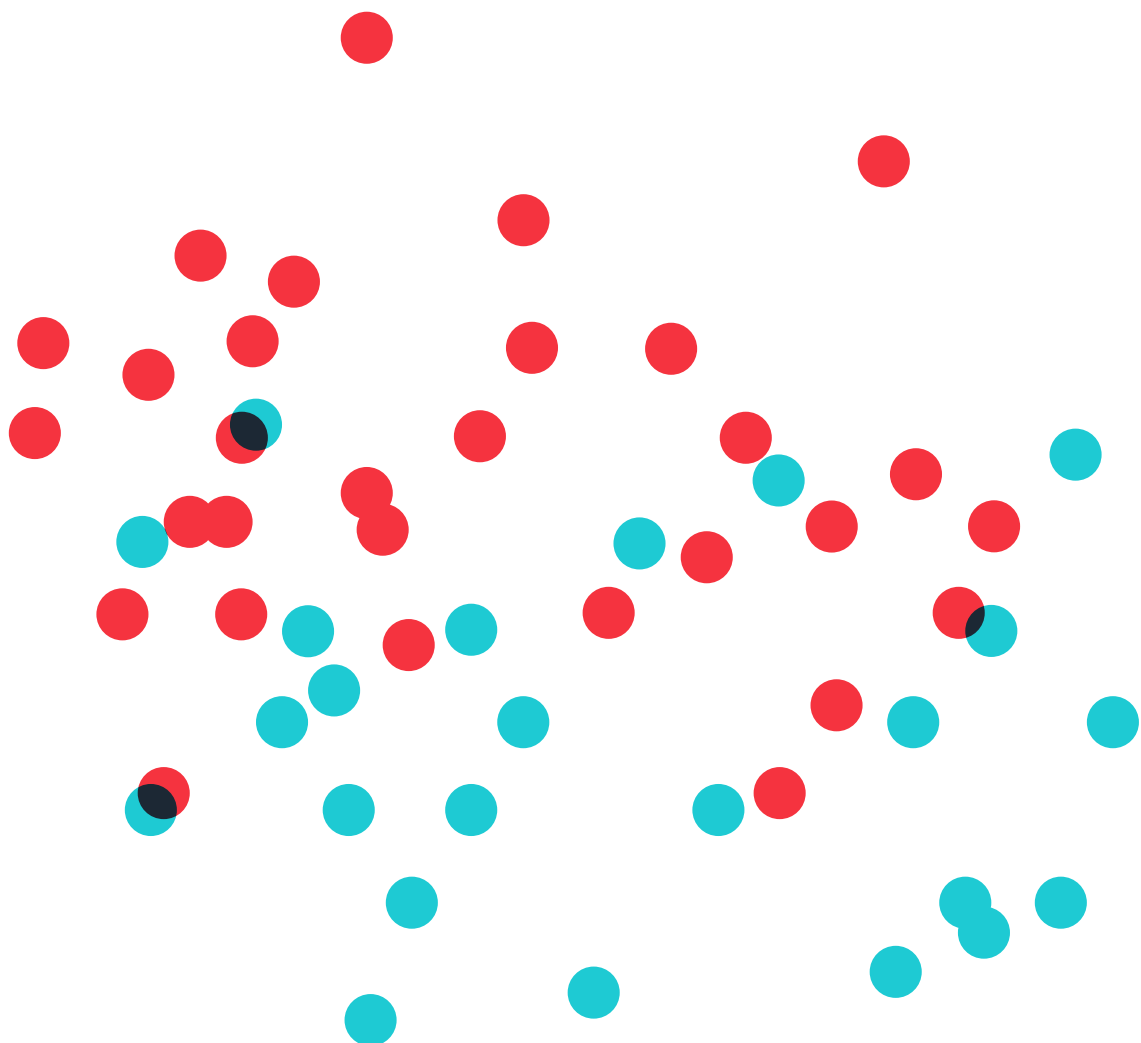


# Искусство статистики

Как находить  
ответы в данных

Дэвид Шпигельхалтер



Купить книгу на сайте [kniga.biz.ua](http://kniga.biz.ua) >>>



# Оглавление

<b>Введение</b>	11
<hr/>	
ГЛАВА 1	
<b>Расчет долей: качественные данные и проценты</b>	29
<hr/>	
ГЛАВА 2	
<b>Числовые характеристики выборки и представление данных</b>	49
<hr/>	
ГЛАВА 3	
<b>Почему мы смотрим на данные? Совокупности и измерение</b>	84
<hr/>	
ГЛАВА 4	
<b>Причины и следствия</b>	108
<hr/>	
ГЛАВА 5	
<b>Моделирование зависимости с помощью регрессии</b>	134
<hr/>	
ГЛАВА 6	
<b>Алгоритмы, аналитика и прогнозирование</b>	155
<hr/>	
ГЛАВА 7	
<b>Насколько мы можем быть уверены в происходящем? Оценки и интервалы</b>	202
<hr/>	
ГЛАВА 8	
<b>Вероятность — язык неопределенности и случайности</b>	218
<hr/>	
ГЛАВА 9	
<b>Объединяем вероятность и статистику</b>	242
<hr/>	

