Practical Statistics for Data Scientists

50 Essential Concepts

Peter Bruce and Andrew Bruce

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Питер Брюс, Эндрю Брюс

Практическая статистика для специалистов Data Science

50 ВАЖНЕЙШИХ ПОНЯТИЙ

Санкт-Петербург
«БХВ-Петербург»
2018

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. — СПб.: БХВ-Петербург, 2018. — 304 с.: ил.

ISBN 978-5-9775-3974-6

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.

Для аналитиков данных

УДК 004.6+519.2
ББК 32.81+22.172

Группа подготовки издания:

Руководитель проекта	*Евгений Рыбаков*
Зав. редакцией	*Екатерина Капалыгина*
Компьютерная верстка	*Ольги Сергиенко*
Оформление обложки	*Марины Дамбиевой*

Подписано в печать 31.05.18.
Формат 70×100 1/16. Печать офсетная. Усл. печ. л. 24,51.
Тираж 1300 экз. Заказ №
"БХВ-Петербург", 191036, Санкт-Петербург, Гончарная ул., 20.

Отпечатано с готового оригинал-макета
ООО "Принт-М", 142300, М.О., г. Чехов, ул. Полиграфистов, д. 1

ISBN 978-1-491-95296-2 (англ.)
ISBN 978-5-9775-3974-6 (рус.)

Оглавление

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

Об авторах

Питер Брюс 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 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.

Предисловие

Книга рассчитана на аналитика данных, обладающего некоторым опытом работы с языком программирования R и имеющего предшествующий (возможно, обрывочный или сиюминутный) контакт с математической статистикой. Мы оба, авторы этой книги, пришли в мир науки о данных из мира статистики, и поэтому у нас есть определенное понимание того вклада, который статистика может привнести в науку о данных, как прикладную дисциплину. В то же время мы хорошо осведомлены об ограничениях традиционного статистического обучения: 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 удобной, 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.

Чего ожидать

> **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 специальной выноске, такой как эта.

Условные обозначения, принятые в книге

В книге используются следующие условные обозначения.

- *Курсив* указывает новые термины.
- **Полужирный шрифт** — URL-адреса, адреса электронной почты.
- `Моноширинный шрифт` используется для распечаток программ, а также внутри абзацев для ссылки на элементы программ, такие как переменные или имена функ-

ций, базы данных, типы данных, переменные окружения, операторы и ключевые слова.

- *Моноширинный шрифт курсивом* показывает текст, который должен быть заменен значениями пользователя либо значениями, определяемыми по контексту.

Данный элемент обозначает подсказку или совет.

Данный элемент обозначает общее замечание.

Данный элемент обозначает предупреждение или предостережение.

Использование примеров кода

Дополнительный материал (примеры кода, упражнения и пр.) доступен для скачивания по адресу **https://github.com/andrewgbruce/statistics-for-data-scientists**.

Эта книга предназначена для того, чтобы помочь вам решить ваши задачи. В целом, если код примеров предлагается вместе с книгой, то 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.

Адаптированный вариант примеров в виде электронного архива вы можете скачать по ссылке **ftp://ftp.bhv.ru/9785977539746.zip**, которая доступна также со страницы книги на сайте **www.bhv.ru**.

Благодарности

Авторы выражают признательность всем, кто помог воплотить эту книгу в реальность.

Герхард Пилхер (Gerhard Pilcher), генеральный директор консалтинговой фирмы Elder Research в области глубинного анализа данных, был свидетелем ранних черновиков книги и предоставил нам подробные и полезные поправки и комментарии. Так же как и Энья Макгверк (Anya McGuirk) и Вей Сяо (Wei Xiao), специалисты

в области статистики в SAS, и Джэй Хилфигер (Jay Hilfiger), член авторского коллектива O'Reilly, которые предоставили полезные рекомендации на первоначальных стадиях работы над книгой.

В издательстве O'Reilly Шэннон 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) гладко провела нашу книгу через производственную стадию. Рэйчел Монагэн (Rachel Monaghan) и Элайю Сассмэн (Eliahu Sussman) бережно и терпеливо проводили корректировку и улучшение нашего текста, тогда как Эллен Траутмэн-Зайг (Ellen Troutman-Zaig) подготовила предметный указатель. Мы также благодарим Мари Богуро (Marie Beaugureau), которая инициировала наш проект в O'Reilly, и Бена Бенгфорда (Ben Bengfort), автора O'Reilly и преподавателя в statistics.com, представившего нас издательству O'Reilly.

Мы извлекли большую пользу из многих бесед, которые Питер провел за последние годы с Галитом Шмуели (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.

Комментарий переводчика

Прилагаемый к настоящей книге программный код протестирован в среде Windows 10 с использованием действующих версий программных библиотек (время перевода книги — октябрь-ноябрь 2017 г.). При тестировании исходного кода за основу взята среда R версии 3.4.2.

