющие распределение Бернулли с вероятностью успеха  $p$ . Тогда их сумма  $R = X_1 + X_2 + \dots + X_n$  имеет биномиальное распределение, при этом

$P(R=r) = C_n^r p^r (1-p)^{n-r}$ , математическое ожидание (среднее) равно  $np$ , а дисперсия  $np(1-p)$ . Наблюдаемое отношение  $R/n$  имеет среднее  $p$  и дисперсию  $p(1-p)/n$ . Поэтому величину  $R/n$  можно рассматривать как оценку для  $p$  со стандартной ошибкой

$$\sqrt{p(1-p)/n};$$

**большие данные:** становящееся все более анахроничным выражение, которое иногда характеризуется четырьмя параметрами: большим объемом данных, разнообразием источников (изображения, аккаунты в социальных сетях, транзакции), большой скоростью получения и возможной нехваткой достоверности из-за шаблонных способов сбора;

**Бонферрони поправка:** метод для регулирования размера критерия (ошибка первого рода) или доверительных интервалов при одновременном тестировании многих гипотез. Более точно, при проверке  $n$  гипотез при общем размере критерия (ошибка первого рода)  $\alpha$  каждую гипотезу проверяют с размером  $\alpha/n$ . Это эквивалентно тому, что для каждой оцениваемой величины указываются доверительные

интервалы  $100(1-\alpha/n)\%$ . Например, если вы проверяете 10 гипотез с общим 5 %, то Р-значения нужно сравнивать с  $0,05/10 = 0,005$  и использовать 99,5-процентные доверительные интервалы;

**Бриера показатель:** мера точности вероятностных прогнозов, основанная на среднеквадратичной ошибке прогноза. Если  $p_1, \dots, p_n$  – это вероятности для двоичных наблюдений  $x_1, \dots, x_n$ , принимающих

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

значение 0 и 1, то показатель Бриера – это число . По сути, это критерий среднеквадратичной ошибки, примененный к бинарным данным;

**бутстрэппинг:** способ генерировать доверительные интервалы и распределения тестовых статистик путем создания повторных выборок из наблюдаемых данных, а не использования вероятностной модели для соответствующей случайной величины. Бутстрэп-выборка из набора данных  $x_1, \dots, x_n$  – это выборка размера  $n$  с возвратом, так что хотя в нее попадают те величины, которые есть в исходной выборке, их доли в бутстрэп-выборке в целом будут отличаться от долей в исходной выборке;

**вероятностное распределение:** общий термин для математического закона, описывающего вероятность, с которой случайная величина принимает то или иное значение. Распределение величины  $X$  описывается функцией распределения  $F(x) = P(X \leq x)$ ,  $-\infty < x < \infty$ ;

**вероятностный прогноз:** прогноз в виде вероятностного распределения для будущего события, а не категорического суждения о том, что оно произойдет;

**вероятность:** формальное математическое выражение неопределенности. Обозначим  $P(A)$  вероятность события  $A$ . Тогда справедливы такие правила для вероятности<sup>[277]</sup>:

1.  $0 \leq P(A) \leq 1$ , при этом вероятность невозможного события равна 0, а достоверного – 1.

2. Вероятность противоположного (дополнительного) события  $\bar{A}$  (которое заключается в том, что  $A$  не произошло):  $P(\bar{A}) = 1 - P(A)$ .

3. Правило сложения: если  $A$  и  $B$  – несовместные события (то есть произойти может только какое-то одно), то  $P(A \text{ или } B) = P(A) + P(B)$ .

4. Правило умножения: для любых событий  $A$  и  $B$ ,  $P(A \text{ и } B) = P(A|B)P(B)$ , где  $P(A|B)$  означает вероятность события  $A$  при условии, что  $B$  произошло.  $A$  и  $B$  независимы тогда и только тогда, когда  $P(A|B)=P(A)$ , то есть наступление события  $B$  не влияет на вероятность события  $A$ . В этом случае мы имеем  $P(A \text{ и } B) = P(A)P(B)$  – правило умножения для независимых событий<sup>[278]</sup>;

**вероятность случайного совпадения:** при судебной экспертизе ДНК – вероятность того, что человек, случайно выбранный из надлежащей популяции, будет соответствовать найденной ДНК в степени, которая связывает подозреваемого и преступление;

**внешняя валидность:** когда заключения исследования можно обобщать на целевую совокупность, которая шире, чем непосредственно исследуемая совокупность. Относится к релевантности исследования;

**внутренняя валидность:** когда заключения какого-либо исследования действительно касаются только изучаемой совокупности. Это относится к строгости, с которой проведено исследование;

**воздействие:** фактор, влияние которого на заболевание, смерть или иной медицинский исход, представляет для нас интерес, например какой-то аспект окружающей среды или поведения;

**возмущающий (искажающий) фактор:** переменная, которая связана и с предикторной переменной, и с переменной отклика и может объяснить часть их видимой взаимосвязи. Например, рост и вес детей сильно коррелированы, но в основном эта взаимосвязь объясняется возрастом ребенка;

**воронкообразный график:** график, где наблюдениям, соответствующим отдельным элементам (учреждения, области или исследования), сопоставляется мера их точности. Часто две «воронки» указывают на то, где можно ожидать месторасположения 95 % и 99,8 % наблюдений, когда между элементами в действительности нет разницы. Если распределение наблюдений приближенно нормальное, то граничные значения для 95 % и 99,8 % примерно соответствуют  $\pm 2$  и  $\pm 3$  стандартным ошибкам;

**выборочное среднее:** см. [среднее 2](#).

**генеральная совокупность (популяция):** группа, из которой, как предполагается, берутся данные в выборке и которая дает вероятностное распределение для отдельного наблюдения. При

проведении измерений или наличии у вас всех возможных данных это понятие становится математической идеализацией;

**глубокое обучение:** метод машинного обучения, который расширяет стандартные модели искусственных нейронных сетей на множество слоев, представляющих различные уровни абстракции, например переход от отдельных пикселей изображения к распознаванию объектов;

**гипергеометрическое распределение:** пусть имеется конечное множество из  $N$  элементов,  $K$  из которых обладают некоторым свойством. Мы выбираем  $n$  элементов без возвращения. Тогда случайная величина  $Y$  – число успехов (выбранных элементов с этим свойством) имеет гипергеометрическое распределение. Формально для  $k = 0, 1, \dots, n$

$$P(Y = k) = \frac{C_K^k C_{N-K}^{n-k}}{C_N^n}$$

**грамотность в работе с данными:** умение понимать принципы, лежащие в основе работы с данными, выполнять базовые анализы данных, критически анализировать качества утверждений, сделанных на основе данных;

**дерево классификации:** форма алгоритма классификации, при котором характеристики проверяются последовательно; ответ на очередной вопрос определяет, какая характеристика проверяется следующей; процедура повторяется до итоговой классификации;

**дилемма смещения – дисперсии:** когда для прогноза используется обучение модели, повышение ее сложности в итоге приводит к тому, что у модели уменьшается смещение (в том смысле, что у нее возрастает потенциал для адаптации к деталям базового процесса), но увеличивается дисперсия, поскольку данных для уверенности в параметрах модели оказывается недостаточно. Чтобы избежать переобучения, нужен компромисс;

**дисперсия выборочная:** если имеется выборка  $x_1, x_2, \dots, x_n$  со

средним  $s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$ , то выборочная дисперсия (хотя

знаменатель может быть равен  $n$ , а не  $n-1$ )<sup>[279]</sup>;

**дисперсия:** характеристика разброса случайной величины; если случайная величина  $X$  имеет математическое ожидание  $E(X) = \mu$ , то дисперсия  $D(X) = E(X-\mu)^2$ . Среднеквадратичное (стандартное) отклонение является корнем из дисперсии, так что

$$SD(X) = \sqrt{D(X)};$$

**доверительный интервал:** оцениваемый интервал, в котором может находиться неизвестный параметр. Например, при наличии наблюдаемого множества данных  $x$  95-процентный доверительный интервал для среднего  $\mu$  – это такой интервал от  $L(x)$  до  $U(x)$ , когда до наблюдения данных вероятность того, что случайный интервал  $(L(x), U(x))$  содержит  $\mu$ , составляет 95 %. Если соединить центральную предельную теорему с тем фактом, что примерно 95 % нормального распределения отклоняется от среднего не более чем на 2 стандартных отклонения, мы получим популярное приближение, что 95-процентный доверительный интервал – это оценка в  $\pm 2$  стандартные ошибки. Предположим, что мы хотим найти доверительный интервал для разности  $\mu_2 - \mu_1$  между двумя параметрами  $\mu_2$  и  $\mu_1$ . Если  $T_1$  – это оценка для  $\mu_1$  со стандартной ошибкой  $SE_1$ , а  $T_2$  – это оценка для  $\mu_2$  со стандартной ошибкой  $SE_2$ , то  $T_2 - T_1$  представляет собой оценку для  $\mu_2 - \mu_1$ . Дисперсия разности между оценками равна сумме их дисперсий, и поэтому стандартная ошибка для  $T_2 - T_1$  определяется формулой

$\sqrt{SE_1^2 + SE_2^2}$ . Отсюда можно найти 95-процентный доверительный интервал для разности  $\mu_2 - \mu_1$ ;

**зависимая переменная (переменная отклика):** переменная, которая представляет основной интерес, которую мы желаем спрогнозировать или объяснить;

**зависимые события:** когда вероятность одного события зависит от наступления другого;

**закон больших чисел:** общее название нескольких теорем о сходимости средних для последовательности случайных величин к истинному математическому ожиданию. На практике это означает, что

выборочное среднее близко к среднему значению всей генеральной совокупности;

**иерархическое моделирование:** в байесовском анализе – когда параметры, определяющие число элементов (например, районов или школ), сами считаются взятыми из общего априорного распределения. Это приводит к уменьшению оценок параметров для отдельных элементов в сторону общего среднего;

**индуктивное поведение:** сделанное в 1930-х годах предложение Ежи Неймана и Эгона Пирсона по проверке гипотез в терминах принятия решений. От него остались идеи размера и мощности критерия, а также ошибок первого и второго рода;

**индукция (индуктивное умозаключение):** построение обобщающего вывода на основании частных примеров;

**интерквартильный размах:** мера разброса выборки или распределения; конкретно – разность между третьим и первым квартилем, то есть между 75-м и 25-м процентилем;

**искусственный интеллект (ИИ):** компьютерные программы, предназначенные для выполнения задачи, обычно связываемой с человеческими способностями;

**исследование «случай – контроль»:** ретроспективное исследование, в котором люди с заболеванием или с интересующей нас характеристикой (случай) сопоставляются с одним или несколькими людьми, не имеющими заболевания (контрольные экземпляры), и сравниваются истории этих групп – чтобы увидеть, дают ли воздействия систематическую разницу между группами. Такая схема может оценивать только относительные риски, связанные с воздействиями;

**калибровка:** требование, чтобы наблюдаемые частоты событий соответствовали вероятностным прогнозам. Например, если вероятность какого-нибудь события 0,7, то оно должно происходить примерно в 70 % случаев;

**качественная (категорийная) переменная:** переменная, принимающая два или несколько дискретных значений, которые могут или не могут быть упорядоченными;

**квартиль (генеральной совокупности):** 25-й, 50-й и 75-й процентиля;

**комбинированные признаки:** когда несколько объясняющих переменных соединяются и производят эффект, отличный от ожидаемого при их отдельном воздействии;

**конструирование признаков:** в машинном обучении процесс уменьшения размерности входных переменных с созданием сводных характеристик, которые содержат информацию о данных в целом;

**контрольная группа:** множество людей, которые не попадали под интересующее нас воздействие;

**контрольные граничные значения:** заранее определенные ограничения для случайной величины, используемые при контроле качества для отслеживания отклонений от предполагаемых стандартов; например, могут отображаться на воронкообразном графике;

**контрфактуальный:** относящийся к сценариям вида «что, если», где рассматривается альтернативная история событий;

**коэффициент регрессии:** оцениваемый параметр в статистической модели, который выражает степень взаимосвязи между объясняющей переменной и результатом во множественной регрессии. Этот коэффициент будет иметь различную интерпретацию в зависимости от того, является ли результирующая переменная непрерывной (множественная линейная регрессия), долей (логистическая регрессия), целым числом (пуассоновская регрессия) или временем выживания (регрессия Кокса);

**кризис воспроизводимости:** утверждение, что многие опубликованные научные выводы основаны на недостаточно качественных работах, поэтому такие результаты не могут воспроизвести другие исследователи;

**критерий независимости хи-квадрат/критерий согласия хи-квадрат:** статистический критерий, показывающий степень несовместимости данных с принятой статистической моделью, заключающей нулевую гипотезу (например, величины независимы или имеют определенное распределение). А именно: критерий сравнивает множества каких-то наблюдаемых величин  $x_1, \dots, x_m$  и ожидаемых при нулевой гипотезе величин  $y_1, \dots, y_m$ . Простейший вариант критерия –

$$\chi^2 = \sum_{i=1}^m \frac{(x_i - y_i)^2}{y_i}.$$

При нулевой гипотезе значение  $\chi^2$  приближенно будет иметь известное  $\chi^2$ -распределение. Это позволяет вычислить соответствующее Р-значение;

