

Оглавление

Предисловие от издательства	14
Предисловие	15
Как построена эта книга	17
Кому адресована эта книга?.....	18
Благодарности	19
Глава 1. Критическое мышление в эпоху данных.....	21
О чем эта глава.....	21
Введение.....	21
Поучительные истории	22
Поспешный диагноз Эйба.....	22
Гражданское сопротивление	24
Теория разбитых окон.....	26
Дополнение или замена?	29
Дополнительное чтение и ссылки.....	31
ЧАСТЬ I. В ПОИСКЕ ОБЩЕГО ЯЗЫКА	33
Глава 2. Корреляция: что это такое и для чего она нужна?.....	35
О чем эта глава.....	35
Введение.....	35
Что такое корреляция?	36
Факт или корреляция?.....	40
Для чего нужна корреляция?	42
Описание.....	43
Прогнозирование	44
Причинный вывод.....	47
Измерение корреляций.....	48
Среднее значение, дисперсия и стандартное отклонение	48
Ковариация	53
Коэффициент корреляции	53

Наклон линии регрессии.....	54
Совокупности и выборки.....	55
Откровенно о линейности.....	55
Подведение итогов.....	59
Ключевые термины.....	59
Упражнения.....	60
Дополнительное чтение и ссылки.....	62

Глава 3. Причинно-следственная связь:

что это такое и для чего она нужна?..... 63

О чем эта глава.....	63
Введение.....	63
Что такое причинно-следственная связь?.....	64
Потенциальные исходы и контрфактические сравнения.....	65
Зачем нужно знать причинно-следственную связь?.....	67
Фундаментальная проблема причинного вывода.....	67
Принципиальные вопросы.....	69
В чем причина?.....	69
Причинность и контрпримеры.....	72
Причинность и закон.....	75
Может ли причинно-следственная связь распространяться вспять во времени?.....	76
Требуется ли причинно-следственная связь физической связи?.....	77
Причинно-следственная связь не обязательно подразумевает корреляцию.....	77
Подведение итогов.....	78
Ключевые термины.....	78
Упражнения.....	79
Дополнительное чтение и ссылки.....	81

ЧАСТЬ II. СУЩЕСТВУЕТ ЛИ ВЗАИМОСВЯЗЬ?..... 83

Глава 4. Не бывает корреляции без вариаций..... 85

О чем эта глава.....	85
Введение.....	85
Выбор зависимой переменной.....	87
Правило 10 000 часов.....	88
Деградация молодежи.....	90
Уход из средней школы.....	93
Атаки смертников.....	94
Мир заставляет нас выбирать зависимую переменную.....	97
Врачи чаще наблюдают за больными людьми.....	97
Анализ постфактум.....	97
Катастрофа «Челленджера».....	100
Финансовый кризис 2008 года.....	101
Жизненные советы.....	103
Подведение итогов.....	103

Ключевые термины	103
Упражнения	104
Дополнительное чтение и ссылки	105
Глава 5. Применение регрессии в описании и прогнозировании.....	107
О чем эта глава.....	107
Введение.....	107
Основы регрессии.....	107
Линейная регрессия при нелинейных данных.....	113
Проблема переобучения	121
Прогнозирование президентских выборов.....	122
Как представляют выводы регрессии	124
Краткая история регрессии.....	125
Подведение итогов	127
Ключевые термины	127
Упражнения	128
Дополнительное чтение и ссылки.....	129
Глава 6. Выборки, неопределенность и статистические выводы.....	130
О чем эта глава.....	130
Введение.....	130
Оценка	130
Почему оценка отличается от оцениваемой величины?	132
Смещение.....	133
Шум.....	134
Как получается хороший оценщик?.....	134
Количественная оценка точности	136
Стандартные ошибки	137
Маленькие выборки и экстремальные наблюдения	139
Доверительные интервалы	140
Статистический вывод и проверка гипотез	141
Проверка гипотез	141
Статистическая значимость	143
Статистический вывод о взаимосвязях	143
Что, если у нас есть данные для всей совокупности?.....	145
Содержательная и статистическая значимость.....	146
Социальные сети и голосование.....	147
Второй закон о реформе	147
Подведение итогов	148
Ключевые термины	149
Упражнения	150
Дополнительное чтение и ссылки.....	151
Глава 7. Завышение значимости и занижение отчетности	153
О чем эта глава.....	153

Введение.....	153
Может ли осьминог быть футбольным экспертом?	153
Предвзятость публикации	159
<i>p</i> -хакинг.....	161
<i>p</i> -скрининг	162
Являются ли большинство научных «фактов» ложными?	163
Экстрасенсорное восприятие	164
Явка избирателей на голосование.....	165
Выявление <i>p</i> -хакинга	166
Возможные решения проблемы	169
Уменьшение порога статистической значимости	169
Корректировка <i>p</i> -значения при многократном тестировании.....	170
Не зацикливайтесь на статистической значимости	170
Предварительная регистрация	171
Проверка важных и правдоподобных гипотез.....	174
За пределами науки.....	175
Суперзвезды.....	176
Подведение итогов	179
Ключевые термины	179
Упражнения	180
Дополнительное чтение и ссылки.....	181

Глава 8. Возврат к среднему значению 183

О чем эта глава.....	183
Введение.....	183
Исчезает ли истина?	183
Фрэнсис Гальтон и возврат к среднему.....	185
Возврат к среднему значению не является силой притяжения	188
Поиск помощи	191
Работает ли операция на колене?.....	193
Возвращение к среднему, эффект плацебо и космическое привыкание ...	194
Эффект плацебо	194
Объяснение космического привыкания	195
Космическое привыкание и генетика.....	197
Убеждения не возвращаются к среднему значению.....	198
Подведение итогов	200
Ключевые термины	200
Упражнения	200
Дополнительное чтение и ссылки.....	203

ЧАСТЬ III. ЯВЛЯЕТСЯ ЛИ СВЯЗЬ ПРИЧИННО-СЛЕДСТВЕННОЙ? 205

Глава 9. Почему корреляция и причинно-следственная связь не одно и то же 207

О чем эта глава.....	207
----------------------	-----

Введение.....	207
Чартерные школы.....	208
Критический анализ потенциальных исходов.....	212
Источники смещения.....	217
Искажающие факторы.....	217
Обратная причинно-следственная связь.....	219
Новый взгляд на правило 10 000 часов.....	220
Диетическая газировка.....	224
Насколько похожи искажающие факторы и обратная причинность?.....	225
Расходы на предвыборную кампанию.....	226
Признаки смещения.....	228
Контрацепция и ВИЧ.....	232
Механизмы или факторы?.....	233
Критические размышления о смещении и шуме.....	236
Подведение итогов.....	240
Ключевые термины.....	241
Упражнения.....	242
Дополнительное чтение и ссылки.....	245

Глава 10. Выявление и ограничение искажающих факторов 247

О чем эта глава.....	247
Введение.....	247
Влияние партии на голосование в Конгрессе.....	247
Примечание о гетерогенных эффектах воздействия.....	252
Анатомия регрессии.....	253
Как регрессия ограничивает влияние искажающего фактора?.....	257
Контроль и причинно-следственная связь.....	265
Вредят ли нам социальные сети?.....	267
Чтение таблицы регрессии.....	268
Чем искажающий фактор отличается от механизма?.....	271
Статистика без волшебства.....	271
Подведение итогов.....	273
Ключевые термины.....	273
Упражнения.....	274
Дополнительное чтение и ссылки.....	275

Глава 11. Рандомизированные эксперименты 276

О чем эта глава.....	276
Введение.....	276
Грудное вскармливание.....	277
Рандомизация и причинно-следственный вывод.....	280
Оценка и вывод в экспериментах.....	283
Стандартные ошибки.....	283
Проверка гипотезы.....	285
Проблемы, возникающие при экспериментах.....	285
Несоблюдение условий и инструментальные переменные.....	285
Случайный дисбаланс.....	294

Нехватка статистической мощности.....	296
Убыль в ходе эксперимента.....	297
Взаимное влияние.....	298
Естественные эксперименты.....	300
Военная служба и будущие доходы.....	301
Подведение итогов.....	302
Ключевые термины.....	302
Упражнения.....	304
Дополнительное чтение и ссылки.....	305
Глава 12. Модели разрывной регрессии.....	307
О чем эта глава.....	307
Введение.....	307
Реализация метода разрывной регрессии.....	312
Какие кандидаты более успешны – радикальные или умеренные?.....	314
Непрерывность в пороговой точке.....	317
Сохраняется ли непрерывность в разрывных регрессиях для анализа выборов?.....	322
Несоблюдение условий и нечеткая разрывная регрессия.....	323
Бомбардировки во Вьетнаме.....	324
Мотивация и успех.....	328
Подведение итогов.....	329
Ключевые термины.....	330
Упражнения.....	330
Дополнительное чтение и ссылки.....	332
Глава 13. Метод разности различий.....	334
О чем эта глава.....	334
Введение.....	334
Параллельность трендов.....	335
Два объекта и два периода.....	337
Безработица и минимальная заработная плата.....	337
N объектов и два периода.....	341
Вредит ли просмотр телевизора детям?.....	342
N объектов и N периодов.....	345
Контрацепция и гендерный разрыв в оплате труда.....	346
Полезные проверки.....	348
Влияет ли поддержка газет на решение по голосованию?.....	349
Заразно ли ожирение?.....	350
Разность различий как проверка достоверности выводов.....	353
Подведение итогов.....	353
Ключевые термины.....	353
Упражнения.....	354
Дополнительное чтение и ссылки.....	356
Глава 14. Механизмы причинно-следственных связей.....	358
О чем эта глава.....	358

Введение.....	358
Анализ причинной медиации	359
Промежуточные результаты	361
Когнитивно-поведенческая терапия и молодежь из группы риска в Либерии	361
Независимые теоретические прогнозы.....	362
Дискриминируют ли избиратели женщин?.....	363
Естественные способы тестирования механизмов	364
Давление общества и голосование	365
Косвенное выявление механизма	365
Скачки цен на сырьевые товары и вооруженные конфликты	365
Подведение итогов	368
Ключевые термины	369
Упражнения	369
Дополнительное чтение и ссылки.....	370
ЧАСТЬ IV. ОТ ИНФОРМАЦИИ К РЕШЕНИЮ	373
Глава 15. Как наделить статистику смыслом.....	375
О чем эта глава.....	375
Введение.....	375
Каков правильный масштаб?.....	376
Мили на галлон или галлоны на милю ?.....	376
Процент или процентный пункт?	379
Визуальное представление данных.....	380
Политические предпочтения и перестройка Юга.....	382
Некоторые ключевые правила визуализации данных	385
От статистики к убеждениям.....	386
Правило Байеса.....	390
Информация, априорные и апостериорные убеждения	391
Возвращаясь к целиакии Эйба.....	391
Поиск террористов в аэропорту	394
Правило Байеса и количественный анализ	398
Ожидаемые затраты и выгоды	403
Скрининг: часто и точно.....	403
Подведение итогов	407
Ключевые термины	407
Упражнения	407
Дополнительное чтение и ссылки.....	410
Глава 16. Измерение показателей вашей миссии	411
О чем эта глава.....	411
Введение.....	411
Оценка неправильного результата или воздействия	412
Частичные измерения.....	412
Промежуточные результаты.....	414
Плохо определенные миссии.....	416

Есть ли у вас подходящая выборка?	419
Внешняя валидность	419
Ограниченная выборка	421
Стратегическая адаптация и изменение отношений	426
Налоги на свет и окна	426
Сдвиг в бейсболе	428
Война с наркотиками	429
Подведение итогов	431
Ключевые термины	431
Упражнения	431
Дополнительное чтение и ссылки	433

Глава 17. О пределах возможностей количественной оценки ... 435

О чем эта глава	435
Введение	435
Принятие решений при ограниченных данных	436
Анализ затрат и выгод и экологическое регулирование	436
Использование зубной нити и ношение маски	438
Количественные данные и ценности	440
Как количественные инструменты крадут наши ценности	440
Как количественная оценка навязывает нам ценности	443
Научитесь мыслить критически и помогите научиться другим	447
Упражнения	448
Дополнительное чтение и ссылки	449
Предметный указатель	451

Предисловие

Наш мир удивительно изменился. Мы утопаем в океане данных, которые сами же и генерируем. Количественная информация пронизывает наши разговоры обо всем: от политики и здравоохранения до поиска работы, спорта, образования, свиданий и национальной безопасности.

