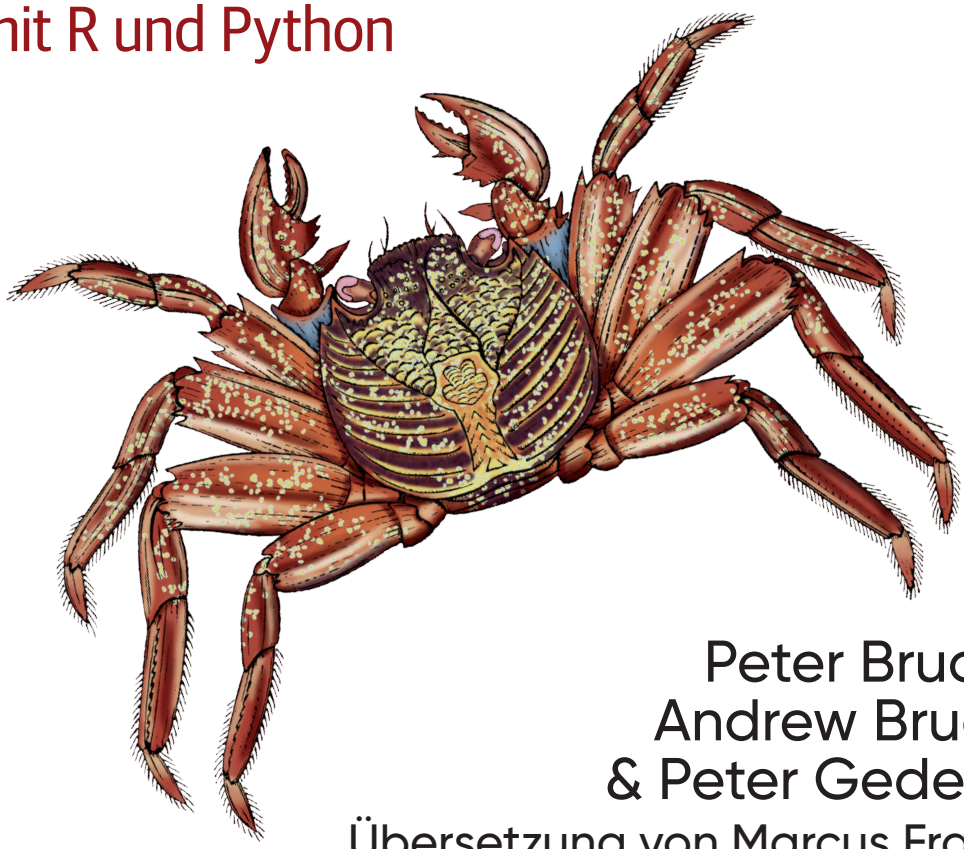
O'REILLY®

Übersetzung der 2. Auflage

Praktische Statistik

für Data Scientists

50+ essenzielle Konzepte mit R und Python



Peter Bruce,
Andrew Bruce
& Peter Gedeck

Übersetzung von Marcus Fraaß

Praktische Statistik für Data Scientists

50+ essenzielle Konzepte mit R und Python

Peter Bruce, Andrew Bruce & Peter Gedeck

Deutsche Übersetzung von Marcus Fraaß

O'REILLY®

Peter Bruce, Andrew Bruce, Peter Gedeck

Lektorat: Alexandra Follenius
Übersetzung: Marcus Fraaß
Korrektorat: Sibylle Feldmann, *www.richtiger-text.de*
Satz: III-satz, *www.drei-satz.de*
Herstellung: Stefanie Weidner
Umschlaggestaltung: Karen Montgomery, Michael Oréal, *www.oreal.de*
Druck und Bindung: mediaprint solutions GmbH, 33100 Paderborn

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über *http://dnb.d-nb.de* abrufbar.

ISBN:
Print 978-3-96009-153-0
PDF 978-3-96010-467-4
ePub 978-3-96010-468-1
mobi 978-3-96010-469-8

1. Auflage

Wieblinger Weg 17
69123 Heidelberg

Authorized German 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.

Dieses Buch erscheint in Kooperation mit O'Reilly Media, Inc. unter dem Imprint »O'REILLY«.
O'REILLY ist ein Markenzeichen und eine eingetragene Marke von O'Reilly Media, Inc. und wird mit Einwilligung des Eigentümers verwendet.

Hinweis:
Dieses Buch wurde auf PEFC-zertifiziertem Papier aus nachhaltiger Waldwirtschaft gedruckt. Der Umwelt zuliebe verzichten wir zusätzlich auf die Einschweißfolie.

PEFC zertifiziert
Das Papier für dieses Buch stammt aus nachhaltig bewirtschafteten Wäldern und kontrollierten Quellen.
PEFC™
PEFC/04-31-0810 www.pefc.de

Schreiben Sie uns:
Falls Sie Anregungen, Wünsche und Kommentare haben, lassen Sie es uns wissen: kommentar@oreilly.de.

5 4 3 2 1 0

Inhalt

Vorwort

Dieses Buch richtet sich an Data Scientists, die mit den Programmiersprachen *R* und/oder *Python* vertraut sind und sich bereits früher (wenn auch nur punktuell oder zeitweise) mit Statistik beschäftigt haben. Zwei der Autoren entstammen der Welt der Statistik, ehe sie sich in den weiten Raum der Data Science begeben haben, und schätzen den Beitrag, den die Statistik zur Datenwissenschaft zu leisten vermag, sehr. Gleichzeitig sind wir uns der Grenzen des traditionellen Statistikunterrichts durchaus bewusst: Statistik als Disziplin ist anderthalb Jahrhunderte alt, und die meisten Statistiklehrbücher und -kurse sind nicht gerade von Dynamik geprägt, sondern erinnern eher an die Trägheit eines Ozeanriesen. Alle Methoden in diesem Buch haben einen gewissen historischen oder methodologischen Bezug zur Disziplin der Statistik. Methoden, die sich hauptsächlich aus der Informatik entwickelt haben, wie z. B. neuronale Netze, werden nicht behandelt.