**логарифмическая шкала:** логарифм по основанию 10 для положительного числа  $x$  обозначается  $y = \log_{10}x$ , что эквивалентно  $x = 10^y$ . В статистическом анализе  $\log x$  обычно обозначает натуральный логарифм  $\log_e x$ , что эквивалентно  $x = e^y$ , где  $e$  – основание натурального логарифма 2,71828...;

**логистическая регрессия:** форма множественной регрессии, когда переменная отклика – это доля, а коэффициенты соответствуют  $\log(\text{отношение шансов})$ . Допустим, мы наблюдаем набор долей  $y_i = r_i/n_i$  в предположении, что у нас биномиальные величины с вероятностями  $p_i$ , а соответствующий набор предикторных переменных –

$(x_{i_1}, x_{i_2}, \dots, x_{i_p})$ . Предполагается, что логарифм шансов с оцениваемой вероятностью  $\hat{p}_i$  определяется линейной регрессией:

$$\log \frac{\hat{p}_i}{(1 - \hat{p}_i)} = b_0 + b_1 x_{i_1} + b_2 x_{i_2} + \dots + b_p x_{i_p}.$$

Допустим, что одна из предикторных переменных, например  $x_1$ , является двоичной, где  $x_1 = 0$  соответствует отсутствию воздействия потенциального риска, а  $x_1 = 1$  соответствует воздействию. Тогда коэффициент  $b_1$  – это  $\log(\text{отношение шансов})$ ;

**ложноположительный:** неверная классификация «отрицательного» случая как «положительного»;

**математическое ожидание (среднее):** среднее значение случайной величины (взвешенное по вероятностям или по плотности). Для дискретной случайной величины это  $\sum xp(x)$ , а для непрерывной случайной величины это  $\int xp(x)dx$ . Например, если случайная величина  $X$  – это число очков, выпавших на симметричной игральной кости,

то есть  $P(X = x) = 1/6$  для  $x = 1, 2, 3, 4, 5, 6$ , то  $E(X) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3,5$ ;

**матрица ошибок:** таблица, где собраны верные и неверные классификации, произведенные каким-либо алгоритмом;

**машинное обучение:** процедуры извлечения алгоритмов (например, для классификации, прогнозирования или кластеризации) из сложных данных;

**медиана (выборки):** значение, которое окажется посередине, если упорядочить числа в выборке. Более строго: упорядочив числа в выборке, обозначим наименьшее число  $x_{(1)}$ , второе по величине  $x_{(2)}$  и так далее (получившийся набор  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  называют вариационным рядом). Если  $n$  – нечетное число, то медиана – число, находящееся

точно посередине вариационного ряда, то есть число  $x_{\left(\frac{n+1}{2}\right)}$ . Если же  $n$  – четное число, то медианой обычно считают полусумму двух средних чисел;

**метаанализ:** формальный статистический метод объединения результатов нескольких исследований;

**метод наименьших квадратов:** предположим, что у нас есть  $n$  пар чисел  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ,  $\bar{x}$ ,  $s_x$  – выборочное среднее и среднее квадратичное отклонение для чисел  $x$  и  $\bar{y}$ ,  $s_y$  – выборочное среднее и среднее квадратичное отклонение для чисел  $y$ . Тогда прямая регрессии, вычисленная по методу наименьших квадратов, определяется уравнением

$$\hat{y} = b_0 + b_1(x - \bar{x}),$$

где  $\hat{y}$  – прогнозируемое значение зависимой переменной для определенного значения независимой переменной  $x$ ;

$$b_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2};$$

коэффициент наклона

отсекаемый отрезок  $b_0 = \bar{y}$ . Прямая по методу наименьших квадратов проходит через центр тяжести  $\bar{x}, \bar{y}$ ;

$i$ -й остаток – разность между  $i$ -м наблюдением и его предсказанным значением  $y_i - \hat{y}_i$ ;

скорректированное значение  $i$ -го наблюдения – это сумма остатка и отсекаемого отрезка, то есть  $y_i - \hat{y}_i + \bar{y}$ . Это значение мы наблюдали бы в «среднем» случае, если бы имели  $x = \bar{x}$ , а не  $x = x_i$ ;

остаточная сумма квадратов – это сумма квадратов всех остатков,

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

то есть  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ . Прямая, построенная по методу наименьших квадратов, определяется как прямая, минимизирующая сумму квадратов разностей;

коэффициент наклона  $b_1$  и коэффициент корреляции Пирсона  $r$  связаны формулой  $b_1 = rs_y / s_x$ . Поэтому в случае, когда стандартные отклонения для  $x$  и  $y$  одинаковы, коэффициент угла наклона в точности равен коэффициенту корреляции;

**множественная линейная регрессия:** предположим, что для каждого отклика  $y_i$  есть набор из  $p$  предикторных переменных  $(x_{i1}, x_{i2}, \dots, x_{ip})$ . Тогда множественная линейная регрессия по методу наименьших квадратов определяется уравнением

$$\hat{y}_i = b_0 + b_1(x_{i1} - \bar{x}_1) + b_2(x_{i2} - \bar{x}_2) + \dots + b_p(x_{ip} - \bar{x}_p),$$

где коэффициенты  $b_0, b_1, \dots, b_p$  выбираются так, чтобы

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

минимизировать сумму остатков  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ . Отсекаемый отрезок

$b_0$  – это просто среднее  $\bar{y}$ , а формулы остальных коэффициентов

сложны, но легко вычисляются. Обратите внимание, что  $b_0 = \bar{y}$  является спрогнозированным значением наблюдения  $y$ , если

предикторные переменные были средними  $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$ , и, как в случае линейной регрессии, скорректированные определяются суммой остатка и отсекаемого отрезка, или  $y_i - \hat{y}_i + \bar{y}$ ;

**многоуровневая регрессия и постстратификация (MRP):** современный способ создания выборки, при котором из многих областей берутся достаточно небольшие количества респондентов с похожими характеристиками. Затем строится регрессионная модель для откликов в соответствии с демографическими факторами, что допускает дополнительный разброс между областями. Знание демографии для всех областей позволяет делать прогнозы на местном и национальном уровне с соответствующей неопределенностью;

**множественная проверка гипотез:** выполнение сразу нескольких проверок, что увеличивает вероятность получения хотя бы одного ложноположительного результата (ошибка первого рода);

**мода (вероятностного распределения):** для дискретного распределения – самое вероятное значение, для непрерывного – точка максимума плотности;

**мода (выборки):** значение, которое встречается в выборке чаще всего;

**мощность критерия:** вероятность правильного отклонения нулевой гипотезы при условии справедливости альтернативной гипотезы. Равна  $1 - \beta$ , где  $\beta$  – вероятность ошибки второго рода для статистического критерия;

**мудрость толпы:** идея, согласно которой характеристика, определяемая групповым мнением, ближе к истине, чем предположения большинства отдельных людей;

**наука о данных:** изучение и применение методов получения информации из данных, включая построение алгоритмов для прогнозов. Традиционная статистика – часть науки о данных, в которую также входят кодирование и управление данными;

**независимая (предикторная) переменная:** переменная, которая фиксируется посредством проекта или наблюдения, чья связь с зависимой переменной может представлять интерес;

**независимые события:** события  $A$  и  $B$  независимы, если наступление  $A$  не влияет на вероятность наступления  $B$ , то есть

$(B|A) = p(B)$ , или, что эквивалентно,  $p(BA) = p(B)p(A)$ <sup>[280]</sup>;

**непрерывная случайная величина:** случайная величина  $X$ , которая может (по крайней мере, в принципе) принимать любое значение в пределах определенного промежутка. Непрерывная величина имеет плотность вероятности<sup>[281]</sup> – такая функция  $f$ , что

$$P(X \leq x) = \int_{-\infty}^x f(t) dt$$

, а ее математическое ожидание определяется формулой

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx$$

. Вероятность

того, что  $X$  попадет в промежуток  $(A, B)$ , равна  $\int_A^B f(x) dx$ ;

**нормальное распределение:** случайная величина имеет нормальное (гауссовское) распределение со средним  $\mu$  и дисперсией  $\sigma^2$ , если ее плотность имеет вид

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty.$$

Математическое ожидание  $E(X) = \mu$ , дисперсия  $D(X) = \sigma^2$ , среднеквадратичное отклонение  $SD(X) = \sigma$ .

$$Z = \frac{X - \mu}{\sigma}$$

Стандартизованная случайная величина  $Z$  имеет среднее 0 и дисперсию 1, и тогда говорят, что у нее стандартное нормальное распределение. Функцию распределения для стандартной нормальной величины  $Z$  обозначают  $\Phi$ . Например,  $\Phi(-1) = 0,16$  – это вероятность того, что стандартная гауссовская случайная величина не превосходит  $-1$ , или (что эквивалентно) вероятность того, что произвольная гауссовская случайная величина с параметрами  $\mu$  и  $\sigma$  принимает значение, которое меньше  $\mu - \sigma \cdot 100p\%$ .% процентиль для стандартного нормального распределения – такое число  $z_p$ , что  $P(Z \leq z_p) = p$ . Как значения функции  $\Phi$ , так и величины  $z_p$  можно найти в таблицах или в стандартных программах: например, 75-й процентиль для стандартного нормального распределения равен  $z_{0,75} = 0,67$ ;

**нулевая гипотеза:** принимаемое по умолчанию теоретическое предположение, как правило, означающее отсутствие эффекта или результата, проверяемое с помощью Р-значения. Обычно обозначается  $H_0$ ;

**обратная причинная зависимость:** когда связь между двумя переменными изначально кажется причинно-следственной, а на деле причинно-следственные отношения оказываются обратными. Например, у людей, которые не употребляют алкоголь, показатели здоровья хуже, чем у умеренно пьющих, однако как минимум частично это объясняется тем, что некоторые ныне непьющие бросили пить из-за ухудшения здоровья;

**обучение без учителя:** определение классов на основании случаев без подтвержденного состава с использованием какой-либо формы процедуры кластеризации;

**обучение с учителем:** построение алгоритма классификации на основании случаев с подтвержденным составом классов;

**объективное априорное распределение:** попытка устранить субъективный компонент в байесовском анализе, заранее определив априорные распределения, которые должны отражать наше незнание параметров, а данные должны говорить за себя. Нет никакой общей процедуры для определения таких априорных распределений;

**односторонние и двусторонние Р-значения:** значения, соответствующие односторонним и двусторонним критериям;

**односторонние и двусторонние критерии:** односторонний критерий для проверки гипотезы используется тогда, когда нулевая гипотеза, например, указывает, что эффект медицинского вмешательства отрицателен. Эта гипотеза отвергается только в случае, если наблюдаются большие положительные значения тестовой статистики, выражающие оценку эффекта вмешательства. Двусторонний критерий уместен, когда нулевая гипотеза говорит, что эффект медицинского вмешательства равен в точности нулю. Тогда к отказу от такой гипотезы ведут и большие положительные, и большие отрицательные значения тестовой статистики;

**ожидаемые частоты:** количество событий, которые должны произойти в будущем в соответствии с принятой вероятностной моделью;

**остаток:** разность между наблюдаемым значением и значением, предсказываемым статистической моделью;

**относительный риск:** если в группе людей, подвергавшихся какому-то воздействию, абсолютный риск равен  $p$ , а в группе людей, не подвергавшихся этому воздействию, абсолютный риск равен  $q$ , то относительный риск определяется как  $p / q$ ;

**отношение показателей:** относительное увеличение ожидаемого числа событий за определенный период времени, связанное с каким-либо воздействием. Пуассоновская регрессия – это форма множественной регрессии, когда переменная отклика представляет собой наблюдаемый показатель, а коэффициенты соответствуют  $\log(\text{отношение показателей})$ ;

**отношение правдоподобия:** мера относительного подтверждения, которое дают данные для двух конкурирующих гипотез. Для гипотез  $H_0$  и  $H_1$  отношение правдоподобия при данных  $x$  определяется формулой  $p(x|H_0) / p(x|H_1)$ ;

**отношение рисков:** при анализе времени выживания – связанный с воздействием относительный риск пережить какое-то событие за определенный промежуток времени. Регрессия Кокса – это форма множественной регрессии, когда переменная отклика – это время выживания, а коэффициенты соответствуют  $\log(\text{отношение рисков})$ ;

**ошибка второго рода:** происходит, когда альтернативная гипотеза верна, но после проверки нулевая гипотеза не отвергается, то есть делается ложноотрицательное утверждение;

**ошибка первого рода:** происходит, когда ошибочно отклоняется верная нулевая гипотеза в пользу альтернативы, то есть делается ложноположительное утверждение;

**ошибка прокурора:** когда малая вероятность факта при условии невиновности ошибочно истолковывается как вероятность невиновности при условии наличия данного факта;

**параметры:** неизвестные величины в статистической модели, обычно обозначаемые греческими буквами;

**перекрестная проверка:** способ оценивания качества алгоритма для прогноза или классификации путем нескольких выделений части случаев в качестве тестового набора;

**переобучение (переподгонка):** построение статистической модели, которая чрезмерно адаптирована к тренировочному набору

данных, из-за чего ее прогнозные возможности начинают ухудшаться;

**пиктографические диаграммы:** графическое отображение величин с помощью небольших изображений, например изображений людей;

