

SECOND EDITION

Practical Statistics for Data Scientists

50+ Essential Concepts Using R and Python

Peter Bruce, Andrew Bruce, and Peter Gedeck

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Питер Брюс, Эндрю Брюс, Питер Гедек

Практическая статистика для специалистов Data Science

50+ важнейших понятий с использованием R и Python

2-е издание

Санкт-Петербург
«БХВ-Петербург»

2021

УДК 004.6+519.2
ББК 32.81+22.172
Б89

Брюс, П.

Б89 Практическая статистика для специалистов Data Science: Пер. с англ. / П. Брюс, Э. Брюс, П. Гедек. — 2-е изд., перераб. и доп. — СПб.: БХВ-Петербург, 2021. — 352 с.: ил.

ISBN 978-5-9775-6705-3

Книга рассчитана на специалистов в области Data Science, обладающих некоторым опытом работы с языком программирования R и имеющих предварительное понятие о математической статистике. В ней в удобной и легкодоступной форме представлены ключевые понятия из статистики, которые относятся к науке о данных, а также объяснено, какие понятия важны и полезны с точки зрения науки о данных, какие менее важны и почему. Подробно раскрыты темы: разведочный анализ данных, распределения данных и выборок, статистические эксперименты и проверка значимости, регрессия и предсказание, классификация, статистическое машинное обучение и обучение без учителя. Во второе издание включены примеры на языке Python, что расширяет практическое применение книги.

Для аналитиков данных

УДК 004.6+519.2
ББК 32.81+22.172

Группа подготовки издания:

Руководитель проекта	<i>Евгений Рыбаков</i>
Зав. редакцией	<i>Людмила Гауль</i>
Перевод с английского	<i>Андрея Логунова</i>
Компьютерная верстка	<i>Ольги Сергиенко</i>
Оформление обложки	<i>Карины Соловьевой</i>

© 2021 БНВ

Authorized Russian translation of the English edition of *Practical Statistics for Data Scientists 2nd edition*

ISBN 9781492072942 © 2020 Peter Bruce, Andrew Bruce, and Peter Gedeck.

This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

Авторизованный перевод с английского языка на русский издания *Practical Statistics for Data Scientists 2nd edition*

ISBN 9781492072942 © 2020 Peter Bruce, Andrew Bruce, Peter Gedeck.

Перевод опубликован и продается с разрешения компании-правообладателя O'Reilly Media, Inc.

"БХВ-Петербург", 191036, Санкт-Петербург, Гончарная ул., 20.

ISBN 978-1-492-07294-2 (англ.)
ISBN 978-5-9775-6705-3 (рус.)

© Peter Bruce, Andrew Bruce, Peter Gedeck, 2020
© Перевод на русский язык, оформление. ООО "БХВ-Петербург",
ООО "БХВ", 2021

Оглавление

Об авторах.....	13
Предисловие	15
Условные обозначения, принятые в книге	15
Использование примеров кода	16
Благодарности.....	16
Комментарии переводчика	17
Глава 1. Разведывательный анализ данных.....	19
Элементы структурированных данных	20
Дополнительные материалы для чтения	22
Прямоугольные данные	23
Кадры данных и индексы.....	24
Непрямоугольные структуры данных.....	25
Дополнительные материалы для чтения.....	26
Оценки центрального положения	26
Среднее	27
Медиана и робастные оценки.....	29
Выбросы	29
Пример: средние оценки численности населения и уровня убийств.....	30
Дополнительные материалы для чтения.....	32
Оценки вариабельности	32
Стандартное отклонение и связанные с ним оценки.....	33
Оценки на основе процентилей.....	35
Пример: оценки вариабельности населения штатов.....	36
Дополнительные материалы для чтения.....	37
Разведывание распределения данных.....	38
Процентили и коробчатые диаграммы	38
Частотные таблицы и гистограммы.....	40
Графики и оценки плотности.....	42
Дополнительные материалы для чтения.....	44
Разведывание двоичных и категориальных данных.....	44
Мода.....	46
Ожидаемое значение	47
Вероятность.....	47
Дополнительные материалы для чтения.....	48
Корреляция.....	48
Диаграммы рассеяния	52
Дополнительные материалы для чтения.....	53

Разведывание двух или более переменных	53
Сетка из шестиугольных корзин и контуры (сопоставление числовых данных с числовыми данными на графике)	54
Две категориальные переменные	57
Категориальные и числовые данные	58
Визуализация многочисленных переменных	60
Дополнительные материалы для чтения	62
Резюме	63
Глава 2. Распределение данных и распределение выборок	65
Случайный отбор и смещенная выборка	66
Смещение	68
Случайный отбор	69
Размер против качества: когда размер имеет значение?	70
Выборочное среднее против популяционного среднего	71
Дополнительные материалы для чтения	71
Систематическая ошибка отбора	72
Регрессия к среднему	73
Дополнительные материалы для чтения	75
Выборочное распределение статистической величины	75
Центральная предельная теорема	78
Стандартная ошибка	79
Дополнительные материалы для чтения	80
Бутстрап	80
Повторный отбор против бутстрапирования	84
Дополнительные материалы для чтения	84
Доверительные интервалы	84
Дополнительные материалы для чтения	87
Нормальное распределение	87
Стандартное нормальное распределение и квантиль-квантильные графики	89
Длиннохвостые распределения	91
Дополнительные материалы для чтения	93
<i>t</i> -Распределение Стьюдента	93
Дополнительные материалы для чтения	95
Биномиальное распределение	95
Дополнительные материалы для чтения	98
Распределение хи-квадрат	98
Дополнительные материалы для чтения	99
<i>F</i> -распределение	99
Дополнительные материалы для чтения	100
Распределение Пуассона и другие связанные с ним распределения	100
Пуассоновские распределения	101
Экспоненциальное распределение	101
Оценивание интенсивности отказов	102
Распределение Вейбулла	102
Дополнительные материалы для чтения	103
Резюме	104
Глава 3. Статистические эксперименты и проверка значимости	105
<i>A/B</i> -тестирование	105
Зачем нужна контрольная группа?	108

