

УДК 004.6:519.25
ББК 16.35
Р83

Алекс Руис де Вилья Роберт

Р83 Причинно-следственный анализ в науке о данных / пер. с англ. А. Н. Киселева. – М.: ДМК Пресс, 2025. – 432 с.: ил.

ISBN 978-5-93700-365-2

Понимание сути и роли причинности имеет большое значение для анализа данных. Эта книга послужит прочной основой для применения методов причинно-следственного вывода в повседневных задачах науки о данных. Вы научитесь правильно выбирать и применять математические и статистические инструменты, необходимые для успешного моделирования причинно-следственных связей.

Книга адресована начинающим и опытным аналитикам данных, а также специалистам в области машинного обучения, экономики и статистики, желающим усовершенствовать процесс принятия решений.

УДК 004.6:519.25
ББК 16.35

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

©DMKPress 2025. Authorized translation of the English edition. © 2025 Manning Publications. This translation is published and sold by permission of Manning Publications, the owner of all rights to publish and sell the same.

ISBN (анг.) 9781633439658
ISBN (рус.) 978-5-93700-365-2

© 2025 Manning Publications Co. All rights reserved
© Оформление, издание, перевод,
ДМК Пресс, 2025

Краткое главление

Часть I. Причинно-следственный анализ и роль искажающих факторов	27
1 ■ <i>Введение в причинность.....</i>	29
2 ■ <i>Первые шаги: работа с искажающими факторами.....</i>	60
3 ■ <i>Применение причинно-следственного анализа</i>	102
4 ■ <i>Как машинное обучение и причинно-следственный анализ могут помочь друг другу</i>	123
Часть II. Формула корректировки на практике	161
5 ■ <i>Поиск сопоставимых случаев с использованием меры склонности.....</i>	163
6 ■ <i>Расчет прямых и косвенных эффектов с помощью линейных моделей.....</i>	198
7 ■ <i>Работа со сложными графами</i>	228
8 ■ <i>Расширенные инструменты в библиотеке DoubleML....</i>	276
Часть III. Другие стратегии помимо формулы корректировки	305
9 ■ <i>Инструментальные переменные.....</i>	307
10 ■ <i>Схема оценки потенциальных исходов</i>	332
11 ■ <i>Эффекты событий во времени.....</i>	350

Оглавление

Предисловие от издательства	14
Предисловие	15
Благодарности	16
О книге.....	18
Об авторе.....	24
Об иллюстрации на обложке	25
ЧАСТЬ I. Причинно-следственный анализ и роль искажающих факторов	27
1 Введение в причинность	29
1.1. Как работает причинно-следственный анализ.....	31
1.1.1. Шаг 1: определение типа данных	31
1.1.2. Шаг 2: изучение задачи	32
1.1.3. Шаг 3: создание модели	32
1.1.4. Шаг 4: представление модели другим.....	33
1.1.5. Шаг 5: применение методов причинно-следственного анализа.....	33
1.2. Различия между причинно-следственными моделями и прогностическими моделями машинного обучения.....	35
1.3. Экспериментальные исследования.....	37
1.3.1. Мотивирующий пример: развертывание нового веб-сайта	37
1.3.2. А/В-тестирование.....	40
1.3.3. Рандомизированные контролируемые исследования	41
1.3.4. Шаги проведения А/В-тестирования.....	42
1.3.5. Ограничения А/В-тестирования и RCT	44
1.4. Наблюдательные исследования.....	45
1.4.1. Моделирование синтетических данных	47
1.4.2. Причинно-следственные связи при наличии искажающих факторов.....	49
1.5. Обзор основных статистических концепций	51
1.5.1. Эмпирические распределения и распределения сгенерированных данных.....	51
1.5.2. Условные вероятности и ожидания	53
1.6. Для дополнительного чтения	57
1.7. Контрольные вопросы	58
Итоги.....	59
2 Первые шаги: работа с искажающими факторами	60
2.1. Основы причинно-следственного анализа и парадокс Симпсона	62
2.1.1. Чем обусловлена эта проблема?.....	64
2.1.2. Развивайте интуицию: как исправить проблему.....	66
2.1.3. Решение парадокса Симпсона	67
2.2. Обобщение на другие задачи	69

2.2.1. Описание проблемы с помощью графа	70
2.2.2. Формулирование того, что хотелось бы узнать	70
2.2.3. Поиск способа вычисления причинно-следственной связи	71
2.2.4. Формулирование того, что хотелось бы узнать: язык вмешательств	71
2.2.5. Поиск способа вычисления причинно-следственной связи: формула корректировки	74
2.2.6. Какие результаты дает метод лечения в каждой ситуации? Предположение о положительности.....	76
2.3. Вмешательства и RCT	78
2.4. Первая встреча со структурным подходом	79
2.4.1. Моделирование примера почечнокаменной болезни	81
2.4.2. Вмешательства в структурном подходе	83
2.5. Когда применять формулу корректировки.....	85
2.5.1. RCT или A/V-тестирование	86
2.5.2. Искажающие факторы	87
2.5.3. Ненаблюдаемые искажающие факторы.....	88
2.5.4. Медиаторы	89
2.5.5. Множество искажающих факторов	90
2.5.6. Переменные, прогнозирующие результат	91
2.5.7. Переменные, прогнозирующие выбор метода лечения.....	93
2.5.8. Условное вмешательство.....	93
2.5.9. Объединение всех предыдущих ситуаций	95
2.5.10. Обобщение различий между вмешательством и применением формулы корректировки	96
2.6. Итак, каков план?	97
2.7. Основные уроки главы	99
2.8. Контрольные вопросы	100
Итоги.....	101
3 Применение причинно-следственного анализа	102
3.1. Когда и зачем использовать графы в причинно-следственном анализе.....	103
3.2. Шаги по формулированию задачи с использованием графов.....	105
3.2.1. Составьте список всех переменных	106
3.2.2. Создайте свой граф.....	108
3.2.3. Сформулируйте свои предположения	112
3.2.4. Сформулируйте свои цели.....	113
3.2.5. Проверьте предположение о положительности	114
3.3. Другие примеры	115
3.3.1. Рекомендательные системы	115
3.3.2. Ценообразование	118
3.3.3. Моделирование	118
3.4. Для дальнейшего чтения.....	121
3.5. Контрольные вопросы	121
Итоги.....	121
4 Как машинное обучение и причинно-следственный анализ могут помочь друг другу.....	123
4.1. Что дает обучение с учителем?	126
4.1.1. Когда следует использовать причинно-следственный анализ, а когда обучение с учителем?.....	128

4.1.2. Цель аппроксимации данных.....	129
4.1.3. Когда будущее и прошлое имеют одинаковое поведение.....	131
4.1.4. Когда причинно-следственный анализ и обучение с учителем совпадают?.....	133
4.1.5. Ошибка прогнозирования – ложный друг	134
4.1.6. Проверка вмешательств.....	140
4.2. Как обучение с учителем участвует в причинно-следственном анализе?	141
4.2.1. Эмпирические распределения и распределения сгенерированных данных в формуле корректировки	143
4.2.2. Гибкость формулы корректировки	144
4.2.3. Формула корректировки для непрерывных распределений	145
4.2.4. Алгоритмы расчета формулы корректировки	145
4.2.5. Перекрестное обучение: предотвращение переобучения	149
4.3. Другие применения причинно-следственного анализа в машинном обучении	153
4.3.1. Обучение с подкреплением.....	153
4.3.2. Справедливость	155
4.3.3. Ложные корреляции.....	155
4.3.4. Обработка естественного языка.....	156
4.3.5. Объяснимость.....	156
4.4. Для дальнейшего чтения.....	156
4.5. Контрольные вопросы.....	158
Итоги.....	158

ЧАСТЬ II. Формула корректировки на практике 161

5 Поиск сопоставимых случаев с использованием меры склонности..... 163

5.1. Знакомство с мерой склонности	166
5.1.1. Поиск соответствий для оценки причинно-следственных связей.....	167
5.1.2. Но есть ли соответствие?.....	168
5.1.3. Почему сопоставление может быть трудным	169
5.1.4. Как меры склонности можно использовать для расчета АТЕ.....	171
5.2. Основные понятия мер склонности	172
5.2.1. С какими случаями мы работаем?	173
5.2.2. Что такое мера склонности?	176
5.2.3. Предположение о положительности – это... предположение	176
5.3. Оценки меры склонности на практике	177
5.3.1. Подготовка данных.....	177
5.3.2. Вычисление меры склонности.....	178
5.3.3. Оценка предположения о положительности	181
5.3.4. Вычисление АТЕ на основе мер склонности	187
5.4. Вычисление корректировки меры склонности: упражнение.....	194
5.4.1. Шаги упражнения	195
5.5. Для дальнейшего чтения.....	196
5.6. Контрольные вопросы.....	196
Итоги.....	197

6 Расчет прямых и косвенных эффектов с помощью линейных моделей 198

6.1. Оценка причинно-следственных связей с помощью линейных моделей	201
6.1.1. Моделирование задачи ценообразования: знакомство	201
6.1.2. Прямые и косвенные эффекты.....	207
6.2. Изучение причинно-следственной динамики с помощью линейных моделей	214
6.2.1. Аналогия с газом, текущим по трубам.....	215
6.2.2. Как корреляция распространяется в графе	215
6.2.3. Расчет причинно-следственной связи и корреляции по коэффициентам при стрелках.....	220
6.2.4. Линейные модели и оператор «do»	222
6.3. Контрольные вопросы	226
Итоги.....	226
7 <i>Работа со сложными графами.....</i>	228
7.1. Изменение корреляции между двумя переменными с учетом третьей.....	232
7.1.1. Пример условной независимости времени прихода на работу.....	233
7.1.2. Математический пример условной независимости	234
7.1.3. Разбиение причинно-следственной модели на независимые модули.....	235
7.1.4. Кирпичики DAG: факторизация распределений вероятностей.....	240
7.1.5. Что такое d-разделение?	247
7.1.6. Определение d-разделения.....	251
7.2. Критерий обходного пути	253
7.2.1. Важность критерия обходного пути	258
7.3. Хорошие и плохие наборы корректируемых переменных	261
7.3.1. Хорошие наборы корректируемых переменных.....	261
7.3.2. Нейтральные наборы корректируемых переменных	262
7.3.3. Плохие наборы корректируемых переменных.....	262
7.4. Возвращаясь к предыдущим главам	264
7.4.1. Эффективные наборы корректируемых переменных.....	264
7.4.2. Оценка склонности.....	268
7.4.3. И снова: не включайте переменные в модель только потому, что они делают ее более точной	269
7.4.4. Следует ли делать поправку на доход?	269
7.5. Дополнительный инструмент для выявления причинно-следственных связей: do-исчисление	272
7.6. Для дальнейшего чтения.....	273
7.7. Контрольные вопросы.....	274
Итоги.....	275
8 <i>Расширенные инструменты в библиотеке DoubleML.....</i>	276
8.1. Двойное машинное обучение.....	279
8.1.1. Теорема FWL: предшественница DML	281
8.1.2. Нелинейные модели в DML	285
8.1.3. DML на практике.....	290
8.1.4. Гетерогенные эффекты воздействия	294
8.2. Доверительные интервалы	296

8.2.1. Моделирование новых наборов данных с помощью бутстрэппинга	297
8.2.2. Аналитические формулы вычисления доверительных интервалов.....	298
8.3. Оценки с двойной надежностью	300
8.3.1. AIPW на практике.....	302
8.4. Для дальнейшего чтения.....	302
8.5. Контрольные вопросы.....	303
Итоги.....	303

ЧАСТЬ III. Другие стратегии помимо формулы корректировки 305

9 Инструментальные переменные 307

9.1. Знакомство с IV на примере.....	309
9.1.1. Граф примера.....	311
9.1.2. Предположения IV.....	312
9.1.3. Инструментальные переменные и RCT	314
9.2. Оценка причинно-следственной связи с помощью IV.....	315
9.2.1. Применение IV с линейными моделями	315
9.2.2. Применение IV с частично линейными моделями.....	318
9.2.3. Альтернативная формула для метода IV.....	319
9.2.4. Отсутствие общей формулы для обобщенного графа IV.....	320
9.3. Инструментальные переменные на практике.....	320
9.3.1. Двухэтапный алгоритм наименьших квадратов (2SLS)	322
9.3.2. Слабые инструменты.....	324
9.3.3. IV и DoubleML	327
9.4. Ссылки	330
9.5. Контрольные вопросы.....	331
Итоги.....	331

10 Схема оценки потенциальных исходов 332

10.1. Что такое потенциальный исход?	333
10.1.1. Индивидуальные исходы	334
10.1.2. Коллективные исходы.....	336
10.1.3. Причинно-следственные эффекты	337
10.1.4. Предположения о потенциальных исходах	338
10.2. Связь метода потенциальных исходов с методом графов.....	339
10.2.1. Первый закон причинно-следственного анализа	339
10.2.2. Выражение предположений PO с помощью DAG	340
10.2.3. Контрфактуалы	341
10.3. Формула корректировки с потенциальными исходами	343
10.4. Инструментальные переменные с потенциальными исходами.....	347
10.5. Контрольные вопросы.....	348
Итоги.....	349