**Пирсона коэффициент корреляции:** если у нас есть  $n$  пар чисел  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  и  $\bar{x}, s_x$  – это выборочное среднее и среднеквадратичное отклонение для чисел  $x$ , а  $\bar{y}, s_y$  – это выборочное среднее и среднеквадратичное отклонение для чисел  $y$ , то коэффициент корреляции Пирсона определяется формулой

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

Предположим, что  $x$  и  $y$  стандартизованы до Z-оценок  $u$  и  $v$  соответственно, то есть  $u_i = (x_i - \bar{x}) / s_x$ , а  $v_i = (y_i - \bar{y}) / s_y$ .

$$\sum_{i=1}^n u_i v_i$$

Тогда коэффициент корреляции Пирсона можно выразить как  $\frac{\sum_{i=1}^n u_i v_i}{n}$ , то есть прямого произведения Z-оценок;

**плацебо:** пустое вещество (например, таблетка с сахаром), которое дают контрольной группе в рандомизированном клиническом испытании под видом реального лечения;

**погрешность:** правдоподобный промежуток, в котором может лежать истинная характеристика популяции. Часто используются 95-процентные доверительные интервалы, которые примерно заключают промежуток  $\pm 2$  стандартных ошибки, но иногда используются «усы» (планки погрешностей), отображающие  $\pm 1$  стандартную ошибку;

**подтверждающие исследования и анализы:** строгие исследования, в идеале выполняющиеся с заранее утвержденным протоколом в целях подтверждения или опровержения гипотез, выдвинутых в ходе «поисковых» исследований или анализов;

**поисковые исследования и анализы:** первоначальные гибкие исследования, которые допускают адаптивные изменения в планах и анализе в целях поиска многообещающих результатов и предназначены

для того, чтобы генерировать гипотезы, которые будут проверяться последующими подтверждающими исследованиями;

**поперечное исследование:** исследование, в котором анализ основан исключительно на текущем состоянии участников, без какого-либо последующего наблюдения в течение долгого времени;

**поправка/стратификация:** включение в регрессионную модель известных возмущающих факторов, которые не представляют прямого интереса, но позволяют провести более сбалансированное сравнение между группами; при этом можно надеяться, что оцененные эффекты, связанные с объясняющими переменными, должны быть ближе к причинной связи;

**последовательное тестирование:** когда какая-либо статистическая проверка повторно проводится на накапливающихся данных, что повышает вероятность появления в какой-то момент ошибки первого рода. Если процесс продолжается достаточно долго, гарантируется «значимый результат»;

**правдоподобие:** мера подтверждения, обеспечиваемая данными для конкретных значений параметра. Когда вероятностное распределение какой-либо случайно величины зависит от параметра, например  $\theta$ , то после наблюдения данных  $x$  правдоподобие для  $\theta$  пропорционально  $p(x|\theta)$ ;

**практическая значимость:** когда какой-нибудь результат имеет реальную важность. Масштабные исследования могут давать результаты, которые статистически значимы, но не имеют практической значимости;

**предсказательная аналитика:** использование данных в целях создания алгоритмов для прогнозов;

**проверка гипотезы:** формальная процедура для оценки подтверждения гипотезы имеющимися данными. Обычно представляет собой сочетание классических фишеровских критериев для проверки нулевой гипотезы с помощью  $P$ -значения и конструкции Неймана – Пирсона, где фигурируют нулевая и альтернативная гипотезы и ошибки первого и второго рода;

**проспективное когортное исследование:** когда выбирается множество испытуемых, измеряются фоновые факторы, а затем за ними следят и наблюдают за соответствующими результатами. Такие

исследования – продолжительные и дорогостоящие и могут не идентифицировать многие редкие события;

**процентиль (выборки):** если взять упорядоченный набор данных (вариационный ряд), то, например, 70-й процентиль – это такая величина, что 70 % наблюдений будут меньше ее. В частности, медиана – это 50-й процентиль. При необходимости используется интерполяция;

**процентиль (генеральной совокупности):** например, 70-й процентиль – это такая величина, что с вероятностью 70 % ваше случайное наблюдение будет меньше ее;

**Пуассона распределение:** случайная величина  $X$  имеет пуассоновское распределение с параметром  $\mu > 0$ , если

$$P(X = n) = \frac{\mu^n}{n!} e^{-\mu}, n = 0, 1, 2, \dots$$

Математическое ожидание  $E(X) = \mu$ , дисперсия  $D(X) = \mu$ ;

**размах (выборки):** разность между максимальным и минимальным значением, то есть  $x_{(n)} - x_{(1)}$ ;

**размер критерия:** величина ошибки первого рода в каком-либо статистическом критерии, обычно обозначается  $\alpha$ ;

**рандомизированное контролируемое исследование (РКИ):** эксперимент, в котором люди или иные объекты случайным образом распределяются по различным вмешательствам, и такая случайность гарантирует, что группы будут сбалансированы в отношении известных и неизвестных факторов. Если в дальнейшем группы демонстрируют различные результаты, то либо вмешательство дало эффект, либо произошло какое-то удивительное событие, вероятность которого выражается через  $P$ -значение;

**распределение выборки:** закономерность в наборе числовых или категориальных наблюдений. Также именуется эмпирическим распределением, или распределением данных;

**распределение генеральной совокупности (распределение популяции):** когда она реально существует – закономерность, описывающая потенциальные наблюдения во всей популяции. Также так называется распределение порождающей случайной величины;

**регрессия к среднему (регресс к среднему):** когда в процессе естественных изменений наблюдается возврат от очень больших или малых наблюдений к более умеренным. Это происходит в силу того,

что первоначальные экстремальные величины получались случайным образом, поэтому повторение в той же степени маловероятно;

**регрессия Кокса:** см. [отношение рисков](#);

**ретроспективное когортное исследование:** исследование, в рамках которого набор испытуемых определяется в какой-то момент в прошлом, а их характеристики прослеживаются вплоть до сегодняшнего дня. Такое исследование не требует продолжительного периода наблюдения, но зависит от надлежащих объясняющих переменных, измеренных в прошлом;

**сигнал и шум:** идея, согласно которой наблюдаемые данные включают два компонента: детерминистский сигнал, который нас действительно интересует, и случайный шум, включающий остаточные ошибки. Задача статистики – правильно идентифицировать оба компонента и не принять шум за сигнал;

**Симпсона парадокс:** когда при учете возмущающего фактора видимое направление взаимосвязи становится обратным;

**систематическая ошибка установки:** происходит, когда вероятность включения в выборку человека или наблюдаемой характеристики зависит от какого-то фактора, например, когда в каком-нибудь рандомизированном испытании наблюдение за людьми в испытуемой группе оказывается более тщательным, чем наблюдение за контрольной группой;

**скрытый фактор:** в эпидемиологии – воздействие, которое не определялось, но может быть возмущающим фактором, ответственным за часть наблюдаемой связи. Например, когда в исследовании изучается связь рациона и заболевания, но не учитывается социально-экономическое положение;

**слепой метод:** чтобы избежать предвзятости в оценивании результатов, участвующие в клиническом исследовании не обладают всей информацией. При слепом методе пациенты не знают, какое лечение получают. При двойном слепом методе люди, наблюдающие за больными, тоже не знают, какое лечение те получают. При тройном слепом методе распределение по методам лечения не знают также и статистики, анализирующие данные;

**случайная величина:** переменная величина, принимающая различные значения с какими-то вероятностями. Случайные величины

обычно обозначаются прописными буквами, например  $X$ , в то время как наблюдаемые значения обозначаются  $x$ ;

**случайный разброс:** неизбежные различия, возникающие при измерениях и наблюдениях; некоторый из них могут объясняться известными факторами, а оставшиеся приписываются случайному шуму;

**специфичность:** доля «отрицательных» случаев, которые правильно определены при классификации или тестировании. Единица минус специфичность – это доля ложноположительных наблюдений (ошибка первого рода);

**Спирмена ранговый коэффициент корреляции:** ранг наблюдения – это его номер в вариационном ряду (упорядоченном наборе), причем равным величинам приписывается одинаковый средний ранг. Например, если у нас есть набор данных (3, 2, 1, 0, 1), то соответствующий набор рангов – (5, 4, 2,5, 1, 2,5). Ранговый коэффициент корреляции Спирмена – это просто коэффициент корреляции Пирсона, в котором наборы  $x$  и  $y$  заменены их соответствующими рангами;

**среднее (выборки):** 1) в широком смысле – общий термин для какой-то одной величины, характеризующей набор чисел, например среднее арифметическое, медиана или мода; 2) в узком смысле – то же, что среднее арифметическое (также говорят выборочное среднее). Предположим, что у нас есть выборка (набор чисел)  $x_1, x_2, \dots, x_n$ . Тогда их выборочное среднее определяется формулой  $m = (x_1 + x_2 + \dots + x_n)/n$ ,

$$m = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

что можно записать в виде  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ . Например, если пять человек сообщили о количестве своих детей и получилась выборка 3, 2, 1, 0, 1, то среднее число детей равно  $(3 + 2 + 1 + 0)/5 = 7/5 = 1,4$ ;

**среднее (популяции):** см. [математическое ожидание](#);

**среднеквадратичная ошибка:** мера качества прогноза; если спрогнозированы значения  $t_1, t_2, \dots, t_n$ , а сделаны наблюдения  $x_1, x_2, \dots,$

$x_n$ , то среднеквадратичная ошибка равна  $\frac{1}{n} \sum_{i=1}^n (x_i - t_i)^2$  ;

**среднеквадратичное (стандартное) отклонение:** квадратный корень из дисперсии выборки или распределения. Для хорошо себя

ведущих разумно симметричных распределений без длинных хвостов можно ожидать, что подавляющее большинство наблюдений будут лежать в пределах двух стандартных отклонений от выборочного среднего;

**стандартная ошибка:** стандартное отклонение выборочного среднего, когда оно рассматривается как случайная величина. Предположим, что  $X_1, X_2, \dots, X_n$  – это независимые одинаково распределенные случайные величины, взятые из распределения со средним  $\mu$  и среднеквадратичным отклонением  $\sigma$ . Тогда их среднее  $Y = (X_1 + X_2 + \dots + X_n)/n$  имеет среднее  $\mu$  и дисперсию  $\sigma^2/n$ . Стандартное отклонение для  $Y$  равно  $\sigma/\sqrt{n}$  и известно как стандартная ошибка. Оценкой будет  $s/\sqrt{n}$ , где  $s$  – выборочное стандартное отклонение для наблюдаемых величин  $X$ ;

**статистика:** 1) дисциплина, занимающаяся изучением мира на основе данных; как правило, включает цикл решения проблем наподобие РРДАС; 2) какая-либо функция от данных. Например, наибольшее значение выборки, выборочное среднее, интерквартильный размах, выборочная дисперсия – различные статистики;

**статистическая значимость:** наблюдаемый эффект считается статистически значимым, когда Р-значение, соответствующее нулевой гипотезе, меньше некоторого заранее установленного уровня, например 0,05 или 0,001. Это означает, что такой экстремальный результат маловероятен при справедливости нулевой гипотезы и всех прочих предположениях при моделировании;

**статистическая модель:** математическое представление вероятностного распределения какого-либо набора случайных величин, содержащее неизвестные параметры;

**статистическое заключение:** процесс использования данных выборки, для того чтобы что-либо узнать о неизвестных параметрах, лежащих в основе статистической модели;

**стохастическая неопределенность:** неизбежная непредсказуемость будущего, также известная как случайность, случай и так далее;

**судебная эпидемиология:** использование знаний о причинах заболеваний в популяциях при вынесении суждений о случаях болезни у отдельных людей;

**счетные переменные:** переменные, которые могут принимать целочисленные значения 0, 1, 2 и так далее или быть взаимнооднозначно сопоставлены с такими значениями;

**тест перестановки/рандомизации:** форма критерия для проверки гипотезы, когда распределение тестовой статистики при нулевой гипотезе получается не с помощью детальной статистической модели для случайных величин, а путем перестановки «меток» данных. Предположим, что нулевая гипотеза такова: какая-то «метка» (например, мужчина это или женщина) не связана с результатом обследования. Тесты рандомизации исследуют все возможные способы перестановки таких меток для отдельных элементов данных, при этом при нулевой гипотезе все они равновероятны. Для каждой перестановки вычисляется тестовая статистика, а Р-значение определяется как доля тех перестановок, где получаются более экстремальные значения тестовой статистики, нежели реально наблюдаемые;

**уровень ложноположительных результатов:** при проверке многих гипотез доля положительных утверждений, которые оказываются ложноположительными;

**фрейминг:** выбор способа подачи информации, влияющего на впечатление аудитории;

**центральная предельная теорема:** общее название нескольких теорем, утверждающих, что при определенных условиях выборочное среднее для множества случайных величин сходится к нормальному распределению вне зависимости (за некоторыми исключениями) от исходного распределения этих случайных величин. Если у нас есть  $n$  независимых наблюдений с математическим ожиданием  $\mu$  и дисперсией  $\sigma^2$ , то при широких условиях их выборочное среднее является оценкой для  $\mu$  и приближенно имеет нормальное распределение со средним  $\mu$ , дисперсией  $\sigma^2/n$  и среднеквадратичным отклонением  $\sigma/\sqrt{n}$  (также известным как стандартная ошибка оценки);

**цикл PPDAC:** предлагаемая структура «цикла данных», куда входят *проблема, планирование, сбор данных, анализ* (поисковый или подтверждающий), *заключение* и *коммуникация*;

**чувствительность:** доля «положительных» случаев, которые правильно определены при классификации или тестировании; часто называется долей истинно положительных наблюдений. Единица минус

чувствительность – это доля ложноотрицательных наблюдений (ошибка второго рода);

**шансы, отношения шансов:** если вероятность какого-то события равна  $p$ , то шансы для такого события определяются как  $\frac{p}{1-p}$ . Если шансы для какого-то события в группе с воздействием равны  $\frac{p}{1-p}$ , а шансы в группе без воздействия –  $\frac{q}{1-q}$ , то отношение шансов составит  $\frac{p}{1-p} / \frac{q}{1-q}$ . Если  $p$  и  $q$  малы, то отношение шансов близко к относительному риску  $p/q$ , но если абсолютные риски значительно превышают 20 %, то отношения шансов и относительные риски начинают различаться;

**эпидемиология:** изучение скорости распространения и причин заболеваемости;

**эпистемическая неопределенность:** недостаток знаний о фактах, числах или научных гипотезах.

## Благодарности

Все идеи, возникающие в ходе долгой карьеры в статистике, – результат вдохновляющих бесед с коллегами. Хотя перечислить всех, у кого я их позаимствовал, сложно даже мне как статистику, короткий список я все же приведу, это Ники Бест, Шейла Бёрд, Дэвид Кокс, Филип Дэвид, Стивен Эванс, Эндрю Гельман, Тим Харфорд, Кевин Макконвей, Уэйн Олдфорд, Сильвия Ричардсон, Этан Шах, Адриан Смит и Крис Вайлд. Я искренне благодарен вам и многим другим людям за поддержку и вдохновение.