Поэтому навыки количественных рассуждений больше не должны быть прерогативой только тех, кто обладает математическими способностями или собирается сделать техническую карьеру. Способности к базовым количественным рассуждениям жизненно необходимы каждому образованному человеку и гражданину. Совершенно очевидно, что обучение людей новым навыкам требует новых методов преподавания.

Именно с этой целью мы решили написать книгу, которую вы сейчас читаете. Но мы начали не с книги. Большая часть материала и идей, которые в конечном итоге нашли отражение в последующих главах, сначала была разработана для курсов, которые помогали учащимся, почти не имеющим технического образования, стать серьезными, вдумчивыми и скептически настроенными потребителями количественной информации. В состав этих курсов входят традиционные темы, такие как введение в количественное мышление, которое преподают как студентам, так и аспирантам Чикагского университета. Но, кроме этого, мы перенесли в них многое из специальных методик обучения руководителей, политиков, военных офицеров, экспертов по национальной безопасности, аналитиков разведки и журналистов.

Занимаясь преподаванием, мы и сами извлекли много уроков, которые повлияли на выбор, сделанный нами при планировании и написании этой книги. Возможно, самым важным было создание *общедоступного* языка. Мы твердо знали, что не хотим преподавать традиционный курс статистики. Для многих студентов (особенно гуманитариев) такие курсы являются слишком техническими и не дают ответов на наиболее интересные вопросы, которые действительно важны для использования количественной информации в повседневной жизни и работе. Поэтому было заманчиво как можно быстрее перейти к интересным темам, например почему корреляция не означает причинно-следственную связь. Но это было бы ошибкой. Человек не может понять, почему корреляция не предполагает причинно-следственной связи, пока не поймет, что такое корреляция и причинность.

Поэтому первая часть книги посвящена поиску общего языка с читателем. Мы начинаем не с математических формул (хотя они тоже будут), а с ответов на вопросы о том, что мы имеем в виду, когда говорим о корреляции и причинно-следственной связи. Что сложного в корреляции и причинно-следственной связи? Почему их нужно разделять? Где и как применяются понятия корреляции и причинно-следственной связи?

А как же проблема мотивации? Считается, что для вовлечения читателей нужно начинать свой рассказ с демонстрации *полезности* обсуждаемой темы. Так вот, понимание сути корреляции – чрезвычайно полезная вещь. Но, что более важно, наш подход таков: если вы хотите, чтобы читатели были вовлечены, сделайте материал *интересным*. Для нас это означает несколько вещей.

Первый способ привлечь внимание – рассказывать истории. Вскоре вы обнаружите, что каждое концептуальное понятие дополнено по крайней мере одним развернутым, подлинным примером из реальной жизни. Некоторые примеры касаются научных исследований. Многие из них будут посвящены нашему личному опыту, когда взвешенное размышление о количественных показателях повлияло на принимаемые нами решения. Другие относятся к использованию данных и доказательств в новостях, спорте, политике, здравоохранении и культуре. От правильного толкования данных напрямую зависит, как люди проживают свою жизнь и принимают решения во всех областях человеческой деятельности. Мы хотим, чтобы вы постоянно имели это в виду. Вот почему, хотя авторы книги в свое время много занимались политологией, многие примеры взяты не из политики.

Второй способ привлечь читателей – сначала сделать акцент на идеях, а затем на технических вопросах. Мы любим технические детали. Но техничность часто бывает врагом понимания. Когда разговор переходит на технические детали, многие люди перестают думать и начинают запоминать. Мы горячо хотим этого избежать. Поэтому всегда сначала говорим об идеях и о том, почему они важны. Мы используем как можно больше визуальных представлений. И как можно меньше математики. Но как можно меньше – это не ноль, по крайней мере, по двум причинам.

Понимание технических деталей является частью критического мышления. Вы не сможете понять возврат к среднему, если не знаете, что такое среднее значение или шум. Вы не сможете понять предвзятость публикаций и кризис репликации, если не знаете, что такое *статистическая значимость* или *p*-значение. И трудно понять проблему смещения оценки эффекта или решения, предлагаемые различными способами исследования, не имея возможности интерпретировать регрессию.

Более того, иногда для ясности и точности требуется немного математики. Мы проводим много времени, рассуждая о контрфактическом подходе и причинно-следственных связях. Но без математики разговор о контрфактичности приобретает мистический оттенок. Строгая запись определений и ожидаемых эффектов вносит дополнительную ясность. Поэтому нам не обойтись без математики. Но мы всегда делаем акцент на понимании сути.

Третий способ вовлечения читателей заключается в том, что после каждой главы или урока мы оставляем небольшую недосказанность для самостоятельных размышлений. Это очень важный момент: принятие правильных решений в наш век, основанный на данных, требует навыка критического мышления от каждого из нас. Мы не можем просто оставить это на усмотрение экспертов, поскольку – и это может шокировать – многих экспертов никогда не учили правильному толкованию количественной информации. Мы вынуждены делать это сами, иначе нас часто будут вводить в заблуждение, и мы можем совершить непоправимые ошибки.

КАК ПОСТРОЕНА ЭТА КНИГА

Как мы уже отмечали, мы начинаем часть I с поиска общего языка, фокусируясь на идеях корреляции и причинно-следственной связи как краеугольных камнях количественного анализа.

Развивая эти идеи, часть II фокусируется на том, как мы используем данные и свидетельства, чтобы выяснить, существует ли причинно-следственная связь между явлениями окружающего мира. Назначение этой части книги – показать читателям, что из анализа доступной информации можно извлечь много полезного даже до причинно-следственных выводов. Глава 4 объясняет невероятно распространенную ошибку выбора зависимой переменной, показывая невозможность установить корреляцию без вариации и приводя в пример ошеломляющее количество случаев, когда эта ошибка действительно имеет значение. Глава 5 посвящена измерению корреляций с упором на графическое объяснение регрессии. В главе 6 рассказывается о статистической значимости и проверке гипотез с точки зрения нашего любимого уравнения, которое повторяется на протяжении всей книги:

$$\text{Оценка} = \text{Оцениваемая величина} + \text{Смещение} + \text{Шум.}$$

Если в главе 4 еще не была достигнута эта цель, то в главе 7 становится ясно, как много зависит от выявления взаимосвязей в данных. Мы обсудим проблемы *p*-манипуляций, публикационного смещения и связанных с ними проблем. Наконец, в главе 8 рассматривается редко обсуждаемая тема возврата к среднему значению, а затем мы объединяем ее с ранее упомянутой темой публикационного смещения, чтобы отразить кризис репликации и распространенное явление, заключающееся в сокращении научных оценок с течением времени.

Часть III посвящена причинному выводу, напоминая читателям о том, насколько важно понимание причинности для принятия решений о том, как взаимодействовать с миром. В главе 9 объясняется, почему корреляция не обязательно подразумевает причинно-следственную связь, обсуждаются как искажающие факторы, так и обратная причинно-следственная связь. Глава 10 посвящена вопросу статистического контроля и содержит графические пояснения относительно регрессии. В главах 11–13 представлен обзор того, как ученые используют специальные исследовательские приемы, чтобы попытаться узнать о причинно-следственных связях. В главе 11 рассматриваются как рандомизированные, так и естественные эксперименты, в которых представлены инструментальные переменные как метод решения проблем несоответствия. Главы 12 и 13 посвящены экспериментальным методам «разрывной регрессии» и «разности разностей» соответственно. Глава 14 завершает эту часть книги обсуждением трудностей изучения причинно-следственных механизмов.

Часть IV указывает на то, что мы еще не закончили работу, когда занялись причинно-следственными связями. Даже надежное знание причинно-следственных связей само по себе недостаточно, чтобы гарантировать ясное понимание того, как использовать количественную информацию для принятия правильных решений. В главе 15 показано, как легко обмануть себя, думая, что часть количественной информации отвечает на один вопрос, когда на самом

деле она отвечает на совершенно другой вопрос. Мы призываем читателей избегать этой ошибки, извлекая из технической информации суть. Далее мы вводим правило Байеса. Глава 16 посвящена вопросам измерения, внешней валидности и экстраполяции, что также приводит нас к обсуждению смещения выборки. И наконец, в главе 17 рассматриваются фундаментальные ограничения, с которыми сталкивается количественный анализ при принятии обоснованных решений независимо от того, насколько четко он продуман.

В конце каждой главы есть упражнения, которые читатели могут выполнить самостоятельно, чтобы убедиться, что они усвоили материал. Некоторые из этих упражнений включают анализ данных, с ним могут справиться читатели и студенты, которые научились (или учатся) использовать статистическое программное обеспечение, такое как Stata или R. В конце каждой главы также есть раздел «Дополнительное чтение и ссылки», где любознательные читатели найдут источники, упомянутые в основном тексте, и смогут более глубоко погрузиться в ту или иную тему.

Кому адресована эта книга?

Мы надеемся, что эта книга будет полезна всем, кто хочет научиться видеть суть данных, фактов и количественных рассуждений. Как уже было сказано, мы использовали эти материалы для обучения очень широкой аудитории: от студентов до высококвалифицированных специалистов.

По нашему мнению, чтобы подготовиться к жизни в наш век данных, каждый студент должен познакомиться с подобным материалом, в идеале в первые пару лет обучения в вузе. Поэтому мы написали книгу в надежде, что она будет полезна преподавателям различных дисциплин, обучающих количественному анализу, будь то общеобразовательные курсы или вводные занятия в рамках специального курса. Мы считаем, что это особенно актуально для преподавателей, которые хотят использовать более концептуальный подход, чем традиционно основанный на статистических или математических методах, но при этом опирающийся на технический фундамент.

Мы считаем, что книга так же хорошо подойдет профессионалам, заинтересованным в повышении квалификации. Например, мы обучаем этому аспирантов, получающих степень магистра в области государственной политики. Некоторые продолжают посещать дополнительные технические курсы по эконометрике или анализу данных. Но для многих важно научиться критически осмысливать количественную информацию. Наш подход соответствует потребностям этих студентов и в то же время обеспечивает понятийную основу, которая понадобится более технически подкованным студентам на будущих курсах.

Коллеги из других университетов также использовали эти материалы в более продвинутых курсах для специалистов по общественным наукам, которые, например, должны изучать количественные методы при подготовке к написанию диссертации. В этом контексте наша книга только выигрывает от совместного использования с другими изданиями, которые являются более техническими или уделяют больше внимания вопросам статистических вычислений. Мы надеемся, что во всех этих случаях будут полезными упражнения в конце

каждой главы. Особенно это относится к прикладному анализу, примеры наборов данных для которого можно скачать в интернете.