11 Эффекты событий во времени 350

11.1. Какие типы данных будут использоваться?	354
11.2. Разрывной регрессионный дизайн	355
11.2.1. Моделирование данных	356
11.2.2. Терминология RDD.....	358
11.2.3. Предположения.....	360

11.2.4. Оценка эффекта	360
11.2.5. RDD на практике	361
11.3. Синтетический контроль	373
11.3.1. Моделирование данных	376
11.3.2. Терминология метода синтетического контроля	377
11.3.3. Предположения	378
11.3.4. Оценка эффекта	379
11.3.5. Синтетический контроль на практике	380
11.3.6. Выбор периодов обучения и прогнозирования	380
11.4. Метод сравнения разностей	386
11.4.1. Моделирование данных	388
11.4.2. Терминология DiD	389
11.4.3. Предположения	392
11.4.4. Оценка эффекта	394
11.4.5. Практика	395
11.5. Контрольные вопросы	401
11.6. Сравнение методов	401
11.7. Ссылки	402
Итого	402
<i>Приложение А. Математика, лежащая в основе формулы коррективки</i>	403
<i>Приложение В. Решения упражнений в главе 2</i>	407
В.1. Решение парадокса Симпсона для метода лечения В	407
В.2. Наблюдать и делать – это не одно и то же	407
В.2.1. Решение	408
В.3. Что нужно скорректировать?	409
В.3.1. RCT	409
В.3.2. Искажающие факторы	410
В.3.3. ненаблюдаемые искажающие факторы	412
В.3.4. Медиаторы	413
В.3.5. Переменные, прогнозирующие результат	414
<i>Приложение С. Техническая лемма о мерах склонности</i>	416
<i>Приложение D. Доказательство двойной надежности оценки \widehat{ATE}_{airwo}</i>	420
D.1. Свойство двойной надежности по отношению к T-learner	420
D.2. Свойство двойной надежности относительно обратного взвешивания вероятности	421
<i>Приложение E. Техническая лемма для альтернативной формулы инструментальной переменной</i>	423
<i>Приложение F. Доказательство формулы неидеального соответствия инструментальной переменной</i>	424
<i>Предметный указатель</i>	427

Предисловие

Впервые проблема причинно-следственного анализа привлекла мое внимание в 2016 году, когда я прочитал статью о причинности и совершенно ничего не понял. Заинтересовавшись этой экзотической темой, я прочитал работы Джуды Перла (Judea Pearl). Поначалу я не думал, что с помощью математических и статистических инструментов можно многого добиться в моделировании причинности. Но, продолжая читать, я понял, что причинность имеет большое значение с прикладной точки зрения.

В 2016 году я уже несколько лет работал специалистом по данным. Мне нравилось исследовать методы машинного обучения; это был новый мир, в котором мое техническое образование давало хорошее преимущество. Машинное обучение открыло двери в различные отрасли и компании, позволяя мне наслаждаться своей работой.

Между 2016 и 2018 годами я начал понимать, что прогресс в машинном и особенно в глубоком обучении в значительной мере был достигнут путем проб и ошибок, без глубокого понимания его внутренних механизмов. Основное внимание уделялось вычислительной мощности и программированию, а не моделированию мира. Я не против такого подхода, но он не удовлетворял меня. В то же время я все глубже исследовал область причинно-следственного анализа. С каждым днем я обнаруживал, что все больше разделяю ее цели – получение ответа на вопрос «почему» – и ее методы, уделяющие большое внимание статистике и математике, но при этом включающие и программирование.

В конце 2018 года я решил взять годовой отпуск, чтобы вплотную заняться темой причинно-следственного анализа. Я понял, что компании часто испытывают трудности с точной оценкой эффекта своих решений, и казалось неизбежным, что они быстро осознают необходимость выявления причинно-следственных связей и создадут вакансии для специалистов в этом направлении. Когда я обсуждал рабочие проблемы со своими коллегами, то заметил, что способен быстрее и точнее понять и проанализировать их проблемы, чем раньше. А на личном уровне изучение причинно-следственного анализа кардинально изменило мой взгляд на мир: я начал находить причинно-следственные связи во многих сферах жизни общества, включая здравоохранение, экономику, журналистику, политику и т. д.

Приятно тратить время на получение новых знаний, которые могут принести мне доход, а также помочь лучше понять мир, в котором мы живем. С 2018 года я объяснял основы причинно-следственного анализа на многих семинарах, занятиях, в статьях, а теперь и в этой книге! Каждый раз, когда я рассказываю об этой теме, я испытываю то же волнение, что и тогда, когда впервые ее открыл.

Эта книга была написана и для начинающих, и для опытных специалистов по данным, для практиков и исследователей машинного обучения, для аналитиков данных, экономистов и статистиков, желающих усовершенствовать процесс принятия решений с использованием результатов наблюдений. Ее цель – дать вам прочную основу для применения методов причинно-следственного анализа в повседневных задачах. Она даст вам интуитивное понимание в выборе правильных инструментов, которое в сочетании с более формальным подходом гарантирует вам уверенность в своих действиях.

Предварительные требования

Для работы с книгой вам понадобятся базовые знания по следующим темам.

- Теория вероятности:
 - основные формулы вероятности, такие как закон полной вероятности и условные вероятности;
 - базовые распределения вероятностей, такие как гауссово и биномиальное;
 - как генерировать случайные числа с помощью компьютера.
- Статистика:
 - линейная и логистическая регрессия;
 - доверительные интервалы;
 - рекомендуется: понимание основ A/B-тестирования и рандомизированных контролируемых испытаний (как выполняется распределение по группам и проверка гипотез).
- Программирование:
 - базовые навыки программирования (чтение/написание базовых программ) как минимум на одном языке программирования, например Python, R или Julia.
- Машинное обучение:
 - перекрестная проверка и настройка гиперпараметров;
- рекомендуется: опыт работы с моделями машинного обучения, такими как kNN, случайные леса, бустинг и глубокое обучение.

Структура книги

Книга разделена на три части. Часть I посвящена основам, необходимым для понимания причин и следствий. Здесь вы узнаете, когда применять причинно-следственный анализ; как определенные переменные, известные как искажающие факторы, могут затруднить анализ; и как оценить причинно-следственные связи, устранив их влияние с помощью метода, называемого формулой корректировки.

- Глава 1 представляет два способа причинно-следственного анализа на основе данных: с помощью экспериментов (этот способ еще называют A/B-тестированием, или рандомизированным контролируемым испытанием) или без них. Здесь также объясняются риски анализа неэкспериментальных данных при наличии искажающих факторов.
- Глава 2 представляет формулу корректировки, которая оценивает причинно-следственное влияние в неэкспериментальных данных путем устранения влияния искажающих факторов.
- Глава 3 приводит примеры моделирования анализа с использованием графов.
- Глава 4 объясняет использование машинного обучения для расчета формулы корректировки и показывает, как причинно-следственный анализ может улучшить некоторые аспекты машинного обучения.

В части II рассматриваются реальные проблемы, с которыми можно столкнуться при использовании формулы корректировки из части I.

- Глава 5 рассказывает, как выявить нехватку данных в оцениваемом вами решении.
- Глава 6 объясняет, как оценить причинно-следственные связи непрерывных переменных с помощью линейных моделей.
- Глава 7 рассказывает, как с помощью критерия обходного пути выбрать переменные для включения в анализ, если причинно-следственный граф слишком сложный.
- Глава 8 описывает прием двойного машинного обучения (продвинутый метод оценки причинно-следственных связей) и доверительные интервалы. В главе также объясняется, как их вычислить с помощью пакета DoubleML.

В части III показаны дополнительные методы изучения причин и следствий помимо формулы корректировки.

- Глава 9 представляет метод инструментальных переменных, использующий независимый источник вариации для оценки причинно-следственной связи, и не требует знания каких-либо искажающих факторов, влияющих на результаты.
- Глава 10 обсуждает схему оценки потенциальных исходов – альтернативу графовым причинно-следственным моделям.
- Глава 11 объясняет методы, связанные со временем, часто используемые в экономике, такие как синтетический контроль, разрывной регрессионный дизайн и метод сравнения разностей.

Путь обучения и философия этой книги

Первым делом вы узнаете, как определить, что имеет место причинно-следственная проблема. Не каждый вопрос имеет причинно-следственную связь, иногда просто нужно описать происходящее или предсказать, что произойдет в будущем.

Затем вы пройдете через этапы осмысления причинно-следственных вопросов, включая следующие важные идеи:

- когда нужно провести эксперимент, когда использовать причинно-следственные связи, а когда – машинное обучение;
- использование причинно-следственных диаграмм для представления происходящего в реальном мире;
- использование этих диаграмм для четкого обозначения своих целей, предположений, рисков, на которые вы идете, и что ваши данные могут и не могут вам рассказать;
- проверка наличия всей необходимой для анализа информации (переменных) и выяснение, чего не хватает;
- оценка причинно-следственных связей с использованием статистических методов и методов машинного обучения.

Различные стили обучения

Не все обучаются одинаково: кто-то предпочитает учиться на *примерах*, другие заглядывают в *код*, а кто-то находит более ясными *математические выкладки*. Я включил в книгу все три аспекта. У вас может быть свой предпочтительный метод обучения, но я предлагаю выйти из зоны комфорта и попробовать другие. Если вы легко читаете и понимаете математические формулы, то попробуйте применить методы с использованием синтетических наборов данных. И наоборот, если вам больше нравится исследовать программный код, то попробуйте внимательно прочитать и осмыслить математические доказательства. Чем больше точек зрения на этот материал у вас будет, тем лучше вы его поймете.

Когда я изучаю что-то новое, то следую подходу «сначала думай, потом читай». Далее я перечислю основные его положения для тех, кто захочет попробовать учиться так же, как я.

- Увидев технический термин, я стараюсь самостоятельно вспомнить его определение и только потом заглядываю в книгу.
- Если в книге утверждается какой-то факт, например о том, что искажающие факторы вызывают ложные корреляции, то я обдумываю его смысл для меня, прежде чем прочитать его объяснение.
- Когда в книге говорится о математической идее, например о линейной регрессии, я придумываю пример в голове. Благодаря этому мои мысли начинают совпадать с тем, что я читаю.

Поначалу этот подход может показаться немного сложным, потому что чтение движется медленнее, но он очень помогает в долгосрочной

перспективе. Вот почему я добавил в книгу разделы «Сначала подумай, потом читай» – они должны поощрять сначала думать.

Как проверить, действительно ли вы поняли описываемую идею?

- Не спрашивайте себя, поняли ли вы идею, а спросите себя, знаете ли вы, как ее использовать. Если вы не сможете четко ответить на этот вопрос, то, вероятно, вы не до конца поняли идею.
- Если вам захочется вернуться к прочитанному тексту еще раз, то потратьте немного времени, перечитайте его и лишь потом двигайтесь дальше.
- Отличный способ убедиться, что вы что-то поняли, – объяснить это простыми словами. Поэтому в каждой главе вы найдете раздел с контрольными вопросами, которые помогут вам проверить себя.
- Еще одна хорошая проверка – способность обобщить основные идеи. Вот почему в конце каждой главы есть раздел «Итоги».

А как быть, если вы на чем-то застрянете? Не торопитесь и двигайтесь постепенно, шаг за шагом. Вам некуда спешить.

Развитие интуиции и освоение методологии

Чтобы начать использовать причинно-следственный анализ, необходимо освоиться с его идеями и методами, а для этого нужна практика решения правильных типов задач.

Развивайте свою интуицию

Причинно-следственный анализ увлекателен, потому что смешивает идеи, которые кажутся как естественными, так и удивительно неожиданными. В нашей повседневной жизни мы часто думаем о причинах и следствиях. Например, мы все согласны, что если идет дождь и земля становится мокрой, то дождь является причиной сырости земли. Все просто. Однако намного сложнее понять, откуда мы знаем, что существует причинно-следственная связь. Мы просто наблюдаем, как одно событие происходит перед другим. Читая эту книгу, вы столкнетесь с идеями, которые могут оказаться вам новыми или неожиданными. Возможно, вам придется взглянуть на знакомые идеи, такие как условные вероятности, линейные модели и даже модели машинного обучения, под новым углом. Я представлю эти точки зрения с примерами и понятиями, которые легко усвоить. Однако, придерживаясь интуитивных объяснений, мы иногда пропускаем формальные детали. Не поймите меня неправильно: определения, теоремы и формулы должны быть на 100 % правильными. Неформальность не является оправданием неточностей. Однако, следуя идее Джорджа Э. П. Бокса (George E.P. Box) о том, что «все модели неверны, но некоторые полезны», эта книга фокусируется на простом объяснении полезных идей, часто с использованием метафор и упрощений.

В данной книге вы столкнетесь с некоторой неформальностью, когда мы будем обсуждать различия между причинным анализом, машинным обучением и статистикой. Например, я могу сказать, что причинный анализ предназначен для поиска причин, а машинное обучение – для прогнозирования. Это не абсолютно точное утверждение, а скорее обобщение, которое поможет вам определить тип задачи, стоящей перед вами, и необходимые для ее решения инструменты. По мере погружения в причинно-следственный анализ вы обнаружите, что, как и в любой области знаний, существует некоторое перекрытие между предметами и подходами, а границы несколько размыты (если вас интересуют формальные основы анализа причинности, то я настоятельно рекомендую прочитать книгу Джуды Перла (Judea Pearl) «Causality»; 2000, Cambridge University Press).