Из-за моей хронической прокрастинации процесс написания книги сильно затянулся. Поэтому я хотел бы поблагодарить Лору Стикни из издательства Penguin не только за заказ книги, но и за проявленное в течение всего этого времени спокойствие, сохраняемое даже тогда, когда работа была закончена, но мы все никак не могли договориться о названии. Спасибо Джонатану Пеггу за создание хороших условий, Джейн Бёрдселл за колоссальное терпение при редактировании и всему персоналу Penguin за кропотливую работу.

Я крайне признателен за разрешение использовать иллюстрации следующим людям и организациям: Крису Вайлду (рис. 0.3), Джеймсу Грайму (рис. 2.1), Кэт Мерсер из Natsal (рис. 2.4 и 2.10), Национальной статистической службе Великобритании (рис. 2.9, 8.5 и 9.4), Службе общественного здравоохранения Англии (рис. 6.7), Полу Бардену (рис. 9.2) и «Би-би-си» (рис. 9.3). Общественные данные о Великобритании используются в соответствии с Открытой государственной лицензией, версия 3.0.

Поскольку я не особо хорошо программирую на R<sup>[282]</sup>, Мэтью Пирс и Мария Сколариду очень помогли мне с выполнением анализов и построением графиков. Я также не особо силен в писательстве, поэтому в неоплатном долгу перед многочисленными людьми, которые читали текст и делали замечания. Среди них Джордж Фармер, Алекс Фримэн, Кэмерон Брик, Майкл Поснер, Сандер ван дер Линден и Симона Варр; отдельное спасибо Джулиану Гилби за поиск ошибок и двусмысленностей.

Кроме того, я должен поблагодарить Кейт Булл не только за важные комментарии по тексту, но и за поддержку как в хорошие (когда

я писал в пляжной хижине на Гоа), так и в плохие (в сыром феврале под давлением чрезмерного количества обязательств) времена. Также я глубоко признателен Дэвиду и Клаудии Хардинг за финансовую поддержку и постоянное подбадривание, что позволило мне заниматься интересными вещами в последние десять лет.

Наконец, как бы мне ни хотелось возложить вину на кого-нибудь другого, я должен взять всю ответственность на себя за неизбежные оставшиеся в книге недостатки.

### ***КОД ПРИМЕРОВ***

Код на языке R и данные для воспроизведения большей части анализов и рисунков можно найти на сайте <https://github.com/dspiegel29/ArtofStatistics>. Я благодарен за помощь при подготовке этих материалов.

# Список иллюстраций

- [0.1. Возраст и год смерти жертв Гарольда Шипмана](#)
- [0.2. Время смерти жертв Гарольда Шипмана](#)
- [0.3. Цикл решения задач PPDAC](#)
- [1.1. Уровень выживаемости за 30 дней после операции на сердце](#)
- [1.2. Процентные доли операций на сердце у детей для разных больниц – круговая диаграмма](#)
- [1.3. Процентные доли операций на сердце у детей для разных больниц – гистограмма](#)
- [1.4. Риск употребления бутербродов с беконом](#)
- [2.1. Банка с жевательными драже](#)
- [2.2. Различные способы показать догадки о количестве драже в банке](#)
- [2.3. Графическое изображение догадок о числе драже в банке в логарифмическом масштабе](#)
- [2.4. Сообщенное количество партнеров противоположного пола за всю жизнь](#)
- [2.5. Показатели выживаемости в зависимости от количества операций на сердце у детей](#)
- [2.6. Нулевые коэффициенты корреляции Пирсона](#)
- [2.7. Тренды мирового населения](#)
- [2.8. Относительный рост населения по странам](#)
- [2.9. Популярность имени Дэвид во времени](#)
- [2.10. Инфографика по сексуальным отношениям и образу жизни](#)
- [3.1. Диаграмма индуктивного умозаключения](#)
- [3.2. Распределение новорожденных по массе](#)
- [5.1. Разброс роста сыновей в зависимости от роста их отцов](#)
- [5.2. Модель логистической регрессии для данных об операциях на сердце](#)
- [6.1. Надгробие жертвы «Титаника»](#)
- [6.2. Сводные данные о выживании для пассажиров «Титаника»](#)
- [6.3. Дерево классификации для данных о «Титанике»](#)
- [6.4. ROC-кривые для алгоритмов, использованных для тренировочного и тестового наборов данных](#)
- [6.5. Вероятности выживания при катастрофе «Титаника»](#)

6.6. Дерево классификации для данных о «Титанике», слишком хорошо подогнанное к данным

6.7. Уровни выживаемости после операции для женщин с раком груди

7.1. Эмпирическое распределение числа сексуальных партнеров для разных размеров выборки

7.2. Бутстрэп-выборки из первоначальной выборки размера 50

7.3. Распределение средних значений для бутстрэп-выборок разных размеров

7.4. Регрессионные прямые для данных Гальтона по росту матерей и дочерей

8.1. Компьютерное моделирование вариантов игр шевалье де Мере

8.2. Дерево ожидаемых количеств для подбрасывания двух монет

8.3. Дерево вероятностей для подбрасывания двух монет

8.4. Дерево ожидаемых количеств для скрининга при раке молочной железы

8.5. Наблюдаемое и ожидаемое количество случаев убийства

9.1. Вероятностное распределение левшей

9.2. Воронкообразный график для уровней смертности от колоректального рака

9.3. График данных опросов общественного мнения «Би-би-си» перед всеобщими выборами 2017 года

9.4. Число ежегодных убийств в Англии и Уэльсе

10.1. Данные о соотношении полов при процедуре крещения в Лондоне

10.2. Эмпирическое распределение разницы между долями женщин и мужчин, которые при скрещивании рук поверх кладут правую/левую руку

10.3. Совокупное количество свидетельств о смерти, подписанных Шипманом

10.4. Статистика для последовательного критерия отношения вероятностей для обнаружения удвоения риска смертности

10.5. Ожидаемые количества для результатов 1000 проверок гипотез

11.1. Дерево ожидаемых количеств для задачи про три монеты

11.2. Дерево ожидаемых количеств для задачи о допинге

11.3. «Обращенное» дерево ожидаемых количеств для задачи о допинге

11.4. «Бильярдный» стол Байеса

12.1. Традиционные информационные потоки для статистических сведений

## Список таблиц

- [1.1. Результаты после детских сердечных операций](#)
- [1.2. Способы выразить риск рака кишечника при ежедневном употреблении сэндвича с беконом](#)
- [2.1. Характеристики выборки для предположений о числе драже в банке](#)
- [2.2. Характеристики выборки числа сексуальных партнеров за всю жизнь](#)
- [4.1. Результаты Исследования защиты сердца](#)
- [4.2. Иллюстрация парадокса Симпсона](#)
- [5.1. Данные о росте родителей и их взрослых детей](#)
- [5.2. Корреляция между ростом взрослых детей и родителей того же пола](#)
- [5.3. Результаты множественной линейной регрессии для роста взрослых потомков в зависимости от роста их отцов и матерей](#)
- [6.1. Матрица ошибок для дерева классификации для тренировочных и тестовых данных о «Титанике»](#)
- [6.2. Вымышленные прогнозы «вероятности осадков»](#)
- [6.3. Результаты логистической регрессии для данных о выживании пассажиров «Титаника»](#)
- [6.4. Эффективность различных алгоритмов для тестового набора данных о «Титанике»](#)
- [6.5. Доля выживаемости женщин с раком молочной железы, определенная с помощью алгоритма Predict 2.1](#)
- [7.1. Характеристики для числа сексуальных партнеров за всю жизнь, которое указывали мужчины](#)
- [7.2. Выборочные средние для числа сексуальных партнеров за всю жизнь, которое указывали мужчины](#)
- [9.1. Сравнение точных доверительных интервалов и для бутстрэппинга](#)
- [10.1. Таблица сопряженности полов и положения рук](#)
- [10.2. Наблюдаемые и ожидаемые количества людей, держащих сверху правую или левую руку, в зависимости от пола](#)
- [10.3. Наблюдаемые и ожидаемые количества дней с определенным числом случаев убийств](#)

10.4. Результаты Исследования защиты сердца с доверительными интервалами и Р-значениями

10.5. Результат, выдаваемый программой R для множественной линейной регрессии по данным Гальтона

10.6. Возможные результаты проверки гипотезы

11.1. Отношения правдоподобия для отдельных фактов-доказательств в отношении скелета Ричарда III

11.2. Рекомендуемые словесные интерпретации для отношений правдоподобия

11.3. Шкала Касса и Рафтери для интерпретации коэффициентов Байеса

13.1. Прогнозы числа мест, полученных каждой партией на трех последних национальных выборах

## Эту книгу хорошо дополняют:

### [Голая статистика](#)

Чарльз Уилан

### [Теория игр](#)

Авинаш Диксит, Барри Нейлбафф

### [Аналитическая культура](#)

Карл Андерсон

### [Удовольствие от \$x\$](#)

Стивен Строгац

---

**notes**

# **СНОСКИ**

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

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

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>.

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

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.*

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

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.

E. J. Evans, *Thatcher and Thatcherism* (Routledge, 2013), p. 30.

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

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

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

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.

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

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.

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>.

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.*

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.*

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

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

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

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

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.

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.

Данные о результатах выживания детей, перенесших операции на сердце, в Соединенном Королевстве Великобритании и Северной Ирландии можно получить на сайте <http://childrenshearturgery.info/>.

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

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).

Индекс массы тела разработан бельгийским статистиком и социологом Адольфом Кетле в 1830-х годах. Он определяется так:  $\text{ИМТ} = \text{масса (кг)} / \text{рост}^2 \text{ (м)}$ . 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 Ансея Киза с соавторами, опубликованной в 1972 году в *Journal of Chronic Diseases*. *Прим. пер.*

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.

Строго говоря, относительное увеличение на 18 % дает  $6 \times 1,18 = 7,08$  процента, но для наших целей округления до 7 % вполне достаточно.

Это было любимое наблюдение Ханса Рослинга, см. [следующую главу](#).

E. A. Akl et al., 'Using Alternative Statistical Formats for Presenting Risks and Risk Reductions', *Cochrane Database of Systematic Reviews* 3 (2011).

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.

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.*

‘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.

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.

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

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.

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

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.

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

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

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

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.

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.

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  $\alpha$ , то говорят об  $\alpha$ -квантиле. *Прим. пер.*

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.*

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.*

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

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.*].

C. H. Mercer et al., 'Changes in Sexual Attitudes and Lifestyles in Britain through the Life Course and Over Time: Findings from the National Surveys of Sexual Attitudes and Lifestyles (Natsal)', *The Lancet* 382 (2013), 1781–94. Красочное рассмотрение статистики о сексе см. в работе: D. Spiegelhalter, *Sex by Numbers* (Wellcome Collection, 2015).

Множество всех мужчин и множество всех женщин имеют одно и то же количество связей, поскольку каждая связь включает одного мужчину и одну женщину. Поэтому, если мужчин и женщин поровну, то и среднее число связей, приходящихся на них, должно быть одинаково. Когда я объясняю это в школах, я использую пример с рукопожатиями или партнерами по танцу.

Хотя общие показатели выживаемости на двух диаграммах напрямую сравнивать нельзя (из-за разных возрастных групп детей), фактически выживаемость детей всех возрастов за эти двадцать лет повысилась с 92 % до 98 %.

Английский математик Карл Пирсон был сторонником всего немецкого: он даже изменил написание своего имени с Carl на Karl. Впрочем, это не помешало ему применять статистику в баллистике во время Первой мировой войны. В 1911 году он основал первый в мире факультет статистики в Университетском колледже Лондона, а также возглавил евгеническую лабораторию, финансируемую по завещанию Гальтона.