Адаптированный и скорректированный исходный код примеров лучше всего разместить в подпапке домашней папки пользователя. Например:

/home/r_projects/statistics-for-data-scientists-master

или

C:\Users\[*ИМЯ_ПОЛЬЗОВАТЕЛЯ*]\r_projects\statistics-for-data-scientists-master

Структура папки с прилагаемыми примерами такова:

data	Наборы данных, используемые в книге и в сценариях R
figures	Графики, полученные в результате выполнения сценариев R
src	Исходный код примеров в виде сценариев R

Разведочный анализ данных

Статистика как дисциплина получила свое развитие главным образом в прошлом столетии. Теория вероятностей — математический фундамент статистики — разрабатывалась с XVII по XIX в. на основе работ Томаса Байеса (Thomas Bayes), Пьера-Симона Лапласа (Pierre-Simon Laplace) и Карла Гаусса (Carl Gauss). В отличие от чисто теоретической природы вероятности, статистика является прикладной наукой, занимающейся анализом и моделированием данных. Современная статистика как строгая научная дисциплина восходит корнями к концу 1800-х гг. — Фрэнсису Гальтону (Francis Galton) и Карлу Пирсону (Karl Pearson). Р. А. Фишер (R. A. Fisher) в начале XX в. был ведущим новатором современной статистики, который ввел в употребление такие ключевые понятия, как *планирование эксперимента* и *оценка максимального правдоподобия*. Эти и многие другие статистические понятия в основном находятся в отдаленных уголках науки о данных. Главная цель настоящей книги состоит в том, чтобы помочь высветить эти понятия и разъяснить их важность — или ее отсутствие — в контексте науки о данных и больших данных.

В данной главе основное внимание уделяется первому шагу в любом проекте науки о данных: разведке данных. *Разведочный анализ данных* (exploratory data analysis, EDA) — это сравнительно новая область статистики. Классическая статистика фокусировалась почти исключительно на *статистическом выводе*, т. е. иногда сложном наборе процедур для получения выводов о популяциях (или генеральных совокупностях) на основе небольших выборок. В 1962 г. Джон У. Тьюки (John W. Tukey, рис. 1.1) в своей концептуальной статье "Будущее анализа данных" [Tukey-1962] призвал к преобразованию статистики. Он предложил новую научную дисциплину под названием *анализ данных*, которая включала статистический вывод в качестве всего лишь одного из компонентов. Тьюки наладил связи с инженерным и вычислительным сообществами (он ввел термины *"бит"* — от англ. *binary digit*, и *"программное обеспечение"*), а его исходные принципы оказались удивительно прочными и формируют часть фундамента науки о данных. Область разведочного анализа данных появилась благодаря книге Тьюри *"Анализ результатов наблюдений"* [Tukey-1977], теперь уже ставшей классической.

Благодаря доступности вычислительных мощностей и выразительному программному обеспечению для анализа данных разведочный анализ данных эволюционировал далеко за пределы своей исходной области. Ключевыми факторами этой дисциплины явились быстрая разработка новой технологии, доступ к более разнообразным и бóльшим по объему данным и более широкое применение количественного анализа во множестве дисциплин. Дэвид Донохо (David Donoho), препода-

ватель статистики в Стэнфордском университете и прежний выпускник Тьюки, написал превосходную статью "50 лет науки о данных" на основе своей презентации на семинаре в честь 100-летия Тьюки, проходившем в Принстоне, шт. Нью-Джерси [Donoho-2015]. В ней Донохо прослеживает возникновение науки о данных к новаторскому вкладу Тьюки в анализ данных.

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

Элементы структурированных данных

Данные поступают из многих источников: показаний датчиков, событий, текста, изображений и видео. *Интернет вещей* (Internet of Things, IoT) извергает потоки информации. Значительная часть этих данных не структурирована: изображения представляют собой набор пикселов, 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, основная задача науки о данных состоит в том, чтобы переработать этот поток сырых данных в информацию, полезную в практической деятельности. Для применения рассмотренных в этой книге статистических понятий неструктурированные сырые данные должны быть обработаны и помещены в структурированную форму — подобно той, которая может появляться из реляционной базы данных — либо быть собранными для статистического исследования.

Ключевые термины

Непрерывные данные (continuous)

Данные, которые могут принимать любое значение в интервале.

Синонимы: интервал, число с плавающей точкой, числовое значение.

Дискретные данные (discrete)

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

Синонимы: целое число, количество.

Категориальные данные (categorical)

Данные, которые могут принимать только определенный набор значений, в частности набор возможных категорий.

Синонимы: перечисления, перечислимые данные, факторы, именованные данные, полихотомические данные.

Двоичные данные (binary)

Особый случай категориальных данных всего с двумя категориями значений (0/1, истина/ложь).

Синонимы: дихотомический, логический, флаг, индикатор, булево значение.

Порядковые данные (ordinal)

Категориальные данные с явно выраженной упорядоченностью.

Синонимы: порядковый фактор.