ГЛАВА 10	
<b>Отвечаем на вопросы и заявляем об открытиях</b>	266
<hr/>	
ГЛАВА 11	
<b>Учимся на опыте — байесовский путь</b>	319
<hr/>	
ГЛАВА 12	
<b>Когда дела идут не так</b>	353
<hr/>	
ГЛАВА 13	
<b>Как можно улучшить статистику</b>	374
<hr/>	
ГЛАВА 14	
<b>В заключение</b>	392
<hr/>	
<b>Глоссарий</b>	394
<hr/>	
<b>Примечания</b>	422
<hr/>	
<b>Алфавитный указатель</b>	434
<hr/>	
<b>Благодарности</b>	440
<hr/>	
<b>Список иллюстраций</b>	442
<hr/>	
<b>Список таблиц</b>	445
<hr/>	

# 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>

## 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).

---

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

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>2</sup>.

Эта книга посвящена применению **статистики\*** 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).

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,

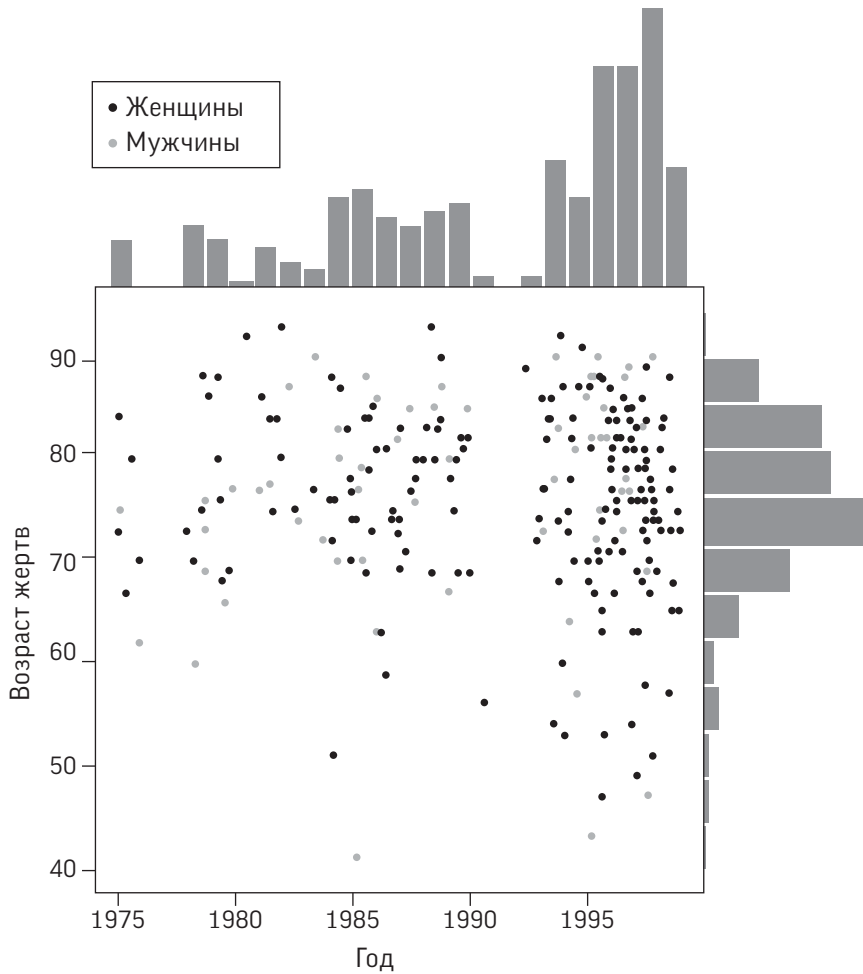
\* Термины, выделенные **полужирным шрифтом**, включены в глоссарий в конце книги.

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 отражено 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.

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».



**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

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

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\*\*. 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.