В этой книге примеры используются не только для объяснения идей причинно-следственного анализа, но и для демонстрации его применения. Было очень важно соблюсти баланс в количестве деталей, включенных в эти примеры. Чем примеры подробнее, тем реалистичнее они становятся. Но если деталей слишком много, то можно «не увидеть леса за деревьями», что сделает примеры менее полезными. Обычно я привожу простые примеры, чтобы вам было легче их понять и найти аналогии с вашей ситуацией позже.

Практикуйте методологию

Чтобы помочь вам освоить методы причинно-следственного анализа, помимо упражнений я часто использую повторение. Вы узнаете, как выявлять причинно-следственные связи с использованием булевых переменных, линейных моделей, различных алгоритмов, моделей машинного обучения и многого другого. Сначала может показаться, что каждая глава посвящена отдельной теме, но спустя какое-то время вы заметите, что мы часто делаем нечто похожее, просто подходим с разных сторон.

В анализе причинно-следственных связей важно знать и понимать, что делать, но часто не менее важно знать, чего не нужно делать. Может быть много потенциальных причин, о которых мы даже не подозреваем. Эти не известные неизвестные жизненно важны. Помимо понимания, какие формулы следует использовать в разных ситуациях, также важно знать, когда стоит сражаться, а когда лучше отступить. Осознание того, что вы знаете, а чего нет, поможет вам избежать многих проблем. Эти знания позволят вам выбирать проекты, в которых вы с большей вероятностью добьетесь успеха.

О примерах программного кода

Эта книга содержит множество примеров исходного кода как в пронумерованных листингах, так и в обычном тексте. В обоих случаях исходный код оформлен моноширинным шрифтом, чтобы визуально отделить его от обычного текста.

Во многих случаях исходный код был дополнительно отформатирован; мы добавили разрывы строк и изменили отступы, чтобы уместить примеры по ширине книжной страницы. В некоторых примерах мы удалили комментарии из листингов, если код достаточно подробно описывается в тексте.

Код в книге был создан с помощью Quarto и преобразован в файлы HTML (для R) и IPYNB (для Python). Фрагменты выполняемого кода можно загрузить на сайте <https://livebook.manning.com/book/causal-inference-for-data-science>. Весь код, показанный в примерах, вместе с ответами к упражнениям доступен на веб-сайте книги www.manning.com/books/causal-inference-for-data-science и в репозитории GitHub <https://github.com/aleixrvr/CausalInference4DataScience>.

Живое обсуждение книги

Приобретая книгу «Причинно-следственный анализ в науке о данных», вы получаете бесплатный доступ к онлайн-платформе для чтения, организованной издательством Manning Publications, где можно оставлять комментарии к книге в целом или к конкретным разделам и абзацам. Там вы можете делать заметки для себя, задавать технические вопросы и отвечать на них, а также получать помощь от автора и других пользователей. Чтобы получить доступ к форуму и зарегистрироваться на нем, откройте в веб-браузере страницу <https://livebook.manning.com/book/causal-inference-for-data-science/discussion>. Узнать больше о форумах Manning и познакомиться с правилами поведения можно по адресу <https://livebook.manning.com/discussion>.

Издательство Manning обязуется предоставить своим читателям место встречи, где может состояться содержательный диалог между отдельными читателями и между читателями и автором. Но со стороны авторов отсутствуют какие-либо обязательства уделять форуму какое-то определенное внимание – их присутствие на форуме остается добровольным (и неоплачиваемым). Мы предлагаем задавать авторам стимулирующие вопросы, чтобы их интерес не угасал! Форум и архив с предыдущими обсуждениями остается доступным на сайте издательства, пока книга продолжает издаваться.

Об авторе

Алекс Руис де Вилья Роберт (Aleix Ruiz de Villa Robert) вот уже более 15 лет работает специалистом по данным. Имеет степени доктора философии по математике и магистра по финансовой математике, полученные в Автономном университете Барселоны. Был руководителем отдела науки о данных в LaVanguardia.com и SCRM (Lidl International Hub) и главным специалистом по данным в Onna.

С 2019 года работает как фрилансер и занимается исследовательскими проектами и преподаванием в различных университетах и бизнес-школах. Он также консультирует стартап Nuclia.

Алекс очень активен в организации открытых мероприятий, связанных с наукой о данных. Помогал управлять Barcelona R Users Group с 2011 по 2017 год, а также основал и был соорганизатором конференций Barcelona Data Science and Machine Learning с 2014 по 2021 год.

Часть I

Причинно-следственный анализ и роль искажающих факторов

Вероятно, вы слышали поговорку «Корреляция – это не причинно-следственная связь». Но что она означает? Помогает ли корреляция понять причинно-следственную связь? Например, у вас может появиться желание узнать, как изменение цены на товар влияет на его продажи. В этой части книги вы узнаете, что лучший способ выяснить, является ли одно явление причиной другого, – провести эксперименты. Однако не всегда есть такая возможность. Вот тут-то и вступает в игру причинно-следственный анализ.

Одна из главных причин, почему простое изучение корреляций в отсутствие экспериментальных данных может не помочь (и даже увести в неправильном направлении), – это *искажающие факторы* (*confounders*). Искажающие факторы – это факторы, влияющие как на решение, которое мы оцениваем, так и на результат, который нас интересует. Они играют важную роль в причинно-следственном анализе. В главе 2 вы узнаете, насколько такие факторы могут исказить анализ и как оценить их влияние на ваше решение, устранив их с помощью *формулы корректировки* (*adjustment formula*).

Глава 3 научит пользоваться графами для моделирования анализа. Графы помогут вам четко сформулировать ваши цели, изложить ваши предположения и выяснить, какие причины и следствия можно оценить на основе имеющихся данных.

Наконец, в главе 4 вы узнаете, как использовать машинное обучение для получения формулы корректировки в случаях, когда имеется множество факторов, искажающих результаты.

Введение в причинность

В этой главе:

- когда и почему необходим причинно-следственный анализ;
- как работает причинно-следственный анализ;
- различия между наблюдаемыми и экспериментальными данными;
- соответствующие статистические концепции.

Во многих компаниях и организациях, использующих машинное обучение, главной целью является возможность делать обоснованные предположения о том, что произойдет в будущем. Например, в больнице могут пожелать выявить группу пациентов с наибольшими рисками, чтобы начать лечить их в первую очередь. Часто достаточно просто уметь делать такие предположения, а понимание сути происходящего не всегда необходимо.

Причинно-следственный анализ заключается в выяснении, *почему* что-то происходит. Более того, речь идет о том, чтобы получить ответ на вопрос: что можно сделать, чтобы изменить результат? Например, в больнице могут пожелать понять, какие факторы вызывают определенную болезнь. Если эти факторы известны, то с целью сокращения числа заболевших можно предпринять такие шаги, как консультирование по вопросам политики общественного здравоохранения или поддержка исследований по разработке препаратов, предотвращающих заболевание.

Почему причинность важна для любого, кто работает с данными? Будучи специалистами по данным или аналитиками, мы больше всего

интересуемся вопросами, связанными с пониманием причин и следствий. Мы говорим, что X обуславливает Y , если при изменении X изменяется и Y . Например, если ваша цель – удержать клиентов, то вам может быть интересно узнать, какие действия заставят их остаться. Это причинно-следственный вопрос: вы пытаетесь выяснить, что стоит за показателями удержания клиентов, чтобы вы могли их улучшить. Эта идея применима во многих областях, таких как разработка маркетинговых стратегий, установление цен, добавление новых функций в приложение, внесение изменений в организацию, внедрение новых политик и разработка лекарств. Понимание причинно-следственных связей помогает нам увидеть последствия наших решений и определить, какие факторы влияют на результаты.

Спроси себя

Подумайте, какие вопросы у вас возникают, когда вы смотрите на данные. Сколько из них касаются причинно-следственной связи? Подсказка: многие причинно-следственные вопросы подразумевают рассмотрение последствий решений или выявление факторов (особенно тех, которые можно изменить), влияющих на ваши результаты.

Понять причинно-следственную связь непросто. Например, представьте, что вы пытаетесь выяснить, почему кто-то болеет чаще, а кто-то реже. При просмотре данных вы замечаете, что люди, живущие в сельской местности, болеют чаще, чем городские жители. Означает ли это, что жизнь в сельской местности заставляет людей болеть? Если бы это было правдой, то после переезда в город вы должны бы болеть реже. Но так ли это? Жизнь в городе имеет свои проблемы, такие как загрязнение воздуха, нехватка свежей еды и стресс. Поэтому тот факт, что жители городов болеют реже, может быть связан с тем, что у городских жителей обычно более высокие доходы и они могут позволить себе лучшие лекарства, более полезную пищу и абонемент в спортзал. Если это действительно так, то переезд из деревни в город не сделает вас здоровее. Фактически без дохода, который позволит вам заботиться о здоровье или смягчить новые городские риски для здоровья, вы можете начать болеть еще чаще и сильнее.

Этот пример демонстрирует распространенную проблему непонимания причины и следствия. Тот факт, что жизнь в городе и более крепкое здоровье часто идут рука об руку, вовсе не означает, что жизнь в городе делает людей более здоровыми. Именно в таких ситуациях мы часто говорим: «Корреляция – это не причинно-следственная связь». Факт совместного происхождения двух явлений не доказывает, что одно из них является причиной другого. Могут быть и другие, более важные причины, объясняющие различия, например величина дохода, как в нашем примере.

Вот почему анализ причинно-следственных связей так важен для специалистов по данным: он дает инструменты для оценки причин-

но-следственных связей, т. е. помогает отличать простые совпадения (корреляции) от истинных причин (причинно-следственных связей) и позволяет определять фактические факторы, приводящие к определенным результатам.

1.1. Как работает причинно-следственный анализ

Продолжим наш пример с попыткой выяснить, что вызывает определенную болезнь. Допустим, у вас есть набор данных, включающий сведения о пациентах (возраст, сколько раз лечился в больнице и другая подобная информация) и о лечении, которое они получили. Что бы вы сделали, чтобы понять, что вызывает болезнь? Давайте посмотрим, как определить причины с помощью пяти шагов, которые обычно используются для решения причинных задач (рис. 1.1). Мы рассмотрим эти шаги подробнее далее в этой книге.



Рис. 1.1. Пять шагов процесса анализа причинно-следственной связи

1.1.1. Шаг 1: определение типа данных

Первое, что нужно выяснить, – это как были созданы данные. Иногда перед сбором данных можно выполнить эксперимент. Это позволяет контролировать среду, чтобы быть уверенными в том, что нас интересует. В таких случаях мы работаем с *экспериментальными данными*. Однако не всегда есть возможность проводить эксперименты. Например, пытаясь выяснить, как курение влияет на подростков, и зная, что

курение может вызывать рак, мы не можем из этических соображений заставить подростков курить. Или если перед нами старые данные и мы не можем вернуться во времени, чтобы поставить эксперимент. Когда нет возможности провести эксперименты для получения данных, то это означает, что перед нами *данные наблюдений*. Эти два типа данных сильно различаются. Как правило, результаты анализа экспериментальных данных вызывают больше доверия, чем данных наблюдений.

Когда нет возможности проводить эксперименты, на помощь приходит причинно-следственный анализ. Если вас волнуют причинно-следственные вопросы, но нет экспериментальных данных, значит, вам нужен причинно-следственный анализ. В нашем примере город-деревня эксперимент не проводился; мы просто собирали данные по мере их поступления. То же касается примера об определении причины некоторых заболеваний.

1.1.2. Шаг 2: изучение задачи

Чтобы выяснить, что заставляет людей болеть, нужно собрать все возможные причины. Помимо основных данных, таких как возраст, пол и место жительства, также нужна история болезни пациента. Хотя поговорка «чем больше информации, тем лучше» может напомнить вам мантру из сферы больших данных (Big Data), здесь она приобретает особое значение. В машинном обучении можно строить точные прогностические модели, не имея всех переменных, но в причинно-следственном выводе пропуск релевантной переменной может привести к сбою.

Например, иногда знание о сопутствующих заболеваниях (других заболеваниях, помимо того, что интересует) может быть очень важным. Допустим, по какой-то причине у вас нет доступа к информации о сопутствующих заболеваниях пациентов. Применение причинно-следственного анализа не поможет вам определить причину изучаемого заболевания. Однако вы сможете создать успешную прогностическую модель с помощью машинного обучения. Это обусловлено тем, что сопутствующие заболевания приводят к более частым визитам в больницу. Вследствие этого даже в отсутствие подробностей о сопутствующих заболеваниях пациентов у вас могут иметься сильно коррелированные данные в виде частоты визитов пациентов, которой может оказаться достаточно для машинного обучения, чтобы предсказать вероятность заболевания пациентов.