Почему только A/B ? Почему не $C, D...?$	109
Дополнительные материалы для чтения.....	110
Проверки гипотез	110
Нулевая гипотеза	112
Альтернативная гипотеза.....	112
Односторонняя проверка гипотезы против двухсторонней	113
Дополнительные материалы для чтения.....	114
Повторный отбор.....	114
Перестановочный тест.....	115
Пример: прилипчивость веб-страниц	115
Исчерпывающий и бутстраповский перестановочные тесты.....	119
Перестановочные тесты: сухой остаток для науки о данных	119
Дополнительные материалы для чтения.....	120
Статистическая значимость и p -значения	120
p -Значение	123
Альфа	124
Разногласия по поводу p -значения.....	124
Практическая значимость	125
Ошибки 1-го и 2-го рода	125
Наука о данных и p -значения	126
Дополнительные материалы для чтения.....	126
Проверки на основе t -статистики.....	127
Дополнительные материалы для чтения.....	129
Множественное тестирование.....	129
Дополнительные материалы для чтения.....	132
Степени свободы	133
Дополнительные материалы для чтения.....	134
Дисперсионный анализ	134
F -статистика	138
Двухсторонний дисперсионный анализ.....	139
Дополнительные материалы для чтения.....	140
Проверка на основе статистики хи-квадрат	140
Проверка хи-квадрат: подход на основе повторного отбора	141
Проверка хи-квадрат: статистическая теория	143
Точный тест Фишера.....	144
Релевантность для науки о данных	146
Дополнительные материалы для чтения.....	147
Алгоритм многорукого бандита.....	147
Дополнительные материалы для чтения.....	150
Мощность и размер выборки.....	151
Размер выборки.....	152
Дополнительные материалы для чтения.....	155
Резюме	155
Глава 4. Регрессия и предсказание	157
Простая линейная регрессия.....	157
Уравнение регрессии.....	158
Подогнанные значения и остатки.....	161
Наименьшие квадраты	162
Предсказание против объяснения (профилирование).....	163
Дополнительные материалы для чтения.....	164

Множественная линейная регрессия	164
Пример: данные жилого фонда округа Кинг.....	165
Оценивание результативности модели.....	167
Перекрестный контроль.....	169
Отбор модели и пошаговая регрессия	170
Взвешенная регрессия.....	173
Дополнительные материалы для чтения.....	175
Предсказание с использованием регрессии	175
Опасности экстраполяции.....	175
Доверительный и предсказательный интервалы	176
Факторные переменные в регрессии.....	178
Представление фиктивных переменных.....	178
Факторные переменные с многочисленными уровнями	181
Упорядоченные факторные переменные.....	183
Интерпретирование уравнения регрессии.....	184
Коррелированные предсказатели	185
Мультиколлинеарность	186
Искажающие переменные.....	187
Взаимодействия и главные эффекты	188
Диагностика регрессии	190
Выбросы	191
Влиятельные значения	193
Гетероскедастичность, ненормальность и коррелированные ошибки	196
Графики частных остатков и нелинейность	199
Многочленная и сплайновая регрессия	201
Многочлены	202
Сплайны.....	203
Обобщенные аддитивные модели	206
Дополнительные материалы для чтения	207
Резюме	208
Глава 5. Классификация	209
Наивный Байес.....	210
Почему точная байесова классификация непрактична?.....	211
Наивное решение	211
Числовые предсказательные переменные	214
Дополнительные материалы для чтения.....	215
Дискриминантный анализ.....	215
Матрица ковариаций	216
Линейный дискриминант Фишера	217
Простой пример	217
Дополнительные материалы для чтения.....	221
Логистическая регрессия	221
Функция логистического отклика и логит	222
Логистическая регрессия и ОЛМ	223
Обобщенные линейные модели.....	225
Предсказанные значения из логистической регрессии	225
Интерпретирование коэффициентов и отношений перевесов.....	226
Линейная и логистическая регрессия: сходства и различия	228
Подгонка модели	228

Оценивание результативности модели	229
Анализ остатков	231
Дополнительные материалы для чтения	232
Оценивание классификационных моделей	233
Матрица путаницы	234
Проблема редкого класса	236
Прецизионность, полнота и специфичность	236
ROC-кривая	238
Площадь под ROC-кривой	240
Лифт	241
Дополнительные материалы для чтения	243
Стратегии для несбалансированных данных	243
Понижающий отбор	244
Повышающий отбор и повышающая/понижающая перевесовка	245
Генерация данных	246
Стоимостная классификация	247
Разведывание предсказаний	247
Дополнительные материалы для чтения	249
Резюме	249
Глава 6. Статистическое машинное обучение	251
<i>k</i> ближайших соседей	252
Небольшой пример: предсказание невыплаты ссуды	253
Метрики расстояния	255
Кодировщик с одним активным состоянием	256
Стандартизация (нормализация, <i>z</i> -оценки)	257
Выбор числа <i>k</i>	260
<i>k</i> ближайших соседей как механизм порождения признаков	261
Древесные модели	263
Простой пример	264
Алгоритм рекурсивного подразделения	267
Измерение однородности или загрязненности	268
Остановка выращивания дерева	270
Контроль за сложностью дерева в R	270
Контроль за сложностью дерева в Python	271
Предсказывание непрерывного значения	271
Каким образом используются деревья	272
Дополнительные материалы для чтения	273
Бэггинг и случайный лес	273
Бэггинг	274
Случайный лес	275
Важность переменных	279
Гиперпараметры	282
Бустинг	283
Алгоритм бустирования	285
XGBoost	286
Регуляризация: предотвращение перепогонки	288
Гиперпараметры и перекрестный контроль	292
Резюме	296