Согласно теории Спирмена, любая интеллектуальная деятельность определяется двумя факторами – общим (G) и специфическим (S).  
Общий фактор – основа всех умственных действий. *Прим. пер.*

A. Cairo, 'Download the Datasaurus: Never Trust Summary Statistics Alone; Always Visualize Your Data', <http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html>.

Альберто Каиро придумал тринадцать наборов точек, которые изображают звезду, динозавра, крест, ряды линий и так далее. При этом для всех рисунков средние значения и среднеквадратичные отклонения для обеих координат этих точек практически одинаковы, а коэффициент корреляции везде примерно равен 0. С помощью этого примера Каиро демонстрирует, что выборочное среднее и среднеквадратичное отклонение не описывают выборку в достаточной степени, поэтому всегда нужно визуализировать данные. Другой известный подобный пример – так называемый квартет Энскомба, предложенный в 1973 году английским статистиком Фрэнком Энскомбом. Это четыре набора из одиннадцати пар чисел с одинаковыми средними значениями переменной  $x$ , переменной  $y$ , дисперсии  $x$ , дисперсии  $y$ , корреляции между  $x$  и  $y$  и прямой линейной регрессии. Однако расположение точек на соответствующих четырех рисунках различно. *Прим. пер.*

Показатели выживаемости соответствуют различным количествам операций и потому подвержены разной степени изменчивости в силу воздействия случайных факторов. Поэтому, хотя для описания какого-то набора данных и можно посчитать коэффициент корреляции, формальный вывод должен учитывать, что эти данные являются долями. В главе 6 я покажу, как это делать.

<https://esa.un.org/unpd/wpp/Download/Standard/Population/>.

Перечень популярных имен, согласно данным Национальной статистической службы, содержится по адресу:  
<https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/babynamesenglandandwales/2015>.

При рождении будущего короля назвали Эдуард Альберт Кристиан Джордж Эндрю Патрик Дэвид, но он предпочитал имя Дэвид, которым всю жизнь называли его друзья. *Прим. пер.*

I. D. Hill, 'Statistical Society of London – Royal Statistical Society: The First 100 Years: 1834–1934', *Journal of the Royal Statistical Society: Series A (General)* 147:2 (1984), 130–39.

<http://www.natsal.ac.uk/media/2102/natsal-infographic.pdf>.

TED (Technology, Entertainment, Design) – американская организация, миссия которой – проведение ежегодных конференций под лозунгом «Идеи, достойные распространения». *Прим. пер.*

H. Rosling, Unveiling the Beauty of Statistics for a Fact-Based World View, доступно на [www.gapminder.org](http://www.gapminder.org).

К сожалению, книга по статистике с иллюстрациями в оттенках серого не подходит для иллюстрации его работы, поэтому я могу только порекомендовать заглянуть на сайт [gapminder.org](http://gapminder.org). Однажды Рослинг спорил на телевидении с датским журналистом, который бездумно повторил какое-то утверждение о мире, которое Ханс пытался опровергнуть всю жизнь. Рослинг просто сказал: «Эти факты не подлежат обсуждению. Я прав, а вы нет». Для статистики это необычно прямое заявление.

Некоторые доказательства такого искажения были получены в рандомизированном эксперименте с участием студентов в США, где женщины, подсоединенные к детектору лжи, как правило, признавали большее количество партнеров, в отличие от женщин, которым гарантировалась анонимность. В то же время у мужчин такого эффекта не наблюдалось. Участникам не сообщали, что детектор лжи ненастоящий.

Артур Конан Дойль ошибался: методы Холмса не имели ничего общего с дедукцией. Его рассуждения – абдукция. Дедукция – это переход от общих посылок к частным следствиям. Классический пример: 1) все люди смертны; 2) Сократ – человек; 3) следовательно, Сократ смертен. Дедукция гарантирует истинность заключения, если истинными были исходные посылки. При абдукции у нас есть заключение, а мы восстанавливаем какую-нибудь посылку. Например, если к нам летит футбольный мяч, мы делаем абдуктивное заключение, что по мячу кто-нибудь ударил. Или пусть у нас есть первая посылка «Все люди смертны» и заключение «Сократ смертен». Тогда мы предполагаем, что вторая посылка – «Сократ – человек». Абдукция вовсе не гарантирует истинности нашего заключения (например, мячом могли выстрелить из специальной пушки, да и если на первой странице детектива какого-то человека застали над трупом с пистолетом, то, скорее всего, окажется, что он как раз и не убийца). Однако абдуктивные рассуждения дают определенный ориентир, позволяя выдвигать разумные гипотезы. Выдающийся логик Чарльз Пирс полагал, что дедукция, индукция и абдукция – три основных вида элементарных рассуждений. *Прим. пер.*

Индукция может быть полной и неполной. Полная индукция гарантирует истинность заключения, неполная – нет. Вот пример полной индукции. Предположим, в классе 30 человек, и все сдавали экзамен. Если у вас есть 30 посылок вида «Александр сдал экзамен», «Мария сдала экзамен» и аналогичные утверждения для всех остальных учеников, то вы можете сделать вывод: «Весь класс сдал экзамен», и это будет истинным заключением. Однако в большинстве случаев индукция является неполной – вам известно, что какой-то признак есть только у части элементов множества, и вы делаете вывод, что он имеется у всех его элементов. В этом случае истинность заключения не гарантируется. Например, если у вас есть информация только о 25 школьниках, сдавших экзамен, то вы можете предположить, что его сдали все 30 учеников, но это заключение уже носит вероятностный характер. *Прим. пер.*

Такая структура из четырех этапов украдена у Уэйна Олдфорда.

После того как кто-то из Королевского статистического общества раскритиковал такие методы опроса, представитель руководства Ryanair Майкл О'Лири заявил: «95 % клиентов Ryanair никогда не слышали о Королевском статистическом обществе, 97 % не волнует, что там говорят, и 100 % сказали, что это звучит так, будто его участникам нужно забронировать недорогой отпуск с Ryanair». В другом современном исследовании Ryanair была признана худшей из двадцати европейских авиакомпаний (но у этого опроса свои проблемы с надежностью, поскольку он проводился как раз в то время, когда Ryanair отменила большое количество рейсов).

Ipsos MORI, What the UK Thinks (2015),  
<https://whatukthinks.org/eu/poll/ipsos-mori-141215>.

Сообщено в More or Less, 5 October 2018;  
<https://www.bbc.co.uk/programmes/p06n2lmp>. Классическая  
демонстрация этого эффекта представлена в британском комедийном  
сериале «Да, господин министр», когда высокопоставленный чиновник  
сэр Хамфри Эпплби показывает, как подходящие наводящие вопросы  
могут привести к любому желаемому ответу. Этот пример сейчас  
используется в учебных методах.  
<https://researchmethodsdataanalysis.blogspot.com/2014/01/leading-questions-yes-prime-minister.html>.

Иисус же говорил: Отче, прости им, ибо не знают, что делают. И делили одежды Его, бросая жребий (Лк 23:34).

Не следует путать с гаданием [в оригинале используются сходные английские слова *sortition* и *sortilege*. *Прим. пер.*], представляющим собой форму предсказания, в которой очевидно случайные явления используются для определения божественной воли или будущего, что также известно как клеромантия. Примеры существуют во многих культурах, включая гадание с помощью чайных листьев, куриных внутренностей, библейское бросание жребия для определения воли Божьей или гадание по «Книге перемен» («И-Цзин»).

Видеозапись лотереи для вьетнамской войны ищите по адресу:  
[https://www.youtube.com/watch?v=-p5X1FjyD\\_g](https://www.youtube.com/watch?v=-p5X1FjyD_g); см. также  
<http://www.historynet.com/whats-your-number.htm>.

Это как раз и означает возможность распространения результатов конкретного исследования на более широкий класс объектов или ситуаций. В реальности обобщать полученный результат на любые популяции, любые условия и любое время вряд ли реально, поэтому говорят только о некоторой степени соблюдения внешней валидности.

*Прим. пер.*

Подробную информацию об «Опросе о преступности в Англии и Уэльсе» и полицейской статистике преступлений можно получить в Национальной статистической службе Великобритании:  
<https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice>.

Информация о весе новорожденных в США:  
[http://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64\\_01.pdf](http://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_01.pdf).

Математик сказал бы, что эта кривая отображает плотность распределения. *Прим. пер.*

Выводы Гаусса не основывались на эмпирических наблюдениях; это была теоретическая форма ошибки измерений, которая оправдывала бы его статистические методы.

Если бы мы записывали массу тела не с шагом в 500 граммов, а более точно, например с шагом в 100 граммов, то гистограмма была бы еще ближе к теоретической плотности распределения. *Прим. пер.*

Для такого отслеживания будут использоваться более сложные распределения, чем нормальное.

‘Why Going to University Increases Risk of Getting a Brain Tumour’, Mirror Online, 20 June 2016. Исходная статья: A. R. Khanolkar et al., ‘Socioeconomic Position and the Risk of Brain Tumour: A Swedish National Population-Based Cohort Study’, *Journal of Epidemiology and Community Health* 70 (2016), 1222–8.

Ошибка обращаемости – это систематическая ошибка, при которой в выборке собраны случаи, не представляющие равным образом все категории генеральной совокупности (популяции). *Прим. пер.*

T. Vigen, <http://www.tylervigen.com/spurious-correlations>.

Размышления о том, что могло бы произойти, но не произошло.  
*Прим. пер.*

Липопротеины низкой плотности (ЛПНП) – белки крови, которые переносят холестерин. Хотя такой белок и называют пренебрежительно «плохим холестерином», на самом деле такие белки вовсе не холестерин, а лишь его переносчики. *Прим. пер.*

‘MRC/BHF Heart Protection Study of Cholesterol Lowering with Simvastatin in 20,536 High-Risk Individuals: A Randomised Placebo-Controlled Trial’, *The Lancet* 360 (2002), 7–22.

Такой метод называется **слепым** (пациенты не знают важных деталей испытания). В нашем случае они не знают, принимают лекарство или плацебо. Если же и экспериментаторы не знают важных деталей испытания, метод называется двойным слепым. *Прим. пер.*

Cholesterol Treatment Trialists' (CTT) Collaborators, 'The Effects of Lowering LDL Cholesterol with Statin Therapy in People at Low Risk of Vascular Disease: Meta-Analysis of Individual Data from 27 Randomised Trials', *The Lancet* 380 (2012), 581–90.

Согласно заключению исследователей, для людей с моим базовым риском и без предшествующего заболевания уменьшение ЛПНП на 1 ммоль/л снижает риск серьезных проблем с сердечно-сосудистой системой на 25 %. После начала приема статинов мой уровень ЛПНП снизился на 2 ммоль/л, а значит, ежедневное употребление статинов изменило мой годовой риск развития инфаркта или инсульта примерно на  $0,75 \times 0,75 = 0,56$ , то есть снизило мой риск на 44 %. Поскольку у меня риск инфаркта или инсульта за 10 лет составлял примерно 13 %, прием статинов уменьшил бы его до 7 %. Это означает, что прописанные мне статины полезны – и хорошо, что я их действительно принимаю.

Мастэктомия – операция по полному удалению молочной железы.  
Лампэктомия – удаление опухоли с частичным удалением здоровой  
ткани. *Прим. пер.*

Возможно, еще более удивительно и отрадно то, что так много людей согласилось участвовать в испытаниях исключительно для пользы будущих пациентов.

Первоначально государственная, а затем частная британская компания, которая разрабатывает методики повышения эффективности управления с помощью психологии. *Прим. пер.*

Испытания организации Behavioural Insights Team описаны на сайтах <http://www.behaviouralinsights.co.uk/education-and-skills/helping-everyone-reach-their-potential-new-education-results/> и <http://www.behaviouralinsights.co.uk/trial-results/measuring-the-impact-of-body-worn-video-cameras-on-police-behaviour-and-criminal-justice-outcomes/>.

H. Benson et al., 'Study of the Therapeutic Effects of Intercessory Prayer (STEP) in Cardiac Bypass Patients: A Multicenter Randomized Trial of Uncertainty and Certainty of Receiving Intercessory Prayer', *American Heart Journal* 151 (2006), 934–42.

Хотя А/В-тестирование широко применяется в веб-дизайне, его можно использовать и в других областях – например, написать два электронных письма с каким-либо предложением двум группам людей.  
*Прим. пер.*

J. Heathcote, 'Why Do Old Men Have Big Ears?', *British Medical Journal* 311 (1995), <https://www.bmj.com/content/311/7021/1668>.  
См. также 'Big Ears: They Really Do Grow as We Age', *The Guardian*, 17 July 2013.

К сожалению, маловероятно, что на подобные исследования найдется финансирование.

На самом деле буква М в аббревиатуре STEM традиционно означает математику (Science, Technology, Engineering and Mathematics). Иногда при добавлении к ним медицины пишут STEMМ. *Прим. пер.*

Назван в честь английского статистика Эдварда Симпсона (1922–2019), описавшего парадокс в 1951 году. Впрочем, об этом эффекте упоминал Карл Пирсон еще в 1899 году и шотландский математик Джордж Удни Юл в 1903-м. *Прим. пер.*

‘Waitrose Adds £36,000 to House Price’, Daily Mail, 29 May 2017.