1.1.3. Шаг 3: создание модели

Собрав все важные факторы, вы создаете причинно-следственную модель, которая попытается выделить истинную причинно-следственную связь путем учета потенциальных искажающих переменных, которые могут влиять как на переменные «причина», так и на переменные «следствие». Причинно-следственные модели – это концептуальные рамки, часто представленные с использованием ориентированных ациклических графов (Directed Acyclic Graph, DAG), которые визуаль-но отображают причинно-следственные связи между переменными,

включая потенциальные промежуточные переменные и искажающие факторы. В причинном графе переменные представлены в виде узлов, а причинно-следственные связи между ними – в виде стрелок. В главе мы детальнее рассмотрим причинное моделирование с помощью DAG и будем использовать их на протяжении большей части этой книги. Другой подход к причинному моделированию – использование уравнений, о которых мы поговорим в главе 10.

Вы можете создавать простые модели с небольшим количеством факторов и допущений. Но если модель слишком проста, она может плохо отражать реальный мир и не поможет вам делать точные прогнозы. С другой стороны, со сложными моделями трудно обращаться с точки зрения математических формул, вычислительных методов или проверки соответствия сложных предположений действительности.

В конце концов, лучший способ узнать, достаточно ли хороша ваша модель, – это определить ее полезность для вашей цели. Например, если модель создается для личного приложения поиска объектов на фотографиях, то некоторые ошибки могут быть допустимы. Но если модель предназначена для использования в беспилотном автомобиле, то небольшая ошибка может привести к аварии. В последнем случае модель должна быть сверхточной, чтобы быть полезной.

1.1.4. Шаг 4: представление модели другим

Ваша модель основывается на определенных допущениях и ясной цели (обычно для оценки причинно-следственных связей). Очень важно доходчиво описать эти предположения и цели и вступить в диалог с экспертами и другими участниками анализа. Этот шаг жизненно важен, поскольку вы можете упустить из виду важные переменные или неправильно понять, как они связаны. В нашем примере вы можете показать модель врачам и биологам, чтобы узнать их мнение.

Общение с другими людьми – хорошая практика, которая помогает вам задавать правильные вопросы, согласовывать общие цели и точно определять возможные переменные и их взаимосвязи. Такой процесс сотрудничества помогает свести к минимуму «слепые зоны» и повысить надежность анализа.

1.1.5. Шаг 5: применение методов причинно-следственного анализа

На этом заключительном этапе наступает пора применить методы причинно-следственного анализа к набору данных и использовать модель для получения ответов на причинные вопросы. Помните, что наличие корреляции между двумя переменными не всегда говорит о том, что одна из них является причиной другой. Часто корреляция в отсутствие прямой причинно-следственной связи обусловлена некоторым третьим фактором, который влияет на обе переменные. Такие третьи факторы называются *искажающими факторами (confounders)*, и мы более детально разберем их в этой главе.

Неформально говоря, в причинном анализе наличие искажающих факторов является корнем всех зол, поскольку они могут приводить к ошибочным выводам. К счастью, причинный анализ предоставляет набор формул, алгоритмов и методологий для их учета и включает ряд шагов:

- 1 Спросите себя, на какие вопросы позволяет ответить имеющаяся у вас информация, и выясните, на какие из ваших причинных вопросов можно ответить, используя вашу модель и ваши данные. Иногда недостаток информации об искажающих факторах может превратиться в проблему. Определите, какие недостатки можно преодолеть, и рассмотрите альтернативные решения для устранения других, такие как сбор новых данных или поиск суррогатных переменных.
- 2 Проведите границу между корреляцией и причинно-следственной связью, используя ваши данные для оценки причинно-следственного влияния. Это делается с помощью определенного набора формул. Большая часть нашей книги посвящена объяснению, как, когда и почему следует использовать эти формулы, как выбирать подходящие формулы для различных видов задач и как эффективно применять их с использованием статистических методов и методов машинного обучения.

В этой книге мы рассмотрим три разных метода оценки причинно-следственных связей и обсудим, когда и как их применять.

- *Формула корректировки.* Формула корректировки – это математическая формула, разработанная для устранения искажающих факторов, которые могут повлиять на решение. Обсуждение этой формулы и ее применения занимает большую часть данной книги. Основная формула объясняется в главе 2, а ее вариации – в главах 4–8.
- *Инструментальные переменные.* Некоторые методы используют дополнительную переменную, называемую *инструментом*. Эти переменные не связаны ни с какими искажающими факторами и напрямую влияют только на обработку, а не на результат. Они служат независимыми источниками вариации. Существует много версий этого метода, но в этой книге, в главе 10, мы рассмотрим только его базовую форму.
- *Методы временных рядов.* Эти методы были открыты эконометристами и хорошо подходят для данных временных рядов. К ним относятся сравнение разностей (*differences in differences*), модели разрывной регрессии и синтетический контроль. Как и метод инструментальных переменных, этот тоже имеет множество вариантов. В этой книге, в главе 11, мы рассмотрим только варианты, применимые к временным рядам.

Помимо оценки причинно-следственных связей, также необходимо рассчитать *доверительные интервалы*, как описано в главе 8. Для оценки доверительных интервалов существуют общие методы, такие

как бутстрэппинг, а также специальные методы для каждого из трех методов, часто включаемые в соответствующие пакеты программного обеспечения.

Как я уже отмечал, существует два подхода к проведению причинного анализа. Первый, популяризированный Джудой Перлом, лауреатом премии Тьюринга, использует ориентированные ациклические графы (DAG). Второй основан на *потенциальных результатах* (*potential outcomes, PO*). Этот подход ближе к статистике и эконометрике и был разработан и популяризирован Дональдом Рубином (Donald Rubin), Джеймсом Робинсом (James Robins), Гвидо Имбенсом (Guido Imbens) и лауреатом Нобелевской премии Джошуа Энгристом (Joshua Angrist). В первой и второй частях этой книги мы будем использовать язык DAG, а в третьей части – работать с PO. Каждый подход использует разные обозначения и базовые концепции, но по своей сути они схожи. Оба дают одинаковые результаты для многих задач, но иногда тот или другой может оказаться более подходящим.

1.2. **Различия между причинно-следственными моделями и прогностическими моделями машинного обучения**

Основной принцип, которым руководствуется эта книга, прост: *чем понятнее инструменты анализа причинно-следственных связей и контекст задачи, тем надежнее будут результаты причинного анализа*. Свою цель я вижу в том, чтобы вооружить вас прочными знаниями, чтобы вы могли уверенно применять причинный анализ в повседневной работе. Поэтому мы углубимся в математику, алгоритмы и статистические расчеты. Если вы пришли из мира машинного обучения, то поначалу эти детали могут показаться вам немного ошеломляющими, и вы зададитесь вопросом: зачем их изучать? В некотором смысле причинный анализ ближе к статистике, чем машинное обучение. Проще говоря, теоретические основы причинно-следственного анализа необходимо знать, потому что в отличие от прогностических моделей машинного обучения *трудно определить, точны ли причинно-следственные выводы*. Поэтому для оценки своих выводов специалистам по причинно-следственному анализу необходимо более глубоко понимать происходящее в их моделях.

Например, для построения успешных моделей прогностического машинного обучения знание тонкостей проектирования моделей или предметной области не является обязательным условием. Конкурс Kaggle можно выиграть, не имея вообще никаких знаний предметной области. Аналогично для создания успешных прогностических моделей можно использовать метод опорных векторов (support vector machines, SVM), не понимая сложной математики, лежащей в его основе. Можно создавать очень точные текстовые классификаторы, используя наборы векторных представлений, не зная, как эти представ-

ления создаются. Можно даже обучать надежные нейронные сети, не понимая математических различий между различными стратегиями градиентного спуска.

Конечно, создание надежного продукта на основе прогностической модели требует понимания особенностей работы этой модели. Без этого понимания ваша модель может страдать предвзятостью или, в более серьезных случаях, нанести вред обществу. Вот почему мы заботимся о справедливости и объяснимости в машинном обучении – они помогают решать эти проблемы и обеспечивать ответственное и этичное использование технологии.

В отличие от прогностического моделирования, причинный анализ требует четкого понимания четырех ключевых областей проблемы, перечисленных ниже.

- *Определения* – такие концепции, как ложные корреляции, искажающие факторы, вмешательства и т. д., сложны, но отражают аспекты реальности. Без их четкого понимания ваши причинно-следственные модели могут получаться некорректными.
- *Предположения* – выбор конкретных методов для применения основан на предположениях. В этой книге многие предположения изначально заложены в направленный ациклический граф (DAG).
- *Риски и ограничения* – понимание, что можно вывести из данных, а что нельзя, а также понимание последствий упущения соответствующей переменной имеет огромное значение в причинном анализе. Эти соображения могут существенно повлиять на выводы.
- *Цели* – четкое определение того, что оценивается, имеет большое значение. Определение целей помогает направлять процесс причинного анализа.

Давайте проиллюстрируем эту разницу между двумя типами моделей на примере. Закончив строительство дома, вы можете осмотреть получившийся результат и убедиться, что каждая деталь (стены, потолки, двери) соответствует плану, составленным архитектором. Так можно узнать, правильно ли построен дом. Аналогично в машинном обучении можно узнать, будет ли обученная модель работать в соответствии с ожиданиями, опробовав ее на тестовом наборе данных с использованием перекрестной проверки. По сути, как только модель будет завершена, предполагая, что будущие данные ведут себя аналогично прошлым, вы сможете оценить, насколько хорошо ваша модель будет работать.

Но предположим, что вы хотите убедиться, что ваш дом выдержит землетрясения. Как оценить устойчивость дома к землетрясениям? Это причинно-следственный вопрос (какое влияние землетрясение окажет на структурную целостность дома?), требующий причинно-следственного подхода. Ждать настоящего землетрясения непрактично, поэтому вы изучаете другие способы повысить свою уверенность.

- *Определение правильного процесса строительства сейсмостойких зданий* – вы полагаетесь на теорию влияния землетрясений на здания и строите свой дом в соответствии с этой теорией. Это

гарантирует, что полученный дом будет построен правильно, поскольку вы тщательно контролируете каждый шаг процесса строительства.

- *Моделирование сценариев* – если теория не идеальна или нужна дополнительная уверенность, то можно создать модель и симулировать небольшие землетрясения, чтобы понаблюдать за реакцией вашего дома. В причинно-следственном анализе это сравнимо с созданием синтетических данных для проверки эффективности метода, который планируется использовать.

Обеспечение правильности процессов строительства и моделирование сценариев повышают уверенность, но они не являются окончательными решениями. Даже если ваши выводы верны, если вы им не доверяете, вы не будете действовать в соответствии со своими выводами.

Вы можете задаться вопросом, есть ли такой инструмент, как перекрестная проверка, для задач причинного анализа. К сожалению, как отмечается далее в книге, *такого инструмента не существует*. Единственный способ узнать, верны ли ваши выводы, – это проверить их в реальном мире или провести эксперимент, например А/В-тестирование или рандомизированное контролируемое испытание. По сути, цель причинно-следственного анализа – понять, что произойдет в сценарии, отличном от того, что произошло на самом деле, – означает, что без тестирования достигнуть 100%-ной уверенности в точности ваших выводов практически невозможно.

1.3. Экспериментальные исследования

Как мы уже говорили, наличие экспериментальных данных определяет необходимость причинно-следственного анализа. В этом разделе мы познакомимся с А/В-тестированием, также известным как рандомизированные контролируемые испытания (randomized controlled trials, RCT), и выясним, почему этот вид тестирования считается золотым стандартом для установления причинно-следственных связей. Несмотря на то что эта книга не фокусируется на них, многие методы причинно-следственного анализа пытаются имитировать то же, что и А/В-тестирование. Вот почему важно его понимать.

1.3.1. Мотивирующий пример: развертывание нового веб-сайта

Представьте, что вы работаете в онлайн-бизнесе, например в электронной коммерции. Веб-сайт имеет решающее значение для вашей компании, поскольку качественный сайт дает пользователям чувство комфорта и способствует увеличению числа продаж. За последние два года было собрано множество отзывов постоянных клиентов, в которых они рассказали, что им не нравится в сайте, поэтому вы имеете четкое представление о том, что нужно улучшить. Компания решает переделать веб-сайт и добавить в него все новые функции, о которых просили клиенты. Через шесть месяцев напряженной работы IT-отдела обновленный веб-сайт был запущен.

Теперь вы начинаете сомневаться: имело ли смысл прикладывать такие усилия? Увеличил ли новый веб-сайт вовлеченность пользователей и продажи? Вы основывали необходимость обновления на отзывах самых лояльных клиентов, которые казались надежными. Но то, что нравится одной группе клиентов, может не понравиться остальным. У ваших постоянных клиентов особые потребности, и они не могут представлять большинство ваших пользователей. Поэтому вы понимаете, что важно проверить показатели веб-сайта, чтобы определить успешность новой версии.

Через месяц наблюдений вы просматриваете данные и видите график изменения посещений веб-сайта с течением времени (рис. 1.2). На первый взгляд кажется, что веб-сайт терял клиентов до запуска новой версии, а после ее внедрения число посещений начало расти. Похоже, что ваши усилия окупились!