Глава 7. Неконтролируемое самообучение.....	297
Анализ главных компонент	298
Простой пример	299
Вычисление главных компонент.....	301
Интерпретирование главных компонент	302
Анализ соответствия	305
Дополнительные материалы для чтения.....	307
Кластеризация на основе K средних	307
Простой пример	308
Алгоритм K средних	310
Интерпретирование кластеров	311
Выбор числа кластеров	313
Иерархическая кластеризация.....	315
Простой пример	316
Дендограмма	317
Агломеративный алгоритм	318
Меры несхожести	319
Модельно-ориентированная кластеризация.....	321
Многомерное нормальное распределение	321
Смеси нормальных распределений	322
Выбор числа кластеров	325
Дополнительные материалы для чтения.....	327
Шкалирование и категориальные переменные	328
Шкалирование переменных	328
Доминантные переменные	330
Категориальные данные и расстояние Говера	332
Проблемы кластеризации смешанных данных	334
Резюме	336
Библиография	337
Предметный указатель.....	339

*Мы хотим посвятить эту книгу памяти наших родителей,
Виктора Г. Брюса и Нэнси С. Брюс, которые воспитали в нас
страсть к математике и точным наукам,
а также нашим первым учителям, Джону У. Тьюки и Джулиану Саймону,
и нашему верному другу, Джеффу Уотсону, который вдохновил нас на то,
чтобы мы посвятили свою жизнь статистике.*

*Питер Гедек хотел бы посвятить эту книгу
Тиму Кларку и Кристиану Крамеру с глубокой благодарностью
за их научное сотрудничество и дружбу.*

Об авторах

Питер Брюс (Peter Bruce) основал и расширил Институт статистического образования Statistics.com, который теперь предлагает порядка 100 курсов в области статистики, из которых примерно половина предназначена для аналитиков данных. Нанимая в качестве преподавателей ведущих авторов и шлифуя маркетинговую стратегию для привлечения внимания профессиональных аналитиков данных, Питер развил широкое представление о целевом рынке и свои собственные экспертные знания для его завоевания.

Эндрю Брюс (Andrew Bruce) имеет более чем 30-летний стаж работы в области статистики и науки о данных в академической сфере, правительстве и бизнесе. Он обладает степенью кандидата наук в области статистики Вашингтонского университета и опубликовал несколько работ в рецензируемых журналах. Он разработал статистико-ориентированные решения широкого спектра задач, с которыми сталкиваются разнообразные отрасли, начиная с солидных финансовых фирм до интернет-стартапов, и располагает глубоким пониманием практики науки о данных.

Питер Гедек (Peter Gedeck) имеет более чем 30-летний опыт работы в области научных вычислений и науки о данных. После 20 лет работы в качестве вычислительного химика в компании Novartis он занимает должность старшего исследователя данных в компании Collaborative Drug Discovery. Питер специализируется на разработке алгоритмов машинного обучения для предсказания биологических и физико-химических свойств препаратов-кандидатов. Соавтор книги "Добыча закономерностей из данных для бизнес-аналитики" (Data Mining for Business Analytics). Имеет докторскую степень по химии, которую он получил в Университете Эрланген-Нюрнберг в Германии, а в Университете Фернуниверситет-Хаген (Германия) изучал математику.

Предисловие

Книга рассчитана на исследователя данных, имеющего некоторый опыт работы с языком программирования R и/или Python и имеющего предшествующий (возможно, обрывочный или сиюминутный) контакт с математической статистикой. Двое из трех авторов пришли в мир науки о данных из мира статистики и поэтому обладают некоторым пониманием того вклада, который статистика может привнести в науку о данных как прикладную дисциплину. В то же время мы хорошо осведомлены об ограничениях традиционного статистического образования: статистика как дисциплина насчитывает полтора столетия, и большинство учебников и курсов по статистике отягощены кинетикой и инерцией океанского лайнера.

В основе настоящей книги лежат две цели:

- ◆ представить в удобоваримой, пригодной для навигации и легкодоступной форме ключевые понятия из статистики, которые относятся к науке о данных;
- ◆ объяснить, какие понятия важны и полезны с точки зрения науки о данных, какие менее важны и почему.

Условные обозначения, принятые в книге

В книге используются следующие типографические условные обозначения:

- ◆ *курсив* указывает на новые термины;
- ◆ **полужирный шрифт** — URL-адреса, адреса электронной почты;
- ◆ моноширинный шрифт используется для листингов программ, а также внутри абзацев для ссылки на элементы программ, такие как переменные или имена функций, базы данных, типы данных, переменные среды, инструкции и ключевые слова.

Ключевые термины

Наука о данных — это сплав многочисленных дисциплин, включая статистику, информатику, информационные технологии и конкретные предметные области. В результате при упоминании конкретной идеи могут использоваться несколько разных терминов. Ключевые термины и их синонимы в данной книге будут выделяться в специальных врезках, таких как показанные ниже.



Данный элемент обозначает подсказку или совет.



Данный элемент обозначает общее замечание.



Данный элемент обозначает предупреждение или предостережение.

Использование примеров кода

Во всех случаях в этой книге даются примеры исходного кода сначала на языке R и потом на Python. Во избежание ненужных повторов мы обычно показываем только результат и графики, создаваемые кодом на R. Мы также пропускаем код, необходимый для загрузки требуемых пакетов и наборов данных. Вы можете найти полный код, а также наборы данных для скачивания по адресу

<https://github.com/gedeck/practical-statistics-for-data-scientists>.

Эта книга предназначена для того, чтобы помочь вам в выполнении вашей работы. В целом, если код примеров предлагается вместе с книгой, вы можете использовать его в своих программах и документации. Вам не нужно связываться с нами с просьбой о разрешении, если вы не воспроизводите значительную часть кода. Например, написание программы, которая использует несколько фрагментов кода из данной книги, официального разрешения не требует. Ответ на вопрос путем цитирования этой книги и цитирования кода примера разрешения не требует. Включение значительного количества примеров кода из этой книги в документацию вашего продукта действительно требует отдельной миссии.