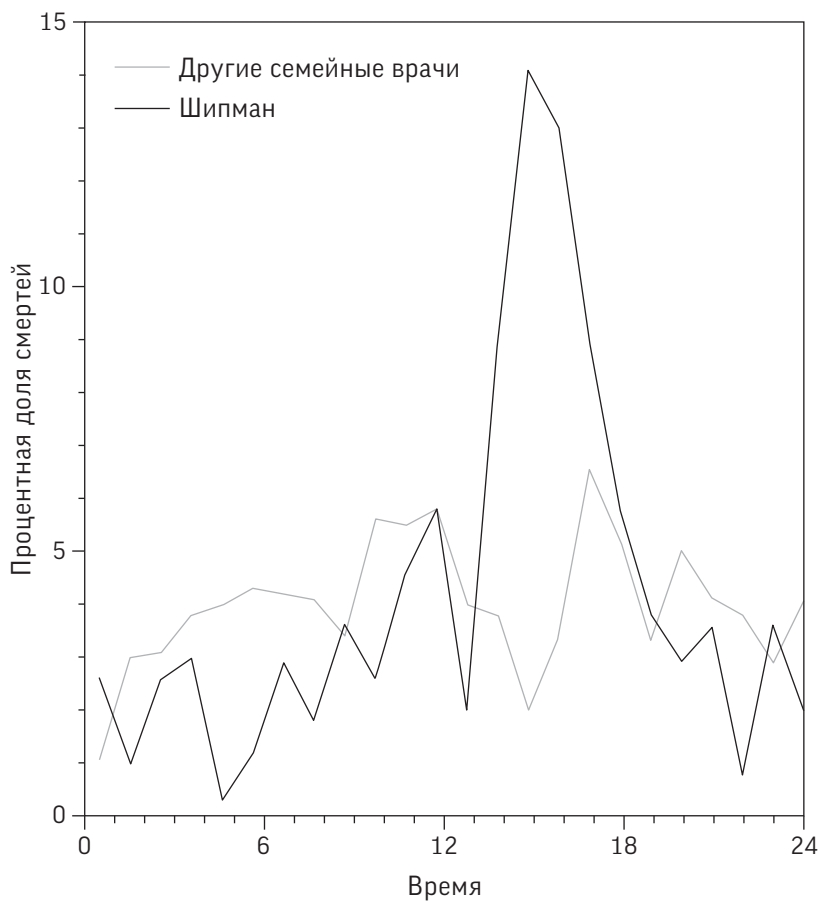
### Превращение мира в набор данных

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.

\* 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.*

\*\* Спойлер: это можно было сделать практически наверняка.





**Рис. 0.2**

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

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)\*. 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,

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

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>3\*</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>4</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>5</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>6</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

\* 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.

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\*.

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 едином образном 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>7</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.

\* Если бы я был типичным среднестатистическим человеком, этот факт давал бы мне основание заранее чему-то радоваться.

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\* 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.

---

\* Викторианская эпоха — время правления королевы Виктории (1837–1901). *Прим. пер.*

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

Еще одну реальную угрозу традиционному взгляду на статистику представляет колоссальный рост количества проводимых исследований, особенно в биомедицине и социальных науках, в сочетании с требованием публикаций в высокорейтинговых журналах. Это привело к сомнениям в надежности определенной части научной литературы и утверждениям о невоспроизводимости многих «открытий» другими исследователями. Как, например, продолжающийся спор, может ли «поза силы» вызвать гормональные и другие изменения у человека<sup>8</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. Большее количество данных означает, что нам надо еще лучше осознавать, чего на самом деле стоят такие доказательства.

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

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

### Преподавание статистики

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

К счастью, все меняется. Наука о данных и грамотность в работе с ними требуют подхода, направленного на решение основных проблем, где применение конкретных статистических

инструментов рассматривается лишь как один из компонентов цикла исследований. Цикл **PPDAC** (Problem, Plan, Data, Analysis, Conclusion) был предложен как модель решения проблем, которую мы будем использовать в этой книге<sup>9</sup>. Рис. 0.3 основан на примере Новой Зеландии, которая считается мировым лидером по преподаванию статистики в школах.

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 делать 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, подобранный на улице.

В курсах статистики основной упор делается на стадию *анализа*, и мы рассмотрим в книге ряд аналитических методов; однако





**Рис. 0.3**

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

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»\*. Поэтому для самосовершенствования мы обращались к книгам издательства Pelican, и их легко узнаваемые синие корешки были обычной приметой каждой студенческой полки\*\*.

\* 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.*

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)\*. 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\*\* или убедительные подтверждения экстрасенсорного восприятия. Другие касаются здравоохранения — например, выше ли показатель выживаемости в более загруженных больницах и полезны ли скрининговые

\* Издана на русском языке: Хафф Д. Как лгать при помощи статистики. М. : Альпина Паблишер, 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.*

исследования\* 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\*\*.

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.

\* 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.*

## 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.