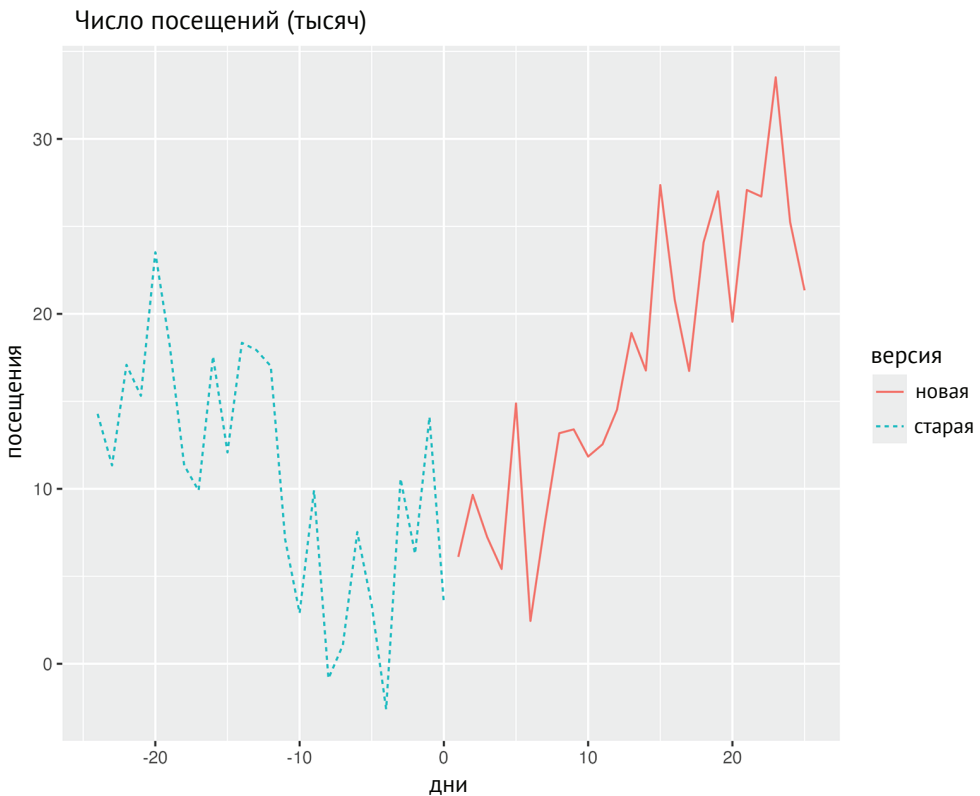


Рис. 1.2. Сравнение трафика веб-сайта до и после запуска новой версии. Ваша цель – выяснить, помогло ли обновление сайта привлечь большее количество посетителей

Вы запустили новый веб-сайт сразу после праздника и понимаете, что это может быть причиной первоначального падения и последующего роста посещений сайта. Спад может быть простым следствием праздничного затишья, а подъем – обычным послепраздничным

всплеском, не обязательно вызванным внедрением новых функций. Но у вас есть предчувствие, что без обновления сайта этот прирост мог быть меньше. Чтобы по-настоящему понять эффект от переделки сайта, вы решаете сравнить текущее число посетителей с тем же периодом прошлого года.

Но сложности не заканчиваются учетом сезонности. На трафик сайта могут влиять и другие факторы. Например, немалую роль в привлечении дополнительных посетителей могла сыграть недавняя маркетинговая кампания. Это предполагает, что для оценки успешности нового веб-сайта нужно учесть множество разных потенциальных влияний, а не только время его запуска.

Сначала подумай, потом читай

Какие еще факторы могут влиять на количество посещений?

На цифры может влиять не только сезонность, но и другие факторы:

- посещаемость сайта могла увеличить маркетинговая кампания;
- возможно, ваши конкуренты совершили ошибку, и их клиенты ушли к вам;
- на доверие клиентов и удобство использования веб-сайта могли повлиять новые правила (например, Общий регламент по защите данных [General Data Protection Regulation, GDPR]);
- большое значение могут иметь успешные новые сотрудники;
- общая социально-экономическая ситуация могла улучшиться по сравнению с прошлым годом.

Чем больше вы углубляетесь, тем очевиднее становится, насколько сложно оценить эффект нового веб-сайта. Чтобы по-настоящему сравнить новый сайт со старым, в идеале необходимо одно из двух условий. Первое – очень стабильная среда, где все, кроме веб-сайта, осталось таким же, как в прошлом году, – сценарий. Такой сценарий редко встречается в нашем переменчивом мире. Второе – выявление и учет всех возможных факторов, способных повлиять на посещаемость сайта, включая те, о которых вы можете даже не подозревать. Например, ваш конкурент мог внести значительные изменения у себя, косвенно повлиявшие на ваш трафик, но вы не можете точно определить, что именно они сделали. А кроме того, есть неизвестные факторы – переменные, которые влияют на трафик вашего сайта, но о которых вы не знаете. Эти неизвестные особенно сложны.

Учитывая сложность воспроизведения любого из этих идеальных сценариев, возникает вопрос: есть ли другой способ уверенно определить, является ли новый веб-сайт успешным? Эта задача подчеркивает сложность причинного анализа связи в реальных условиях, где контроль или даже идентификация всех переменных часто невозможны.

1.3.2. А/В-тестирование

Самый эффективный способ решения проблем, описанных выше, – провести А/В-тестирование: особый вид эксперимента, призванный обеспечить необходимую ясность. Однако перед погружением в А/В-тестирование важно понять, какие события вызываются теми или иными другими событиями. Такое понимание помогает оценить, как с помощью А/В-тестирования изолировать влияние изменения одной переменной (например, дизайна веб-сайта) на результат (например, количество посещений или продаж) и получить более ясную картину причины и следствия.

Развивайте интуицию

Наша главная цель – выяснить, какая из двух версий наберет больше посещений. Для этого мы сначала рассмотрим сложности изучения данных из рис. 1.2. Затем разберем различные решения и будем улучшать их шаг за шагом. Этот процесс естественным образом приведет нас к объяснению сути А/В-тестирования.

Обычно мы считаем, что X является причиной Y , если изменение X приводит к изменению Y . В нашем сценарии X – это версия веб-сайта, а Y – количество его посещений. Чтобы эмпирически измерить влияние X на Y , исходя из этого определения, нам нужно изменить X и посмотреть, изменится ли Y . Но проблема в том, что изменение X не должно сопровождаться изменением других факторов, которые могут повлиять на Y . Например, если новый веб-сайт запускается одновременно с новой маркетинговой кампанией, то будет трудно понять, связано изменение Y с изменением веб-сайта или с проведением кампании.

В идеальном мире для точной оценки влияния нового дизайна веб-сайта на количество посещений достаточно создать две идентичные вселенные: одну со старым веб-сайтом и одну с новым. Тогда любую разницу в посещениях можно напрямую отнести к версии веб-сайта.

К сожалению, создание таких параллельных вселенных невозможно, поэтому вам придется искать другой подход.

Вообще говоря, на количество посещений веб-сайта в любой момент времени могут влиять два основных фактора: окружающая среда (или контекст) и тип посетителей. Для справедливого сравнения старой и новой версий веб-сайта желательно, чтобы эти факторы были одинаковыми для обеих версий, насколько это возможно.

Один из способов гарантировать одинаковые условия для обеих версий – провести эксперимент, в котором оба веб-сайта будут работать одновременно. В этом случае любые внешние изменения, такие как новые законы или конкурентные стратегии, будут одинаково влиять на обе версии.

Следующая сложность состоит в том, чтобы решить, какие пользователи увидят ту или иную версию веб-сайта. Технически пользо-

вателей можно идентифицировать с помощью файлов cookie, что позволит переадресовывать их к определенной версии сайта. С точки зрения бизнеса может показаться логичным направлять самых частых посетителей на новый сайт, поскольку именно их отзывы послужили толчком к изменениям. Такое условие выглядит более разумным для бизнеса, но усложняет эксперимент. Если только частые пользователи увидят новую версию, то будет трудно сказать, связано изменение общего количества посещений с улучшениями веб-сайта или оно обусловлено тем, что эти конкретные пользователи более склонны посещать его чаще.

Это критически важный момент: *какую бы характеристику вы ни использовали для оценки версий веб-сайта, результаты вашего эксперимента все равно будут вызывать сомнения.* Вам придется ответить на вопрос о том, действительно ли изменения в количестве посещений вызваны модификациями веб-сайта или они являются результатом определенного поведения выбранной вами группы пользователей.

Единственный способ гарантировать, что ваш выбор не внесет никакой предвзятости, – выбирать пользователей случайно, наугад. Представьте, что для каждого человека (напомню, что по нашим условиям каждый имеет уникальный идентификатор) проводится виртуальное подбрасывание монеты и выбирается версия сайта, к какой его направить, – А или В. Выбранная версия связывается с идентификатором пользователя, и при следующем посещении пользователь будет направлен к той же версии, т. е. версии не будут меняться при каждом следующем посещении. Эта процедура гарантирует, что при большом количестве пользователей группы А и В будут иметь одинаковое распределение характеристик населения, и такие параметры, как частота посещения ими вашего сайта, возраст, место жительства и другие важные детали, которые могут влиять на посещаемость сайта, будут равномерно распределены между двумя группами.

А/В-тестирование – это эксперимент по тестированию двух разных вариантов, А и В, чтобы увидеть, какой из них лучше подходит для конкретной цели. Оба варианта тестируются одновременно, и пользователям случайным образом назначается вариант А или В без учета их личных качеств. Этот метод гарантирует тестирование обоих вариантов в одинаковых условиях, и если у вас достаточно большое число пользователей, то обе группы будут схожи с точки зрения важных характеристик людей, которые посещают ваш веб-сайт.

1.3.3. Рандомизированные контролируемые исследования

А/В-тестирование известно уже очень давно. В здравоохранении А/В-тестирование называется *рандомизированные контролируемые исследования* (randomized controlled trials, RCT) и используется уже сотни лет. Самый ранний из известных нам случаев тестирования имел место в 1747 году, когда Джеймс Линд (James Lind) искал способ лечения цинги.

Основная идея та же самая. Предположим, у вас есть новый *способ лечения*, который, по вашему мнению, может вылечить определенную болезнь. Вы хотите проверить, так ли это. С этой целью вы создаете две группы: одна (*экспериментальная*, или *интервенционная*, группа) получает новое лекарство, а другая (*контрольная* группа) получает что-то еще для сравнения, например старое лекарство или плацебо (вещество под видом лекарства, которое не оказывает влияния на болезнь). Это – группы А и В. Важно, чтобы люди не назначались ни в одну из групп на основе таких факторов, как возраст, потому что тогда вы не узнаете, сработало ли лечение само по себе или из-за конкретных особенностей включенных в нее людей. Поэтому назначение лечения производится случайным образом, как в А/В-тестировании. Официально А/В-тестирование считается разновидностью RCT. Но вообще термин *RCT* используется в основном в здравоохранении, а термин *А/В-тестирование* – в разного рода компаниях, особенно осуществляющих свою деятельность в интернете. В этой книге оба термина будут использоваться для обозначения одного и того же: у нас есть лечение или решение, которое нужно принять, и мы хотим увидеть, как это повлияет на конкретный результат.

1.3.4. Шаги проведения А/В-тестирования

Метод проведения А/В-тестирования похож на научный метод: сначала выдвигается гипотеза, а затем проводится эксперимент, чтобы проверить ее истинность. Для более подробных объяснений ознакомьтесь с разделом «Дополнительная литература». Здесь мы просто приводим краткий обзор.

Шаг 1: гипотеза и план эксперимента

Эксперимент начинается с выдвижения гипотезы, которая должна описывать интересующую вас причинно-следственную связь между лечением или решением и желаемым результатом. Например, если вы придумали новый способ лечения болезни, то первоначальная гипотеза может заключаться в утверждении, что этот новый метод работает так же хорошо, как и старый (то есть новый метод не обязательно лучше). Согласно этому предположению, применение нового метода не даст никаких дополнительных преимуществ. Однако если новый способ лечения действительно лучше, то эксперимент должен предоставить доказательства, чтобы опровергнуть эту гипотезу.

После формулирования гипотезы планируется эксперимент. План эксперимента может значительно различаться, особенно между А/В-тестированием веб-сайтов и RCT новых медицинских методов лечения. RCT, как правило, более сложны, потому что имеют дело со здоровьем людей (что подразумевает соблюдение множества правил и международных стандартов и уважение конфиденциальности). Они также должны учитывать психологические эффекты использования плацебо и поиск добровольцев для исследования. С другой стороны,

A/B-тестирование цифровых платформ может быть проще, так как при этом этические проблемы обычно отсутствуют, и пользователи могут участвовать в тестировании, даже не осознавая этого.

Шаг 2: исполнение

Проводя эксперимент, будьте очень осторожны, чтобы не допустить никаких ошибок, особенно в процессе рандомизации, потому что ошибки могут привести к неверным выводам.