Наконец, мы также считаем, что эта книга будет полезна многим аспирантам. Часто в аспирантуре статистику преподают быстро и на высоком техническом уровне. Это может быть продуктивно; овладение передовыми методами одновременно сложно и важно. Но, по нашему опыту, даже лучшие аспиранты могут упустить из виду то, что действительно важно – то, как мы узнаем о мире из данных, – поскольку они сосредотачиваются на доказательстве теорем и программировании алгоритмов. Мы очень надеемся, что эта книга послужит руководством для таких студентов, помогая четко видеть общую картину, даже когда они усердно работают над техническими деталями.

БЛАГОДАРНОСТИ

Как мы уже упоминали, некоторые материалы в этой книге были разработаны совместными усилиями для студенческого курса Энтони и учебного курса для руководителей, который совместно разработали Джейк Шапиро, Лиам Коллинз, Кэти Фетелл и Итан. Мы выражаем огромную благодарность и глубокую признательность Джейку, Лиаму и Кэти.

Хотелось бы поблагодарить Скотта Эшворта, Криса Берри, Криса Блаттмана, Мэтта Бремса, Брюса Буэно де Мескиту, Кервина Чарльза, Девина Чесни, Линдси Кормак, Энди Эггерса, Натана Фаверо, Алекса Фуирне, Мэтта Гейбелла, Джеффа Гроггера, Энди Холла, Косуке Имаи, Ренана Левина, Эндрю Литтла, Йенса Людвига, Мордехая Магенси, Эндрю Минса, Пабло Монтаня, Эмили Риттер, Стива Шваба, Майка Спагата, Дастина Тингли, Стефана Уолтона и Остина Райта за их невероятно полезные отзывы.

Том Будеску, Гаутам Наир, Том Насет, Джефф Рафф, Ванита Вируудачалам, Бекки Ван и Синъюй Инь оказали потрясающую помощь в исследованиях на ранних этапах создания этого проекта. Было очень интересно работать с ними, и мы благодарны за их вклад.

Мы хотели бы поблагодарить наших студентов за то, что они заметили многочисленные ошибки и опечатки в черновиках книги. А. К. Алилон, Дениз Азаде, Элли Ратки и Аль Шах обнаружили их поистине смущающее количество. Спасибо!

Команда издательства Princeton University Press была потрясающей. Мы особенно признательны Бриджит Флэннери-Маккой за веру в проект и ее руководство, а также Алене Чекановой за контроль над процессом. И мы в долгу перед Данной Локвуд за ее невероятную работу, которая помогла улучшить процесс написания и структуру книги. Мы также очень ценим потрясающий контроль производства Мелоди Негрон и всегда отличную индексацию терминов Дэвида Люляка.

Итан благодарит своих коллег из школы Харриса и многочисленных сотрудников, соавторов и учеников, которые являются интеллектуальным источником вдохновения и за годы героических усилий улучшили ясность его собственного мышления. Он глубоко благодарен своей жене Ребекке, которая мирилась с нервозностью автора, неизбежно сопровождающей завершение книги, и воспринимала ее с неизменной поддержкой, любовью и терпением,

с которыми она встречает все радости жизни с Итаном. Итан посвятил эту книгу своим детям, Ханне и Эйбу, доставляющим ему только радость и удовольствие. Его самая искренняя надежда состоит в том, что книга широко разоидется по учебным заведениям и однажды его дети будут учиться по учебнику, посвященному им самим. Это был бы статистически значимый успех.

Энтони благодарит своих консультантов, соавторов и коллег, которые помогают ему мыслить яснее. Он благодарен своим родителям, поощряющим и поддерживающим его на протяжении всей жизни, даже когда он не планировал поступать на юридический или медицинский факультет. И самое главное, он благодарит Глорию, свою жену и лучшую подругу, которая прочитала бесчисленные черновики научных статей, выдержала слишком много разговоров о регрессиях, развлекалась вводом данных, проверяла каждую идею мужа, вносила непропорционально большую долю своих собственных идей и украшала его жизнь способами, которые не поддаются количественной оценке.

Глава 1

Критическое мышление в эпоху данных

О ЧЕМ ЭТА ГЛАВА

- Навык ясного и концептуального восприятия количественной информации важен по многим причинам, даже если вы не заинтересованы в карьере аналитика данных.
- Даже хорошо обученные профессионалы часто допускают серьезные ошибки при обработке данных.
- Мышление и данные дополняют, а не заменяют друг друга.
- Навыки, которые вы приобретете в этой книге, помогут вам использовать фактические данные для принятия более эффективных решений в личной и профессиональной жизни и стать более вдумчивыми и хорошо информированными людьми.

ВВЕДЕНИЕ

Мы живем в век данных. По словам бывшего генерального директора Google Эрика Шмидта, современный мир каждые два дня создает столько же новых данных, сколько было создано с начала времен до 2003 г. Предполагается, что вся эта информация призвана улучшить нашу жизнь, но, чтобы использовать ее скрытую мощь, мы должны научиться искусству здравомыслия в мире, основанном на данных. Это сложная задача, особенно если не уметь абстрагироваться от технических деталей, которыми обычно окружены данные и их анализ.

Ясное мышление в эпоху данных – это, прежде всего, сосредоточенность на идеях и вопросах. Формулы и алгоритмы должны служить этим идеям и вопросам. К сожалению, курсы по статистике и количественному анализу, на которых большинство людей изучают науку о данных, делают прямо противоположное – т. е. фокусируются на технических деталях. Студенты изучают математические формулы, запоминают названия статистических процедур и начинают манипулировать числами, хотя никто не позаботился о ясном и принципиальном понимании того, что они делают и почему они это делают. Такой подход может сработать среди людей, для которых математическое мышление является естественным. Но мы считаем, что это контрпродуктивно для подавляющего большинства из нас. Когда формальные курсы подталкива-

ют учащихся к тому, чтобы перестать думать и начать запоминать, они теряют лес за деревьями. И это совсем не весело.

Эта книга, напротив, сосредоточена на концептуальном понимании. Какие признаки мира вы сравниваете, когда анализируете данные? На какие вопросы отвечают различные виды сравнений? У вас есть правильный вопрос и сравнение для задачи, которую вы пытаетесь решить? Почему ответ, звучащий убедительно, на самом деле может вводить в заблуждение? Какие творческие подходы можно использовать, чтобы дать более информативный ответ?

Мы не стремимся преуменьшить значение технических знаний, но полагаем, что технические навыки без концептуального понимания или критического мышления – это путь к катастрофе. По нашему мнению, как только у вас появится интуитивное понимание количественного анализа и как только вы поймете, почему так важно задавать аккуратные и точные вопросы, технические навыки придут сами собой. К тому же этот способ веселее.

Вдохновленные этими соображениями, мы написали книгу, которая не требует предварительного знакомства с анализом данных, статистикой или количественными методами. Считая, что концептуальное мышление важнее, мы минимизировали (хотя, конечно, не исключили) технический материал в пользу объяснений на простом разговорном языке, где это возможно. Надеемся, что эта книга будет использоваться в качестве введения и руководства к тому, как правильно мыслить и проводить количественный анализ. Мы считаем, что любой из нас может стать искушенным потребителем (и даже производителем) количественной информации. Требуется лишь немного терпения, настойчивости, тяжелой работы и твердой решимости никогда не позволять технике заменять критическое мышление.

Большинство людей не становятся профессионалами количественного анализа. Но, независимо от того, встанете вы на этот путь или нет, мы уверены, что вы будете использовать навыки, полученные в этой книге, самыми разными способами. У многих из вас будут количественные аналитики, работающие на вас или вместе с вами. И все вы будете читать исследования, новостные репортажи и брифинги, на которых кто-то пытается убедить вас в правильности выводов, используя количественный анализ. Эта книга даст вам навыки критического мышления, необходимые для того, чтобы задавать правильные вопросы, проявлять скептицизм, когда это необходимо, и отличать полезные свидетельства от вводящих в заблуждение.

ПОУЧИТЕЛЬНЫЕ ИСТОРИИ

Чтобы разжечь ваш аппетит к предстоящей тяжелой работе, давайте начнем с нескольких поучительных историй, которые подчеркивают важность критического мышления в эпоху данных.

Поспешный диагноз Эйба

Первый ребенок Итана, Эйб, родился в июле 2006 г. В младенчестве он кричал и плакал почти без перерыва по ночам в течение первых пяти месяцев. В остальном Эйб был счастлив и здоров, хотя и весил немного меньше нормы. Когда ему исполнился год, семья переехала в Чикаго, а иначе вы не читали бы

эту книгу. (Последнее предложение представляет собой пример особого вида утверждений, называемых *контрфактическими* (counterfactual). Контрфактические утверждения действительно важны, и вы узнаете о них больше в главе 3.) Заметив, что Эйб был маловат для своего возраста и рос медленнее, чем полагалось, его педиатр решил провести несколько анализов.

После ряда лабораторных исследований врачи были почти уверены, что у Эйба целиакия – заболевание пищеварительной системы, характеризующееся непереносимостью глютена. Хорошая новость: целиакия не опасна для жизни и даже не очень серьезна, если ее правильно лечить. Плохая новость: в 2007 г. варианты безглютеновой диеты для маленьких детей были довольно убогими.

Оказывается, Эйбу на самом деле провели два анализа крови на целиакию. Один результат оказался положительным (что указывает на наличие заболевания), другой – отрицательным (указывает на отсутствие заболевания). По словам врачей, точность положительного теста превышает 80 %. «Это достоверный диагноз», – сказали они. Предложенный план действий заключался в том, чтобы посадить Эйба на безглютеновую диету на пару месяцев и посмотреть, не увеличится ли его вес. При таком исходе врачи могли бы сделать более точные дополнительные анализы или просто оставить Эйба без глютена на всю оставшуюся жизнь.

Итан захотел взглянуть на распечатку анализов крови Эйба. Врачи сказали, что вряд ли это будет полезно, поскольку Итан не врач. В таком ответе нет ничего удивительного, и его легко объяснить. Люди, особенно эксперты и авторитетные лица, часто не любят признавать ограниченность своих знаний. Но Итан хотел принять правильное решение для своего сына, поэтому настаивал на информации. Одна из важных целей этой книги – дать вам уверенность в себе, чтобы вы могли защищать себя, используя информацию для принятия жизненно важных решений.

Эффективность любого диагностического теста характеризуют два числа. Во-первых, это доля *ложноотрицательных* результатов, т. е. насколько часто тест показывает, что больной человек здоров. Во-вторых, это доля *ложноположительных* результатов, т. е. как часто тест показывает, что здоровый человек болен. Чтобы правильно интерпретировать результаты диагностического теста, вам необходимо знать *оба* числа. Поэтому заявление врачей Эйба о том, что положительный анализ крови был точным на 80 %, было не очень информативным. Означало ли это, что у теста 20 % ложноотрицательных результатов? 20% ложноположительных результатов? Или тест говорит о том, что у 80 % людей с положительным результатом теста есть целиакия?

К счастью, беглый поиск в Google позволил найти значения ложноположительных и ложноотрицательных результатов для обоих тестов Эйба. Вот что узнал Итан. Тест, который у Эйба дал положительный результат на целиакию, имеет долю ложноотрицательных результатов около 20 %. То есть, если 100 человек с целиакией пройдут тест, около 80 из них получают правильный положительный результат, а остальные 20 – ошибочно отрицательный. Мы предполагаем, что именно на этом факте и основано заявление врача о 80-процентной точности. Тест, однако, имеет 50 % ложноположительных результатов! Люди, у которых нет целиакии, с одинаковой вероятностью могут получить как положительный, так и отрицательный результат. (Следует отметить, что этот тест больше не рекомендуется для диагностики целиакии.) Напротив, тест, соглас-

но которому у Эйба нет целиакии, имел гораздо меньшую долю ложноотрицательных и ложноположительных результатов.