'Fizzy Drinks Make Teenagers Violent', Daily Telegraph, 11 October 2011.

S. Coren and D. F. Halpern, 'Left-Handedness: A Marker for Decreased Survival Fitness', *Psychological Bulletin* 109 (1991), 90–106. Критику см. в работе 'Left-Handedness and Life Expectancy', *New England Journal of Medicine* 325 (1991), 1041–3.

J. A. Hanley, M. P. Carrieri and D. Serraino, 'Statistical Fallibility and the Longevity of Popes: William Farr Meets Wilhelm Lexis', *International Journal of Epidemiology* 35 (2006), 802–5.

J. Howick, P. Glasziou and J. K. Aronson, 'The Evolution of Evidence Hierarchies: What Can Bradford Hill's "Guidelines for Causation" Contribute?', *Journal of the Royal Society of Medicine* 102 (2009), 186–94.

Менделевская рандомизация использовалась, например, для проверки спорного вопроса о том, приносит ли умеренное потребление алкоголя пользу здоровью. У людей, которые никогда не употребляли алкоголь, как правило, более высокий уровень смертности, чем у умеренно пьющих, но есть разногласия во мнениях, обусловлено это алкоголем или трезвенники менее здоровы по каким-то иным причинам. Одна версия гена связана с пониженной переносимостью алкоголя, поэтому унаследовавшие его люди пьют меньше. Носители гена и те, у кого его нет, должны быть сбалансированы по всем остальным факторам, из-за чего какая-либо систематическая разница в их здоровье может быть приписана именно этому гену – словно в рандомизированном испытании. Исследователи обнаружили, что люди с этим геном, как правило, более здоровы, и пришли к заключению, что в целом алкоголь вреден. Однако для подтверждения такого вывода нужны дополнительные предположения, и споры не утихают. См. Y. Cho et al., ‘Alcohol Intake and Cardiovascular Risk Factors: A Mendelian Randomisation Study’, Scientific Reports, 21 December 2015.

Как писал Гальтон, «я ставлю Лондон на первое место по красоте, а Абердин на последнее».

M. Friendly et al., 'HistData: Data Sets from the History of Statistics and Data Visualization' (2018), <https://CRAN.R-project.org/package=HistData>.

Вот для сравнения российские данные примерно того же времени. Дмитрий Николаевич Анучин приводит величины среднего роста новобранцев в Российской империи для второй половины XIX века: Московская губерния – 164,4 см; Нижегородская – 164,0 см; Варшавская – 162,1 см. Выше всего средний рост призывников в Курляндской губернии – 167,0 см. «Очень высокорослыми рекрутами» именуется люди с ростом выше 177,8 см. (Д. Н. Анучин, «О географическом распределении роста мужского населения России [по данным о всеобщей воинской повинности в Империи за 1874–1883 гг.]: сравнительно с распределением роста в других странах: (с десятью раскрашенными картами)» / [соч.] проф. Д. Н. Анучина. – С.-Петербург: в тип. В. Безобразова и К°, 1889). *Прим. пер.*

Можно было бы проводить прямую, которая минимизирует сумму абсолютных величин этих остатков, а не сумму их квадратов, однако без современных компьютеров это практически невозможно.

То есть возвратом. *Прим. пер.*

Например, мы можем предсказать рост дочери, используя формулу: средний рост всех дочерей +  $0,33 \times (\text{рост матери} - \text{средний рост всех матерей})$ .

См. статью о методе наименьших квадратов в глоссарии.

J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books, 2018), p. 471.

Прекрасное обсуждение рисков моделирования см. в работе: A. Aggarwal et al., ‘Model Risk – Daring to Open Up the Black Box’, *British Actuarial Journal* 21:2 (2016), 229–96.

По сути, мы говорим, что изменения будут коррелировать с исходными измерениями, даже если в реальности никаких изменений в базовом процессе не происходит. Мы можем выразить это математически. Предположим, я беру случайное наблюдение  $X$  из генеральной совокупности с каким-то распределением. Потом беру другое независимое наблюдение  $Y$  с тем же распределением и смотрю на их разность  $Y - X$ . Справедливо замечательное утверждение: коэффициент корреляции между величиной  $X - Y$  и первым наблюдением  $X$  равен  $-1/\sqrt{2} = -0,71$ , причем вне зависимости от распределения генеральной совокупности. Например, если у какой-то женщины есть ребенок, а затем ребенок появляется у ее подруги, то они начинают сравнивать вес детей, вычитая вес второго ребенка из веса первого. Тогда эта разность будет иметь корреляцию  $-0,71$  с весом первого ребенка. Объяснение тут простое: если первый ребенок легкий, то мы, по всей вероятности, можем ожидать, что второй будет тяжелее, поэтому разница будет положительной. А если первый ребенок тяжелый, то мы ожидаем, что второй будет легче, и разница между их весом будет отрицательной.

L. Mountain, 'Safety Cameras: Stealth Tax or Life-Savers?', *Significance* 3 (2006), 111–13.

Слово «линейный» отражает тот факт, что в итоговое уравнение входит линейная комбинация независимых переменных (то есть сумма переменных, умноженных на какие-то коэффициенты). Такая модель называется линейной.

Независимые переменные стандартизированы путем вычитания выборочного среднего. Таким образом, чтобы спрогнозировать рост сына, мы используем формулу:  $69,2 + 0,33 (\text{рост матери} - \text{средний рост матерей}) + 0,41 (\text{рост отца} - \text{средний рост отцов})$ .

Следующая таблица показывает виды множественной регрессии, используемые для различных типов зависимых переменных, а также интерпретацию коэффициента для каждой независимой переменной. (Для просмотра таблицы перейти в текст сноски.)

<b>Тип зависимой переменной</b>	<b>Тип регрессии</b>	<b>Интерпретация коэффициента</b>
Непрерывные переменные	Множественная линейная	Угловой коэффициент
События или доли	Логистическая	Log (отношение шансов)
Количества	Пуассоновская	Log (отношение показателей)
Срок выживаемости	Регрессия Кокса	Log (отношение рисков)

Коэффициент 0,001 логистической регрессии означает, что логарифм для величины шансов смерти понижается примерно на 0,001 на каждого дополнительного пациента в год, то есть на 0,1 на каждых 100 дополнительных пациентов. Это соответствует примерно 10-процентному понижению риска.

Данные о «Титанике» можно найти здесь: <https://bio304-class.github.io/bio304-fall2017/data-story-titanic.html>.

Сюда входили дона (Dona), леди (Lady), графиня (Countess), капитан (Capt), полковник (Col), дон (Don), доктор (Dr), майор (Major), преподобный (Rev.), сэр (Sir), йонкхеер (Jonkheer). [Йонкхеер – в Нидерландах обращение к дворянину без титула. *Прим. пер.*].

Мастер – форма обращения к мальчику или юноше. *Прим. пер.*

Чтобы не заставлять всех ждать окончания конкурса (для данных по «Титанику» это 2020 год), Kaggle делит тестовый набор на две части – открытую и закрытую. Создается таблица лидеров, где отображаются только результаты конкурсантов в открытой части, и этот предварительный рейтинг могут видеть все. Однако итоговым рейтингом участников после окончания конкурса станет эффективность, показанная в закрытой части тестового набора.

В общем случае чувствительность – это доля истинно положительных наблюдений; специфичность – доля истинно отрицательных наблюдений. *Прим. пер.*

Receiver Operating Characteristic – рабочая характеристика приемника. *Прим. пер.*

Может возникнуть соблазн использовать «абсолютную ошибку», а не квадратичную, то есть если мы указываем вероятность 0,1 для несостоявшегося события, то теряем 0,1 (в то время как для квадратов мы теряем 0,01). Но это, казалось бы, невинное изменение будет очень большим просчетом. Довольно простые теоретические рассуждения показывают, что такое «абсолютное» наказание приведет людей к рациональному преувеличению своей уверенности ради минимизации ожидаемой ошибки и указыванию вероятности 0 % для дождя, даже если на самом деле они считают, что она равна 10 %.

Оценка качества работы вычисляется так:  $(BC - B) / BC = 1 - B / BC$ . Отсюда получаем  $1 - 0,11 / 0,28 = 0,61$ .

Проверка вероятности осадков:  
<http://www.cawcr.gov.au/projects/verification/POP3/POP3.html>.

Здесь приставка «пере-» означает избыточность, а не повторность.  
*Прим. пер.*

‘Electoral Precedent’, xkcd, <https://xkcd.com/1122/>.

Настольная интеллектуальная игра в слова, в России больше известна под названием «Эрудит». *Прим. ред.*

В общем случае имеющиеся данные разбивают на  $N$  частей, а затем обучают алгоритм с помощью  $N - 1$  части, а одну часть используют для тестирования. Процедуру повторяют  $N$  раз, каждый раз выбирая в качестве тестовой части одну из  $N$  частей. *Прим. пер.*

Англ. boosting – усиление. *Прим. пер.*

Для перевода оценки  $S$  в вероятность выживания  $p$  используется формула  $p = 1 / (1 + e^{-S})$ . Это преобразование уравнения логистической регрессии  $\log^e p / (1 - p) = S$ .

По сути, метод LASSO (Least absolute shrinkage and selection operator) отбирает наиболее информативные признаки – те, которые оказывают большее влияние на отклик, поскольку для остальных ставит нулевые коэффициенты. *Прим. пер.*

Bagging – сокращение от Bootstrap aggregating, бутстрэп-агрегирование. *Прим. пер.*

Карл Даль родился в Норвегии в 1866 году, но в 1892 году эмигрировал в Австралию. В 1912 году решил вернуться в Норвегию, но, добравшись до Лондона, поменял планы и отправился в США к родственникам. *Прим. пер.*

Из книги *Кэти О'Нил* «Убийственные большие данные. Как математика превратилась в оружие массового поражения», в которой приведено множество примеров неправильного использования алгоритмов. [Издана на русском языке: *О'Нил К.* Убийственные большие данные. Как математика превратилась в оружие массового поражения. М.: АСТ, 2018. *Прим. пер.*].

<http://innovation.uci.edu/2017/08/husky-or-wolf-using-a-black-box-learning-model-to-avoid-adoption-errors/>.

Использование алгоритмов COMPAS и MMR критикуется в книге С. О'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Penguin, 2016).

Также болезнь Гентингтона, хоря Гентингтона или хоря Хантингтона. Названа по имени американского врача Джорджа Хантингтона (1850–1916). При этом заболевании в мозге происходят изменения, которые ведут к изменениям в личности. *Прим. пер.*

Обратный инжиниринг (обратная разработка, обратное проектирование) – исследование некоей системы (устройства, алгоритма, программы), для того чтобы понять схему ее работы. *Прим. пер.*

Иначе – вспомогательная терапия. *Прим. пер.*

NHS, Predict: Breast Cancer (2.1):  
[http://www.predict.nhs.uk/predict\\_v2.1/](http://www.predict.nhs.uk/predict_v2.1/).

Возможно, исключительно для того, чтобы привлечь финансирование.

Когда однажды я предложил группе журналистов четко указывать это в своих статьях, то столкнулся с полнейшим непониманием.

Статистика рынка труда в Великобритании, январь 2018 года:  
<https://www.ons.gov.uk/releases/uklabourmarketstatisticsjan2018>. Bureau  
of Labor Statistics, 'Employment Situation Technical Note 2018',  
<https://www.bls.gov/news.release/empsit.tn.htm>.

Изменения в уровне безработицы, определенные по зарплатным ведомостям, основаны на налоговых декларациях работодателей и несколько более точны, их погрешность составляет  $\pm 100\ 000$ .

Часто их называют псевдовыборками. *Прим. пер.*

Слово *bootstraps* означает ремешки в виде ушка, которые прикрепляются к верхней части обуви, чтобы ее было проще натягивать. В английском языке есть выражение *To pull oneself over a fence by one's bootstraps* (буквально – перетащить себя через ограду за ушки своей обуви), которое означает «выпутаться из своих проблем самому». Отсюда и название статистического метода. *Прим. пер.*

Писатель Антуан Гомбо (1607–1684) не был дворянином, а имя шевалье де Мере использовал в своих литературных сочинениях для персонажа, который выражал мысли автора. Впоследствии друзья стали так называть и его самого. *Прим. пер.*

Де Мере считал, что в Варианте 1, когда кость бросают четыре раза с вероятностью успеха  $1/6$ , общая вероятность победы равняется  $4 \times 1/6 = 2/3$ . Аналогично для Варианта 2 он полагал, что при 24 подбрасываниях с вероятностью успеха  $1/36$ , вероятность победы составит  $24 \times 1/36 = 2/3$ . Студенты часто прибегают к подобным ложным рассуждениям, но ошибку легко заметить: если бы в Варианте 1 у игрока было 12 бросков, то вероятность выигрыша равнялась бы  $12 \times 1/6 = 2$ . Правильная аргументация приведена в примечании 2 к этой главе в конце книги.