Шаг 3: анализ

После завершения эксперимента вы получаете данные, показывающие, как отреагировала каждая группа. Например, вы можете обнаружить, что новый метод излечивает 80 % пациентов, а старый – только 70 %. В упрощенном виде эти данные могли бы выглядеть так:

- старый метод лечения: 1, 0, 0, 1, 1, 1, 1, 0, 1, 1;
- новый метод лечения: 1, 1, 1, 0, 1, 1, 1, 1, 1, 0.

Здесь «1» означает излечение пациента, а «0» – нет. При необходимости можно проанализировать и другие результаты. Теперь главный вопрос: «Действительно ли новое лечение лучше старого?» С показателем выздоровления на 10 % выше кажется, что так оно и есть.

Разница в средних результатах между двумя группами известна как *средний лечебный эффект* (average treatment effect, ATE). Он говорит нам, насколько эффективнее один метод лечения по сравнению с другим. ATE очень важен для причинно-следственной связи и будет подробно рассматриваться в главе 2.

Важная интерпретация

Поскольку люди по группам распределяются случайным образом, то при достаточно большой выборке обе группы будут иметь схожие характеристики, такие как возраст, пол и т. д. Это означает: вычисляя средний результат для группы, вы оцениваете, что произойдет, если все получат это лечение. Таким образом, вычисление ATE – это оценка разницы в эффективности нового и старого способов лечения, если бы они были применены ко всей популяции.

Следующий вопрос после вычисления ATE: «Если провести эксперимент снова тем же способом, то получатся ли те же результаты?» Ответ «Нет» в таком случае означает, что результатам нельзя доверять.

На самом деле нельзя ожидать получить те же результаты при повторении эксперимента. Каждый раз, когда проводится эксперимент, ситуация у каждого участника может немного меняться, что может повлиять на результаты. Это заставляет задуматься, какое количество результатов обусловлено случайностью. Отделение сигнала от шума является ключевой частью статистики. Обычно случайность A/B-тестирования анализируется с помощью проверки гипотез и p -значений, но некоторые предпочитают байесовскую проверку

гипотез. Мы не будем здесь погружаться в эти методы, а желающие смогут найти дополнительную информацию в разделе «Дополнительная литература».

Важно отметить, что RCT и A/B-тестирование могут сказать нам, приводят ли два варианта к разным результатам, но они не объясняют причин этого. Например, если новый дизайн веб-сайта имеет много изменений по сравнению со старым и наблюдается разница в количестве посещений, то вы не сможете точно сказать, какое конкретное изменение привело к этой разнице. С другой стороны, если ваш тест изменяет что-то незначительное, и это не оказывает большого эффекта, то вам может понадобиться действительно большая группа людей, чтобы заметить разницу. Поэтому важно найти хороший баланс.

Контрольная точка

Давайте подчеркнем, насколько важны A/B-тестирование или RCT для установления причинно-следственной связи со статистической достоверностью. Это самый надежный метод подтверждения причинно-следственной связи. Если вы в этом сомневаетесь, то прочитайте еще раз раздел 1.3.2.

Выполняя каждый шаг, подумайте, что можно было бы сделать иначе. Запишите свои вопросы и размышления, чтобы при чтении следующих разделов увидеть, прояснятся ли они.

1.3.5. Ограничения A/B-тестирования и RCT

A/B-тестирование и RCT являются золотым стандартом для обнаружения причинно-следственной связи: по возможности старайтесь всегда проводить A/B-тестирование или RCT. Однако, как и любые инструменты, они также имеют свои ограничения. Ниже перечислены некоторые ситуации, когда A/B-тестирование или RCT невозможны и вместо них можно использовать причинно-следственный анализ.

- *Экспериментирование невозможно.* Например, если вы хотите узнать, как новый продукт вашего конкурента повлиял на ваши продажи, то можно, скрестив пальцы на удачу, попросить конкурентов сообщить вам статистику своих продаж.
- *Экспериментировать неэтично.* Если вы хотите узнать, вызывает ли курение рак у детей, то было бы неэтично заставлять курить детей в экспериментальной группе.
- Экспериментирование сопряжено со значительными затратами времени или средств. Чтобы узнать наверняка, имеет ли лечение долгосрочные побочные эффекты, эксперимент должен проводиться длительное время.
- *Недостаток внешней валидности.* Обычно для проведения RCT необходимо набирать людей-добровольцев. Добровольцы могут иметь личную заинтересованность к участию в эксперименте (деньги, возможное выздоровление и т. д.). Поэтому выборка может получиться нерепрезентативной. В этом случае мы говорим, что недостает внешней валидности.

Вопрос

Какие эксперименты вы могли бы провести в своей организации для выявления интересующих вас причинно-следственных связей? Как бы вы их спроектировали?

Подсказка: попробуйте спроектировать эксперименты, связанные с ответами, которые вы дали в этой главе.

1.4. Наблюдательные исследования

Эксперименты – хорошая штука, потому что ясно показывают причинно-следственную связь между двумя переменными. Но не всегда есть возможность проводить эксперименты. Когда такая возможность отсутствует, вы попадаете в область наблюдаемых данных: данных, которые не были получены с помощью А/В-тестирования или RCT. В этом разделе рассказывается о том, что может пойти не так без А/В-тестирования. Наблюдаемые данные обычно охватывают большую часть данных, с которыми мы работаем.

Спроси себя

Какие виды наблюдаемых данных использует ваша организация?

Вот некоторые примеры наблюдаемых данных:

- база данных клиентов, которая связывает маркетинговые компании с реакцией клиентов;
- отслеживание изменений продаж на основе решений, принимаемых компанией;
- данные о том, как разные страны справились с COVID-19, и последствиях распространения вируса.

Как правило, большинство данных используются для оценки социальной политики без проведения экспериментов. Мы привыкли делать выводы из наблюдаемых данных (иногда правильные, иногда не очень). Далее мы рассмотрим, почему с наблюдаемыми данными может быть сложно работать.

Когда нужен причинный анализ?

Причинный анализ необходим, когда приходится принимать решения на основе наблюдаемых данных. Если вы продолжите погружение в сферу причинного анализа, то в какой-то момент обнаружите, что этот вид анализа также помогает извлечь информацию из результатов RCT или А/В-тестирования, которую нельзя извлечь иным способом. Однако для простоты мы сначала рассмотрим применение причинного анализа только к наблюдаемым данным.

Как уже упоминалось, «корреляция – это не причинно-следственная связь». Это известное высказывание в статистике. Здесь под *корреляцией* понимается не математическая формула, а тот факт, что два явления происходят вместе. Но насколько обманчивой она может быть? Рассмотрим диаграмму (рис. 1.3), показывающую связь между доходом от игровых автоматов и количеством присуждений степени доктора философии в области информатики. Математическая корреляция между этими показателями очень высокая и равна 0,98. Это может создать впечатление, что одно является причиной другого. Но если подумать, то нетрудно понять, что эта идея не имеет смысла.

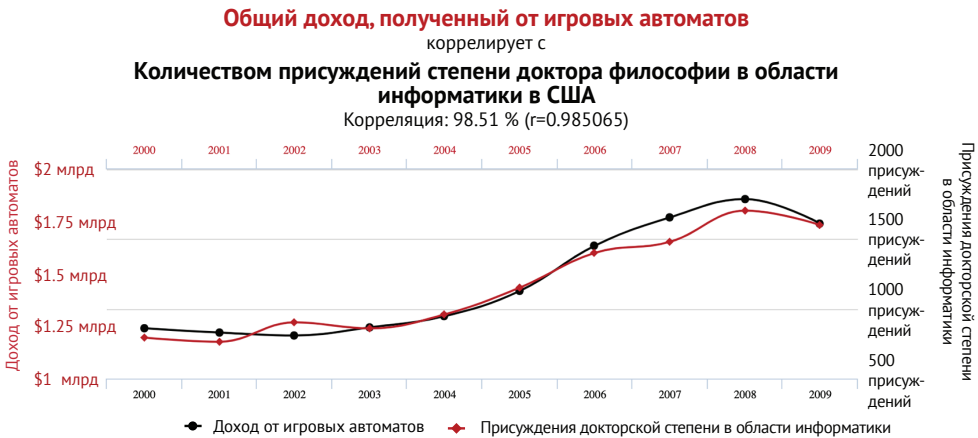


Рис. 1.3. Связь между доходами от игровых автоматов и количеством присуждений степени доктора философии в области информатики. Между этими показателями существует сильная корреляция, но это не является доказательством наличия причинно-следственной связи. Источник: www.tylervigen.com/spurious-correlations

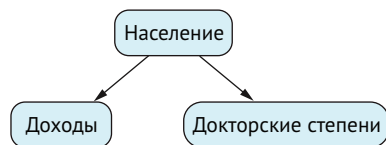
И еще

Посетите замечательную веб-страницу Spurious Corrections (www.tylervigen.com/spurious-correlations) Тайлера Вигена (Tyler Vigen), где приводятся графики корреляций. В некоторых случаях можно дать правдоподобное объяснение наблюдаемой корреляции, но есть и такие, которые явно не имеют смысла.

Наличие корреляции между двумя явлениями не означает, что одно из них является причиной другого. Но почему они коррелируют? Обычно этому есть причина! Часто корреляция обусловлена влиянием общей причины. Например, рост доходов от игровых автоматов и количества докторов философии в области информатики можно объяснить ростом численности населения. Чем больше людей, тем больше клиентов игровых автоматов и больше студентов, получающих высшее образование.

Люди часто думают, что если два явления происходят вместе, то у них должна быть общая причина. Это не всегда так, но иногда такой подход может быть полезен. Идею общей причины, влияющей на оба явления, можно показать с помощью графика, например как на рис. 1.4.

Рис. 1.4. Доходы от игровых автоматов и количество присуждений степени доктора философии в информатике растут с ростом населения, поэтому население является потенциальной общей причиной этих двух явлений



Можно сказать, что простого наличия корреляции недостаточно, чтобы говорить о причине и следствии. Это объясняется тем, что причинность имеет направление. Например, когда вы зажигаете огонь на газовой плите со стоящей на ней кастрюлей с водой, то понимаете, что огонь заставит воду вскипеть. В причинно-следственном анализе мы выражаем это так: *огонь* → *кипение*. Эта связь не работает в обратном направлении, она не симметрична. Очевидно, что кипение воды, например в микроволновке, не зажжет огонь на плите. Поэтому хотя корреляция рассматривает два явления так, будто они могут влиять друг на друга в равной степени ($\text{corr}(x, y) = \text{corr}(y, x)$), причинно-следственная связь так не работает. Корреляция слепа относительно причинно-следственной направленности.

1.4.1. Моделирование синтетических данных

В предыдущем разделе вы узнали, что фактор может связывать другие факторы вместе. В этом есть определенный смысл, не так ли? Но можно ли проверить верность этой идеи? Конечно. Для этого можно создать *синтетические наборы данных* и опробовать на них свои идеи.

Мы воспользуемся выдуманными данными, чтобы показать, как два явления могут казаться связанными, когда на самом деле это не так. Давайте создадим число, отражающее численность населения США, и назовем его *population*. Затем создадим две новые переменные, *revenue* и *doctorates*. Их значения будут основаны только на *population*, без учета какой-либо информации друг о друге. В конце вы заметите, что *doctorates* и *revenue* выглядят так, как будто они связаны. Это пример так называемой *ложной корреляции*.

Код в листингах 1.1 и 1.2 показывает, что корреляция между *doctorates* и *revenue* составляет 0,95. Этот пример не доказывает, что численность населения является фактической причиной связи между доходами от игровых автоматов и количеством присвоений степени доктора философии в информатике. Он просто предполагает, что причиной связи может быть третий фактор, такой как численность населения, влияющий и на доходы, и на количество докторов философии.

Листинг 1.1. (R) Создание синтетического набора данных с ложными корреляциями

```

set.seed(1234)

time <- 2000:2009
population <- 280 + 3 * (time - 2000) +
  rnorm(n=length(time), sd=0.1)
revenue <- 1.25 + (population - 280) * 0.015 +
  rnorm(n=length(time), sd=0.05)
)
doctorates <- 700 + (population - 280) * 30 +
  rnorm(n=length(time), sd=10)
cor(doctorates, revenue)

```

В миллионах человек
 В миллиардах долларов
 В единицах человек

Листинг 1.2. (Python) Создание синтетического набора данных с ложными корреляциями

```

from numpy.random import uniform, seed, normal
from numpy import arange, corrcoef
import pandas as pd

seed(1234)

time = arange(2000, 2010)
population = 280 + 3 * (time - 2000) + \
  normal(size=len(time), scale=0.1)
revenue = 1.25 + (population - 280) * 0.015 + \
  normal(size=len(time), scale=0.05)
doctorates = 700 + (population - 280) * 30 + \
  normal(size=len(time), scale=10)
corrcoef(doctorates, revenue)[0][1]

```

В миллионах человек
 В миллиардах долларов
 В единицах человек

В этой книге мы будем использовать много синтетических наборов данных. Причина в том, что когда мы работаем с данными как аналитики, наша главная цель – создать модель реального мира. Есть процесс, который создает данные, например законы физики или реальные явления. Мы пытаемся выяснить, как работает реальность, на основе имеющихся у нас данных.