Diesem Buch liegen zwei Ziele zugrunde:

- Schlüsselbegriffe aus der Statistik, die für die Data Science relevant sind, in zugänglicher, übersichtlich gegliederter und leicht referenzierbarer Form darzulegen.
- Eine Erläuterung dazu zu geben, welche Konzepte aus datenwissenschaftlicher Sicht wichtig und nützlich sind, welche weniger wichtig sind und warum.

In diesem Buch verwendete Konventionen

Die folgenden typografischen Konventionen werden in diesem Buch verwendet:

Kursiv
: Kennzeichnet neue Begriffe, URLs, E-Mail-Adressen, Dateinamen und Dateiendungen.

`Konstante Zeichenbreite`
: Wird für Programmlistings und für Programmelemente in Textabschnitten wie Namen von Variablen und Funktionen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter verwendet.

`Konstante Zeichenbreite, fett`
: Kennzeichnet Befehle oder anderen Text, den der Nutzer wörtlich eingeben sollte.

> **Schlüsselbegriffe**
>
> Die Data Science baut auf mehreren Disziplinen auf, darunter Statistik, Informatik, Informationstechnologie und domänenspezifische Bereiche. Infolgedessen können mehrere unterschiedliche Begriffe verwendet werden, um auf ein bestimmtes Konzept zu verweisen. Schlüsselbegriffe und ihre Synonyme werden im gesamten Buch in einem Kasten wie diesem hervorgehoben.

Dieses Symbol steht für einen Tipp oder eine Empfehlung.

Dieses Symbol steht für einen allgemeinen Hinweis.

Dieses Symbol warnt oder mahnt zur Vorsicht.

Verwenden von Codebeispielen

Zu sämtlichen Beispielen zeigen wir in diesem Buch die entsprechenden Codebeispiele – zuerst immer in *R* und dann in *Python*. Um unnötige Wiederholungen zu vermeiden, zeigen wir im Allgemeinen nur Ausgaben und Diagramme, die durch den *R*-Code erzeugt wurden. Wir klammern auch den Code aus, der zum Laden der erforderlichen Pakete und Datensätze erforderlich ist. Den vollständigen Code sowie die Datensätze zum Herunterladen finden Sie unter *https://github.com/gedeck/practical-statistics-for-data-scientists*.

Dieses Buch dient dazu, Ihnen beim Erledigen Ihrer Arbeit zu helfen. Im Allgemeinen dürfen Sie die Codebeispiele aus diesem Buch in Ihren eigenen Programmen und der dazugehörigen Dokumentation verwenden. Sie müssen uns dazu nicht um Erlaubnis bitten, solange Sie nicht einen beträchtlichen Teil des Codes reproduzieren. Beispielsweise benötigen Sie keine Erlaubnis, um ein Programm zu schreiben, in dem mehrere Codefragmente aus diesem Buch vorkommen. Wollen Sie dagegen eine CD-ROM mit Beispielen aus Büchern von O'Reilly verkaufen oder verteilen,

brauchen Sie eine Erlaubnis. Eine Frage zu beantworten, indem Sie aus diesem Buch zitieren und ein Codebeispiel wiedergeben, benötigt keine Erlaubnis. Eine beträchtliche Menge Beispielcode aus diesem Buch in die Dokumentation Ihres Produkts aufzunehmen, bedarf hingegen unserer ausdrücklichen Zustimmung.

Wir freuen uns über Zitate, verlangen diese aber nicht. Ein Zitat enthält Titel, Autor, Verlag und ISBN, zum Beispiel: »*Praktische Statistik für Data Scientists* von Peter Bruce, Andrew Bruce und Peter Gedeck (O'Reilly). Copyright 2020 Peter Bruce, Andrew Bruce und Peter Gedeck, ISBN 978-3-96009-153-0.«

Wenn Sie glauben, dass Ihre Verwendung von Codebeispielen über die übliche Nutzung hinausgeht oder außerhalb der oben vorgestellten Nutzungsbedingungen liegt, kontaktieren Sie uns bitte unter *kontakt@oreilly.de*.

Danksagungen

Die Autoren danken den zahlreichen Menschen, die dazu beigetragen haben, dieses Buch Wirklichkeit werden zu lassen.

Gerhard Pilcher, CEO des Data-Mining-Unternehmens Elder Research, sah frühe Entwürfe des Buchs und half uns mit detaillierten und hilfreichen Korrekturen sowie Kommentaren. Ebenso gaben Anya McGuirk und Wei Xiao, Statistiker bei SAS, und Jay Hilfiger, ebenfalls Autor von O'Reilly, hilfreiches Feedback zu den ersten Entwürfen des Buchs. Toshiaki Kurokawa, der die erste Auflage ins Japanische übersetzte, leistete dabei umfassende Überarbeitungs- und Korrekturarbeit. Aaron Schumacher und Walter Paczkowski haben die zweite Auflage des Buchs gründlich überarbeitet und zahlreiche hilfreiche und wertvolle Anregungen gegeben, für die wir sehr dankbar sind. Es versteht sich von selbst, dass alle noch verbleibenden Fehler allein auf uns zurückzuführen sind.

Bei O'Reilly begleitete uns Shannon Cutt mit guter Laune und der richtigen Portion Nachdruck durch den Publikationsprozess, während Kristen Brown unser Buch reibungslos durch den Produktionsprozess geführt hat. Rachel Monaghan und Eliahu Sussman korrigierten und verbesserten unser Buch mit Sorgfalt und Geduld, während Ellen Troutman-Zaig den Index erarbeitete. Nicole Tache übernahm das Lektorat der zweiten Auflage und hat den Prozess effektiv geleitet sowie viele gute redaktionelle Vorschläge gemacht, um die Lesbarkeit des Buchs für ein breites Publikum zu verbessern. Wir danken auch Marie Beaugureau, die unser Projekt bei O'Reilly initiiert hat, sowie Ben Bengfort, Autor von O'Reilly und Ausbilder bei Statistics.com, der uns O'Reilly vorgestellt hat.