Существует два основных типа структурированных данных: числовой и категориальный. Числовые данные поступают в двух формах: *непрерывной*, как например скорость ветра или продолжительность времени, и *дискретной*, как например количество возникновений события. *Категориальные* данные принимают только фиксированный набор значений, например тип экрана телевизора (плазма, LCD, LED и т. д.) или название штата (Алабама, Аляска и т. д.). *Двоичные* данные представляют собой важный особый случай 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 представлен *порядковыми* данными, в которых категории упорядочены; их примером является числовая характеристика (1, 2, 3, 4 или 5).

Зачем заморачиваться таксономией типов данных? Оказывается, что в целях анализа данных и предсказательного моделирования тип данных играет важную роль для определения типа визуального отображения, анализа данных либо статистической модели. По сути дела, в программных системах для науки о данных, таких как R и Python, эти типы данных используются для улучшения вычислительный производительности. А более важно, что тип данных переменной определяет то, каким образом программная система будет обращаться с вычислениями для этой переменной.

У разработчиков программного обеспечения (ПО) и программистов баз данных (БД) может возникнуть вопрос: зачем в аналитике нужны понятия *"категориальные"* и *"порядковые"* данные? В конце концов, категории являются просто набором текстовых (либо числовых) значений, и основная БД автоматически работает с их внутренним представлением. Однако четкая идентификация данных как категориальных, в отличие от текстовых, действительно предлагает некоторые преимущества.

- Знание, что данные категориальные, может служить сигналом для программной системы о том, каким образом должны вести себя статистические процедуры, такие как создание графика или подгонка модели. В частности, в R и Python порядковые данные могут быть представлены как порядковый фактор `ordered.factor`, сохраняя определенную пользователем упорядоченность в графиках, таблицах и моделях.
- Могут быть оптимизированы хранение и индексация данных (как в реляционной базе данных).
- Возможные значения, которые принимает конкретная категориальная переменная, реализуются в ПО (как, например, перечисление `enum`).

Третье "преимущество" может привести к непреднамеренному или неожиданному поведению: по умолчанию функции импорта данных в R (например, read.csv) ведут себя таким образом, что автоматически преобразуют столбец текста в `factor`. Последующие операции на этом столбце будут исходить из предположения, что единственно допустимыми значениями для этого столбца являются значения, которые были импортированы первоначально, и присвоение нового текстового значения выдаст предупреждение и сообщение об отсутствии значения NA (not available).

Ключевые идеи для структурированных данных

- В программной системе данные обычно классифицируются по типу.
- Тип данных может быть непрерывным, дискретным, категориальным (который включает двоичный тип) и порядковым.
- Типизация данных в программной системе сигнализирует программной системе, каким образом обрабатывать данные.

Дополнительные материалы для чтения

- Типы данных могут вызывать путаницу, поскольку одни и те же данные можно отнести к разным типам, и таксономия в одной программной системе может отличаться от таксономии в другой. Веб-сайт R-tutorial знакомит с таксономией языка R (**http://www.r-tutor.com/r-introduction/basic-data-types**).
- В БД применяется более подробная классификация типов данных, включая учет уровней прецизионности, фиксированную либо переменную длину полей и мно-

гое другое (см. руководство по SQL W3Schools guide for SQL — **http://www.w3schools.com/sql/sql_datatypes_general.asp**).

Прямоугольные данные

В науке о данных типичной опорной конструкцией для анализа является объект с *прямоугольными данными* наподобие электронной таблицы или таблицы базы данных.

Ключевые термины

Кадр данных (data frame)

Прямоугольные данные (как в электронной таблице) — это типичная структура данных для статистических и машинно-обучаемых моделей.

Признак (feature)

Столбец в таблице принято называть *признаком.*

Синонимы: атрибут, вход, предиктор, переменная.

Исход (outcome)

Многие проекты науки о данных сопряжены с предсказанием *исхода* — нередко в формате да/нет (например, в табл. 1.1 это ответ на вопрос "были ли торги состязательными или нет?"). Для предсказания *исхода* в эксперименте или статистическом исследовании иногда используются *признаки*.

Синонимы: зависимая переменная, отклик, цель, выход.

Записи (records)

Строку в таблице принято называть *записью*.

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

Прямоугольные данные по существу представляют собой двумерную матрицу, в которой строки обозначают записи (случаи), а столбцы — признаки (переменные). Исходно данные поступают в такой форме не всегда: неструктурированные данные (например, текст) необходимо обработать и привести к такому виду, чтобы их можно было представить как набор признаков в прямоугольных данных *(см. разд. "Элементы структурированных данных" ранее в этой главе)*. Данные в реляционных БД должны быть извлечены и помещены в единственную таблицу для большинства аналитических и модельных задач.

В табл. 1.1 показана смесь измерительных и 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 или 0/1), которую можно увидеть в самом правом столбце табл. 1.1 — это индикаторная переменная, показывающая, были ли торги состязательными.