Но вот в чем загвоздка: у нас нет прямого способа увидеть эти правила. Мы можем только догадываться о них на основе данных и экспериментов. Поэтому, даже хорошо потрудившись, мы не можем быть уверены, что наши выводы в точности соответствуют происходящему в реальности. Всегда есть некоторая неопределенность. На рис. 1.5 показано, как мы используем данные для представления реальности.

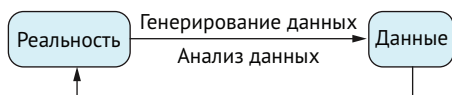


Рис. 1.5. Стандартные шаги в анализе данных: мы собираем некоторые данные, генерируемые средой. Процесс генерации данных скрыт. Наша цель – изучить эти данные и выяснить, как в действительности работает процесс. Рассматривайте это как разновидность реверс-инжиниринга

Когда мы изучаем новую тему, нам важно знать, делаем ли мы это правильно. В нашем случае при работе с реальными данными, как уже отмечалось выше, мы не всегда можем быть уверены в правильности наших результатов. Нам нужен процесс для проверки нашей методологии.

Вот как можно решить эту проблему: вместо использования реальных данных мы создаем синтетические данные, а затем используем методы, которые, по нашему мнению, лучше всего подходят для их анализа. Мы точно знаем, как создавались эти данные, а это все равно, что знать «истину» или правильные ответы. Это помогает нам увериться в точности наших выводов.

Сравните это с обучением плаванию. Сначала вы учитесь в безопасном мелком бассейне, где можно коснуться дна. Это похоже на анализ данных, когда заранее известны правильные ответы. Получив основные навыки и закрепив их в безопасной среде, вы сможете переместиться в более сложные условия, например начать плавать в открытом море. Это похоже на анализ реальных данных, когда правильные ответы заранее неизвестны.

1.4.2. Причинно-следственные связи при наличии искажающих факторов

В этом разделе представлен базовый граф, который вы часто будете видеть в данной книге. Разберем простейший пример, требующий причинно-следственного анализа: оценку влияния одной переменной на другую при общем влиянии на них третьей переменной. Этот третий фактор может создавать ложные корреляции между переменными, запутывать общую картину и приводить к классической проблеме «корреляция – это не причинно-следственная связь». Первый большой шаг в этой книге – научиться выделять это смещение, называемое *смещением искажающего фактора*, с помощью так называемой *формулы коррективки*, которую мы рассмотрим в главе 2.

Вот сценарий для размышления. Вы – врач, и у вас появилось новое лекарство от некоторой болезни. Первичные исследования показали, что оно эффективнее лекарств, применяемых в настоящее время. Вы почти год испытываете это новое лекарство на пожилых пациентах, которые чаще болеют этой болезнью. Однако новый препарат получают не только пожилые пациенты, но и молодые. Теперь, основываясь на своем собственном опыте, вы хотите выяснить, действительно ли новый препарат так эффективен, как утверждают исследования. Ситуацию можно представить с помощью графа на рис. 1.6.

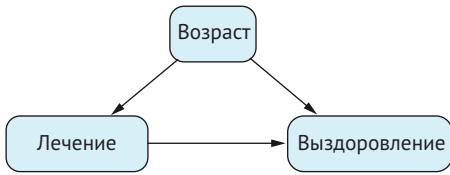


Рис. 1.6. Помимо препарата, на выздоровление может повлиять третий фактор. Такие третьи факторы называют искажающими факторами. Например, искажающим фактором в этом примере является возраст

Возраст влияет на выбор лекарства (лечения), поскольку вы учитываете возраст пациента, прежде чем принять решение о конкретном препарате. Кроме того, возраст влияет на скорость выздоровления, поскольку пожилые люди обычно выздоравливают медленнее, чем молодые. В анализе влияния решения на результат, когда имеется третий фактор, влияющий и на решение, и на результат, мы называем этот фактор *искажающим*. В нашем сценарии искажающим фактором является возраст. В реальной жизни мы часто сталкиваемся не с одним, а сразу с несколькими искажающими факторами. Фактор, влияние которого мы пытаемся измерить, часто называют *переменной лечения* или *решения*. Результат, который нас интересует, или что мы хотим увидеть в результате лечения, называется *исходом*. См. рис. 1.7.

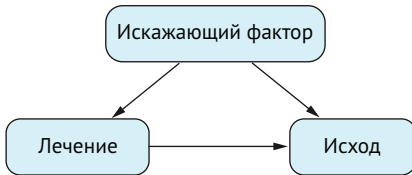


Рис. 1.7. Этот граф играет важную роль в причинном анализе, и вы будете часто с ним сталкиваться. При анализе влияния лечения на исход вам могут мешать факторы, влияющие и на то, и на другое. Такие факторы называются искажающими, и они могут изменить связь между лечением и исходом

Искажающий фактор получил такое название, потому что он может затруднить анализ данных. Вам будет трудно сказать, вызваны ли изменения в исходе изменением лечения или влиянием искажающего фактора. В этом примере новый препарат давался в основном пожилым людям, поэтому вы не можете уверенно утверждать, чем обусловлено ускоренное выздоровление – препаратом или возрастом пациента. Эта ситуация означает, что формула АТЕ, используемая вами для измерения эффективности влияния нового препарата на всех пациентов, в этом случае не работает.

Сначала подумай, потом читай

Какие факторы могут возникнуть при проведении А/В-тестирования или RCT?

Чтобы понять эту идею, давайте еще раз посмотрим, что такое искажающий фактор: это фактор, который влияет как на выбор лечения, так и на исход. Но в экспериментах выбор лечения основан только на случайности (т. е. лечение назначается случайным образом). Эта случайность не меняет исход, поэтому можно сказать, что *в экспериментальных данных нет искажающих факторов*. Это ключевая причина, почему экспериментальные данные надежнее наблюдаемых.

Причинно-следственный анализ дает нам инструменты для исследования наблюдаемых данных, а не экспериментальных. На рис. 1.7 показан типичный граф независимо от того, содержит он один или несколько искажающих факторов. Первые несколько глав данной книги посвящены этому сценарию. В частности, вы узнаете, как получить непредвзятые оценки АТЕ, когда это возможно.

1.5. Обзор основных статистических концепций

В этом разделе мы рассмотрим некоторые статистические идеи, важные для этой книги. Вы могли познакомиться с ними на вводных лекциях по статистике. Если вы уже знакомы с этими концепциями, то можете смело пропустить данный раздел. Однако имейте в виду, что для дальнейшего изучения причинного-следственного анализа очень важно понимать концепции *условной вероятности* и *ожидания*, которые мы будем часто использовать в книге.

1.5.1. Эмпирические распределения и распределения сгенерированных данных

В практических задачах обычно есть два различных распределения. Первое – это распределение, которое генерирует данные. В действительности за создание данных отвечает физический механизм. Это то, что мы называем процессом генерирования данных. Этот процесс может иметь некоторую неопределенность и, следовательно, присущее ему распределение вероятностей, которое мы будем называть *распределением сгенерированных данных*. За редким исключением, мы ничего не знаем об этом распределении. Законы, которые создают наши данные, обычно определяются природой, и у нас нет доступа к этой информации.

Возьмем в качестве примера подбрасывание монеты. Подбрасывая монету n раз с вероятностью выпадения орла $P(H) = p$, мы ожидаем, что в конечном итоге, когда n стремится к бесконечности, мы получим долю p выпадений орлов и долю $1 - p$ выпадения решек. Дело в том, что мы не знаем точного значения p . Даже если процесс изготовления монеты был тщательно настроен, в нем все равно могут сохраняться некоторые неточности. Обычно мы предполагаем, что $p = 1/2$, но чтобы подтвердить это предположение, потребуется бесконечное число подбрасываний. Поэтому на самом деле мы не знаем истинного значения p .

Другое распределение – это то, что мы называем *эмпирическим распределением*, полученным из выборки. Предположим, что мы подбросили монету пять раз и получили О, О, Р, Р, Р. Мы можем обобщить результаты, как показано в табл. 1.1. Мы ожидаем, что если подбросить монету не пять, а большее количество раз, то вероятность получить О (орел) в нашей выборке будет близка к p .

Таблица 1.1. Эмпирическое распределение по выборке O, O, P, P, P

Исход	Вероятность
O	2/5
P	3/5

Формально мы можем так поступить, потому что *выборка сама по себе является распределением*: эмпирическим распределением. Если предположить, что каждое наблюдение имеет одинаковый вес и у нас есть размер выборки n , то вероятность каждого наблюдения равна $1/n$. В нашем случае каждое наблюдение имеет вес $1/5$, а распределение вероятностей в точности соответствует показанному в табл. 1.1. Тогда математическое ожидание этого распределения совпадает со средним значением по выборке:

$$1 \times \frac{2}{5} + 0 \times \frac{3}{5} = \bar{x}.$$

Эмпирическое распределение и распределение, генерирующее данные, тесно связаны. Когда n стремится к бесконечности, эмпирическое распределение стремится к распределению, генерирующему данные. Этот результат известен как теорема Гливенко–Кантелли, которая является сугубо технической и не будет рассматриваться в этой книге. Эта теорема также справедлива для различных ситуаций, например когда переменная непрерывна или когда вместо случайной величины у нас есть случайный вектор.

Предположим, мы хотим, чтобы выиграл орел. Обозначим успех (орел) как 1, а неудачу (решку) как 0. В таком случае предыдущая выборка (O, O, P, P, P) превратится в 1, 1, 1, 0, 0, а доля выпадений орла совпадет со средним значением, обозначенным как $\bar{x} = 2/5$. В то же время это среднее значение можно вычислить как

$$\bar{x} = 1 \times \frac{2}{5} + 0 \times \left(1 - \frac{2}{5}\right) = 1 \times \frac{2}{5} + 0 \times \frac{3}{5}$$

Аналог среднего для распределения, генерирующего данные, называемый *ожиданием*, вычисляется с использованием того факта, что $P(O) = p$. Для заданной бинарной случайной величины X ожидание вычисляется как

$$E[X] = 1 \times p + 0 \times (1 - p) = p.$$

Итак, мы получаем вероятность выпадения орла $\bar{x} = 2/5$, но если бы у нас была очень большая выборка, то эта вероятность была бы близка к p .

Обратите внимание на обозначение, использованное для различения выборочного среднего (\bar{x}) и ожидания ($E[X]$). В статистике нас часто интересуют ответы на вопросы, связанные с различиями между выборкой и базовым процессом, сгенерировавшим данные, например достаточен ли размер выборки, чтобы быть уверенными в результате, получаемом из выборки.

Имейте в виду

При выяснении причинно-следственных связей прежде всего нужно определить правильную формулу для использования. После этого можно выбирать наилучший статистический подход. Углубленное обсуждение оптимальных статистических методов и определения доверительных интервалов вы найдете в главе 8.

По этой причине, подойдя к выбору формулы, мы фокусируемся на распределении, генерирующем данные, а выбирая статистические подходы, переходим к эмпирическому распределению.

1.5.2. Условные вероятности и ожидания

Для работы с этой книгой необходимо твердое понимание условных вероятностей и условных ожиданий. Если вы уже знакомы с этими понятиями, то можете пропустить данный раздел. Однако, как показывает мой опыт, условные вероятности сложнее, чем кажется. Я начал изучать их с первого курса университета, но только годы спустя почувствовал, что понимаю их суть. Так что если какая-то из этих концепций вам не до конца понятна, то продолжайте читать этот раздел.

Условные вероятности

Начнем с предположения, что у нас есть некоторые данные, как в табл. 1.2 с переменными X , Y и Z . Значение a используется намеренно, чтобы выработать привычку к абстрактной записи условных вероятностей.

Таблица 1.2. Смоделированные данные

X	Y	Z
3	0,03	A
6	-24,08	A
a	7,01	A
-2	-3,00	B
a	10,89	B

В этом случае обусловливание $X = a$ означает получение новой таблицы и выбор тех случаев, когда переменная X равна a . С точки зрения программирования это равносильно фильтрации данных, т. е. выбору строк, как в табл. 1.3. Переменные Y и Z , при ограничении $X = a$, обозначаются как $Y | X = a$ и $Z | X = a$ соответственно. Обусловливание $X = a$ может изменить распределение Y и Z . Например, если в табл. 1.2 переменная Z принимает значение A в трех случаях из пяти, то в табл. 1.3 она принимает значение A в одном случае из двух. Это можно записать как $P(Z = A) = 3/5$ и $P(Z = A | X = a) = 1/2$.