Wir und dieses Buch haben auch von den vielen Gesprächen profitiert, die Peter im Laufe der Jahre mit Galit Shmueli, Mitautorin bei anderen Buchprojekten, geführt hat.

Schließlich möchten wir besonders Elizabeth Bruce und Deborah Donnell danken, deren Geduld und Unterstützung dieses Vorhaben möglich gemacht haben.

KAPITEL 1

Explorative Datenanalyse

Dieses Kapitel erläutert Ihnen den ersten Schritt in jedem datenwissenschaftlichen Projekt: die Datenexploration.

Die klassische Statistik konzentrierte sich fast ausschließlich auf die *Inferenz*, einen manchmal komplexen Satz von Verfahren, um aus kleinen Stichproben Rückschlüsse auf eine größere Grundgesamtheit zu ziehen. Im Jahr 1962 forderte John W. Tukey (*https://oreil.ly/LQw6q*) (siehe Abbildung 1-1) in seinem bahnbrechenden Aufsatz »The Future of Data Analysis« [Tukey-1962] eine Reform der Statistik. Er schlug eine neue wissenschaftliche Disziplin namens *Datenanalyse* vor, die die statistische Inferenz lediglich als eine Komponente enthielt. Tukey knüpfte Kontakte zu den Ingenieurs- und Informatikgemeinschaften (er prägte die Begriffe *Bit*, kurz für Binärziffer, und *Software*). Seine damaligen Ansätze haben bis heute überraschend Bestand und bilden einen Teil der Grundlagen der Data Science. Der Fachbereich der explorativen Datenanalyse wurde mit Tukeys im Jahr 1977 erschienenem und inzwischen als Klassiker geltendem Buch *Exploratory Data Analysis* [Tukey-1977] begründet. Tukey stellte darin einfache Diagramme (z.B. Box-Plots und Streudiagramme) vor, die in Kombination mit zusammenfassenden Statistiken (Mittelwert, Median, Quantile usw.) dabei helfen, ein Bild eines Datensatzes zu zeichnen.

Abbildung 1-1: John Tukey, der bedeutende Statistiker, dessen vor über 50 Jahren entwickelte Ideen die Grundlage der Data Science bilden

Mit der zunehmenden Verfügbarkeit von Rechenleistung und leistungsfähigen Datenanalyseprogrammen hat sich die explorative Datenanalyse weit über ihren ursprünglichen Rahmen hinaus weiterentwickelt. Die wichtigsten Triebkräfte dieser Disziplin waren die rasche Entwicklung neuer Technologien, der Zugang zu mehr und umfangreicheren Daten und der verstärkte Einsatz der quantitativen Analyse in einer Vielzahl von Disziplinen. David Donoho, Professor für Statistik an der Stanford University und ehemaliger Student Tukeys, verfasste einen ausgezeichneten Artikel auf der Grundlage seiner Präsentation auf dem Workshop zur Hundertjahrfeier von Tukey in Princeton, New Jersey [Donoho-2015]. Donoho führt die Entwicklung der Data Science auf Tukeys Pionierarbeit in der Datenanalyse zurück.

Strukturierte Datentypen

Es gibt zahlreiche unterschiedliche Datenquellen: Sensormessungen, Ereignisse, Text, Bilder und Videos. Das *Internet der Dinge* (engl. *Internet of Things* (IoT)) produziert ständig neue Informationsfluten. Ein Großteil dieser Daten liegt unstrukturiert vor: Bilder sind nichts anderes als eine Zusammenstellung von Pixeln, wobei jedes Pixel RGB-Farbinformationen (Rot, Grün, Blau) enthält. Texte sind Folgen von Wörtern und Nicht-Wortzeichen, die oft in Abschnitte, Unterabschnitte usw. gegliedert sind. Clickstreams sind Handlungsverläufe eines Nutzers, der mit einer Anwendung oder einer Webseite interagiert. Tatsächlich besteht eine große Herausforderung der Datenwissenschaft darin, diese Flut von Rohdaten in verwertbare Informationen zu überführen. Um die in diesem Buch behandelten statistischen Konzepte in Anwendung zu bringen, müssen unstrukturierte Rohdaten zunächst aufbereitet und in eine strukturierte Form überführt werden. Eine der am häufigsten vorkommenden Formen strukturierter Daten ist eine Tabelle mit Zeilen und Spalten – so wie Daten aus einer relationalen Datenbank oder Daten, die für eine Studie erhoben wurden.

Es gibt zwei grundlegende Arten strukturierter Daten: numerische und kategoriale Daten. Numerische Daten treten in zwei Formen auf: *kontinuierlich*, wie z.B. die Windgeschwindigkeit oder die zeitliche Dauer, und *diskret*, wie z.B. die Häufigkeit des Auftretens eines Ereignisses. *Kategoriale* Daten nehmen nur einen bestimmten Satz von Werten an, wie z.B. einen TV-Bildschirmtyp (Plasma, LCD, LED usw.) oder den Namen eines Bundesstaats (Alabama, Alaska usw.). *Binäre* Daten sind ein wichtiger Spezialfall kategorialer Daten, die nur einen von zwei möglichen Werten annehmen, wie z.B. 0 oder 1, ja oder nein oder auch wahr oder falsch. Ein weiterer nützlicher kategorialer Datentyp sind *ordinalskalierte* Daten, bei denen die Kategorien in einer Reihenfolge geordnet sind; ein Beispiel hierfür ist eine numerische Bewertung (1, 2, 3, 4 oder 5).

Warum plagen wir uns mit der Taxonomie der Datentypen herum? Es stellt sich heraus, dass für die Zwecke der Datenanalyse und der prädiktiven Modellierung der Datentyp wichtig ist, um die Art der visuellen Darstellung, der Datenanalyse

oder des statistischen Modells zu bestimmen. Tatsächlich verwenden datenwissenschaftliche Softwareprogramme wie *R* und *Python* diese Datentypen, um die Rechenleistung zu optimieren. Noch wichtiger ist es, dass der Datentyp einer Variablen ausschlaggebend dafür ist, wie das Programm die Berechnungen für diese Variable handhabt.