Марен Мерсенн (1588–1648) – французский математик, богослов и философ, организатор кружка, где обсуждались различные научные проблемы. Уже после смерти Мерсенна на основе его кружка была создана Парижская академия наук. *Прим. пер.*

Обсуждение и инструменты для методов моделирования при преподавании статистики см. в работе: M. Pfannkuch et al, 'Bootstrapping Students' Understanding of Statistical Inference', TLRI (2013), and K. Lock Morgan et al., 'STATKEY: Online Tools for Bootstrap Intervals and Randomization Tests', ICOTS 9 (2014).

**160**

Спойлер: правильный ответ –  $1/4$ , или 25 %, или 0,25.

Рассмотрим Вариант 1. В данном случае легче найти вероятность проигрыша (это обычный прием в теории вероятностей). Единственный случай, когда мы проигрываем, – если четыре раза подряд выпадает нешестерка. Вероятность выпадения нешестерки равна  $1 - 1/6 = 5/6$ . Поэтому вероятность выпадения четырех нешестерок подряд составляет  $5/6 \times 5/6 \times 5/6 \times 5/6 = (5/6)^4 = 625/1296 \approx 0,48$ . Поэтому вероятность победы равна  $1 - 0,48 = 0,52$ . Аналогичные рассуждения для Варианта 2 дают вероятность победы, равную  $1 - (35/36)^4 = 0,49$ . Поэтому Вариант 1 чуть более выгоден. Рассуждения также показывают, где ошибся шевалье де Мере – он складывал несовместные вероятности событий. Его рассуждения дают для 12 бросков вероятность  $12 \times 1/6 = 2$ , что не особо осмысленно.

Ошибка также известна как «закон переставленной условной вероятности». Звучит восхитительно непонятно, но на деле просто означает, что вероятность события А при условии, что произошло событие В, смешивается с вероятностью события В при условии, что произошло событие А.

Я благодарен Филиппу Дэвиду за, по-видимому, изобретение этого термина.

Англ. propensity – склонность, предрасположенность, тенденция. Вероятность представляется как физическая склонность ситуации приводить к какому-то результату. То есть здесь вероятность становится свойством ситуации, а не последовательности событий (склонности – это не частоты событий, а предполагаемые причины частот). Концепцию обсуждали, в частности, философы Чарльз Пирс и Карл Поппер. *Прим. пер.*

Предполагается, что генератор псевдослучайных чисел хорошо спроектирован, а получаемые числа предназначены для статистического моделирования или аналогичной цели. Они не особо пригодны для криптографических приложений, где предсказуемость можно использовать для взлома кода.

«Случай убийства» – когда одно лицо (или группа лиц) подозревается в совершении одного или нескольких связанных убийств. Поэтому массовый расстрел или террористический акт будет считаться одним случаем.

Сравнение количества ежедневных убийств с пуассоновским распределением:

<https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/compendium/focusonviolentcrimeandsexualoffences/yearendingmarch2016/homicide#statistical-interpretation-of-trends-in-homicides>.

Чтобы получить такое распределение, нужно определить вероятность двух левшей:  $0,2 \times 0,2 = 0,04$ , вероятность двух правшей:  $0,8 \times 0,8 = 0,64$ , а вероятность последнего варианта можно найти как  $1 - 0,04 - 0,64 = 0,32$ .

Вот блог Пола: <https://pb204.blogspot.com/2011/10/funnel-plot-of-uk-bowel-cancer.html>. Сами данные можно загрузить с сайта <http://pb204.blogspot.co.uk/2011/10/uploads.html>.

Одна из крупных агломераций Великобритании, расположенная в Шотландии, включает 48 населенных пунктов, примыкающих к Глазго. *Прим. пер.*

Абрахам де Муавр в юности переехал в Лондон, где прожил 66 лет, так что он скорее английский математик французского происхождения. *Прим. пер.*

Есть важные исключения – у некоторых распределений такие длинные и тяжелые хвосты, что математических ожиданий у них не существует, поэтому выборочным средним не к чему сходиться.

Если мы предположим, что все наши наблюдения независимы и имеют одинаковое распределение, то стандартная ошибка их среднего равна среднеквадратичному отклонению исходного распределения, деленному на квадратный корень из числа элементов в выборке.

От др.-греч. στόχος – цель, предположение. Такую неопределенность называют также алеаторной, или объективной. *Прим. пер.*

От др.-греч. ἐπιστήμη – научное знание, наука, достоверное знание. Такую неопределенность называют также гносеологической, или субъективной. *Прим. пер.*

В главе 12 мы увидим, что сторонники байесовской статистики рады использовать вероятности для эпистемической неопределенности в отношении параметров.

Строго говоря, 95-процентный доверительный интервал не означает наличия 95-процентной вероятности, что этот конкретный интервал содержит истинное значение, хотя на практике люди часто неверно интерпретируют именно так.

Я имел честь знать обоих, когда они уже были в преклонном возрасте.

Более точно, 95-процентные доверительные интервалы при точном нормальном распределении для статистики устанавливаются на уровне плюс-минус 1,96 стандартной ошибки.

Погрешность равна  $\pm 2\sqrt{[p(1-p)/n]}$ , и максимум этой величины, равный  $\pm 1/\sqrt{n}$ , достигается при  $p = 0,5$ . Следовательно, каково бы ни было истинное значение  $p$ , погрешность не превзойдет  $\pm 1/\sqrt{n}$ .

При 1000 участниках погрешность (в процентах) будет максимум  $\pm 100 / \sqrt{1000} \approx 3\%$ . Опросы могут обладать более сложной схемой, чем простая случайная выборка из генеральной совокупности, однако на погрешность это влияет не сильно.

Диаграмма «Би-би-си» для опросов находится на сайте:  
<http://www.bbc.co.uk/news/election-2017-39856354>.

Погрешности для статистических данных об убийствах:  
[https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/compedium/focusonviolentcrimeandsexualoffences/yearendingmarch2016/homicide#statistical-interpretation-of-trends-in-homicides](https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/compendium/focusonviolentcrimeandsexualoffences/yearendingmarch2016/homicide#statistical-interpretation-of-trends-in-homicides).

J. Arbuthnot, 'An Argument for Divine Providence...', *Philosophical Transactions* 27 (1710), 186–90.

Англиканство – протестантское направление христианства, преобладающее в Великобритании. *Прим. пер.*

R. A. Fisher, *The Design of Experiments* (Oliver and Boyd, 1935), p. 19.

Институт в Майзенберге (пригороде Кейптауна) в ЮАР. *Прим. пер.*

Возможно, более естественным вопросом была бы связь между скрещиванием рук и праворукостью (леворукостью), но для изучения этой проблемы было слишком мало левшей.

Мы могли бы выбрать другую статистику, которая отражает связь (например, отношение шансов), однако фактически получили бы тот же результат.

Всего существует  $54 \times 53 \times 52 \dots \times 2 \times 1$  перестановок. Это число обозначается  $54!$  (читается «54 факториал»). Примерно это число равно  $2$  и  $71$  нуль после двойки. Обратите внимание, что число способов перетасовать колоду карт равно  $52!$  так что, даже если бы мы перебирали миллион миллионов комбинаций в секунду, число лет, которое бы потребовало, чтобы перебрать все возможные комбинации, имело бы  $48$  нулей, в то время как возраст Вселенной сейчас оценивается всего в  $14\,000\,000\,000$  лет. Вот почему мы можем быть абсолютно уверены, что за всю историю карточных игр не было двух колод, перетасованных в точности одинаково.

Иначе – Р-уровень значимости. *Прим. пер.*

Для вычисления можно воспользоваться числом сочетаний. У нас есть 8 чашек, из них нужно выбрать те четыре, в которых молоко наливали в чай (тогда остальные четыре чашки автоматически будут идентифицированы верно). Если мы действуем наугад, то это значит, что нам требуется случайно вытащить 4 конкретных предмета из 8.

Общее число способов сделать это равно  $C_8^4 = 70$ . Нас устроит  $C_4^4 = 1$  способ. Следовательно, вероятность просто угадать равна  $1/70$ .

*Прим. пер.*

На самом деле только в одном случае из 70 мы можем получить результат не хуже этого. Точнее, в нашем конкретном эксперименте «не хуже» означает «столько же», поскольку результат лучше невозможен (верно определены все чашки). *Прим. пер.*

Метод состоит в вычислении для каждого из 412 человек в тестовом наборе разницы между среднеквадратичными ошибками прогноза для двух алгоритмов; получившееся множество из 412 разностей имеет среднее  $-0,0027$  и стандартное отклонение  $0,1028$ . Поэтому стандартная ошибка для оценки «истинной» разности составляет  $0,1028 / \sqrt{412} = 0,0050$ , и тогда  $t$ -статистика = оценка / стандартная ошибка =  $-0,0027/0,0050 = -0,54$ . Это также называют парным  $t$ -критерием Стьюдента, поскольку он основан на множестве разностей между парами чисел.

Вероятность того, что из двух испытаний хотя бы одно даст значимый результат, равна единице минус вероятность того, что оба результата незначимы  $= 1 - 0,95 \times 0,95 = 0,0975$ , что примерно равно 0,10.

С помощью этого экстравагантного эксперимента ученые хотели не столько проверить чудесные свойства лосося, сколько продемонстрировать высокий риск получения ложноположительных результатов при многократном тестировании. Эксперимент подтвердил такой риск. *Прим. пер.*

Исследование мертвого лосося описано на странице  
<http://prefrontal.org/files/posters/Bennett-Salmon-2009.jpg>.

Карло Эмилио Бонферрони (1892–1960) – итальянский математик.  
*Прим. пер.*

ЦЕРН (от фр. CERN – Conseil Européen pour la Recherche Nucléaire, Европейский совет по ядерным исследованиям) – европейская организация, занимающаяся ядерными исследованиями. *Прим. пер.*

Объявление ЦЕРН о бозоне Хиггса можно найти по адресу:  
<http://cms.web.cern.ch/news/observation-new-particle-mass-125-gev>.

В русском языке нет устоявшегося перевода этого термина. Его можно передать как «Эффект поиска в другом месте». Суть эффекта – наблюдение оказывается значимым только по случайности, потому что пространство проверяемых параметров очень велико. Аналогия: если в группе из 25 человек у кого-то день рождения совпадает с вашим, то это довольно удивительно, поскольку такая вероятность достаточно мала. Но если вы в той же группе станете искать какие-нибудь совпадающие дни рождения, то в таком совпадении не будет ничего удивительного – шансы на это больше 0,5. В первом случае вы сравниваете один конкретный день (свой), во втором – пространство сравнений резко расширяется (для 25 человек можно составить множество пар), поэтому вероятность обнаружить какое-то совпадение сильно увеличивается. Вспомните также пример с поисками неожиданных корреляций в одной из предыдущих глав: в очень большом наборе данных что-нибудь да найдется. *Прим. пер.*

Первоначальная теория Неймана и Пирсона включала идею «принятия» нулевой гипотезы, но эта часть их теории теперь не используется.

D. Spiegelhalter, O. Grigg, R. Kinsman and T. Treasure, 'Risk-Adjusted Sequential Probability Ratio Tests: Applications to Bristol, Shipman and Adult Cardiac Surgery', *International Journal for Quality in Health Care* 15 (2003), 7–13.

Здесь Р-значение – одностороннее, потому что нас интересует только повышение уровня смертности, а не снижение. Поэтому Р-значение – это вероятность того, что пуассоновская случайная величина со средним 22,5 примет значение не меньше 40. Стандартное программное обеспечение даст для такой вероятности 0,004.

Руководили этими статистиками Абрахам Вальд в США и Джордж Барнард в Соединенном Королевстве. Барнард до войны был чистым математиком (и коммунистом), а во время войны, как и многие ученые, начал заниматься военными применениями статистики. Позднее он разрабатывал официальный британский стандарт для презерватива (BS 3704).

Статистика имеет простую форму:  $\text{ПКОВ} = 0.69 \times \text{кумулятивное количество наблюдаемых смертей} - \text{кумулятивное количество ожидаемых смертей}$ . Пороговые значения определяются величиной  $\log((1 - \beta)/\alpha)$ .

D. Szucs and J. P. A. Ioannidis, 'Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature', *PLOS Biology* 15:3 (2 March 2017), e2000797.

J. P. A. Ioannidis, 'Why Most Published Research Findings Are False', PLOS Medicine 2:8 (August 2005), e124.

Стандартная доза алкоголя отличается в разных странах, например в США это 14 граммов, в Великобритании – 8 граммов. Многие страны приняли вариант Всемирной организации здравоохранения – 10 граммов. *Прим. пер.*

C. S. Knott et al., 'All Cause Mortality and the Case for Age Specific Alcohol Consumption Guidelines: Pooled Analyses of up to 10 Population Based Cohorts', *British Medical Journal* 350 (10 February 2015), h384. Об этом было сообщено под заголовком: 'Alcohol Has No Health Benefits After All', *The Times*, 11 February 2015.

D. J. Benjamin et al., 'Redefine Statistical Significance', *Nature Human Behaviour* 2 (2018), 6–10.

Артур Боули (1869–1957) – британский статистик и экономист.  
*Прим. пер.*

Томас Байес умер в 1761 году, понятия не имея о своем научном наследии: его основополагающая статья была опубликована только после его смерти, в 1763-м, а имя не связывалось с этим подходом вплоть до XX века.