Таблица 1.1. *Типичный формат данных*

Категория	Валюта	Рейтинг продавца	Длительность	Дань закрытия	Цена закрытия	Цена открытия	Конкурентноспособность?
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

Кадры данных и индексы

Традиционные таблицы БД имеют один или несколько столбцов, называемых *индексом*. Он может значительно повысить эффективность определенных SQL-запросов. В Python при использовании библиотеки `pandas` основной прямоугольной структурой данных является объект `DataFrame`, содержащий таблицу данных. По умолчанию для `DataFrame` создается автоматический целочисленный индекс, который основывается на порядке следования строк. В программной библиотеке `pandas` для повышения эффективности определенных операций также можно задавать многоуровневые/иерархические индексы.

В R основной прямоугольной структурой данных является объект `data.frame`, кадр данных. Объект `data.frame` тоже имеет неявный целочисленный индекс на основе порядка следования строк. Хотя посредством атрибута `row.names`[1] можно создать пользовательский ключ, нативный (родной) для R, объект `data.frame` не поддерживает задаваемые пользователем или многоуровневые индексы. Для преодоления этого недостатка широкое распространение получили два новых программных пакета: `data.table` и `dplyr`. Оба пакета 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 предикторные переменные, которые используются для предсказания отклика либо зависимой переменной. В отличие от них для аналитика данных существуют признаки,

[1] Атрибут кадра данных `row.names` — это символьный вектор длиной, которая соответствует числу строк в кадре данных, без дубликатов и пропущенных значений. — *Прим. пер.*

которые применяются для предсказания целевой переменной. В особенности сбивает с толку один синоним: специалисты по информатике используют термин "выборка" для обозначения одиночной строки, тогда как для статистика выборка означает набор строк.

Непрямоугольные структуры данных

Помимо прямоугольных данных существуют и другие структуры данных.

Временно́й ряд содержит последовательные данные измерений одной и той же переменной. Эти данные представляют собой сырой материал для статистических методов предсказания, и они также являются ключевым компонентом данных, производимых устройствами — Интернет вещей.

Пространственные структуры данных, которые используются в картографической и геопространственной аналитике, более сложны и вариативны, 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 небольшим единицам пространства и значению соответствующего метрического показателя (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 определенных типах задач, таких как оптимизация сети и рекомендательные системы.

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

Графы и графики в статистике

В информатике и информационных технологиях термин "граф" (graph) обычно обозначает описание связей среди объектов и основную структуру данных. В статистике термин "график" (graph) используется для обозначения самых разных графиков и визуализаций, а не только связей между объектами, и этот термин применяется исключительно для обозначения визуализаций, а не структуры данных.

Ключевые идеи для прямоугольных данных

- В науке о данных базовой структурой данных является прямоугольная матрица, в которой строки — это записи, а столбцы — переменные (признаки).
- Терминология может вызывать путаницу, поскольку существует множество синонимов, вытекающих 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 R (**https://stat.ethz.ch/R-manual/R-devel/library/base/html/data.frame.html**).
- Документация по кадрам (таблицам) данных в Python (**http://pandas.pydata.org/pandas-docs/stable/dsintro.html#dataframe**).

Оценки центрального положения

Переменные с измерительными или количественными данными могут иметь тысячи различных значений. Основной этап исследования данных состоит в получении "типичного значения" для каждого признака (переменной): оценки того, где расположено большинство данных (т. е. их центральной тенденции).

Ключевые термины

Среднее (mean)

Сумма всех значений, деленная на количество значений.

Синоним: среднее арифметическое.

Среднее взвешенное (weighted mean)

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

Синоним: среднее арифметическое взвешенное.

Медиана (median)

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

Синоним: 50-й процентиль.

Медиана взвешенная (weighted median)

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

Среднее усеченное (trimmed mean)

Среднее число всех значений после отбрасывания фиксированного числа предельных значений.

Синоним: обрезанное среднее.

Робастный (robust)

Не чувствительный к предельным значениям.

Синоним: устойчивый.

Выброс (outlier)

Значение данных, которое сильно отличается от большинства данных.

Синоним: предельное значение.

На первый взгляд обобщить данные довольно тривиально: просто взять *среднее арифметическое* данных *(см. разд. "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 *"оценки"* часто используется для значений, вычисляемых из данных, которые находятся под рукой, чтобы отличить то, что мы видим, исходя из этих данных, от теоретически истинного или точного положения дел. Аналитики 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: учитывается неопределенность, которая лежит в основе статистики, тогда как центром внимания науки о данных являются конкретные деловые или организационные цели. Следовательно, статистики оценивают, а аналитики измеряют.