Schlüsselbegriffe zu Datentypen

Numerisch
: Daten, die auf einer numerischen Skala abgebildet sind.

Kontinuierlich
: Daten, die innerhalb eines Intervalls einen beliebigen Wert annehmen können.

Synonyme
: intervallskaliert, Gleitkommazahl, numerisch

Diskret
: Daten, die nur ganzzahlige Werte annehmen können, wie z. B. Häufigkeiten bzw. Zählungen.

Synonyme
: Ganzzahl, Zählwert

Kategorial
: Daten, die nur einen bestimmten Satz von Werten annehmen können, die wiederum einen Satz von möglichen Kategorien repräsentieren.

Synonyme
: Aufzählungstyp, Faktor, faktoriell, nominal

Binär
: Ein Spezialfall des kategorialen Datentyps mit nur zwei möglichen Ausprägungen, z. B. 0/1, wahr/falsch.

Synonyme
: dichotom, logisch, Indikatorvariable, boolesche Variable

Ordinalskaliert
: Kategoriale Daten, die eine eindeutige Reihenfolge bzw. Rangordnung haben.

Synonym
: geordneter Faktor

Softwareingenieure und Datenbankprogrammierer fragen sich vielleicht, warum wir überhaupt den Begriff der *kategorialen* und *ordinalskalierten* Daten für unsere Analyse benötigen. Schließlich sind Kategorien lediglich eine Sammlung von Text- (oder numerischen) Werten, und die zugrunde liegende Datenbank übernimmt au-

tomatisch die interne Darstellung. Die explizite Bestimmung von Daten als kategoriale Daten im Vergleich zu Textdaten bietet jedoch einige Vorteile:

- Die Kenntnis, dass Daten kategorial sind, kann als Signal dienen, durch das ein Softwareprogramm erkennen kann, wie sich statistische Verfahren wie die Erstellung eines Diagramms oder die Anpassung eines Modells verhalten sollen. Insbesondere ordinalskalierte Daten können als `ordered.factor` in *R* angegeben werden, wodurch eine benutzerdefinierte Ordnung in Diagrammen, Tabellen und Modellen erhalten bleibt. In *Python* unterstützt `scikit-learn` ordinalskalierte Daten mit der Methode `sklearn.preprocessing.OrdinalEncoder`.
- Das Speichern und Indizieren kann optimiert werden (wie in einer relationalen Datenbank).
- Die möglichen Werte, die eine gegebene kategoriale Variable annehmen kann, werden in dem Softwareprogramm erzwungen (wie bei einer Aufzählung).

Der dritte »Vorteil« kann zu unbeabsichtigtem bzw. unerwartetem Verhalten führen: Das Standardverhalten von Datenimportfunktionen in *R* (z.B. `read.csv`) besteht darin, eine Textspalte automatisch in einen `factor` umzuwandeln. Bei nachfolgenden Operationen auf dieser Spalte wird davon ausgegangen, dass die einzigen zulässigen Werte für diese Spalte die ursprünglich importierten sind und die Zuweisung eines neuen Textwerts eine Warnung verursacht sowie einen Eintrag mit dem Wert `NA` (ein fehlender Wert) erzeugt. Das `pandas`-Paket in *Python* nimmt diese Umwandlung nicht automatisch vor. Sie können jedoch in der Funktion `read_csv` eine Spalte explizit als kategorial spezifizieren.

Kernideen

- Daten werden in Softwareprogrammen typischerweise in verschiedene Typen eingeteilt.
- Zu den Datentypen gehören numerische (kontinuierlich, diskret) und kategoriale (binär, ordinalskaliert).
- Die Datentypisierung dient als Signal für das Softwareprogramm, wie die Daten zu verarbeiten sind.

Weiterführende Literatur

- Datentypen können verwirrend sein, da sich Typen überschneiden und die Taxonomie in einem Softwareprogramm von der in einem anderen abweichen kann. Auf der R-Tutorial-Webseite (*https://oreil.ly/2YUoA*) können Sie die Taxonomie in *R* nachvollziehen. Die `pandas`-Dokumentation (*https://oreil.ly/UGX-4*) beschreibt die verschiedenen Datentypen in *Python* und wie sie verändert werden können.

- Datenbanken sind in ihrer Einteilung der Datentypen detaillierter und berücksichtigen Präzisionsniveaus, Datenfelder fester oder variabler Länge und mehr (siehe den W3Schools-SQL-Leitfaden (*https://oreil.ly/cThTM*).)

Tabellarische Daten

Der typische Bezugsrahmen für eine Analyse in der Data Science ist ein *tabellarisches Datenobjekt* (engl. *Rectangular Data Object*), wie eine Tabellenkalkulation oder eine Datenbanktabelle.

»Tabellarische Daten« ist der allgemeine Begriff für eine zweidimensionale Matrix mit Zeilen für die Beobachtungen (Fälle) und Spalten für die Merkmale (Variablen); in *R* und *Python* wird dies als *Data Frame* bezeichnet. Die Daten sind zu Beginn nicht immer in dieser Form vorhanden: Unstrukturierte Daten (z.B. Text) müssen zunächst so verarbeitet und aufbereitet werden, dass sie als eine Reihe von Merkmalen in tabellarischer Struktur dargestellt werden können (siehe »Strukturierte Datentypen« auf Seite 2). Daten in relationalen Datenbanken müssen für die meisten Datenanalyse- und Modellierungsaufgaben extrahiert und in eine einzelne Tabelle überführt werden.

Schlüsselbegriffe zu tabellarischen Daten

Data Frame
: Tabellarische Daten (wie ein Tabellenkalkulationsblatt) sind die grundlegende Datenstruktur für statistische und maschinelle Lernmodelle.

Merkmal
: Eine Spalte innerhalb einer Tabelle wird allgemein als *Merkmal* (engl. *Feature*) bezeichnet.