До изучения результатов анализов разумная оценка вероятности заболевания Эйба целиакией с учетом его отставания в наборе веса составляла примерно 1 к 100. То есть примерно у одного из каждых 100 детей с отставанием набора веса есть целиакия. Вооружившись лабораторными отчетами и показателями ложноположительных и ложноотрицательных результатов, Итан смог подсчитать вероятность того, что при имеющихся результатах анализов и наличии отставания веса у Эйба действительно есть целиакия. Удивительно, но сочетание положительного результата менее точного теста и отрицательного результата более точного теста на самом деле означало, что вероятность заболевания целиакией у Эйба гораздо ниже, чем 1 из 100. Фактически, как мы покажем вам в главе 15, наибольшая оценка вероятности заболевания Эйба целиакией, учитывая результаты анализов, составляла примерно 1 из 1000. Анализы крови, которые, как были уверены врачи Эйба, подтверждали диагноз целиакии, на самом деле убедительно указывали на противоположный вывод. Эйб почти наверняка не был болен целиакией.

Итан позвонил врачам, чтобы объяснить, что он выяснил, и предположить, что перевод его сына, одержимого макаронами, на безглютеновую диету, возможно, на всю жизнь не является разумным следующим шагом. На это врачи ответили: «Просто вам трудно принять диагноз». Тогда Итан нашел нового педиатра.

Каков же результат? У Эйба не было целиакии. Это был просто маленький ребенок, который в силу индивидуальных особенностей рос немного медленнее, чем другие. Сегодня это обычный ребенок с ненасытным аппетитом. Но если бы его отец не умел работать с количественными показателями или ему не хватало уверенности, чтобы бросить вызов самоуверенному эксперту, бедняга Эйб провел бы свое детство, питаясь рисовыми лепешками. Рисовые лепешки отвратительны, так что Эйб рисковал никогда не набрать вес.

Гражданское сопротивление

Многим людям на своем опыте довелось хоть раз испытать глубокое несогласие со своим правительством. Когда дела идут особенно плохо, они иногда решают провести акции протеста. Если вы когда-нибудь займетесь организацией подобной акции протеста, то столкнетесь с необходимостью принимать важные решения. Например, вам придется сделать выбор, какое движение предпочесть – с мягкой ненасильственной стратегией или допускающее более жесткие формы конфронтации? Размышляя над этим затруднительным положением, вы наверняка прислушаетесь к своим внутренним этическим убеждениям. Но вам будет полезно узнать, что говорят фактические данные о недостатках и преимуществах каждого подхода. Какая организация, скорее всего, добьется изменений в политике правительства? Какой из этих подходов с большей вероятностью приведет вас в тюрьму, больницу или морг?

Существуют количественные данные, которые вы можете использовать для обоснования своих решений. Во-первых, если сравнить антиправительственные движения по всему миру на протяжении достаточно длительного времени, то станет видно, что правительства чаще идут на уступки полностью не-

насильственным группам, чем группам, применяющим насилие. И даже если рассматривать только группы, применяющие насилие, можно сделать вывод, что правительства чаще идут на уступки группам, применяющим насилие против военных и правительственных объектов, а не против гражданского населения. Во-вторых, личные риски, связанные с насильственным протестом, выше, чем риски, связанные с ненасильственным протестом. Правительства подавляют насильственные протесты чаще, чем ненасильственные, что делает опасения по поводу тюрьмы, больницы и морга еще более острыми.

Эти аргументы звучат весьма убедительно. Ненасильственная стратегия кажется очевидным выбором. Очевидно, что это и более эффективно, и менее рискованно. И действительно, на основе такого рода данных политологи Эрика Ченовет и Эван Перкоски приходят к выводу, что «планирование и подготовка ненасильственных акций протеста имеют ключевое значение, особенно (и это парадоксально) при противостоянии жестоким режимам».

Но давайте присмотримся к доказательствам. Начнем с вопроса: «В какой ситуации группа активистов выберет ненасильственный протест вместо насильственных действий?» Нам приходит в голову несколько мыслей. Возможно, люди охотнее примут участие в ненасильственном протесте, когда имеют дело с правительством, которое, по их мнению, с большей вероятностью прислушается к требованиям своих граждан. Или, возможно, люди скорее выберут ненасильственный протест, если они имеют широкую поддержку среди своих сограждан, представляют влиятельную группу в обществе, которая может привлечь внимание средств массовой информации, или сталкиваются с менее жестким правительством.

Если что-то из сказанного верно, нам следует беспокоиться по поводу утверждения о том, что проведение ненасильственных акций является ключом к построению успешного антиправительственного движения. (Это вовсе не значит, что мы призываем к насилию!) Давайте разберемся, в чем дело.

Эмпирические исследования показывают, что в среднем правительства чаще идут на уступки в тех местах, где прошли именно ненасильственные протесты. Данный вывод основан на буквальной интерпретации разницы, а именно на том, что более высокая частота уступок со стороны государства вызвана использованием ненасильственной тактики. Иными словами, при прочих равных условиях, если бы некое движение, использующее насильственные методы, перешло на использование ненасильственных методов, правительство с большей вероятностью пошло бы на уступки. Но действительно ли такая причинная интерпретация оправдана фактами?

Предположим, что протестные движения с большей вероятностью прибегнут к насилию, если они не имеют широкой поддержки среди своих сограждан. Далее, когда мы сравниваем места, где были насильственные протесты, с местами проведения мирных акций, все остальные условия (кроме тактики протеста) нельзя считать равными. Эти места отличаются как минимум по двум причинам. Во-первых, они различаются по факту наличия насильственных и ненасильственных протестов. Во-вторых, они различаются по степени поддержки протестного движения общественностью.

Это второе отличие представляет собой проблему для причинной интерпретации. Вполне логично предположить, что общественное мнение оказывает

независимое влияние на готовность правительства пойти навстречу протестующим. То есть при прочих равных условиях (включая тактику протеста) правительство с большей готовностью пойдет на уступки протестным движениям, пользующимся широкой общественной поддержкой. Если это так, то мы не можем однозначно утверждать, что правительства идут на уступки из-за разницы в тактике протеста. Возможно, все дело в широкой общественной поддержке более миролюбивых движений. Это классическая проблема ошибочного приятия корреляции за причинно-следственную связь.

Стоит отметить несколько моментов. Во-первых, если уступки правительства на самом деле обусловлены общественным мнением, то с учетом этого влияния может оказаться, что мирные протесты не эффективнее насильственных (они могут быть даже менее эффективными). Располагая ограниченными доказательствами, мы просто не можем знать этого наверняка.

Во-вторых, если не заставлять себя мыслить критически, то в этом примере мы приходим к выводу, который нам больше нравится. Кто из нас не хотел бы жить в мире, где миролюбие всегда предпочтительнее насилия? Но весь смысл использования доказательств, помогающих нам принимать решения, состоит в том, чтобы заставить нас признать, что мир не всегда устроен так, как нам хочется или как мы верим. На самом деле именно в ситуациях, когда кажется, что факты говорят то, что вы хотели бы услышать, особенно важно заставить себя сохранить ясность мышления.

В-третьих, мы указали лишь на одну проблему в оценке последствий мирного и насильственного протеста, но есть и другие. Например, подумайте о другом эмпирическом утверждении, которое мы обсуждали: насильственные протесты с большей вероятностью спровоцируют правительство на репрессивные меры, чем мирные акции. Напомним, мы предположили, что люди с большей вероятностью будут участвовать в мирных протестах, когда они меньше злятся на свое правительство, возможно, потому что правительство ведет себя менее жестко. Спросите себя, почему, если это правда, у нас возникает аналогичная проблема интерпретации? Почему тот факт, что после насильственных протестов правительство применяет больше репрессий, чем после мирных акций, не обязательно означает, что переход от насилия к ненасилию снизит риск репрессий? Этот аргумент следует той же логике, что и ранее рассмотренная в отношении уступок. Если вы пока не понимаете, как работает этот аргумент, ничего страшного. Все станет ясно к концу главы 9.

Теория разбитых окон

В 1982 г. криминолог Джордж Келлинг и социолог Джеймс Уилсон опубликовали в *The Atlantic* статью, где предложили новую теорию преступности и полицейской деятельности, которая оказала огромное и долгосрочное влияние на криминальную политику в Соединенных Штатах и за их пределами.

Это знаменитая *теория разбитых окон*. Она была вдохновлена успехом программы в Ньюарке, штат Нью-Джерси, в соответствии с которой полицейские вышли из машин и патрулировали улицы пешком. По словам Келлинга и Уилсона, программа снизила уровень преступности за счет повышения «уровня общественного порядка». Общественный порядок важен, утверждают они, потому что его отсутствие запускает порочный круг:

«Участок заброшен и зарос сорняками, окна в доме разбиты. Родители перестают ругать непослушных детей... Семьи выезжают, заселяются одинокие взрослые. Перед магазином на углу собираются подростки. Владелец просит их уйти; они отказываются. Происходят драки. Растут горы мусора. Люди начинают пить алкоголь прямо у входа в продуктовый магазин... Жители замечают рост количества преступлений, особенно насильственных... Они стараются реже выходить на улицы... Такой район подвержен дальнейшей криминализации».

Идея о том, что работа полиции, сосредоточенная на минимизации любых проявлений беспорядка, может снизить уровень насильственных преступлений, оказала большое влияние на тактику полиции. Теория разбитых окон легла в основу стратегии муниципалитета Нью-Йорка в 1990-е гг. В своей речи 1998 г. тогдашний мэр Нью-Йорка Рудольфо Джулиани сказал:

«Мы сделали теорию разбитых окон неотъемлемой частью нашей правоохранительной стратегии...

Мы концентрируемся на мелочах и посылаем всем четкий сигнал о том, что этот город заботится о поддержании закона и порядка... в итоге город в целом станет безопаснее».

И действительно, преступность в Нью-Йорке ощутимо снизилась, когда полиция начала концентрироваться «на мелочах». Согласно исследованию Хоуп Корман и Наси Мокана, в 1990-е гг. количество арестов за правонарушения увеличилось на 70 %, а количество насильственных преступлений снизилось более чем на 56 %, что вдвое лучше, чем средний показатель по стране.

Чтобы оценить, в какой степени политика борьбы с разбитыми окнами повлияла на падение уровня преступности, Келлинг и Уильям Соуза изучили взаимосвязь между насильственными преступлениями и применением стратегии борьбы с «разбитыми окнами» на территории различных полицейских участков Нью-Йорка. Если наведение порядка ведет к сокращению насильственных преступлений, утверждали они, то нам следует ожидать, что наибольшее снижение преступности произойдет в округах, где полиция больше всего сосредоточилась на стратегии «разбитых окон». Именно это они и обнаружили. На участках, где арестов за мелкие правонарушения было больше, насильственная преступность снизилась сильнее. Они подсчитали, что «для среднего полицейского округа Нью-Йорка... снижение на одно насильственное преступление приходится примерно на 28 дополнительных арестов за мелкие правонарушения».

Звучит довольно убедительно. Но давайте не будем спешить с выводом, что арест людей за мелкие правонарушения – это рецепт прекращения насильственных преступлений. Два других ученых, Бернард Харкорт и Йенс Людвиг, призывают нас более взвешенно отнестись к тому, что говорят данные.