Среднее

Самой элементарной 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 "x с чертой"), который обозначает среднее значение выборки из популяции, или генеральной совокупности. Формула среднего значения для ряда из n значений $x_1, x_2, ..., x_n$ следующая:

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

N (или *n*) обозначает общее число записей или наблюдений. В статистике это обозначение используется с заглавной буквы, если оно относится к популяции, и строчной, если оно относится к выборке 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 $x_{(1)}, x_{(2)}, ..., x_{(n)}$, где $x_{(1)}$ — самое маленькое значение, а $x_{(n)}$ — самое большое, то формула для вычисления усеченного среднего с пропуском p самых малых и самых больших значений будет следующей:

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

Среднее усеченное устраняет влияние предельных значений. Например, в международных состязаниях по прыжкам в воду верхние и нижние баллы пяти судей отбрасываются, и итоговым баллом является среднеарифметический балл трех оставшихся судей [Wikipedia-2016]. 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 предпочтительны вместо обычного среднего *(см. разд. "Медиана и робастные оценки" далее в этой главе)*.

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

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

Существует два главных побудительных мотива для использования среднего взвешенного.

- Некоторые значения внутренне более переменчивы, чем другие, и сильно переменчивым наблюдениям придается более низкий вес. Например, если мы берем среднее 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 проводится онлайновый эксперимент, у нас может не быть набора данных, который точно отражает все группы в базе пользователей. 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 сортированном списке данных ровно посередине. Если число данных четное, срединным значением является то, которое не находится в наборе данных фактически, а является средним арифметическим двух значений, которые делят сортированные данные на верхнюю и нижнюю половины. По сравнению со средним, в котором используются абсолютно все наблюдения, медиана зависит только от значений в центре сортированных данных. Хотя это может выглядеть как недостаток, поскольку среднее значение намного более чувствительно к данным, существует много примеров, в которых медиана является

лучшим метрическим показателем центрального положения. Скажем, мы хотим взглянуть на типичные доходы домохозяйств в округах вокруг 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 станем 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 сортированного списка. Как и медиана, взвешенная медиана устойчива 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 удалено от других значений в наборе данных. Точное определение выброса несколько субъективно, несмотря на то, что в различных сводных данных и графиках используются определенные правила *(см. разд. "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). Однако выбросы часто являются результатом ошибок данных, таких как смешивание данных с разными единицами измерения (километры против метров) или плохие показания датчика. Когда выбросы являются результатом неправильных данных, среднее значение приводит к плохой оценке центрального положения, в то время как медиана будет по-прежнему допустимой. В любом случае выбросы должны быть идентифицированы и обычно заслуживают дальнейшего обследования.

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 соразмеряются аномалии.

Медиана не единственная робастная оценка центрального положения. На самом деле, для того чтобы предотвратить влияние 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 устойчиво к предельным значениям в данных, но использует больше данных для вычисления оценки центрального положения.

Другие робастные метрические показатели центрального положения

В статистике было разработало множество других инструментов оценки, так называемых оценщиков, или эстиматоров, центрального положения преимущественно с целью разработки более робастных инструментов оценки, чем среднее, и более *эффективных* (т. е. способных лучше обнаруживать небольшие разницы 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: оценки центрального положения численности населения и уровня убийств

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

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

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

Вычислим среднее, среднее усеченное и медиану численности населения, используя R:

```
> state <- read.csv(file="/Users/andrewbruce1/book/state.csv")
> mean(state[["Population"]])
[1] 6162876
> mean(state[["Population"]], trim=0.1)
[1] 4783697
> median(state[["Population"]])
[1] 4436370
```

Среднее больше среднего усеченного, которое больше медианы.

Это вызвано тем, что среднее усеченное исключает самые большие и самые малые пять штатов (`trim=0.1` отбрасывает по 10% с каждого конца). Если мы захотим вычислить среднестатистическое количество убийств в стране, то должны использовать среднее взвешенное или медиану, чтобы учесть разную численность населения в штатах. Поскольку базовый R не имеет функции для взвешенной медианы, то мы должны установить программный пакет, в частности `matrixStats`:

```
> weighted.mean(state[["Murder.Rate"]], w=state[["Population"]])
[1] 4.445834
> library("matrixStats")
> weightedMedian(state[["Murder.Rate"]], w=state[["Population"]])
[1] 4.4
```

В этом случае среднее взвешенное и медиана почти одинаковы.

Ключевые идеи для оценок центрального положения

- Основным метрическим показателем центрального положения является среднее, но оно может быть чувствительным к предельным значениям (выбросам).
- Другие метрические показатели (медиана, среднее усеченное) более робастны.

Дополнительные материалы для чтения

- Майкл Левин (Michael Levine, Университет Пердью) разместил несколько полезных слайдов, посвященных основным расчетам мер центрального положения (**http://www.stat.purdue.edu/~mlevins/STAT511_2012/Lecture2standard.pdf**).
- "Анализ результатов наблюдений" [Tukey-1977] — классическая книга Джона Тьюки, которая по-прежнему пользуется спросом.

Оценки вариабельности

Центральное положение — это всего одна из размерностей в обобщении признака. Вторая размерность, *вариабельность*, именуемая также *дисперсностью*, показывает, сгруппированы ли значения данных плотно, или же они разбросаны. В основе статистики лежит вариабельность: ее измерение, уменьшение, различение произвольной вариабельности от реальной, идентификация разных источников реальной вариабельности и принятие решений в условиях ее присутствия.