Synonyme
: Attribut, Eingabe, Prädiktorvariable, Prädiktor, unabhängige Variable

Ergebnis
: Viele datenwissenschaftliche Projekte zielen auf die Vorhersage eines *Ergebnisses* (engl. *Outcome*) ab – oft in Form eines Ja-oder-Nein-Ergebnisses (ob beispielsweise in Tabelle 1-1 eine »Auktion umkämpft war oder nicht«). Die *Merkmale* werden manchmal verwendet, um das *Ergebnis* eines statistischen Versuchs oder einer Studie vorherzusagen..

Synonyme
: Ergebnisvariable, abhängige Variable, Antwortvariable, Zielgröße, Ausgabe, Responsevariable

Eintrag
: Eine Zeile innerhalb einer Tabelle wird allgemein als *Eintrag* (engl. *Record*) bezeichnet.

Synonyme
: Fall, Beispiel, Instanz, Beobachtung

Tabelle 1-1: Ein typisches Data-Frame-Format

Kategorie	Währung	Verkäufer-Rating	Dauer	Schluss-tag	Schluss-preis	Eröffnungs-preis	umkämpft?
Musik/Film/Spiel	USD	3249	5	Mon	0.01	0.01	0
Musik/Film/Spiel	USD	3249	5	Mon	0.01	0.01	0
Automobil	USD	3115	7	Die	0.01	0.01	0
Automobil	USD	3115	7	Die	0.01	0.01	0
Automobil	USD	3115	7	Die	0.01	0.01	0
Automobil	USD	3115	7	Die	0.01	0.01	0
Automobil	USD	3115	7	Die	0.01	0.01	1
Automobil	USD	3115	7	Die	0.01	0.01	1

In Tabelle 1-1 gibt es eine Kombination aus Mess- oder Zähldaten (z. B. Dauer und Preis) und kategorialen Daten (z. B. Kategorie und Währung). Wie bereits erwähnt, ist eine besondere Form der kategorialen Variablen eine binäre Variable (ja/nein oder 0/1), wie in der Spalte ganz rechts in Tabelle 1-1 – eine Indikatorvariable, die angibt, ob eine Auktion umkämpft war (mehrere Bieter hatte) oder nicht. Diese Indikatorvariable ist zufällig auch eine *Ergebnisvariable*, wenn das Modell vorhersagen soll, ob eine Auktion umkämpft sein wird oder nicht.

Data Frames und Tabellen

Klassische Datenbanktabellen haben eine oder mehrere Spalten, die als Index bezeichnet werden und im Wesentlichen eine Zeilennummer darstellen. Dies kann die Effizienz bestimmter Datenbankabfragen erheblich verbessern. In *Pythons* `pandas`-Bibliothek wird die grundlegende tabellarische Datenstruktur durch ein Data-Frame-Objekt umgesetzt. Standardmäßig wird automatisch ein ganzzahliger Index für ein Data-Frame-Objekt basierend auf der Reihenfolge der Zeilen erstellt. In `pandas` ist es auch möglich, mehrstufige bzw. hierarchische Indizes festzulegen, um die Effizienz bestimmter Operationen zu verbessern.

In *R* ist die grundlegende tabellarische Datenstruktur mittels eines `data.frame`-Objekts implementiert. Ein `data.frame` hat auch einen impliziten ganzzahligen Index, der auf der Zeilenreihenfolge basiert. Der standardmäßige `data.frame` in *R* unterstützt keine benutzerdefinierten oder mehrstufigen Indizes. Jedoch kann über das Argument `row.names` ein benutzerdefinierter Schlüssel erstellt werden. Um diesem Problem zu begegnen, werden immer häufiger zwei neuere Pakete eingesetzt: `data.table` und `dplyr`. Beide unterstützen mehrstufige Indizes und bieten erhebliche Beschleunigungen bei der Arbeit mit einem `data.frame`.

Unterschiede in der Terminologie

Die Terminologie bei tabellarischen Daten kann verwirrend sein. Statistiker und Data Scientists verwenden oftmals unterschiedliche Begriffe für ein und denselben Sachverhalt. Statistiker nutzen in einem Modell *Prädiktorvariablen*, um eine *Antwortvariable* (engl. *Response*) oder eine *abhängige Variable* vorherzusagen. Ein Datenwissenschaftler spricht von *Merkmalen* (engl. *Features*), um eine *Zielgröße* (engl. *Target*) vorherzusagen. Ein Synonym ist besonders verwirrend: Informatiker verwenden den Begriff *Stichprobe* (engl. *Sample*) für eine einzelne Datenzeile, für einen Statistiker ist eine *Stichprobe* hingegen eine Sammlung von Datenzeilen.

Nicht tabellarische Datenstrukturen

Neben tabellarischen Daten gibt es noch andere Datenstrukturen.

Zeitreihendaten umfassen aufeinanderfolgende Messungen derselben Variablen. Sie sind das Rohmaterial für statistische Prognosemethoden und auch eine zentrale Komponente der von Geräten – dem Internet der Dinge – erzeugten Daten.

Räumliche Daten- bzw. Geodatenstrukturen, die bei der Kartierung und Standortanalyse verwendet werden, sind komplexer und vielfältiger als tabellarische Datenstrukturen. In der *Objektdarstellung* (engl. *Object Representation*) stehen ein Objekt (z.B. ein Haus) und seine räumlichen Koordinaten im Mittelpunkt der Daten. Die *Feldansicht* (engl. *Field View*) hingegen konzentriert sich auf kleine räumliche Einheiten und den Wert einer relevanten Metrik (z.B. Pixelhelligkeit).

Graphen- (oder Netzwerk-)Datenstrukturen werden verwendet, um physikalische, soziale oder abstrakte Beziehungen darzustellen. Beispielsweise kann ein Diagramm eines sozialen Netzwerks wie Facebook oder LinkedIn Verbindungen zwischen Menschen im Netzwerk darstellen. Ein Beispiel für ein physisches Netzwerk sind Vertriebszentren, die durch Straßen miteinander verbunden sind. Diagrammstrukturen sind für bestimmte Arten von Fragestellungen nützlich, wie z.B. bei der Netzwerkoptimierung und bei Empfehlungssystemen.