Проблема, на которую указывают Харкорт и Людвиг, – это так называемый *возврат к среднему значению* (о котором мы подробнее поговорим в главе 8). В любой конкретный год количество преступлений на участке определяется множеством факторов, включая работу полиции, наркотики, экономику, погоду и т. д. Многие из этих факторов нам неизвестны. Некоторые из них мимолетны;

они возникают и исчезают из года в год. Поэтому мы можем предположить, что на любом конкретном участке существует некоторый «базовый» уровень преступности, причем в некоторые годы преступность случайным образом выше, а в некоторые годы – ниже (относительно базового уровня для данного участка).

Если в определенный год на участке был высокий уровень преступности (относительно его базового уровня), то ему не везло с неизвестными и мимолетными факторами, которые способствуют совершению преступлений. Вероятно, следующий год будет не таким плохим (вот что значит мимолетность), так что на этом участке, скорее всего, снизится преступность. А если в этом году на участке уровень преступности был ниже базового уровня, то ему повезло с неизвестными и мимолетными факторами, а в следующем году, вероятно, будет хуже (преступность снова пойдет вверх). Таким образом, год за годом уровень преступности на участке имеет тенденцию возвращаться к среднему (т. е. базовому уровню).

Теперь представьте себе участок, в котором в конце 1980-х гг. был действительно высокий уровень насильственных преступлений. В отношении этого участка, вероятно, справедливы две вещи. Во-первых, это, вероятно, участок с высоким базовым уровнем насильственных преступлений. Во-вторых, это также, вероятно, участок, в котором год или два были неудачными, т. е. по уникальным и мимолетным причинам уровень преступности в конце 1980-х гг. был высоким по сравнению с базовым уровнем этого участка. То же самое, конечно, верно и наоборот для участков, где в конце 1980-х гг. был низкий уровень преступности. У них, вероятно, низкий базовый уровень преступности, и вдобавок выдалась пара хороших лет.

Почему это проблема для выводов Келлинга и Соузы? Благодаря возвращению к среднему значению мы ожидаем, что в наиболее криминальных районах конца 1980-х гг. в среднем будет наблюдаться снижение насильственных преступлений, даже без каких-либо изменений в работе полиции. И вполне логично с точки зрения полиции, но, к несчастью для исследования, именно на участках, где в конце 1980-х гг. был высокий уровень преступности, чаще всего в начале 1990-х гг. полицейские работали по программе борьбы с «разбитыми окнами». Итак, когда мы видим снижение количества насильственных преступлений на участках, где активно велась полицейская деятельность по борьбе с мелкими преступлениями, мы не знаем, что сработало на самом деле – полицейская стратегия или возврат к среднему значению.

Харкорт и Людвиг пошли еще дальше, пытаясь найти более убедительные доказательства. Они изучили то, как изменения в количестве арестов за мелкие правонарушения связаны с изменениями в уровне насильственных преступлений на участках, где в конце 1980-х гг. наблюдался аналогичный уровень насильственных преступлений. Сравнивая участки с аналогичным начальным уровнем насильственных преступлений, они в некоторой степени устраняют проблему возврата к среднему значению. Удивительно, но это простое изменение на самом деле меняет отношения! Вместо того чтобы подтвердить вывод Келлинга и Соузы о том, что аресты за мелкие правонарушения ведут к снижению количества насильственных преступлений, Харкорт и Людвиг обнаружили, что на участках, которые больше внимания уделяли арестам за мелкие правонарушения, на самом деле, похоже, наблюдался рост

насильственных преступлений. Это полная противоположность нашим ожиданиям от теории разбитых окон.

Впрочем, это изменение не опровергает эффективность борьбы с разбитыми окнами. Связь между арестами за мелкие проступки и насильственными преступлениями, которую обнаружили Харкорт и Людвиг, может существовать по множеству причин. Например, возможно, районы с увеличением количества правонарушений в целом становятся менее безопасными и в них будет больше насильственных преступлений независимо от стратегий полиции. На самом деле все эти рассуждения свидетельствуют лишь об одном – что данные, если их правильно рассмотреть, определенно не дают однозначного подтверждения теории разбитых окон, как можно было подумать, исходя из открытия Келлинга и Соузы. И увидеть это можно только в том случае, если у вас есть способность критически анализировать некоторые тонкие нюансы.

Ошибочное мышление сыграло свою роль. Основанные на фактических данных выводы Келлинга и Соузы убедили политиков и городскую администрацию в том, что работа полиции с мелкими правонарушениями является правильным путем вперед, хотя на самом деле она могла отвлечь ресурсы от предотвращения и расследования насильственных преступлений и, возможно, создать более враждебные отношения между полицией и непропорционально бедным населением и национальными меньшинствами, которых часто обвиняли в мелких нарушениях.

ДОПОЛНЕНИЕ ИЛИ ЗАМЕНА?

Наш количественный мир наполнен новыми интересными данными и аналитическими инструментами для анализа этих данных с причудливыми названиями, такими как алгоритмы машинного обучения, искусственный интеллект, случайные леса и нейронные сети. Мы все чаще слышим, что новые технологии позволят машинам думать за нас. Но это не так. Как подчеркивают наши поучительные истории, никакой анализ данных, каким бы футуристическим ни было его название, не будет работать, если мы не задаем правильные вопросы, если мы не проводим правильные сравнения, если основные предположения неверны или используемые данные не подходят. Тот факт, что аргумент опирается на сложный количественный анализ данных, не означает, что этот аргумент строгий или правильный. Чтобы использовать возможности данных для принятия более эффективных решений, мы должны сочетать количественный анализ с критическим мышлением.

Наши истории также показывают, как интуиция может сбить нас с пути. Требуется много внимания и практики, чтобы научиться критически воспринимать «очевидные» доказательства. Интуитивные выводы врачей о том, что у Эйба была целиакия, из-за теста с 80-процентной точностью и выводы исследователей о том, что полицейская стратегия борьбы с «разбитыми окнами» работает, потому что преступность снизилась в тех местах, где она была применена, кажутся разумными. Но оба интуитивных вывода оказались ошибочными, и мы предположили, что необходимо скептически относиться к первоначальным догадкам. Хорошая новость заключается в том, что критическое мышление можно довести до автоматизма, если его регулярно применять.

Данные и инструменты количественного анализа не заменяют критического мышления. На самом деле навыки работы с количественными данными без критического мышления весьма опасны. Читая следующие главы, вы наверняка будете потрясены тем, в какой степени косное мышление влияет даже на самые важные решения, принимаемые людьми. В ходе прочтения этой книги вы увидите, как неверно истолкованная информация ведет к ошибочным решениям, от которых зависит жизнь больного, как она искажает национальную и международную политику борьбы с терроризмом, вносит путаницу в деловые и филантропические решения, принимаемые самыми богатыми людьми мира, как мы ошибочно устанавливаем приоритеты в образовании наших детей и заблуждаемся во множестве других вопросов, от банальных до глубоких. По сути, ни один аспект жизни не застрахован от критических ошибок в понимании и интерпретации количественной информации.

По нашему опыту, это происходит потому, что искаженное восприятие доказательств глубоко укоренилось в человеческой психологии. Разумеется, наша собственная интуиция, оставленная без контроля, часто подвержена фундаментальным ошибкам. Мы предполагаем, что и ваша тоже. Самое тревожное – что эксперты, от чьих советов вы зависите – будь то врачи, бизнес-консультанты, журналисты, учителя, финансовые консультанты, ученые или кто-то еще, – склонны к совершению таких же ошибок. Поскольку они считаются экспертами, мы безоговорочно доверяем их суждениям, а они не сомневаются в собственных умозаключениях. Вот почему так важно научиться критически воспринимать количественные доказательства в первую очередь для себя. Это единственный способ научиться задавать правильные вопросы, которые приведут вас и тех, от чьих советов вы зависите, к наиболее надежным и продуктивным выводам.

Как получается, что эксперты в столь многих областях так часто допускают существенные ошибки? Экспертное знание в любой области приходит в результате обучения, практики и опыта. Никто не рассчитывает стать экспертом в области инженерии, финансов, медицины или сантехники без обучения и многих лет работы. Но, несмотря на фундаментальную значимость вопроса, почти никто не прикладывает аналогичные усилия, чтобы научиться корректно и непредвзято работать с данными. И даже когда люди стремятся к этому, их преподаватели склонны преувеличивать технические аспекты и недооценивать концептуальные, хотя фундаментальные проблемы почти всегда связаны с концептуальными ошибками в мышлении, а не с техническими ошибками в расчетах.

Отсутствие опыта критического мышления ставит перед нами две проблемы. Во-первых, если столько экспертных советов и выводов ненадежны, откуда нам знать, чему верить? Во-вторых, как выделить мнения экспертов, которые действительно отражают критическое мышление?

В этой книге мы закладываем основу решения упомянутых проблем. Каждая из последующих глав объясняет и иллюстрирует на множестве примеров фундаментальные принципы критического мышления в мире, управляемом данными. Первая часть помогает найти общий язык с читателем, поясняя, что мы подразумеваем под корреляцией и причинно-следственной связью и в каких случаях применяются эти понятия. Во второй части рассказано, как мож-

но определить, является ли статистическая взаимосвязь подлинной. В третьей части вы узнаете, как установить, отражают ли эти отношения причинность. А в четвертой части обсуждается, как нам следует и не следует включать количественную информацию в процесс принятия решений.

Мы надеемся, что чтение этой книги поможет вам усвоить принципы критического мышления настолько глубоко, что они станут вашей второй натурой. Вы поймете, что находитесь на правильном пути, когда обнаружите, что замечаете основные ошибки в том, как люди думают и говорят о значении доказательств, куда бы вы ни обратились – когда вы смотрите новости, листаете журналы, разговариваете с деловыми партнерами, посещаете врача, слушаете комментарии во время спортивных соревнований, читаете научные статьи или участвуете в общественных мероприятиях. Мы подозреваем, что поначалу будет трудно поверить, сколько чепухи вам регулярно говорят самые разные эксперты. Когда у вас наступит прозрение, постарайтесь оставаться скромными и конструктивными в своей критике. Но не стесняйтесь поделиться этой книгой с теми, чьи аргументы, по вашему мнению, особенно в ней нуждаются. Или, еще лучше, предложите им купить собственный экземпляр!

ДОПОЛНИТЕЛЬНОЕ ЧТЕНИЕ И ССЫЛКИ

Эссе о ненасильственном протесте Эрики Ченоуэт и Эвана Перкоски, которое мы цитируем, можно найти по адресу <https://politicalviolenceataglance.org/2018/05/08/states-are-far-less-likely-to-engage-in-mass-violence-against-nonviolent-uprisings-than-violent-uprisings/>.

Следующая книга содержит дополнительные исследования взаимосвязи между миролюбием и эффективностью протестов:

Erica Chenoweth and Maria J. Stephan. 2011. *Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict*. Columbia University Press.

Следующие статьи упоминались в указанном порядке на тему проверки теории разбитых окон:

George L. Kelling and James Q. Wilson. 1982. Broken Windows: The Police and Neighborhood Safety. *The Atlantic*. March. <https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/>.

Archives of Rudolph W. Giuliani. 1998. *The Next Phase of Quality of Life: Creating a More Civil City*. February 24. <http://www.nyc.gov/html/rwg/html/98a/quality.html>.

Hope Corman and H. Naci Mocan. 2005. *Carrots, Sticks, and Broken Windows*. *Journal of Law and Economics* 48 (1): 235–66.

George L. Kelling and William H. Sousa, Jr. 2001. *Do Police Matter? An Analysis of the Impact of New York City's Police Reforms*. Civic Report for the Center for Civic Innovation at the Manhattan Institute.