Ключевые термины

Отклонения (deviations)

Разница между наблюдаемыми значениями и оценкой центрального положения.

Синонимы: ошибки, остатки.

Дисперсия (variance)

Сумма квадратических отклонений от среднего, деленная на $n-1$, где n — число значений данных.

Синонимы: среднеквадратическое отклонение, среднеквадратическая ошибка.

Стандартное отклонение (standard deviation)

Квадратный корень из дисперсии.

Синонимы: норма l_2, евклидова норма.

Среднее абсолютное отклонение (mean absolute deviation)

Среднее абсолютных значений отклонений от среднего[2].

Синонимы: норма l_1, манхэттенская норма.

Медианное абсолютное отклонение от медианы (median absolute deviation from the median)

Медиана абсолютных значений отклонений от медианы.

Размах (range)

Разница между самым большим и самым малым значениями в наборе данных.

Порядковые статистики (order statistics)

Метрические показатели на основе значений данных, отсортированных от самых малых до самых больших.

Синоним: ранг.

Процентиль (percentile)

Такое значение, что P процентов значений принимает данное значение или меньшее и $(100-P)$ процентов значений принимает данное значение или большее.

Синоним: квантиль.

Межквартильный размах (interquartile range)

Разница между 75-м и 25-м процентилями.

Синонимы: MKP, IQR.

Так же как и в случае центрального положения, которое можно измерить разными способами (среднее, медиана и т. д.), существуют различные способы измерить вариабельность.

[2] Абсолютным оно является потому, что суммируются отклонения по модулю, т. к. в противном случае сумма всех разбросов будет равна нулю. — *Прим. пер.*

Стандартное отклонение и связанные с ним оценки

Наиболее широко используемые оценки вариабельности основаны на разницах, или *отклонениях*, между оценкой центрального положения и наблюдаемыми данными. Для набора данных {1, 4, 4}, среднее равняется 3, и медиана — 4. Отклонения от среднего представляют собой разницы: $1-3=-2$, $4-3=1$, $4-3=1$. Эти отклонения говорят о том, насколько данные разбросаны вокруг центрального значения.

Один из способов измерить вариабельность состоит в том, чтобы оценить типичное значение этих отклонений. Усреднение самих отклонений мало, поэтому отрицательные отклонения нейтрализуют положительные. Фактически сумма отклонений от среднего как раз равна нулю. Вместо этого простой подход заключается в том, чтобы взять среднее абсолютных значений отклонений от среднего значения. В предыдущем примере абсолютное значение отклонений равно {2, 1, 1}, а их среднее — $(2+1+1)/3=1{,}33$. Это и есть среднее абсолютное отклонение, которое вычисляется по следующей формуле:

$$\text{Среднее абсолютное отклонение} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n},$$

где $\bar{x}$ — среднее значение в выборке, или выборочное среднее.

Самыми известными оценками вариабельности являются *дисперсия* и *стандартное отклонение*, которые основаны на квадратических отклонениях. Дисперсия — это среднее квадратических отклонений, а стандартное отклонение — квадратный корень из дисперсии.

$$\text{Дисперсия} = s^2 = \frac{\sum (x-\bar{x})^2}{n-1};$$

$$\text{Стандартное отклонение} = s = \sqrt{\text{Дисперсия}}.$$

Стандартное отклонение интерпретируется намного проще, чем дисперсия, поскольку оно находится на той же шкале измерения, что и исходные данные. Однако, учитывая его более сложную и интуитивно менее понятную формулу, может показаться странным, что в статистике стандартному отклонению отдается предпочтение по сравнению со средним абсолютным отклонением. Такое преобладание обязано статистической теории: математически работа с квадратическими значениями намного более удобна, чем с абсолютными, в особенности со статистическими моделями.

Степени свободы и *n* или *n* – 1?

В книгах по статистике всегда так или иначе обсуждается вопрос, почему в формуле дисперсии у нас в знаменателе $n-1$, вместо n, который приводит к понятию *степеней свободы*. Это различие не является важным, поскольку n обычно настолько велико, что уже не имеет большого значения, будет ли деление выполняться на n или $n-1$. Однако в случае если вам интересно, то вот объяснение. Оно основывается на предпосылке, что вы хотите получить оценки популяции исходя из вынутой из нее выборки.

Если в формуле дисперсии применить интуитивно понятный знаменатель *n*, то истинное значение дисперсии и стандартного отклонения в популяции будет недооценено. Это называется *смещенной* оценкой. Однако если поделить на $n-1$ вместо *n*, то стандартное отклонение становится *несмещенной* оценкой.