Jeder dieser Datentypen hat seine eigene spezifische Methodologie in der Data Science. Der Schwerpunkt dieses Buchs liegt auf tabellarische Daten, dem grundlegenden Baustein der prädiktiven Modellierung.

Graphen in der Statistik

In der Informatik und der Informationstechnologie bezieht sich der Begriff *Graph* typischerweise auf die Darstellung von Verbindungen zwischen Entitäten und auf die zugrunde liegende Datenstruktur. In der Statistik wird der Begriff *Graph* verwendet, um sich auf eine Vielzahl von Darstellungen und Visualisierungen zu beziehen, nicht nur von Verbindungen zwischen Entitäten. Zudem bezieht er sich ausschließlich auf die Visualisierung und nicht auf die Datenstruktur.

Kernideen

- Die grundlegende Datenstruktur in der Data Science ist eine rechteckige Matrix, in der die Zeilen den Beobachtungen entsprechen und die Spalten den Variablen (Merkmalen).
- Die Terminologie kann verwirrend sein; es gibt eine Vielzahl von Synonymen, die sich aus den verschiedenen Disziplinen ergeben, die zur Data Science beitragen (Statistik, Informatik und Informationstechnologie).

Weiterführende Literatur

- Dokumentation zu Data Frames in *R* (*https://oreil.ly/NsONR*)
- Dokumentation zu Data Frames in *Python* (*https://oreil.ly/oxDKQ*)

Lagemaße

Variablen für Mess- oder Zähldaten können Tausende von unterschiedlichen Werten haben. Ein grundlegender Schritt bei der Erkundung Ihrer Daten ist die Ermittlung eines »typischen Werts« für jedes Merkmal (Variable) – ein sogenanntes Lagemaß (engl. *Estimates of Location*): eine Schätzung darüber, wo sich die Mehrheit der Daten konzentriert (d.h. ihre zentrale Tendenz).

Schlüsselbegriffe zu Lagemaßen

Mittelwert
: Die Summe aller Werte dividiert durch die Anzahl der Werte.

Synonyme
: arithmetisches Mittel, Durchschnitt

Gewichteter Mittelwert
: Die Summe aller Werte, die jeweils mit einem Gewicht bzw. einem Gewichtungsfaktor multipliziert werden, geteilt durch die Summe aller Gewichte.

Synonym
: gewichteter Durchschnitt

Median
: Der Wert, bei dem die Hälfte der Daten oberhalb und die andere Hälfte unterhalb dieses Werts liegt.

Synonym
: 50%-Perzentil

Perzentil
: Der Wert, bei dem *P* % der Daten unterhalb dieses Werts liegen.

Synonym
: Quantil

Gewichteter Median
: Der Wert, bei dem die Summe der Gewichte der sortierten Daten exakt die Hälfte beträgt und der die Daten so einteilt, dass sie entweder oberhalb oder unterhalb diesen Werts liegen.

Getrimmter Mittelwert
: Der Mittelwert aller Werte, nachdem eine vorgegebene Anzahl von Ausreißern entfernt wurde.

Synonym
: gestutzter Mittelwert

Robust
: Nicht sensibel gegenüber Ausreißern.

Ausreißer
: Ein Datenwert, der sich stark von den übrigen Daten unterscheidet.

Synonym
: Extremwert

Auf den ersten Blick mag für Sie die Ermittlung einer zusammenfassenden Größe, die Aufschluss über einen vorliegenden Datensatz gibt, ziemlich trivial erscheinen: Sie nehmen einfach den *Mittelwert*, der sich für den Datensatz ergibt. Tatsächlich ist der Mittelwert zwar leicht zu berechnen und relativ zweckmäßig, aber er ist nicht immer das beste Maß zur Bestimmung eines Zentralwerts. Aus diesem Grund haben Statistiker mehrere alternative Schätzer zum Mittelwert entwickelt und befürwortet.

Metriken und Schätzwerte

Statistiker verwenden oft den Begriff *Schätzwert* für einen aus den vorliegenden Daten berechneten Wert, um zwischen dem, was wir aus den Daten ziehen, und der theoretisch wahren oder tatsächlichen Sachlage zu unterscheiden. Data Scientists und Geschäftsanalysten sprechen bei einem solchen Wert von einer *Metrik*. Der Unterschied spiegelt den Ansatz der Statistik im Vergleich zur Datenwissenschaft wider: Die Berücksichtigung von Unsicherheit steht im Mittelpunkt der statistischen Disziplin, währnd in der Datenwissenschaft konkrete geschäftliche oder organisatorische Ziele im Fokus stehen. Daher kann man sagen, dass Statistiker Schätzungen durchführen und Data Scientists Messungen vornehmen.

Mittelwert

Das grundlegendste Lagemaß ist der *Mittelwert* (genauer, das arithmetische Mittel) oder auch der *Durchschnitt*. Der Mittelwert entspricht der Summe aller Werte dividiert durch die Anzahl von Werten. Betrachten Sie die folgende Zahlenfolge: {3 5 1 2}. Der Mittelwert beträgt (3 + 5 + 1 + 2) / 4 = 11 / 4 = 2,75. Sie werden auf das Symbol $\bar{x}$ (ausgesprochen als »x quer«) stoßen, das verwendet wird, um den Mittel-

wert einer Stichprobe, die aus einer Grundgesamtheit gezogen wurde, darzustellen. Die Formel zur Berechnung des Mittelwerts für eine Menge von Werten x_1, x_2, ..., x_n lautet:

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

N (oder *n*) bezieht sich auf die Gesamtzahl aller Einträge bzw. Beobachtungen. In der Statistik wird es großgeschrieben, wenn es sich auf eine Grundgesamtheit bezieht, und kleingeschrieben, wenn es auf eine Stichprobe aus einer Grundgesamtheit abzielt. In der Data Science ist diese Unterscheidung nicht von Relevanz, weshalb Sie beide Möglichkeiten in Betracht ziehen können.