Bernard E. Harcourt and Jens Ludwig. 2006. *Broken Windows: New Evidence from New York City and a Five-City Social Experiment*. *University of Chicago Law Review* 73: 271–320. Опубликованная версия имеет опечатку в таблице ключей. Исправление см. в *Errata*, 74 *U. Chi. L. Rev.* 407 (2007).

ЧАСТЬ I

В поиске общего языка

Глава 2

Корреляция: что это такое и для чего она нужна?

О ЧЕМ ЭТА ГЛАВА

- Корреляция говорит нам о том, в какой степени два явления мира имеют тенденцию возникать вместе.
- Чтобы измерить корреляцию двух величин, мы должны иметь данные с изменяющимися значениями двух переменных.
- Корреляции могут быть потенциально полезны для описания, прогнозирования и выявления причинно-следственных связей. Но мы должны четко понимать, когда они подходят для каждой из этих задач.
- Корреляции представляют собой линейные отношения, но это не такое строгое ограничение, как вы думаете.

ВВЕДЕНИЕ

Корреляция не подразумевает причинно-следственную связь. Предельно понятно. Однако, по нашему опыту, от этого краткого высказывания мало пользы. Хотя многие люди запомнили, что корреляция не то же самое, что причинно-следственная связь, на самом деле у них нет четкого понимания определений того и другого.

В первой части книги мы собираемся потратить некоторое время на создание общего словарного запаса. Абсолютно важно убедиться, что мы с вами одинаково понимаем и используем эти и некоторые другие ключевые термины для обозначения одного и того же понятия, если хотим подвергнуть их критическому обсуждению в последующих главах.

Эта глава посвящена изучению корреляции. Мы постараемся пояснить, что это такое и для чего она нужна. Корреляция – это основной инструмент, с помощью которого аналитики количественных данных описывают мир, прогнозируют будущие события и отвечают на научные вопросы. Аккуратные аналитики не избегают и не игнорируют корреляции. Но они должны хорошенько подумать о том, на какие вопросы корреляции могут дать ответы, а от каких лучше воздержаться.

Что такое корреляция?

Корреляция между двумя явлениями мира – это степень, в которой они склонны происходить вместе. Из этого определения следует, что *корреляция* – это связь между двумя явлениями (которые иногда называют *переменными*). Если два явления мира имеют тенденцию проявляться вместе, они *положительно коррелированы*. Если в рассматриваемом мире возникновение одного события никак не связано с возникновением другого, они *некоррелированы*. А если при возникновении одного события мира другое имеет тенденцию не проявляться, такие события называются *отрицательно коррелированными*.

Давайте уточним, что означает высказывание «два явления мира проявляются вместе». Начнем с самого простого примера. Предположим, мы хотим оценить корреляцию между двумя признаками мира, и для каждого из них есть только два возможных значения (мы называем их *бинарными* переменными). Например, деление суток на «после полудня» и «до полудня» характеризуется бинарной переменной (напротив, время, измеряемое в часах, минутах и секундах, не является бинарным; оно может принимать намного больше двух значений).

Политологи и экономисты иногда говорят о *ресурсном проклятии* или *парадоксе изобилия*. Идея состоит в том, что страны с обилием природных ресурсов часто менее экономически развиты и менее демократичны, чем страны с меньшим количеством природных ресурсов. Природные ресурсы могут сделать страну менее склонной инвестировать в другие формы развития или более склонной к насилию и автократии¹.

Чтобы убедиться в наличии или отсутствии ресурсного проклятия, нам придется оценить корреляцию между природными ресурсами и некоторыми особенностями экономической или политической системы. Этот процесс начинается со сбора данных, что мы и сделали. Чтобы измерить природные ресурсы, мы посмотрели, какие страны являются крупнейшими производителями нефти. Мы относим страну к крупному производителю нефти, если она экспортирует более сорока тысяч баррелей в день на миллион человек. Что касается политической системы, мы рассмотрели, какие страны считаются автократиями, а какие демократиями, по мнению авторов проекта Polity IV. В табл. 2.1 показано, сколько стран попадает в каждую из четырех возможных категорий: демократия и крупный производитель нефти, демократия и некрупный производитель нефти, автократия и крупный производитель нефти, автократия и некрупный производитель нефти.

Таблица 2.1. Объем добываемой нефти и тип политической системы

	Некрупный производитель	Крупный производитель	Итого
Демократия	118	9	127
Автократия	29	11	40
Итого	147	20	167

¹ Здесь и далее подобные оценочные суждения выражают личное мнение авторов цитируемых исследований и приведены исключительно в иллюстративных целях. – *Прим. издат.*

Мы можем выяснить, коррелируют ли эти две бинарные переменные (крупный/некрупный производитель и автократия/демократия), путем сравнения. Например, мы могли бы задаться вопросом, являются ли крупные производители нефти более склонными к автократии, чем страны, которые относятся к небольшим производителям. Или аналогичным образом мы могли бы задаться вопросом, являются ли автократии более крупными производителями нефти, чем демократии. Если одно из этих утверждений верно, то и другое должно быть верным. И эти сравнения говорят нам, имеют ли эти два свойства – крупный производитель нефти и автократия – тенденцию встречаться вместе.

В табл. 2.1 добыча нефти и автократия действительно положительно коррелируют. 55 % крупнейших производителей нефти являются автократиями ($11/20 = 0.55$), тогда как среди стран, не являющихся крупными производителями нефти, только около 20 % относятся к автократиям ($29/147 \approx 0.20$). Аналогично крупными производителями нефти являются 27.5 % автократий ($11/40 = 0.275$) и лишь около 7 % демократий ($9/127 \approx 0.07$). Другими словами, крупные производители нефти с большей вероятностью будут автократиями, чем страны, которые не являются крупными производителями нефти, и тогда, естественно, автократии с большей вероятностью будут крупными производителями нефти, чем демократии.

Эта положительная корреляция интересна с описательной точки зрения. Кроме того, она может пригодиться для прогнозирования. Предположим, что за пределами наших данных существуют другие страны, в политической системе правления которых мы не уверены. Знание того, являются ли они крупными производителями нефти, поможет предсказать наиболее вероятную форму правления.

В определенном смысле такие знания могут быть полезны для причинно-следственных выводов. Возможно, в какой-то стране обнаружены большие запасы нефти, и политологам интересно, какое влияние это может оказать на политическую систему страны. Однако, как будет подробно рассмотрено в главе 9, мы должны быть очень осторожны, давая корреляциям такого рода причинную интерпретацию.

Мы можем оценить корреляции, даже если имеющиеся данные не позволяют составить таблицу всех возможных комбинаций, как мы это сделали выше. Предположим, например, что мы хотим оценить взаимосвязь между преступностью и температурой в Чикаго. Мы могли бы составить электронную таблицу, в которой каждая строка соответствует дню, а каждый столбец – признаку каждого дня. Мы часто называем строки *наблюдениями*, а признаки, перечисленные в столбцах, – *переменными*. В этом случае наблюдения проводятся в разные дни. Первой переменной может быть средняя температура в тот день, измеренная в аэропорту Мидуэй. Второй переменной может быть количество преступлений, зарегистрированных в тот день во всем Чикаго. Третья переменная может указывать на то, была ли в тот день на первой полосе газеты Chicago Tribune статья о преступлениях. Как видите, переменные могут принимать бинарные значения (была статья или нет), дискретные, но не бинарные (количество преступлений) или непрерывные (средняя температура). Мы собрали подобные данные для Чикаго в 2018 г. и хотели бы оценить корреляцию между преступностью и температурой. Но как мы можем оценить корреляцию между двумя небинарными переменными?

Одной из отправных точек является построение простого графика, называемого *диаграммой рассеяния*. На рис. 2.1 показаны наши данные по Чикаго за 2018 г. Каждая точка соответствует одному наблюдению в наших данных; здесь это означает, что каждая точка – это день в Чикаго в 2018 г. Горизонтальная ось нашего рисунка – средняя температура в аэропорту Мидуэй в этот день. По вертикальной оси указано количество преступлений, зарегистрированных в городе в этот день. Таким образом, расположение каждой точки отражает среднюю температуру и уровень преступности в данный день.

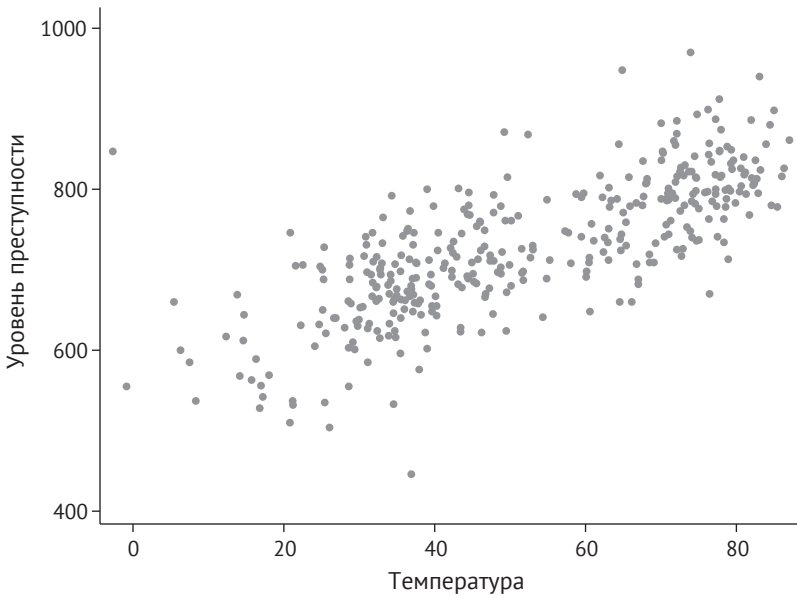


Рис. 2.1. Преступность и температура (в градусах по Фаренгейту) в Чикаго по дням в 2018 г.

Достаточно взглянуть на рисунок, чтобы сделать вывод о наличии положительной корреляции между температурой и преступностью. Точки в левой части диаграммы рассеяния (более холодные дни) также имеют тенденцию располагаться довольно низко по вертикальной оси (низкий уровень преступности), а точки в правой части диаграммы (более теплые дни) расположены довольно высоко по вертикальной оси (более высокий уровень преступности).

Но как нам количественно оценить это визуальное первое впечатление? На самом деле в количественной оценке корреляции можно задействовать различные статистические показатели. Один из таких показателей называется наклоном. Предположим, мы нашли линию, наилучшим образом отражающую расположение точек данных. Под *линией наилучшего соответствия* мы упрощенно подразумеваем линию, для которой среднее расстояние между точками данных и линией является минимальным. (Подробнее об этом мы поговорим в главе 5.) Наклон линии наилучшего соответствия – это один из способов описания корреляции между двумя непрерывными переменными.

На рис. 2.2 показана диаграмма рассеяния с добавленной линией наилучшего соответствия. Наклон линии говорит нам о характере взаимосвязи между этими двумя переменными. Если наклон отрицательный, корреляция отрицательная. Если наклон равен нулю, температура и преступность не коррелируют. Если наклон положительный, корреляция положительная. А крутизна наклона говорит нам о силе корреляции между этими двумя переменными. Здесь мы видим, что переменные положительно коррелируют: в теплые дни преступность обычно выше. В частности, наклон равен 3.1, т. е. в среднем на каждый дополнительный градус температуры (по Фаренгейту) приходится на 3.1 больше преступлений.