Адаптированный вариант примеров в виде электронного архива вы можете скачать по ссылке <ftp://ftp.bhv.ru/9785977567053.zip>. Эта ссылка доступна также со страницы книги на сайте www.bhv.ru.

Благодарности

Авторы признают усилия многих людей, которые помогли сделать эту книгу реальностью.

Герхард Пилчер (Gerhard Pilcher), генеральный директор компании по добыче регулярностей из данных Elder Research, ознакомился с ранними черновиками книги и дал нам подробные и полезные исправления и комментарии. Помимо этого, Аня

МакГирк (Аnya McGuirk) и Вэй Сяо (Wei Xiao), статистики из компании SAS, а также Джей Хилфигер (Jay Hilfiger), автор издательства O'Reilly, предоставили полезные отзывы о первоначальных набросках книги. Тосиаки Курокава (Toshiaki Kurokawa), который перевел первое издание на японский язык, проделал всестороннюю работу по пересмотру и исправлению этого процесса. Аарон Шумахер (Aaron Schumacher) и Вальтер Пачковский (Walter Paczkowski) тщательно рассмотрели второе издание книги и предоставили целый ряд полезных и ценных предложений, за которые мы им чрезвычайно признательны. Излишне говорить, что все оставшиеся ошибки принадлежат только нам.

В издательстве O'Reilly Шеннон Катт (Shannon Cutt) в дружеской атмосфере сопровождал нас в течение всего процесса публикации и в меру подстегивал нашу работу, в то время как Кристен Браун (Kristen Brown) плавно провела нашу книгу через стадию производства. Рэйчел Монаган (Rachel Monaghan) и Элиаху Сассман (Eliahu Sussman) старательно и терпеливо исправляли и улучшали наш почерк, а Эллен Траутман-Зайг (Ellen Troutman-Zaig) готовила предметный указатель. Николь Таш (Nicole Tache) взяла на себя бразды правления вторым изданием и одновременно эффективно руководила этим процессом, а также предоставила ряд хороших редакторских предложений для улучшения читаемости книги для широкой аудитории. Мы также благодарим Мари Бо-Гуро (Marie Beaugureau), которая инициировала наш проект в O'Reilly, а также Бена Бенгфорта (Ben Bengfort), автора O'Reilly и инструктора в Statistics.com, который познакомил нас с O'Reilly.

Мы, как и эта книга, также извлекли пользу из многочисленных бесед Питера с Галитом Шмуэли (Galit Shmueli), соавтором других книжных проектов.

Наконец, мы хотели бы особо поблагодарить Элизабет Брюс (Elizabeth Bruce) и Дебору Доннелл (Deborah Donnell), чьи терпение и поддержка сделали это начинание возможным.

Комментарии переводчика

В центре внимания машинного обучения и его подобласти, глубокого обучения, находится автоматически обучающаяся система, т. е. система, способная с течением времени приобретать новые знания и улучшать свою работу, используя поступающие данные. В зарубежной специализированной литературе *передача* знаний и *получение* знаний выражаются отдельными терминами: *teach/train* (*обучать/тренировать*), с одной стороны, и *learn* (*заучивать, усваивать, учиться*) — с другой.

В контексте строительства автоматически обучающихся моделей *тренировка* — это подача исследователем на вход обучающегося алгоритма нормализованных данных с целью получения обучившейся модели, которую можно применять к ранее не встречавшимся данным с целью предсказания или классифицирования. *Самообучение* же (*learning*) — это работа, которую выполняет обучающийся алгоритм по усвоению связей и регулярностей в данных или изменению и закреплению своего поведения.

В зарубежной литературе появление термина learning подразумевает исключительно вышесказанное — *самообучение*, т. е. самостоятельное *заучивание, усвоение* алгоритмом регулярностей или параметров. Иными словами, учитель как таковой отсутствует. Усвоение происходит под контролем (контролируемое самообучение, когда данные, на которых модель тренируется, были заранее помечены) и бесконтрольно (неконтролируемое самообучение, когда данные поступают непомеченными). Использование терминов "обучение с учителем" и "обучение без учителя" уходит в сторону от сути дела и затрудняет вразумительный перевод, скажем, таких разновидностей самообучения, как *semi-supervised learning* или *self-supervised learning*.

Опираясь на такое размежевание терминов, зонтичный термин *machine learning* (машинное обучение) следует всегда понимать как *усвоение знаний алгоритмической машиной*, а следовательно, более соответствовать оригиналу будет термин *машинное самообучение* или *автоматическое усвоение [регулярностей]*.

Весомым аргументом в пользу этих вариантов термина является и то, что с начала 1960-х и до середины 1980-х годов в ходу был похожий термин — "обучающиеся машины" (см. работы А. Тьюринга, К. Шеннона, Н. Винера, Н. Нильсона, Я. З. Цыпкина и др.).

Следуя принципам здравого смысла и бритвы Оккама, в настоящем переводе за основу принят зарубежный подход.

Разведывательный анализ данных

Эта глава посвящена первому шагу в любом проекте науки о данных — разведыванию данных.

Классическая статистика фокусировалась почти исключительно на *статистическом выводе* — иногда сложном наборе процедур для получения заключения о крупных популяциях, основываясь на малых выборках. В 1962 году Джон У. Тьюки (рис. 1.1) призвал к реформированию статистики в своей концептуальной работе "Будущее анализа данных" [Tukey-1962]¹. Он предложил новую научную дисциплину под названием "*анализ данных*", которая включала статистический вывод как всего лишь один из компонентов. Тьюки наладил связи с инженерным и вычислительным сообществами (он придумал термины *bit* от англ. *binary digit*, т. е. *бит*, и *software*, т. е. *программно-информационное обеспечение, вычислительная система*), и его первоначальные принципы оказались удивительно прочными и составляют часть фундамента науки о данных. Область разведывательного анализа данных была сформулирована во многом благодаря ставшей уже классической книге Тьюки 1977 года "Разведывательный анализ данных" [Tukey-1977]. Тьюки представил простые диаграммы (например, коробчатые диаграммы, диаграммы рассеяния), которые, наряду со сводной статистикой (среднее, медиана, квантили и др.), помогали рисовать картину набора данных.