Eine Variante des Mittelwerts ist der *getrimmte Mittelwert*, den Sie berechnen, indem Sie eine feste Anzahl sortierter Werte an jedem Ende weglassen und dann den Mittelwert der verbleibenden Werte bilden. Für die sortierten Werte $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$, wobei $x_{(1)}$ der kleinste Wert und $x_{(n)}$ der größte ist, wird der getrimmte Mittelwert mit p kleinsten und größten weggelassenen Werten durch folgende Formel berechnet:

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

Durch die Verwendung des getrimmten Mittelwerts wird der Einfluss von Extremwerten beseitigt. Zum Beispiel werden bei internationalen Tauchmeisterschaften die höchste und die niedrigste Punktzahl der fünf Kampfrichter gestrichen, und als Endpunktzahl wird der Durchschnitt der Punktzahlen der drei verbleibenden Kampfrichter gewertet (*https://oreil.ly/uV4P0*). Dies macht es für einen einzelnen Kampfrichter schwierig, das Ergebnis zu manipulieren, etwa um den Kandidaten seines Landes zu begünstigen. Getrimmte Mittelwerte sind sehr verbreitet und in vielen Fällen der Verwendung des gewöhnlichen Mittelwerts vorzuziehen (siehe »Median und andere robuste Lagemaße« auf Seite 11 für weitere Erläuterungen).

Eine weitere Möglichkeit der Mittelwertbildung ist der *gewichtete Mittelwert*. Zur Berechnung multiplizieren Sie jeden Datenwert x_i mit einem benutzerdefinierten Gewicht w_i und dividieren die daraus resultierende Summe durch die Summe der Gewichte. Die Formel für den gewichteten Mittelwert lautet dementsprechend:

$$\text{gewichteter Mittelwert} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Den gewichteten Mittelwert verwendet man hauptsächlich aus zwei Gründen:

- Einige Werte weisen von sich aus eine größere Streuung auf als andere – um den Einfluss stark streuender Beobachtungen zu verringern, erhalten sie ein

geringeres Gewicht. Wenn wir z. B. den Mittelwert von mehreren Sensoren bilden und einer der Sensoren weniger genau misst, können wir die Daten dieses Sensors niedriger gewichten.

- Unsere erhobenen Daten repräsentieren die verschiedenen Gruppen, an deren Messung wir interessiert sind, nicht gleichmäßig. Beispielsweise ist es möglich, aufgrund der Art und Weise, wie ein Onlineversuch durchgeführt wurde, einen Datensatz zu gewinnen, der nicht alle Gruppen in der Nutzerbasis wahrheitsgemäß abbildet. Zur Korrektur können wir den Werten der Gruppen, die unterrepräsentiert sind, ein höheres Gewicht beimessen.

Median und andere robuste Lagemaße

Der *Median* entspricht dem mittleren Wert der sortierten Liste eines Datensatzes. Wenn es eine gerade Anzahl von Datenpunkten gibt, ist der mittlere Wert eigentlich nicht im Datensatz enthalten, weshalb der Durchschnitt der beiden Werte, die die sortierten Daten in eine obere und eine untere Hälfte teilen, verwendet wird. Verglichen mit dem Mittelwert, bei dem alle Beobachtungen berücksichtigt werden, beruht der Median nur auf den Werten, die sich in der Mitte des sortierten Datensatzes befinden. Dies mag zwar nachteilig erscheinen, da der Mittelwert wesentlich empfindlicher in Bezug auf die Datenwerte ist, aber es gibt viele Fälle, in denen der Median ein besseres Lagemaß darstellt. Angenommen, wir möchten die durchschnittlichen Haushaltseinkommen in den Nachbarschaften um den Lake Washington in Seattle unter die Lupe nehmen. Beim Vergleich der Ortschaft Medina mit der Ortschaft Windermere würde die Verwendung des Mittelwerts zu sehr unterschiedlichen Ergebnissen führen, da Bill Gates in Medina lebt. Wenn wir stattdessen den Median verwenden, spielt es keine Rolle, wie reich Bill Gates ist – die Position der mittleren Beobachtung bleibt unverändert.

Aus den gleichen Gründen wie bei der Verwendung eines gewichteten Mittelwerts ist es auch möglich, einen *gewichteten Median* zu ermitteln. Wie beim Median sortieren wir zunächst die Daten, obwohl jeder Datenwert ein zugehöriges Gewicht hat. Statt der mittleren Zahl ist der gewichtete Median ein Wert, bei dem die Summe der Gewichte für die untere und die obere »Hälfte« der sortierten Liste gleich ist. Wie der Median ist auch der gewichtete Median robust gegenüber Ausreißern.

Ausreißer

Der Median wird als *robustes* Lagemaß angesehen, da er nicht von *Ausreißern* (Extremfällen) beeinflusst wird, die die Ergebnisse verzerren könnten. Ausreißer sind Werte, die sehr stark von allen anderen Werten in einem Datensatz abweichen. Die genaue Definition eines Ausreißers ist etwas subjektiv, obwohl bestimmte Konventionen in verschiedenen zusammenfassenden Statistiken und Diagrammen verwendet werden (siehe »Perzentile und Box-Plots« auf Seite 21). Nur weil ein Datenwert einen Ausreißer darstellt, macht es ihn nicht ungültig oder fehlerhaft (wie im

vorherigen Beispiel mit Bill Gates). Dennoch sind Ausreißer oft das Ergebnis von Datenfehlern, wie z. B. von Daten, bei denen verschiedene Einheiten vermischt wurden (Kilometer gegenüber Metern), oder fehlerhafte Messwerte eines Sensors. Wenn Ausreißer das Ergebnis fehlerhafter bzw. ungültiger Daten sind, wird der Mittelwert zu einer falschen Einschätzung der Lage führen, wohingegen der Median immer noch seine Gültigkeit behält. Ausreißer sollten in jedem Fall identifiziert werden und sind in der Regel eine eingehendere Untersuchung wert.

Anomalieerkennung

Im Gegensatz zur gewöhnlichen Datenanalyse, bei der Ausreißer manchmal informativ sind und manchmal stören, sind bei der *Anomalieerkennung* die Ausreißer von Interesse, und der größere Teil der Daten dient in erster Linie dazu, den »Normalzustand« zu definieren, an dem die Anomalien gemessen werden.