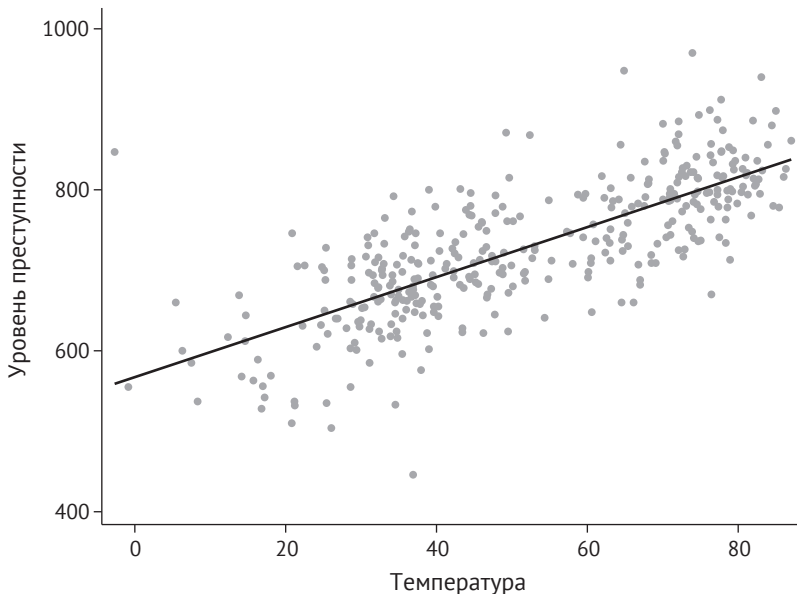


Рис. 2.2. Линия наилучшего соответствия, обобщающая взаимосвязь между преступностью и температурой (в градусах по Фаренгейту) в Чикаго по дням в 2018 г.

Обратите внимание: интерпретация наклона зависит от того, какая переменная находится на вертикальной оси, а какая – на горизонтальной. Если бы мы нарисовали график наоборот (как на рис. 2.3), мы бы описывали взаимосвязь между теми же двумя переменными. Но на этот раз мы бы узнали, что на каждое дополнительное зарегистрированное преступление температура в среднем на 0.18 °F выше. Знак наклона (положительный или отрицательный) не зависит от того, какая переменная находится на горизонтальной или вертикальной оси, поскольку перестановка переменных между осями не меняет их положительную или отрицательную корреляцию. Но числовое значение наклона и его содержательная интерпретация изменились.

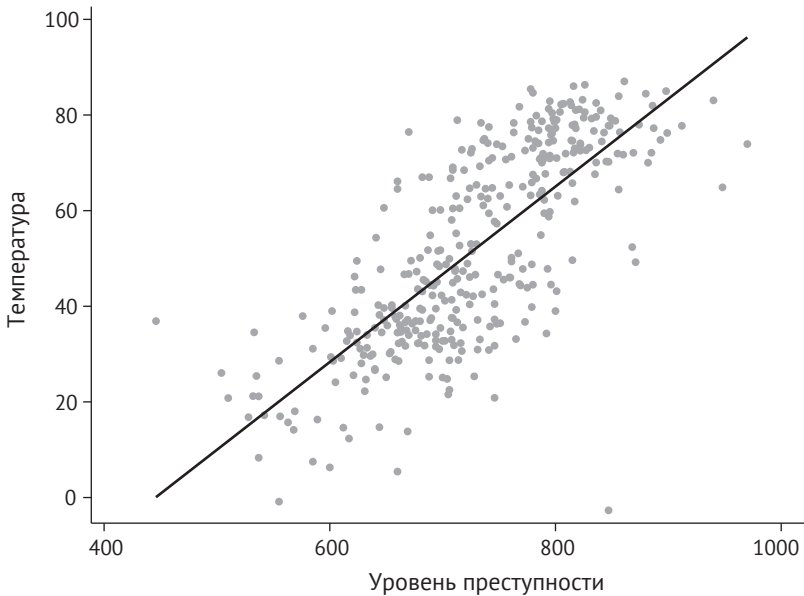


Рис. 2.3. Линия наилучшего соответствия, обобщающая взаимосвязь между температурой и преступностью в Чикаго по дням в 2018 г.

Факт или корреляция?

Чтобы установить, существует ли корреляция, мы всегда должны проводить какое-то сравнение. Например, чтобы узнать о корреляции между температурой и преступностью, нам нужно сравнить теплые и холодные дни и посмотреть, различаются ли уровни преступности, или, как вариант, мы можем сравнить дни с высоким и низким уровнем преступности, чтобы увидеть, различаются ли их температуры. Это означает, что для оценки корреляции между двумя переменными нам необходимо иметь вариации обеих переменных. Например, если бы мы собирали данные только в те дни, когда средняя температура составляла 0 °F, у нас не было бы возможности оценить корреляцию между температурой и преступностью. И то же самое верно, если бы мы рассматривали только дни, в которые зарегистрировано 500 преступлений.

Теперь давайте остановимся и проверим, насколько хорошо вы понимаете, что такое корреляция и как о ней можно узнать. Не волнуйтесь, если почувствуете неуверенность. Обнаружение истинной корреляции – непростая задача. Этой теме мы посвятим всю главу 4. Тем не менее полезно провести предварительную проверку знаний прямо сейчас. Итак, давайте попробуем.

Подумайте над следующими утверждениями. Какие из них описывают корреляцию, а какие – нет?

1. Люди, дожившие до 100 лет, обычно принимают витамины.
2. В городах с высоким уровнем преступности, как правило, нанимают больше полицейских на душу населения.

3. Успешные люди потратили не менее десяти тысяч часов на оттачивание своего мастерства.
4. Большинство политиков, столкнувшихся со скандалом, переизбираются на следующий срок.
5. Пожилые люди голосуют чаще, чем молодые.

Хотя каждое из этих утверждений сообщает о факте, не все они описывают корреляцию, т. е. свидетельство того, что два явления наблюдаемого мира имеют тенденцию проявляться вместе. В частности, утверждения 1, 3 и 4 не описывают корреляцию, а утверждения 2 и 5 – описывают. Давайте разберемся почему.

Утверждения 1, 3 и 4 являются *фактами*. Они исходят из данных. Они звучат научно. И если бы мы добавили к этим утверждениям конкретные цифры, мы могли бы назвать их *статистикой*. Но не все факты и статистика описывают корреляции. Ключевой момент заключается в том, что эти утверждения не описывают, склонны ли два явления мира возникать вместе, т. е. они не сравнивают разные значения двух переменных.

Чтобы лучше понять это, обратите внимание на утверждение 4:

Большинство политиков, столкнувшихся со скандалом, переизбираются на второй срок.

Здесь фигурируют два явления. Первое – участие политика в скандале. Второе – победа политика на следующих выборах. Утверждение намекает на положительную корреляцию между участием в скандале и победой на выборах. Но на самом деле из этой констатации факта мы не можем узнать, имеют ли эти два явления тенденцию возникать вместе, т. е. мы не сравнивали частоту переизбрания тех, кто столкнулся со скандалом, с частотой переизбрания тех, кто не замешан в скандале. Да, мы могли бы оценить эту корреляцию, но только не с помощью данных, упомянутых в утверждении 4. Чтобы оценить корреляцию, нам нужны вариации обеих переменных – количества скандалов и побед на последующих выборах. Просто ради интереса давайте рассмотрим эту корреляцию на реальных данных о действующих членах Палаты представителей США, претендовавших на переизбрание в период с 2006 по 2012 г. Скотт Бейсингер из Университета Хьюстона систематически собирал данные о скандалах в Конгрессе. Используя его данные, давайте посмотрим, сколько случаев попадает в четыре соответствующие категории: политики, которые столкнулись со скандалом и были переизбраны, политики, которые столкнулись со скандалом и не были переизбраны, политики без скандалов, которые были переизбраны, и политики без скандалов, которые не были переизбраны.

Из табл. 2.2 мы видим, что утверждение 4 действительно является фактом: 62 из 70 (около 89 %) членов Конгресса, столкнувшихся со скандалом и добившихся переизбрания, победили на выборах. Но мы также видим, что большинство членов Конгресса не замешаны в скандалах и также добились переизбрания. Фактически были переизбраны 1192 из 1293 (около 92 %) политиков, которых не коснулись скандалы. Теперь мы видим, что на самом деле существует небольшая отрицательная корреляция между участием в скандале и победой на следующих выборах.

Таблица 2.2. Большинство членов Конгресса, столкнувшихся со скандалом, побеждают на выборах, но между скандалом и переизбранием существует отрицательная корреляция

	Без скандала	Со скандалом	Итого
Не переизбраны	101	8	109
Переизбраны	1192	62	1254
Итого	1293	70	1363

Мы надеемся, что теперь понятно, почему утверждение 4 не содержит достаточно информации, чтобы понять, существует ли корреляция между скандалом и переизбранием. Проблема в том, что это утверждение касается только политиков, столкнувшихся со скандалом. Оно говорит нам лишь о том, что среди этой части политиков победителей больше, чем проигравших. Но чтобы выяснить, существует ли корреляция между скандалом и победой на выборах, нам нужно сравнить долю политиков, столкнувшихся со скандалом и выигравших переизбрание, с долей политиков, не сталкивавшихся со скандалом и тоже выигравших переизбрание. Если бы только 85 % членов Конгресса, не замешанных в скандалах, выиграли переизбрание, между скандалом и переизбранием была бы положительная корреляция. Если бы выиграли 89 %, корреляции не было бы. Но поскольку теперь мы знаем, что доля политиков, добившихся переизбрания и не столкнувшихся со скандалом, составляет 92 %, мы видим, что существует отрицательная корреляция. Аналогичный анализ покажет, что утверждения 1 и 3 сами по себе также не несут достаточно информации для оценки корреляции.

Утверждения 2 и 5 действительно описывают корреляции. Обратите внимание, что оба они содержат сравнение. Утверждение 2 говорит нам, что в городах с более высоким уровнем преступности в среднем более крупные полицейские силы, чем в городах с меньшим уровнем преступности. А утверждение 5 говорит, что пожилые люди (большего возраста), как правило, голосуют чаще, чем молодые (меньшего возраста). В обоих случаях мы сравниваем различия в одной переменной (численность полиции или уровень голосования) с различиями в другой переменной (уровень преступности или возраст). Это та информация, которая вам нужна для установления корреляции.

Как мы говорили вначале, не волнуйтесь, если вы почувствуете замешательство. Иногда бывает сложно понять, какая информация необходима для установления корреляции, а не просто факта. Мы посвятим четвертую главу тому, чтобы убедиться, что вы действительно все поняли.

Для чего нужна корреляция?

Теперь, когда у вас есть общее понимание того, что такое корреляция, давайте поговорим о том, для чего она нужна. Мы отметили, что корреляции – пожалуй, самый важный инструмент количественного анализа. Но почему? В общих чертах – потому что корреляция позволяет нам предсказать изменение какого-либо показателя или свойства на основании известных изменений других показателей или свойств.

Есть как минимум три варианта использования такого рода знаний: (1) описание, (2) прогнозирование и (3) причинно-следственная связь. Каждый раз, когда вы собираетесь использовать корреляцию, следует четко понимать, какую из этих трех задач вы пытаетесь решить и какими должны быть достоверные знания о мире, чтобы корреляция была применима в конкретных условиях.

Описание

Описание отношений между свойствами или признаками объектов мира – самый простой способ использования корреляций.

Почему у нас может возникнуть необходимость в описании взаимосвязи между свойствами объектов? Предположим, вы опасаетесь, что молодые люди недостаточно представлены в опросах на конкретных выборах относительно их доли в населении. В таком случае может пригодиться описание взаимосвязи между возрастом и голосованием. На рис. 2.4 показана диаграмма рассеяния данных о возрасте и средней явке избирателей на выборах в Конгресс США в 2014 г. На этом рисунке наблюдение представляет собой возрастную когорту. Для каждого возраста на вертикальной оси показана доля избирателей, имеющих право голоса и принявших участие в голосовании.