Это значительное преувеличение. Фактически в своей работе «Очерки к решению проблемы доктрины шансов» Байес дал только определение условной вероятности, и никакого утверждения, которое мы называем сейчас теоремой Байеса, у него нет. Теорема была сформулирована Лапласом в начале XIX века. *Прим. пер.*

Некоторые даже могут заявить, что это была идеологическая обработка.

Еще раз подчеркнем, что  $1/5$  – это не вероятность, поэтому запись  $1/5$  двусмысленна и лучше писать 1 к 5 или хотя бы 1:5. В русском языке слово «шансы» имеет два значения. Во-первых, оно часто синонимично слову «вероятность» (мы скажем, что вероятность выпадения шестерки равна  $1/6$  и шансы на выпадение шестерки равны  $1/6$ ). Во-вторых, мы можем сказать, что шансы на выпадение шестерки 1:5 (1 к 5). Именно в этом смысле употребляет термин автор. *Прим. пер.*

В нашей литературе теорема Байеса обычно формулируется на языке вероятностей, а не шансов. *Прим. пер.*

Лат. a priori «от предшествующего», a posteriori «от последующего». Априорные знания получены заранее, до опыта, апостериорные – после опыта. Теорема Байеса дает возможность пересчитать априорные вероятности гипотез в апостериорные, поскольку произошедшие события их изменили. *Прим. пер.*

T. E. King et al., 'Identification of the Remains of King Richard III',  
Nature Communications 5 (2014) 5631.

Указания по отношениям правдоподобия: [http://enfsi.eu/wp-content/uploads/2016/09/m1\\_guideline.pdf](http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf).

Статья об использовании байесовской теории в суде: 'A Formula for Justice', The Guardian, 2 October 2011.

В покере – туз, король, дама, валет и десятка одной масти. *Прим. пер.*

Автор ошибается. Вероятность получить роял-флеш составляет  $4 / C_{52}^5 = 1 / 649740$ . Видимо, подразумевается стрит-флеш (пять карт одной масти, идущие подряд): вероятность такой комбинации действительно близка к  $1 / 72\,000$  (точнее, примерно  $1 / 72\,193$ ). *Прим. пер.*

Архиепископ Кентерберийский – глава государственной Церкви Англии. *Прим. пер.*

Его точные слова: «Дано количество раз, когда неизвестное событие случилось и не случилось: требуются шансы на то, что вероятность его появления в одном испытании лежит между любыми двумя степенями вероятности, которые можно указать». В целом тут все довольно понятно, за исключением того, что в современной терминологии мы поменяли бы слова «шансы» и «вероятность».

Будучи пресвитерианским священником, он называл его просто «стол».

Формула для такого распределения –  $60p^2(1-p)^3$ , то есть частный случай бета-распределения –  $B(3,4)$ . Если считать априорное распределение равномерным, то апостериорное распределение для положения белого шара, при условии, что бросили  $n$  красных шаров, из которых  $r$  оказались левее белого, задается формулой

$$\frac{(n+1)!}{r!(n-r)!} p^r (1-p)^{n-r}$$

то есть это бета-распределение  $B(r+1, n-r+1)$ .

Интуиция не должна давать  $2/5$ . Бросим на стол всего один красный шар. Если вам сказали, что он слева, то доля красных шаров слева от белого равна  $1 / 1 = 1$ , но вряд ли ваша интуиция согласится, что среднее положение белого шара равно  $1$ , то есть он лежит у правого края стола. *Прим. пер.*

Онлайн-панель – это группа людей, которые согласились участвовать в онлайн-исследованиях. Они регистрируются на каком-либо сайте и регулярно предоставляют информацию о себе. *Прим. пер.*

В английском языке есть пословица «Нельзя сделать шелковый кошелек из свиного уха». *Прим. пер.*

Подвешенный парламент – парламент, в котором ни одна из партий не имеет большинства. *Прим. пер.*

D. K. Park, A. Gelman and J. Bafumi, 'Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls', *Political Analysis* 12 (2004), 375–85; Результаты опросов YouGov взяты с сайта: <https://yougov.co.uk/news/2017/06/14/how-we-correctly-called-hung-parliament/>.

K. Friston, 'The History of the Future of the Bayesian Brain', *Neuroimage* 62:2 (2012), 1230–33.

N. Polson and J. Scott, *AIQ: How Artificial Intelligence Works and How We Can Harness Its Power for a Better World* (Penguin, 2018), p. 000.

R. E. Kass and A. E. Raftery, 'Bayes Factors', *Journal of the American Statistical Association* 90 (1995), 773–95.

Напоминаем, что это означает, что при многократном повторении 95 % таких интервалов будут включать истинное значение, но мы ничего не можем сказать о каком-то конкретном интервале.

J. Cornfield, 'Sequential Trials, Sequential Analysis and the Likelihood Principle', *American Statistician* 20 (1966), 18–23.

И тем не менее я по-прежнему предпочитаю байесовский подход.

Проект в 2011 году запустил Брайан Нозек из Виргинского университета. Он и его коллеги взялись проверить 100 исследований, выполненных в 2008 году, результаты которых были опубликованы в трех психологических журналах. Они старались максимально придерживаться условий оригинальных экспериментов и даже консультировались с их авторами. Результаты проекта были опубликованы в 2015 году. *Прим. пер.*

Open Science Collaboration, 'Estimating the Reproducibility of Psychological Science', *Science* 349:6251 (28 August 2015), aac4716.

A. Gelman and H. Stern, 'The Difference Between "Significant" and "Not Significant" Is Not Itself Statistically Significant', *American Statistician* 60:4 (November 2006), 328–31.

Падение началось вскоре после запуска Facebook, но данные не могут сказать нам, что это – корреляция или причинно-следственная связь.

Ronald Fisher, Presidential Address to the first Indian Statistical Congress, 1938, *Sankhyā* 4(1938), 14–17.

В английском языке термин *post mortem* (патологоанатомическое вскрытие трупа) имеет переносное значение «разбор действий», «обсуждение причин неудачи», «итоговый анализ». *Прим. пер.*

СМ. ‘The Reinhart and Rogoff Controversy: A Summing Up’,  
New Yorker, 26 April 2013.

Эта ошибка в сочетании с другими критическими замечаниями, как утверждается, повлияла на выводы в исследовании, однако это активно оспаривается авторами.

‘AXA Rosenberg Finds Coding Error in Risk Program’, Reuters, 24 April 2010.

История с Харконеном изложена в статье: 'The Press-Release Conviction of a Biotech CEO and its Impact on Scientific Research', Washington Post, 13 September 2013.

D. Fanelli, 'How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data', PLOS ONE 4:5 (29 May 2009), e5738.

U. Simonsohn, 'Just Post It: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone', *Psychological Science* 24:10 (October 2013), 1875–88.

P-хакинг (P-hacking) – это изменение результатов исследований с целью добиться нужного P-значения. Такая практика манипулирования данными служит для того, чтобы показать статистические значения, подтверждающие желаемый результат, который обычно в чем-то выгоден либо исследователям, либо спонсорам. *Прим. ред.*

J. P. Simmons, L. D. Nelson and U. Simonsohn, 'False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant', *Psychological Science* 22:11 (November 2011), 1359–66.

Студентов спрашивали, сколько времени они наслаждаются едой в закусочной, чему равен квадратный корень из 100, согласны ли они с утверждением, что «компьютеры – это сложные машины», сколько лет их отцу и матери, какова их политическая ориентация, пользуются ли они утренними специальными предложениями в ресторанах, кто из четырех канадских квотербеков выиграет приз, как часто они ссылаются на прошлое как «на старые добрые времена» и так далее.

АНГЛ. HARKing образовано от Hypotheses After the Results are Known. *Прим. пер.*

L. K. John, G. Loewenstein and D. Prelec, 'Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling', *Psychological Science* 23:5 (May 2012), 524–32.

D. Spiegelhalter, 'Trust in Numbers', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180:4 (2017), 948–65.

P. Sumner et al., 'The Association Between Exaggeration in Health Related Science News and Academic Press Releases: Retrospective Observational Study', *British Medical Journal* 349 (10 December 2014), g7015.

‘Nine in 10 People Carry Gene Which Increases Chance of High Blood Pressure’, Daily Telegraph, 15 February 2010.

‘Why Binge Watching Your TV Box-Sets Could Kill You’, Daily Telegraph, 25 July 2016.

Я иногда следую тому, что можно назвать «принципом Граучо», – из-за парадоксального заявления комика Граучо Маркса, что он никогда не вступит в клуб, который согласится его принять. Поскольку тексты прошли сквозь множество фильтров, поощряющих искажения и отбор, уже сам факт того, что я слышу какое-то утверждение, основанное на статистике, – повод не доверять ему.

Слова Бема взяты из 'Daryl Bem Proved ESP Is Real: Which Means Science Is Broken', Slate, 17 May 2017.

В одной онлайн-статье приводятся такие слова Бема: «Я за строгость... но предпочитаю, чтобы ею занимались другие. Я понимаю ее важность – пусть некоторые развлекаются, но у меня нет терпения на это... Если вы посмотрите на мои прошлые эксперименты, они всегда были риторическими приемами. Я собрал данные, чтобы показать, как формировалась моя точка зрения. Я использовал эти данные для убеждения и никогда по-настоящему не заботился о том, будет кто-нибудь это повторять или нет».

Краткая характеристика Гельмана была такой: «Исследование Бема – дерьмо».

I. J. Jacobs et al., 'Ovarian Cancer Screening and Mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): A Randomised Controlled Trial', *The Lancet* 387:10022 (5 March 2016), 945–56.

‘Ovarian Cancer Blood Tests Breakthrough: Huge Success of New Testing Method Could Lead to National Screening in Britain’, Independent, 5 May 2015.

M. R. Munafò et al., 'A Manifesto for Reproducible Science', *Nature Human Behaviour* 1 (2017), a0021.

Адрес Open Science Framework: <https://osf.io/>.

Кристи Ашванден – американская журналистка и популяризатор науки. Сайт [fivethirtyeight.com](http://fivethirtyeight.com) объединяет блоги, посвященные анализу опросов общественного мнения, экономике, политике, спорту. *Прим. пер.*

История Ашванден взята из статьи 'Science Won't Settle the Mammogram Debate', FiveThirtyEight, 20 October 2015.

J. P. Simmons, L. D. Nelson and U. Simonsohn, 'False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant', *Psychological Science* 22:11 (November 2011), 1359–66.

A. Gelman and D. Weakliem, 'Of Beauty, Sex and Power', *American Scientist* 97:4 (2009), 310–16.

U. Simonsohn, L. D. Nelson and J. P. Simmons, 'P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results', *Perspectives on Psychological Science* 9:6 (November 2014), 666–81.

Больше о разумной открытости смотрите в работе: Royal Society, *Science as an Open Enterprise* (2012). Взгляды Оноры О'Нил на степень доверия блестяще объяснены в ее выступлении на конференции TedX 'What We Don't Understand About Trust' (June 2013).

Дэвид Фёрт объясняет методологию для экзитполов здесь:  
<https://warwick.ac.uk/fac/sci/statistics/staff/academic-research/firth/exit-poll-explainer/>.

Нет подтверждений, что Пэдди Эшдаун выполнил свое обещание, хотя ему до сих пор припоминают эти слова. После одной радиопередачи, на которой обсуждались те выборы, нам вынесли большую шоколадную шляпу и предложили поделить на всех.

R. E. Kass et al., 'Ten Simple Rules for Effective Statistical Practice', PLOS Computational Biology 12:6 (9 June 2016), e1004961.

Автор смешивает воедино аксиомы вероятности и ее свойства. Аксиомы вероятности – неотрицательность, ограниченность всего вероятностного пространства единицей и аддитивность (в общем случае – счетная аддитивность). Другие приведенные свойства вытекают из аксиом. *Прим. пер.*

Как уже отмечалось выше, за определение независимости обычно берется равенство  $P(AB) = P(A)P(B)$ , поскольку в этом случае не надо требовать  $P(B) \neq 0$ , что необходимо, чтобы иметь право писать  $P(A|B)$ .  
*Прим. пер.*

В определении выборочной дисперсии используется  $n$ , то есть

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

$$s_1^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2,$$

. Если взять  $n-1$ , то есть получится так называемая исправленная (несмещенная) выборочная дисперсия. Такое исправление делается, поскольку выборочная дисперсия представляет собой смещенную оценку реальной дисперсии

$D(X)$ , то есть  $E(s^2) \neq D(X)$ . Если же умножить  $s^2$  на  $\frac{n}{(n-1)}$ , то есть

получить  $s_1^2$ , то исправленная выборочная дисперсия будет уже несмещенной оценкой дисперсии, то есть  $E(s_1^2) = D(X)$ . При больших  $n$

величины  $s^2$  и  $s_1^2$  почти не отличаются между собой, так что на практике имеет смысл использовать исправленную дисперсию только при небольших  $n$  (примерно при  $n < 30$ ). *Прим. пер.*

Строго говоря, это не эквивалентно. Равенство  $p(BA) = p(B)p(A)$  может быть верно даже для событий с нулевой вероятностью, в то время как для того, чтобы написать условную вероятность  $p(B|A)$ , мы должны предварительно принять, что  $p(A) \neq 0$ . *Прим. пер.*

В отечественной терминологии случайные величины, имеющие плотность, принято называть абсолютно непрерывными. *Прим. пер.*

R – язык программирования для статистической обработки данных. *Прим. пер.*