Рис. 1.1. Джон Тьюки, выдающийся статистик, чьи идеи, разработанные более чем 50 лет назад, формируют фундамент науки о данных

¹ См. <https://oreil.ly/LQw6q>.

Благодаря доступности вычислительных мощностей и выразительному программному обеспечению для анализа данных разведывательный анализ данных эволюционировал далеко за пределы своей исходной области. Ключевыми факторами развития этой дисциплины стали быстрое развитие новых технологий, доступ ко все большему объему данных и более широкое использование количественного анализа в различных дисциплинах. Дэвид Донохо, профессор статистики Стэнфордского университета и бывший студент-старшекурсник Тьюки, написал превосходную статью "50 лет науки о данных", основанную на его выступлении на семинаре в честь 100-летия Тьюки, проходившем в Принстоне, шт. Нью-Джерси [Donoho-2015]. В ней Донохо прослеживает генезис науки о данных вплоть до новаторской работы Тьюки в области анализа данных.

Элементы структурированных данных

Данные поступают из многочисленных источников: показаний датчиков, событий, текста, фотоснимков и видео. *Интернет вещей* (Internet of Things, IoT) извергает потоки информации. Значительная часть этих данных не структурирована: фотоснимки представляют собой набор пикселей, при этом каждый пиксел содержит информацию о цвете в формате RGB (красный, зеленый, синий). Тексты состоят из последовательностей словарных и несловарных символов, часто разбитых на разделы, подразделы и т. д. Потоки нажатий клавиш представляют собой последовательности действий пользователя, взаимодействующего с приложением или веб-страницей. По сути дела, главенствующая задача науки о данных состоит в том, чтобы перерабатывать этот поток сырых данных в информацию, имеющую практическое значение. Для того чтобы применить охваченные в этой книге статистические понятия, неструктурированные сырые данные нужно переработать в структурированную форму. Одной из наиболее часто встречающихся форм структурированных данных является таблица со строками и столбцами, т. к. данные могут поступать из реляционной базы данных либо собираться для исследования.

Существуют два базовых типа структурированных данных: числовой и категориальный. Числовые данные имеют две формы: *непрерывную*, как, например, скорость ветра или продолжительность времени, и *дискретную*, как, например, количество возникновений события. *Категориальные* данные принимают только фиксированный набор значений, как, например, тип экрана телевизора (плазма, LCD, LED и т. д.) или название штата (Алабама, Аляска и т. д.). *Двоичные* данные являются важным особым случаем категориальных данных. Эти данные принимают только одно из двух значений, таких как 0/1, да/нет или истина/ложь. Еще один полезный тип категориальных данных — *порядковые* данные, в которых категории упорядочены; их примером является числовой рейтинг (1, 2, 3, 4 или 5).

Зачем нам вообще вникать в таксономию типов данных? Оказывается, что для целей анализа данных и предсказательного моделирования тип данных играет важную роль, помогая определять тип визуального изображения, анализа данных либо статистической модели. По сути дела, в вычислительных системах науки о данных, таких как *R* и *Python*, эти типы данных используются для улучшения вычислитель-

ный производительности. Еще важнее, что тип переменной определяет то, каким образом вычислительная система будет трактовать вычисления для этой переменной.

Ключевые идеи для типов данных

Числовые (numeric)

Данные, которые выражаются на числовой шкале.

Непрерывные (continuous)

Данные, которые могут принимать любое значение в интервале.

Синонимы: интервал, число с плавающей точкой, числовая величина.

Дискретные (discrete)

Данные, которые могут принимать только целочисленные значения, такие как количества.

Синонимы: целое число, количество, счетная величина.

Категориальные (categorical)

Данные, которые могут принимать только конкретное множество значений, представляющих множество возможных категорий.

Синонимы: перечисления, пронумерованные и номинальные данные, факторы.

Двоичные (binary)

Частный случай категориальных данных с двумя категориями значений, например, 0/1, истина/ложь.

Синонимы: дихотомическая, логическая, индикаторная, булева величина.

Порядковые (ordinal)

Категориальные данные, которые имеют явно заданное упорядочение.

Синоним: упорядоченный фактор.

Разработчики вычислительных систем и программисты баз данных могут задаться вопросом: а зачем нам в аналитике вообще нужно разделение на *категориальные* и *порядковые* данные? В конце концов, категории являются просто коллекцией текстовых (либо числовых) значений, и опорная база данных автоматически работает с их внутренним представлением. Однако четкая идентификация данных в качестве категориальных, в отличие от текстовых, действительно предлагает ряд преимуществ.

- ◆ Знание о том, что данные являются категориальными, может служить сигналом для вычислительной системы о том, каким образом должны вести себя статистические процедуры, такие как продуцирование графика или подгонка модели. В частности, в R порядковые данные могут быть представлены как упорядоченный фактор `ordered.factor`, сохраняя заданную пользователем упорядоченность

в графиках, таблицах и моделях. В Python пакет `scikit-learn` поддерживает порядковые данные с помощью класса `sklearn.preprocessing.OrdinalEncoder`.

- ◆ Хранение и индексация данных могут быть оптимизированы (как в реляционной базе данных).
- ◆ Возможные значения, принимаемые конкретной категориальной переменной, поддерживаются в вычислительной системе (как, например, структура данных `enum`).

Третья "выгода" может привести к непреднамеренному или неожиданному поведению: дефолтное² поведение функций импорта данных в R (например, `read.csv`) состоит в автоматическом конвертировании текстового столбца в `factor`. Последующие операции на этом столбце будут исходить из предположения о том, что единственно допустимыми значениями для этого столбца являются те, которые были первоначально импортированы, и присвоение нового текстового значения выдаст предупреждение и сообщение `NA (not available)` об отсутствии значения. Пакет `pandas` в Python не будет выполнять такую конвертацию автоматически. Однако вы можете явно указать столбец как категориальный в функции `read_csv`.