Полное объяснение, почему использование *n* приводит к смещенной оценке, сопряжено с понятием степеней свободы, которое принимает во внимание число ограничений при вычислении оценки. В данном случае существуют $n-1$ степеней свободы, поскольку существует одно ограничение: стандартное отклонение зависит от вычисления среднего в выборке. В большинстве задач аналитикам данных не нужно беспокоиться по поводу степеней свободы, но в отдельных случаях это понятие имеет особое значение *(см. разд. "Выбор K" главы 6)*.

Ни дисперсия и стандартное отклонение, ни среднее абсолютное отклонение не устойчивы к выбросам и предельным значениям (*см. разд. "Медиана и робастные оценки" ранее в этой главе*, где обсуждаются робастные оценки центрального положения). Дисперсия и стандартное отклонение чувствительны к выбросам больше всего, поскольку они основаны на квадратических отклонениях.

Робастной оценкой вариабельности является *медианное абсолютное отклонение от медианы* (MAO — median absolute deviation, MAD):

$$\text{Медианное абсолютное отклонение} =$$
$$= \text{Медиана}\left(\left|x_1 - m\right|,\ \left|x_2 - m\right|,\ ...,\ \left|x_N - m\right|\right),$$

где *m* — это медиана. Как и в случае с медианой, MAO не находится под влиянием предельных значений. Можно также вычислить усеченное стандартное отклонение по аналогии со средним усеченным *(см. разд. "Среднее" ранее в этой главе)*.

Дисперсия, стандартное отклонение, среднее абсолютное отклонение и медианное абсолютное отклонение от медианы не являются эквивалентными оценками, даже в случае, когда данные поступают из нормального распределения. На деле стандартное отклонение всегда больше среднего абсолютного отклонения, которое в свою очередь больше медианного абсолютного отклонения. Иногда медианное абсолютное отклонение умножается на постоянный поправочный коэффициент (который часто сводится к 1,4826), чтобы в случае нормального распределения привести MAO к той же шкале измерения, что и стандартное отклонение.

Оценки на основе процентилей

Другой подход к оценке дисперсности основывается на рассмотрении разброса сортированных данных, или их спреда. Статистические показатели на основе сортированных (ранжированных) данных называются *порядковыми статистиками*. Элементарная мера — это *размах*, т. е. разница между самым большим и самым малым числом. Минимальные и максимальные значения как таковые полезно знать, поскольку они помогают идентифицировать выбросы, но размах чрезвычайно чувствителен к выбросам и не очень полезен в качестве общей меры дисперсности в данных.

Для того чтобы предотвратить чувствительность к выбросам, можно обратиться к размаху данных после отбрасывания значений с каждого конца. Эти типы оценок формально основываются на разницах между *процентилями.* В наборе данных *P*-й процентиль является таким значением, что, по крайней мере, *P* процентов значений принимает это значение или меньшее и, по крайней мере, (100 – *P*) процентов значений принимает это значение или большее. Например, для нахождения 80-го процентиля надо отсортировать данные. Затем, начиная с самого малого значения продолжить 80% вверх к самому большому значению. Отметим, что медиана — это то же самое, что и 50-й процентиль. Процентиль по существу аналогичен *квантилю*, при этом квантили индексируются долями (так, квантиль 0,8 — это то же самое, что и 80-й процентиль).

Общепринятой мерой вариабельности является разница между 25-м и 75-м процентилями, которая называется *межквартильным размахом* (interquartile range, IQR). Вот простой пример: 3, 1, 5, 3, 6, 7, 2, 9. Эти числа мы сортируем, получив 1, 2, 3, 3, 5, 6, 7, 9. 25-й процентиль находится в 2,5, и 75-й процентиль — в 6,5, поэтому межквартильный размах будет $6,5 - 2,5 = 4$. Программная система может иметь немного другие подходы, которые дают отличающиеся ответы (см. приведенное далее примечание); как правило, эти отличия небольшие.

Для очень больших наборов данных расчет точных процентилей может быть вычислительно очень затратным, поскольку он требует сортировки всех значений данных. В программных системах для машинного обучения и статистического анализа используются специальные алгоритмы, такие как [Zhang-Wang-2007], которые получают приблизительный процентиль, вычисляя его очень быстро и гарантированно обеспечивая определенную точность.

Процентиль: точное определение

Если имеется четное число данных (*n* — четное), то исходя из предыдущего определения процентиль неоднозначен. На деле можно взять любое значение между порядковыми статистиками $x_{(j)}$ и $x_{(j+1)}$, где *j* удовлетворяет:

$$100 \cdot \frac{j}{n} \le P \le 100 \cdot \frac{j+1}{n}.$$

В формальном плане процентиль — это средневзвешенное значение:

$$\text{Процентиль}\left(P\right)=\left(1-w\right)x_{(j)}+wx_{(j+1)}$$

для некоторого веса w между 0 и 1. В статистических программных системах содержатся слегка отличающиеся подходы к выбору значения w. На самом деле, R-функция `quantile` предлагает девять разных способов вычисления квантиля. За исключением небольших наборов данных, вам, как правило, не придется беспокоиться по поводу точного метода, которым вычисляется процентиль.