Таблица 1.3. Смоделированные данные, обусловленные $X = a$

X	Y	Z
a	7,01	A
a	10,89	B

Возможно, в прошлом вы уже познакомились с условными вероятностями на примере математической формулы

$$P(Z = A | X = a) = \frac{P(Z = A, X = a)}{P(X = a)},$$

где запятая (,) обозначает союз *и*: $P(Z = A, X = a)$ означает вероятность одновременного выполнения условий $Z = A$ и $X = a$. Преимущество этой формулы в том, что она показывает, как можно рассчитать условные вероятности на основе исходной табл. 1.2 и частот (вероятностей) появления событий $Z = A$ и $X = a$. Формула получена на основе шагов, которые мы выполнили для вычисления $P(Z = A | X = a) = 1/2$, разделив следующие величины:

$$\begin{aligned} P(Z = A | X = a) &= \frac{1}{2} \\ &= \frac{\text{число вхождений } Z = A \text{ в табл. 1.3}}{\text{число строк в табл. 1.3}} \\ &= \frac{\text{число вхождений } Z = A \text{ и } X = a \text{ в табл. 1.1}}{\#X = a \text{ в табл. 1.1}}. \end{aligned}$$

Выражение остается неизменным при делении числителя и знаменателя на 5 (количество строк в табл. 1.1), поэтому приходим к выражению

$$P(Z = A | X = a) = \frac{P(Z = A, X = a)}{P(X = a)}.$$

Обуславливать можно сразу несколько переменных. Например, обуславливая $X = a$ и $Z = B$, мы получаем табл. 1.4. Выбор переменной Y при ограничениях $X = a$ и $Z = B$ обозначается как $Y | X = a, Z = B$. Так же, как и раньше, мы можем вычислить $P(Y = 10,89 | X = a, Z = B) = 1$, поскольку 10,89 – это уникальное значение, которое принимает переменная Y при ограничениях $X = a$ и $Z = B$.

Таблица 1.4. Смоделированные данные, обусловленные условиями $X = a$ и $Z = B$

X	Y	Z
a	10,89	B

В общем случае, если есть две переменные X и Y , то условие $X = x$, вероятно, изменит поведение Y , а частота, с которой Y принимает свои значения, будет отличаться от частоты до введения условия:

$$P(Y = y | X = x).$$

Иногда, чтобы сделать формулы более удобными для чтения, мы будем слегка злоупотреблять записью и писать $P(Y | X)$ или даже $P(Y | X = x)$ вместо полного правильного выражения $P(Y = y | X = x)$. Условную вероятность можно вычислить из исходной вероятности P по формуле

$$P(Y = y | X = x) = \frac{P(Y = y, X = x)}{P(X = x)}.$$

Условные ожидания

Теперь, когда посредством обусловливания мы получили новую переменную ($Y | X = a$ в предыдущем примере), из нее можно вычислить некоторые типичные величины, такие как $P(Y = 7,01 | X = a) = 1/2$ (потому что у нас осталось только два наблюдения) или $P(Y = 1 | X = a) = 0$. В частности, мы можем вычислить ожидание этой новой переменной, называемое *условным ожиданием*: $E[Y | X = a]$. В нашем примере, поскольку для $Y | X = a$ каждое наблюдение имеет вес $1/2$, ожидание $E[Y | X = a] = 7,01 \times 1/2 + 10,89 \times 1/2 = 8,95$. Обратите внимание, что если переменная Y является бинарной, то $E[Y | X] = P(Y | X)$. Чтобы убедиться в этом, достаточно применить определение условного ожидания. Поскольку Y категориальна, она принимает только два значения: 0 и 1. То есть

$$E[Y | X] = 1 \times P(X = 1 | X) + 0 \times P(X = 0 | X) = P(X = 1 | X).$$

На рис. 1.8 мы смоделировали данные, чтобы на еще одном примере показать разницу между распределением переменной и ее условными распределением и ожиданием. Слева показана выборка переменной Y , которая является комбинацией выборок из двух различных гауссовых распределений: одно основано на значении $x = 0$ с ожиданием 0, а другое – на значении $x = 2$ с ожиданием 2. Как видите, среднее значение выборки безусловного распределения равно 1. Но справа те же значения Y разделены на две группы, каждая для соответствующего значения x : левая группа – это распределение $Y | X = 0$, а правая группа – условное распределение $Y | X = 2$. Черные точки представляют ожидание $E[Y | X = 0]$ и $E[Y | X = 2]$. Точки были смещены по оси x для наглядности и избежания наложения.

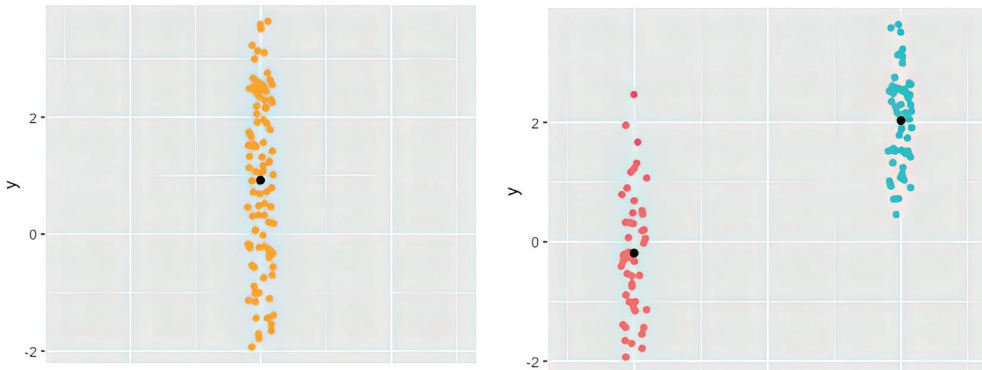


Рис. 1.8. Слева – выборка переменной Y . Справа – те же данные, разделенные на два разных значения x , что дает условные распределения для разных x . Центральные черные точки – это средние значения для каждой группы, поэтому слева – безусловное ожидание, а справа – два условных ожидания

На протяжении всей книги мы будем говорить о $E[Y | X]$ как об абстрактной величине. Это означает, что она отделена от любого конкретного набора данных или задачи. В машинном обучении или причинно-следственном анализе нас интересует связь между некоторой переменной X и результатом Y . Итак, как можно представить $E[Y | X]$ в целом? Математическая нотация $E[Y | X]$ определяет простой рецепт:

- 1 $E[Y | X]$ – это сокращенная запись для выбора конкретных значений x и y : т. е. на самом деле подразумевается $E[Y | X = x]$.
- 2 Эта запись предполагает выборку из данных случаев, где $X = x$.
- 3 Расчет среднего значения переменной Y для этой конкретной группы.

В общем случае, поскольку выражение $E[Y | X = x]$ зависит только от значения x , $E[Y | X = x]$ можно рассматривать как функцию от x .

До сих пор мы говорили о вычислении условных ожиданий с использованием числовых (табл. 1.4) и визуальных (рис. 1.8) примеров. Чтобы упростить использование абстрактной величины $E[Y | X = x]$, давайте теперь рассмотрим пример, в котором нужно представить условные ожидания на основе данного здесь описания. Возьмите численность населения страны по своему выбору. Мысленно создайте две группы: группу A , включающую людей моложе 30 лет, и группу B , включающую людей старше 30 лет. Дадим переменной, определяющей группу, имя X . Итак, $X = A$ обозначает группу людей моложе 30 лет, а $X = B$ – группу людей старше 30 лет. Предположим, что вас интересуют закономерности, наблюдаемые в росте людей (обозначим эту переменную как Y). Вы знаете, как вычислить средний рост для группы A . Эту величину нетрудно выразить в математических терминах. Как вы только что видели, выбор группы A и последующее вычисление их среднего роста обозначается как $E[Y | X = A]$. Соответственно, процесс выбора группы B и вычисления среднего роста для нее обозначается как $E[Y | X = B]$. Обратите внимание, что мы говорим об абстрактных величинах в отсутствие каких-либо данных.

Другая ситуация, когда может понадобиться вычислить условные вероятности и ожидания, – когда имеется явная функциональная связь между переменными. Предположим, что вместо данных у вас есть следующая линейная модель:

$$Y = 1 + 2X + \varepsilon,$$

где ε – центрированное гауссово распределение. Это означает, что X может изменяться независимо, а Y зависит от X и некоторого случайного фактора. Обусловливание $X = x$ просто означает, что значение X будет установлено равным x , тогда как переменная Y все еще может изменяться из-за случайности ε . В общем случае, если Y , X и ε связаны некоторой функцией $Y = f(X, \varepsilon)$, обусловленной $X = x$, то означает, что теперь Y будет меняться при фиксированном X :

$$Y = f(X = x, \varepsilon).$$

Таким образом, *условное ожидание* $E[Y | X = x]$ необходимо рассчитывать только с учетом случайности, поскольку x – фиксированное значение.

В главе 2 вы увидите, что при наличии искажающих факторов расчет условных вероятностей не всегда дает хороший ответ на причинные вопросы и нужно найти формулу, чтобы удалить влияние искажающих факторов из расчетов.

Обратите внимание, что понятие условной вероятности является статистическим или вероятностным понятием, *но не причинным*. Условная вероятность $P(Y | X)$ просто означает вычисление частоты события Y среди случаев, когда также происходит событие X . Таким образом, она описывает частоты или вероятности событий.

1.6. Для дополнительного чтения

- A/B-тестирование:
 - «Refuted Causal Claims from Observational Studies»: https://experimentguide.com/refuted_observational_studies;
 - Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne «Controlled Experiments on the Web: Survey and Practical Guide» (Springer, 2009): www.robotics.stanford.edu/~ronnyk/2009controlledExperimentsOnTheWebSurvey.pdf;
 - глава «A/B Testing» в книге¹ Дэвида Свита (David Sweet) «Experimentation for Engineers» (Manning, 2023): www.manning.com/books/experimentation-for-engineers;
 - Stefan Thomke «Experimentation Works: The Surprising Power of Business Experiments» (Harvard Business Review Press, 2020): <https://store.hbr.org/product/experimentation-works-the-surprising-power-of-business-experiments/10248>;

¹ Свит Д. Тюнинг систем: экспериментирование для инженеров от A/B-тестирования до байесовской оптимизации. Питер, 2024. ISBN: 978-5-4461-2157-1. – *Прим. перев.*

- Ron Kohavi, Diane Tang, and Ya Xu, «Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing»¹ (Cambridge University Press, 2020): <https://experimentguide.com>.
- Причинно-следственный анализ:
 - Judea Pearl and Dana Mackenzie «The Book of Why: The New Science of Cause and Effect»² (Basic Books, 2018) – здесь вы найдете описание причинно-следственной связи с философской точки зрения, без формул и инструментов, которые можно было бы использовать для решения задач;
 - Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell «Causal Inference In Statistics: A Primer» (Wiley, 2016) – техническая вводная книга, более техническая, чем та, которую вы сейчас читаете;
 - Judea Pearl «Causality: Models, Reasoning, and Inference» (Cambridge University Press, 2020) – потрясающая книга, но ее трудно читать;
 - Joshua Angrist and Jörn-Steffen Pischke «Mostly Harmless Econometrics» (Princeton University Press, 2009) – классика эконометрики. Авторы также написали более вводную версию под названием «Mastering 'Metrics: The Path from Cause to Effect» (Princeton University Press, 2014);
 - Guido Imbens and Donald Rubin «Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction» (Cambridge University Press, 2015). В этой книге основное внимание уделяется границе между RCT и причинно-следственным анализом.

1.7. Контрольные вопросы

Завершая главу, важно убедиться в наличии четкого понимания ключевых концепций. Ниже перечислены контрольные вопросы, на которые вы должны уметь ответить четко и кратко. Если это не так, то я рекомендую еще раз прочитать разделы в ссылках, сопровождающих вопросы.

- 1 В чем разница между наблюдаемыми и экспериментальными данными? Ответ в разделе 1.1.1.
- 2 Когда следует проводить A/B-тестирование или RCT? Ответ в разделе 1.3.5.
- 3 При проведении A/B-тестирования, если выборка достаточно большая, то обе группы будут иметь одинаковые выборочные характеристики. Почему? Ответ в разделе 1.3.4, «Шаг 3: анализ» .

¹ Кохави Рон, Тан Диана, Сюй Я. Доверительное A/B-тестирование: Практическое руководство по контролируемым экспериментам. ДМК-Пресс, 2021. ISBN: 978-5-97060-913-2. – Прим. перев.

² Перл Джуда, Маккензи Дана. Думай «почему?». Причина и следствие как ключ к мышлению. АСТ, 2023. ISBN: 978-5-17-123140-8. – Прим. перев.

- 4 Назовите одну причину, почему корреляции не всегда могут служить доказательством причинно-следственной связи. Ответ в разделе 1.4.
- 5 Какие факторы можно обнаружить при проведении А/В-тестирования? Ответ в разделе 1.4.2.

Итоги

- Причинно-следственный анализ использует методы и инструменты для выявления причинно-следственных связей между лечением и исходом путем исследования данных.
- Искажающие факторы влияют как на лечение, так и на исход, усложняя определение причинно-следственных связей на основе одних только корреляций. Выявление искажающих факторов имеет решающее значение.
- Всегда проводите эксперименты (RCT или А/В-тестирование), если это возможно. Случайное назначение лечения ослабляет влияние искажающих факторов и позволяет получить самые четкие оценки.
- Когда эксперименты невозможны, приходится иметь дело с наблюдаемыми данными. Причинно-следственный анализ помогает принимать обоснованные решения на основе таких данных.
- Ориентированные ациклические графы (directed acyclic graph, DAG) помогают проиллюстрировать взаимосвязи между переменными и выявить возможные пути искажений.

Средний лечебный эффект (average treatment effect, ATE) количественно определяет причинно-следственное влияние одной переменной на другую. Расчет ATE зависит от искажающих факторов и будет подробно описан в следующих главах.