Ключевые идеи для структурированных данных

- В вычислительной системе данные, как правило, классифицируются по типу.
- Типы данных включают числовые (непрерывные, дискретные) и категориальные (двоичные, порядковые).
- Ввод данных в вычислительной системе действует как сигнал для вычислительной системы о том, как обрабатывать данные.

Дополнительные материалы для чтения

- ◆ Документация пакета `pandas`³ описывает разные типы данных и то, как ими можно манипулировать на языке Python.
- ◆ Типы данных могут вызывать путаницу, поскольку типы могут накладываться друг на друга, а таксономия в одной вычислительной системе может отличаться от таксономии в другой. Веб-сайт `R-tutorial`⁴ с практическими занятиями по языку R охватывает таксономию для R. Документация `pandas`⁵ описывает разные типы данных и то, как ими можно манипулировать в языке Python.

² То есть такое поведение, которое срабатывает автоматически, если не указано иное. Термин `default` в атрибутивной позиции встречается в зарубежной информатике слишком часто, чтобы его игнорировать. — *Прим. перев.*

³ См. <https://oreil.ly/UGX-4>.

⁴ См. <https://oreil.ly/2YUoA>.

⁵ См. <https://oreil.ly/UGX-4>.

- ◆ Базы данных более детализированы в том, как они классифицируют типы данных, инкорпорируя уровни прецизионности, поля фиксированной или переменной длины и многое другое; см. руководство W3Schools по SQL⁶.

Прямоугольные данные

В науке о данных типичным опорным каркасом для анализа является объект с *прямоугольными данными* наподобие электронной таблицы или таблицы базы данных.

Прямоугольные данные — это общий термин для двумерной матрицы, в которой строки обозначают записи (случай), а столбцы — признаки (переменные). *Кадр данных* — это специфичный формат, присущий языкам R и Python. Исходно данные не всегда поступают в такой форме: неструктурированные данные (например, текст) необходимо обработать и привести к такому виду, чтобы их можно было представить как множество признаков в прямоугольных данных (см. раздел "*Элементы структурированных данных*" ранее в этой главе). Данные в реляционных базах данных должны быть извлечены и помещены в одну-единственную таблицу для большинства задач по анализу данных и моделированию.

Ключевые термины для прямоугольных данных

Кадр данных (data frame)

Прямоугольные данные (подобно электронной таблице) — это базовая структура данных для статистических и автоматически обучающихся моделей.

Признак (feature)

Столбец в таблице обычно называется признаком.

Синонимы: атрибут, вход, предсказатель, предиктор, переменная.

Исход (outcome)

Многие проекты науки о данных предусматривают с предсказание исхода — нередко в формате да/нет (например, в табл. 1.1 это ответ на вопрос "Были ли торги состязательными или нет?"). Признаки иногда используются для предсказания исхода в эксперименте или статистическом исследовании.

Синонимы: результат, зависимая переменная, отклик, цель, выход.

Записи (records)

Строка в таблице обычно называется записью.

Синонимы: случай, пример, прецедент, экземпляр, наблюдение, шаблон, паттерн, образец.

⁶ См. <https://oreil.ly/cThTM>.

Таблица 1.1. Типичный формат данных

Категория	Валюта	Рейтинг продавца	Длительность	День закрытия	Цена закрытия	Цена открытия	Состязательность?
Music/Movie/Game	US	3249	5	Mon	0,01	0,01	0
Music/Movie/Game	US	3249	5	Mon	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	1
Automotive	US	3115	7	Tue	0,01	0,01	1

В табл. 1.1 показана смесь измеряемых или количественных данных (например, длительность и цена) и категориальных данных (например, категория и валюта). Как уже упоминалось ранее, специальной формой категориальной переменной является двоичная переменная (да/нет либо 0/1), которую можно увидеть в самом правом столбце табл. 1.1, — индикаторная переменная, показывающая, были ли торги состязательными (было несколько претендентов или нет). Эта индикаторная переменная также является переменной *результата*, когда сценарий заключается в том, чтобы предсказать состязательность или несостязательность аукциона.

Кадры данных и индексы

Традиционные таблицы базы данных имеют один или несколько столбцов, измеряемых *индексом*. Он значительно повышает эффективность некоторых запросов к базе данных. В Python при использовании библиотеки `pandas` основной прямоугольной структурой данных является объект `DataFrame`, содержащий таблицу данных. По умолчанию для объекта `DataFrame` создается автоматический целочисленный индекс, который основывается на порядке следования строк таблицы. В `pandas` также существует возможность задавать многоуровневые/иерархические индексы с целью повышения эффективности некоторых операций.

В R базовой прямоугольной структурой данных является объект `data.frame`, который тоже имеет неявный целочисленный индекс, основанный на порядке следования строк. Нативный для R объект `data.frame` не поддерживает определяемые пользователем либо многоуровневые индексы, хотя собственный ключ может быть создан посредством атрибута `row.names`⁷. Для преодоления этого недостатка широкое

⁷ Атрибут `row.names` — это символьный вектор длиной, которая соответствует числу строк в кадре данных, без дубликатов и пропущенных значений. — Прим. перев.

распространение получили два новых пакета: `data.table` и `dplyr`. Оба поддерживают многоуровневые индексы и обеспечивают значительное ускорение в работе с объектом `data.frame`.



Различия в терминологии

Терминология для прямоугольных данных может вызывать путаницу. В статистике и науке о данных используются разные термины, которые говорят об одном и том же. Для статистиков в модели существуют *предсказательные переменные*, которые применяются для предсказания *отклика* либо *зависимой переменной*. В отличие от них для исследователя данных существуют признаки, которые используются для предсказания цели. В особенности сбивает с толку один синоним: специалисты по информатике употребляют термин *sample* для обозначения одной-единственной строки, имея под ним в виду образец, тогда как для статистика *sample* означает коллекцию строк, т. е. выборку.