Пример: оценки вариабельности населения штатов

В табл. 1.3 (для удобства взятой повторно из ранее приведенной табл. 1.2) показаны первые несколько строк из набора данных, содержащего численность населения и уровни убийств для каждого штата.

Таблица 1.3. *Несколько строк из кадра* `data.frame` *с данными о численности населения и уровне убийств по каждому штату*

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

Используя встроенные функции R для стандартного отклонения (`sd`), межквартильного размаха (`IQR`) и медианного абсолютного отклонения из медианы (`mad`), можно вычислить оценки вариабельности данных о населении штатов:

```
> sd(state[["Population"]])
[1] 6848235
> IQR(state[["Population"]])
[1] 4847308
> mad(state[["Population"]])
[1] 3849870
```

Стандартное отклонение почти вдвое больше МАО (в R по умолчанию показатель МАО корректируется, чтобы быть на той же шкале измерения, что и среднее). И это не удивительно, поскольку стандартное отклонение чувствительно к выбросам.

Ключевые идеи для оценок вариабельности

- Дисперсия и стандартное отклонение — наиболее широко распространенные и в рутинном порядке регистрируемые статистики вариабельности.
- Оба показателя чувствительны к выбросам.
- Более робастные метрические показатели включают среднее абсолютное отклонение, медианное абсолютное отклонение от медианы и процентили (квантили).

Дополнительные материалы для чтения

- Онлайновый статистический ресурс Дэвида Лэйна (David Lane) содержит раздел по процентилям (**http://onlinestatbook.com/2/introduction/percentiles.html**).
- Кевин Дэйвенпорт (Kevin Davenport) на R-bloggers предлагает полезный пост, посвященный отклонениям от медианы и их робастным свойствам (**http://www.r-bloggers.com/absolute-deviation-around-the-median/**).

Обследование распределения данных

Все рассмотренные нами оценки обобщают данные в одном числе с целью описания центрального положения либо вариабельности данных. Помимо этого, также полезно обследовать характер распределенных данных в целом.

Ключевые термины

Коробчатая диаграмма (boxplot)

График, введенный в употребление Тьюки, в качестве быстрого способа визуализации распределения данных.

Синоним: диаграмма типа "ящик с усами".

Частотная таблица (frequency table)

Сводка количеств числовых значений, которые разбиты на серию частотных интервалов (корзин, бинов).

Гистограмма (histogram)

График частотной таблицы, где частотные интервалы откладываются на оси *x*, а количества (или доли) — на оси *y*.

График плотности (density plot)

Сглаженная версия гистограммы, часто на основе *ядерной оценки плотности*.

Процентили и коробчатые диаграммы

В *разд. "Оценки на основе процентилей" ранее в этой главе* мы рассмотрели, каким образом процентили могут использоваться для измерения разброса данных. Процентили также важны для обобщения всего распределения в целом. Общепринято сообщать о квартилях (25-й, 50-й и 75-й перцентили) и децилях (10-й, 20-й, ..., 90-й процентили). Процентили особенно важны для обобщения *хвостов* (внешнего размаха) распределения. Массовая культура ввела в обиход термин *"однопроцентовики"*, который относится к людям в верхнем 99-м процентиле богатства.

В табл. 1.4 показаны некоторые процентили уровня убийств по штатам. В R их можно получить при помощи функции `quantile`:

```
quantile(state[["Murder.Rate"]], p=c(.05, .25, .5, .75, .95))
   5%   25%   50%   75%   95%
1.600 2.425 4.000 5.550 6.510
```

Таблица 1.4. *Процентили уровня убийств по штатам*

5%	25%	50%	75%	95%
1,60	2,42	4,00	5,55	6,51

Медиана равна 4 убийствам на 100 тыс. человек, несмотря на то, что присутствует довольно большая вариабельность: 5-й процентиль составляет всего 1,6, тогда как 95-й процентиль — 6,51.

Коробчатые диаграммы, введенные в употребление Тьюки [Tukey-1977], основаны на процентилях и обеспечивают быстрый способ визуализации распределения данных. На рис. 1.2 представлена коробчатая диаграмма населения по штатам, полученная в R:

```
boxplot(state[["Population"]]/1000000, ylab="Население, млн человек")
```

Верх и низ коробки представляют собой соответственно 75-й и 25-й процентили. Медиана показана в коробке горизонтальной линией. Пунктирные линии, называемые *усами*, выходят из верха и низа и говорят о размахе основной части данных. Существует много вариантов коробчатой диаграммы; например, обратитесь к документации по R-функции `boxplot` [R-base-2015]. По умолчанию данная функция R простирает усы к самой далекой точке вне коробки, за исключением того, что она не выходит за пределы межквартильного размаха (МКР или IQR), умноженного на 1,5 (в других программных системах могут использоваться иные правила). Все данные за пределами усов отображаются как одиночные точки.