На рисунке также изображена линия наилучшего соответствия данным. Эта линия имеет наклон 0.006. Другими словами, в среднем на каждый дополнительный год возраста вероятность того, что человек проголосовал в 2014 г., увеличивается на 0.6 процентных пункта. Молодые люди действительно выглядят недостаточно представленными на выборах, поскольку их показатели ниже, чем у пожилых людей.

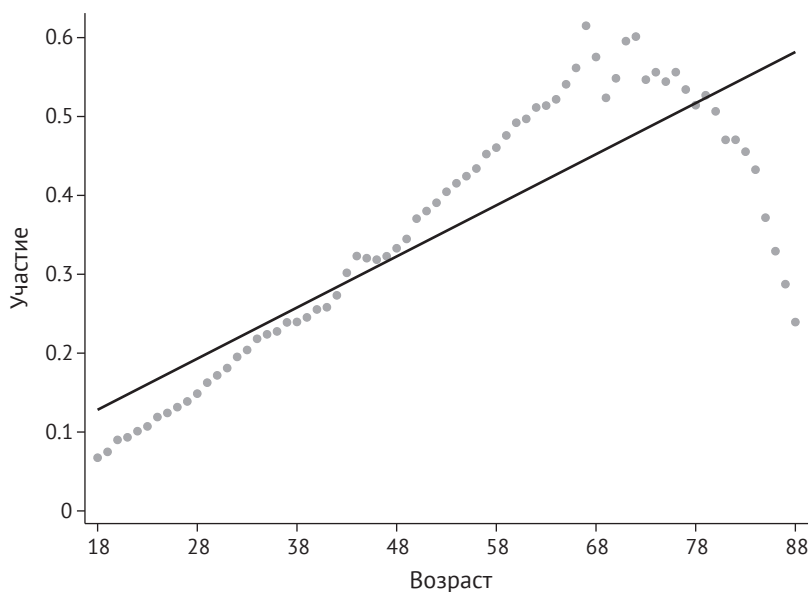


Рис. 2.4. Явка и возраст избирателей на выборах 2014 г.

Этот вид описательного анализа может быть интересен сам по себе. Важно знать, что молодые люди с меньшей вероятностью голосовали в 2014 г., чем по-

жилые, и поэтому были недостаточно представлены в избирательном процессе. Эти отношения могут повлиять на то, как вы рассматриваете исход данных выборов. Более того, знание этой корреляции может побудить вас к дальнейшему изучению причин и последствий феномена низкой рождаемости среди поколения более молодых людей.

Конечно, эта описательная связь не обязательно означает, что молодые люди продолжат меньше голосовать на предстоящих выборах. Поэтому вы вряд ли сможете использовать эти знания для прогнозирования будущей явки избирателей. И это также не означает, что с возрастом бывшие молодые люди обязательно начнут чаще голосовать. Поэтому здесь вряд ли получится обнаружить причинно-следственную связь. Этот описательный анализ всего лишь говорит нам, что на выборах 2014 г. пожилые избиратели в среднем голосовали с большей вероятностью, чем молодые. Чтобы развить интерпретацию дальше, вам нужно сделать более сильные предположения о мире, который мы сейчас исследуем.

Прогнозирование

Другой причиной, побуждающей нас изучать корреляцию, является *прогнозирование* или *предсказание* – два термина, которые мы будем использовать как синонимы. Прогнозирование предполагает использование информации из некоторой выборочной совокупности для прогнозирования другой совокупности.

Например, вы можете использовать данные об избирателях на прошлых выборах, чтобы делать прогнозы об избирателях на будущих выборах. Или вы можете использовать данные избирателей в одном штате, чтобы делать прогнозы об избирателях в другом штате. Предположим, вы проводите избирательную кампанию, у вас ограниченные ресурсы и вы пытаетесь выяснить, на кого из ваших сторонников следует нацелиться, стуча в дверь и напоминая им, что пора идти проголосовать. Если вы заранее уверены в том, что человек пойдет голосовать без вашего вмешательства, незачем тратить время своих волонтеров на посещение его дома. Таким образом, точное прогнозирование явки избирателей может повысить эффективность вашей кампании.

Для такого рода прогнозирования могут быть полезны корреляции, подобные приведенной выше связи возраста и явки избирателей. Поскольку возраст сильно коррелирует с явкой, он может быть полезной переменной для прогнозирования того, кто намерен проголосовать, а кто – нет. Например, если вы можете предсказать на основе возраста, что некоторая группа избирателей практически наверняка явится на выборы даже без усилий ваших волонтеров, можете сосредоточить свои мобилизационные ресурсы на других избирателях.

Чтобы использовать таким способом корреляцию между возрастом и явкой избирателей для прогнозирования, вам незачем знать, *почему* они коррелируют. Но, в отличие от случая, когда вы просто хотите описать взаимосвязь между возрастом и явкой избирателей на выборах 2014 г., если вы хотите делать прогнозы, нужно быть готовым сделать некоторые дополнительные предположения о мире.

Это поднимает две важные проблемы, о которых вы должны хорошенько подумать, чтобы ответственно использовать корреляцию для прогнозирования. Во-первых, является ли взаимосвязь, которую вы обнаружили в своей выборке, отражением более широкого явления, или она является результатом случайных изменений в ваших данных? Ответ на этот вопрос требует статистического

вывода, который будет рассмотрен в главе 6. Во-вторых, даже если вы убеждены, что обнаружили реальную взаимосвязь в своей выборке, следует подумать о том, является ли ваша выборка репрезентативной для генеральной совокупности, относительно которой вы пытаетесь сделать прогнозы. Мы рассмотрим репрезентативность более подробно при обсуждении выборок и внешней валидности в главах 6 и 16.

Давайте вернемся к использованию информации о возрасте и явке избирателей на одних выборах, чтобы сделать прогнозы насчет следующих выборов. Это имеет смысл только в том случае, если можно предположить, что взаимосвязь между этими двумя переменными не меняется слишком быстро. То есть корреляция между возрастом и явкой избирателей, например, на выборах 2014 г. будет полезна для выяснения того, на каких избирателей ориентироваться на выборах 2016 г., если кажется вероятным, что связь между возрастом и явкой в 2016 г. будет приблизительно такая же, как в 2014 г. Аналогично, если бы у вас были данные только о возрасте и явке избирателей на выборах 2014 г. в 25 штатах, вы могли бы использовать корреляцию между возрастом и явкой в этих штатах, чтобы получить информацию стратегии в остальных 25 штатах. Но это было бы разумно только при наличии оснований полагать, что взаимосвязь между возрастом и явкой в остальных штатах будет приблизительно такой же.

Соответственно, при использовании для прогнозирования некоторых статистических данных, таких как наклон линии наилучшего соответствия, нам нужно подумать о том, является ли зависимость на самом деле линейной. В противном случае линейное представление отношений может ввести в заблуждение. Мы обсудим это более подробно ниже.

Стоит отметить, что на практике было бы странно пытаться делать прогнозы на основании корреляции всего лишь между двумя переменными. Более разумно попытаться предсказать явку избирателей, используя ее связь с множеством переменных, таких как пол, раса, доход, образование и явка на предыдущие выборы. Мы обсудим такие многомерные и условные корреляции в главе 5.

Использование данных для прогнозирования и предсказания – быстро растущая область деятельности аналитиков в политике, бизнесе, полиции, спорте, правительстве, разведке и многих других областях. Например, предположим, что вы руководите отделом общественного здравоохранения вашего города. Каждый раз, когда вы отправляете санитарного инспектора в ресторан, это стоит времени и денег. Но нарушения санитарных норм в ресторанах наносят вред жителям вашего города. Поэтому вам бы очень хотелось направить инспекторов именно в те рестораны, которые с наибольшей вероятностью нарушают санитарные нормы, чтобы не тратить время и деньги на проверки, которые в конечном итоге не улучшат общественную безопасность. Чем точнее вы сможете спрогнозировать, какие рестораны нарушают правила, тем эффективнее вы сможете задействовать своих инспекторов. Допустим, можно использовать данные о ресторанах, которые нарушали и не нарушали нормы здравоохранения в прошлом, чтобы попытаться предсказать такие нарушения на основе их корреляции с другими наблюдаемыми характеристиками¹ ресторана. Вероятно,

¹ В машинном обучении и анализе данных такие характеристики называют *признаками* (feature). – Прим. перев.

полезные признаки ресторана могут включать обзоры Yelp, информацию о посещениях больниц при пищевых отравлениях, местоположении, ценах и т. д. Затем, имея на руках эти корреляции, вы можете использовать будущие обзоры Yelp и другую информацию, чтобы предсказать, какие рестораны, скорее всего, нарушают нормы здравоохранения, и направить в эти рестораны проверку.

Этот пример указывает еще на одну сложную проблему. Сам факт использования корреляций для прогнозирования иногда может привести к тому, что корреляции, существовавшие в прошлом, перестанут действовать в будущем. Например, предположим, что департамент здравоохранения наблюдает сильную корреляцию между ресторанами, открытыми 24 часа в сутки, и нарушениями санитарных норм. На основе этой корреляции они могут начать непропорционально часто отправлять инспекторов здравоохранения в круглосуточные рестораны. Наблюдательный владелец ресторана, заметивший эту закономерность, может приспособиться, чтобы обмануть департамент здравоохранения, – скажем, закрывать свой ресторан с 2:00 до 3:00 каждую ночь. Это небольшое изменение в часах работы вряд ли поможет навести порядок в ресторане. Однако оно позволяет вывести ресторан из категории круглосуточных и тем самым обмануть систему, основанную на корреляции. Мы обсудим эту общую проблему адаптации более подробно в главе 16.

Прогнозирование также пригодится политику, который хотел бы знать ожидаемую продолжительность экономического спада для планирования госбюджета, банкиру, который хочет знать кредитоспособность потенциальных заемщиков, или страховой компании, желающей знать, какова вероятность попадания конкретного клиента в аварию. Менеджеры наших любимых «Чикаго Беарс» хотели бы предсказать, какие футболисты колледжа смогут увеличить шансы команды на победу в Суперкубке. Но, учитывая их прошлый послужной список, мы не питаем особых надежд. Данные не могут творить чудеса.

Также стоит подумать о потенциальных этических последствиях использования прогнозов для управления поведением. Например, исследования показывают, что жалобы потребителей на чистоту в онлайн-обзорах ресторанов положительно коррелируют с нарушениями санитарного кодекса. Это потенциально полезная прогностическая информация: правительственные органы могли бы использовать данные, собранные с обзорных сайтов, чтобы выяснить, куда направить проверяющих. Основываясь на этом предположении, статья в *The Atlantic* заявила: «Yelp может навести порядок в ресторанной индустрии». Но исследование Кристен Альтенбургер и Дэниела Хо показывает, что онлайн-рецензенты предвзято относятся к азиатским ресторанам: сравнивая рестораны, получившие одинаковую оценку от инспекторов по безопасности общественного питания, они обнаружили, что рецензенты с большей вероятностью жалуются на чистоту в азиатских ресторанах. Это означает, что, если правительства будут использовать прогнозирующую корреляцию между онлайн-обзорами и нарушениями санитарного кодекса, это приведет к непреднамеренной дискриминации азиатских ресторанов, непропорционально часто подвергая их проверкам. Вы действительно хотите, чтобы ваше правительство использовало такую информацию? Или этические издержки проверки ресторанов перевешивают преимущества более точных прогнозов? Мы вернемся к некоторым из этих этических вопросов в конце книги.