Непрямоугольные структуры данных

Помимо прямоугольных данных существуют и другие структуры данных.

Данные временного ряда записывают поочередные замеры одной и той же переменной. Эти данные представляют собой сырой материал для методов статистического прогнозирования, и они также являются ключевым компонентом данных, производимых устройствами — Интернет вещей.

Пространственные структуры данных, которые используются в картографической и геопространственной аналитике, более сложны и вариабельны, чем прямоугольные структуры данных. В их *объектном* представлении центральной частью данных являются объект (например, дом) и его пространственные координаты. В *полевой* проекции, в отличие от него, основное внимание уделяется малым единицам пространства и значению соответствующей метрики (яркости пиксела, например).

Графовые (или сетевые) структуры данных используются для представления физических, социальных и абстрактных связей. Например, граф социальной сети, такой как Facebook или LinkedIn, может представлять связи между людьми в сети. Соединенные дорогами центры распределения являются примером физической сети. Графовые структуры широко применяются в некоторых типах задач, таких как оптимизация сети и рекомендательные системы.

Каждый из этих типов данных имеет в науке о данных свою специализированную методологию. В центре внимания настоящей книги находятся прямоугольные данные — основополагающий структурный элемент в предсказательном моделировании.



Графы и графики в статистике

В англоязычной информатике и информационных технологиях термин *graph* (граф) обычно обозначает описание связей среди сущностей и опорную структуру данных. В статистике термин *graph* (график) чаще используется для обозначения самых разных диаграмм, графиков и визуализаций, а не только связей среди сущностей.

Ключевые идеи для прямоугольных данных

- В науке о данных базовой структурой данных является прямоугольная матрица, в которой строки являются записями, а столбцы — переменными (признаками).
- Терминология может вызывать путаницу, поскольку существуют разнообразные синонимы, вытекающие из разных дисциплин, которые вносят свой вклад в науку о данных (статистика, информатика и информационные технологии).

Дополнительные материалы для чтения

- ◆ Документация по кадрам данных в R⁸.
- ◆ Документация по кадрам (таблицам) данных в Python⁹.

Оценки центрального положения

Переменные с измеряемыми или количественными данными могут иметь тысячи четко различимых значений. Базовый шаг в разведывании данных состоит в получении "типичного значения" для каждого признака (переменной): оценки того, где расположено большинство данных (т. е. их центральной тенденции).

Ключевые термины оценок центрального положения

Среднее (mean)

Сумма всех значений, деленная на число значений.

Синонимы: среднее арифметическое.

Среднее взвешенное (weighted mean)

Сумма произведений всех значений на их веса, деленная на сумму весов.

Синонимы: среднее арифметическое взвешенное.

Медиана (median)

Такое значение, что половина сортированных данных находится выше и ниже данного значения.

Синоним: 50-й перцентиль.

Медиана взвешенная (weighted median)

Такое значение, что половина суммы весов находится выше и ниже сортированных данных.

⁸ См. <https://oreil.ly/NsONR>.

⁹ См. <https://oreil.ly/oxDKQ>.

Среднее усеченное (trimmed mean)

Среднее число всех значений после отбрасывания фиксированного числа предельных значений.

Синонимы: обрезанное среднее.

Робастный (robust)

Не чувствительный к предельным значениям.

Синоним: устойчивый.

Выброс (outlier)

Значение данных, которое сильно отличается от большинства данных.

Синонимы: предельное, экстремальное или аномальное значение.

На первый взгляд задача обобщения данных выглядит довольно тривиальной: надо просто взять *среднее арифметическое* данных (см. раздел "*Среднее*" далее в этой главе). На самом деле несмотря на то, что среднее вычисляется довольно просто и его выгодно использовать, оно не всегда бывает лучшей мерой центрального значения. По этой причине в статистике были разработаны и популяризированы несколько альтернативных оценок среднего значения.



Метрические и оценочные показатели

В статистике термин "*оценка*" часто используется для значений, вычисляемых из данных, которые находятся под рукой, для того чтобы провести различие между тем, что мы видим из этих данных, и теоретически истинным или точным положением дел. Исследователи данных и бизнес-аналитики с большей вероятностью будут называть такие значения метрическими показателями, или *метриками*. Эта разница отражает подходы, принятые в статистике, в отличие от науки о данных: учет неопределенности лежит в основе статистики, тогда как центром внимания науки о данных являются конкретные деловые или организационные целевые критерии. Следовательно, статистики оценивают, а исследователи данных измеряют.

Среднее

Самой базовой оценкой центрального положения является *среднее значение*, или *среднее арифметическое*. Среднее — это сумма всех значений, деленная на число значений. Рассмотрим следующий ряд чисел: $\{3, 5, 1, 2\}$. Среднее составит $(3 + 5 + 1 + 2) / 4 = 11 / 4 = 2,75$. Вы часто будете встречать символ \bar{x} (произносится "икс с чертой"), который обозначает среднее значение выборки из популяции, или генеральной совокупности. Формула среднего значения для ряда из n значений x_1, x_2, \dots, x_n такова:

$$\text{среднее} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}.$$



N (или n) обозначает общее число записей или наблюдений. В статистике это обозначение используется с заглавной буквы, если оно обозначает популяцию, и строчной, если оно обозначает выборку из популяции. В науке о данных это различие не является принципиальным, и поэтому можно увидеть и то и другое.

Разновидностью среднего является *среднее усеченное*, которое вычисляется путем отбрасывания фиксированного числа сортированных значений с каждого конца последовательности и затем взятия среднего арифметического оставшихся значений. Если представить сортированные значения как $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, где $x_{(1)}$ — это наименьшее значение и $x_{(n)}$ — наибольшее, то формула для вычисления усеченного среднего с пропуском p самых малых и самых крупных значений будет следующей:

$$\text{среднее усеченное} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}.$$

Усеченное среднее устраняет влияние предельных значений. Например, в международных состязаниях по прыжкам в воду верхние и нижние баллы пяти судей отбрасываются, и итоговым баллом считается среднеарифметический балл трех оставшихся судей¹⁰. Такой подход не дает одному судье манипулировать баллом, возможно, чтобы оказать содействие спортсмену из своей страны. Усеченные средние получили широкое распространение и во многих случаях являются предпочтительными вместо обычного среднего (продолжения обсуждения этой темы см. в разделе "*Медиана и робастные оценки*" далее в этой главе).

Еще один вид среднего значения — это *средневзвешенное значение*, которое вычисляется путем умножения каждого значения данных x_i на свой вес w_i и деления их суммы на сумму весов. Формула средневзвешенного выглядит так:

$$\text{среднее взвешенное} = \bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_i w_i}.$$

Существуют два главных побудительных мотива для использования средневзвешенного значения.

- ◆ Некоторые значения внутренне более переменчивы, чем другие, и сильно переменчивым наблюдениям придается более низкий вес. Например, если мы берем среднее арифметическое данных, поступающих от многочисленных датчиков, и один из датчиков менее точен, тогда вес данных от этого датчика можно понизить.
- ◆ Собранные данные не одинаково представляют разные группы, которые мы заинтересованы измерить. Например, в зависимости от того, каким образом про-

¹⁰ См. <https://oreil.ly/uV4P0>.

водится онлайн-эксперимент, у нас может не быть набора данных, который точно отражает все группы в пользовательской базе. Для того чтобы это исправить, можно придать более высокий вес значениям из тех групп, которые были представлены недостаточно.

Медиана и робастные оценки

Медиана — это число, расположенное в отсортированном списке данных ровно по середине. Если имеется четное число данных, то срединным значением является то, которое не находится в наборе данных фактически, а является средним арифметическим двух значений, которые делят отсортированные данные на верхнюю и нижнюю половины. По сравнению со средним, в котором используются абсолютно все наблюдения, медиана зависит только от значений в центре отсортированных данных. Хотя это может выглядеть как недостаток, поскольку среднее значение намного чувствительнее к данным, существует много примеров, в которых медиана является более подходящей метрикой центрального положения. Скажем, мы хотим взглянуть на типичные доходы домохозяйств в округах, расположенных на побережье озера Вашингтон в Сиэтле. При сравнении округа Медина с округом Уиндермир использование среднего значения дало бы совершенно разные результаты, потому что в Медине живет Билл Гейтс. Если же мы будем использовать медиану, то уже не будет иметь значения, насколько богатым является Билл Гейтс, — позиция срединного наблюдения останется той же.

По тем же самым причинам, по которым используется среднее взвешенное, можно вычислить и *медиану взвешенную*. Как и с медианой, мы сначала выполняем сортировку данных, несмотря на то что с каждым значением данных связан вес. В отличие от срединного числа медиана взвешенная — это такое значение, в котором сумма весов равна для нижней и верхней половин отсортированного списка. Как и медиана, взвешенная медиана робастна к выбросам.

Выбросы

Медиана называется *робастной* оценкой центрального положения, поскольку она не находится под влиянием *выбросов* (предельных случаев), которые могут исказить результаты. Выброс — это любое значение, которое сильно дистанцировано от других значений в наборе данных. Точное определение выброса является несколько субъективным, несмотря на то что в различных сводных данных и графиках используются некоторые правила (см. раздел "*Процентили и коробчатые диаграммы*" далее в этой главе). Выброс как таковой не делает значение данных недопустимым или ошибочным (как в предыдущем примере с Биллом Гейтсом). Вместе с тем выбросы часто являются результатом ошибок данных, таких как смешивание данных с разными единицами измерения (километры с метрами) или плохие показания от датчика. Когда выбросы являются результатом неправильных данных, среднее значение будет показывать плохую оценку центрального положения, тогда как медиана будет по-прежнему допустимой. В любом случае выбросы должны быть выявлены и обычно заслуживают дальнейшего расследования.



Обнаружение аномалий

В отличие от типичного анализа данных, где выбросы иногда информативны, а иногда — досадная помеха, в *обнаружении аномалий* целевыми объектами являются именно выбросы, и значительный массив данных преимущественно служит для определения "нормы", с которой соизмеряются аномалии.

Медиана не единственная робастная оценка центрального положения. На самом деле в целях предотвращения влияния выбросов широко используется и среднее усеченное. Например, усечение нижних и верхних 10% данных (общепринятый выбор) обеспечит защиту от выбросов во всех, кроме самых малых, наборах данных. Среднее усеченное может считаться компромиссом между медианой и средним: оно робастно к предельным значениям в данных, но использует больше данных для расчета оценки центрального положения.



Другие робастные метрики центрального положения

В статистике была разработана масса других оценщиков, так называемых эстиматоров, центрального положения преимущественно с целью разработки более робастных инструментов оценки, чем среднее, и более эффективных (т. е. способных лучше обнаруживать небольшие различия в центральном положении между наборами данных). Эти методы потенциально полезны для небольших наборов данных. Вместе с тем они едва дают дополнительные выгоды в условиях крупных или даже умеренно размерных наборов данных.

Пример: средние оценки численности населения и уровня убийств

В табл. 1.2 показаны первые несколько строк из набора данных, содержащего сведения о численности населения и уровне убийств (в единицах убийств на 100 тыс. человек в год) по каждому штату.

Таблица 1.2. Несколько строк кадра данных `data.frame` о численности населения и уровне убийств по штатам

№	Штат	Население	Уровень убийств	Аббревиатура
1	Alabama	4 779 736	5,7	AL
2	Alaska	710 231	5,6	AK
3	Arizona	6 392 017	4,7	AZ
4	Arkansas	2 915 918	5,6	AR
5	California	37 253 956	4,4	CA
6	Colorado	5 029 196	2,8	CO
7	Connecticut	3 574 097	2,4	CT
8	Delaware	897 934	5,8	DE