Der Median ist nicht das einzige robuste Lagemaß. Tatsächlich wird häufig der getrimmte Mittelwert verwendet, um den Einfluss von Ausreißern zu vermeiden. So bietet z. B. die Entfernung der unteren und oberen 10 % der Daten (eine übliche Wahl) Schutz vor Ausreißern, es sei denn, der Datensatz ist zu klein. Der getrimmte Mittelwert kann als Kompromiss zwischen dem Median und dem Mittelwert gesehen werden: Er ist robust gegenüber Extremwerten in den Daten, verwendet jedoch mehr Daten zur Berechnung des Lagemaßes.

Weitere robuste Lagemaße

Statistiker haben eine Vielzahl anderer Lagemaße entwickelt, und zwar in erster Linie mit dem Ziel, einen Schätzer zu entwickeln, der robuster und auch effizienter als der Mittelwert ist (d. h. besser in der Lage, kleine Unterschiede hinsichtlich der Lage zwischen Datensätzen zu erkennen). Während diese Methoden für kleine Datensätze durchaus nützlich sein können, dürften sie bei großen oder selbst bei mittelgroßen Datensätzen keinen zusätzlichen Nutzen bringen.

Beispiel: Lagemaße für Einwohnerzahlen und Mordraten

Tabelle 1-2 zeigt einen Auszug der ersten paar Zeilen eines Datensatzes, der Informationen zu den Einwohnerzahlen und Mordraten für jeden US-Bundesstaat enthält (Zensus 2010). Die Einheit für die Mordrate wurde mit »Morde pro 100.000 Personen pro Jahr« gewählt.

Tabelle 1-2: Die ersten Zeilen des data.frame, der Auskunft über die Einwohnerzahlen und Mordraten der einzelnen Bundesstaaten gibt

	Bundesstaat	Einwohnerzahl	Mordrate	Abkürzung
1	Alabama	4.779.736	5,7	AL
2	Alaska	710.231	5,6	AK
3	Arizona	6.392.017	4,7	AZ

Tabelle 1-2: Die ersten Zeilen des data.frame, der Auskunft über die Einwohnerzahlen und Mordraten der einzelnen Bundesstaaten gibt (Fortsetzung)

	Bundesstaat	Einwohnerzahl	Mordrate	Abkürzung
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

Berechnen Sie den Mittelwert, den getrimmten Mittelwert und den Median für die Einwohnerzahlen in *R*:[1]

```
> state <- read.csv('state.csv')
> mean(state[['Population']])
[1] 6162876
> mean(state[['Population']], trim=0.1)
[1] 4783697
> median(state[['Population']])
[1] 4436370
```

In *Python* können wir zur Berechnung des Mittelwerts und des Medians die pandas-Methoden des Data Frame verwenden. Den getrimmten Mittelwert erhalten wir durch die Funktion `trim_mean` aus dem Modul `scipy.stats`:

```
state = pd.read_csv('state.csv')
state['Population'].mean()
trim_mean(state['Population'], 0.1)
state['Population'].median()
```

Der Mittelwert ist größer als der getrimmte Mittelwert, der wiederum größer als der Median ist.

Dies liegt daran, dass der getrimmte Mittelwert die fünf größten und fünf kleinsten Bundesstaaten ausschließt (`trim=0.1` entfernt 10% an beiden Enden der Verteilung). Wenn wir die durchschnittliche Mordrate für das Land berechnen wollen, müssen wir dazu den gewichteten Mittelwert oder den Median heranziehen, um die unterschiedlich hohe Anzahl an Einwohnern in den Bundesstaaten zu berücksichtigen. Da *R* in seiner Standardbibliothek keine Funktion für den gewichteten Median umfasst, müssen wir zu diesem Zweck zunächst das Paket `matrixStats` installieren:

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

1 Der *R*- und der *Python*-Code sind auf das Wesentliche reduziert. Den vollständigen Code sowie die Datensätze zum Herunterladen finden Sie unter *https://github.com/gedeck/practical-statistics-for-data-scientists*.

Bei *Python* ist die Funktion zur Berechnung des gewichteten Mittelwerts im `NumPy`-Paket enthalten. Für den gewichteten Median können wir speziell das Paket wquantiles (*https://oreil.ly/4SIPQ*) verwenden:

```
np.average(state['Murder.Rate'], weights=state['Population'])
wquantiles.median(state['Murder.Rate'], weights=state['Population'])
```

Im vorliegenden Fall sind der gewichtete Mittelwert und der gewichtete Median in etwa gleich groß.

Kernideen

- Das wesentliche Lagemaß ist der Mittelwert, der jedoch empfindlich auf Extremwerte (Ausreißer) reagiert.
- Andere Maße (Median, getrimmter Mittelwert) sind weniger empfindlich gegenüber Ausreißern und ungewöhnlich verteilten Daten und daher robuster.

Weiterführende Literatur

- In dem Wikipedia-Artikel zur zentralen Tendenz (*https://oreil.ly/qUW2i*) werden verschiedene Lagemaße ausführlich erläutert.
- John Tukeys Standardwerk aus dem Jahr 1977, *Exploratory Data Analysis* (Pearson), erweist sich nach wie vor als eine beliebte Lektüre.

Streuungsmaße

Die Lage ist nur eine Dimension bei der Zusammenfassung eines Merkmals. Eine zweite Dimension, die *Streuung* (engl. *Variability*) – auch *Variabilität* oder *Dispersion* genannt –, misst, ob die Datenwerte eng zusammenliegen oder weit gestreut sind. Die Streuung ist das Herzstück der Statistik: Sie wird gemessen, reduziert, es kann unterschieden werden zwischen zufälliger und tatsächlicher Streuung, die verschiedenen Quellen der wahren Streuung können identifiziert und Entscheidungen in Gegenwart der Streuung können getroffen werden.

Schlüsselbegriffe zu Streuungsmaßen

Abweichung
: Die Differenz zwischen den beobachteten Werten und dem Lagemaß (engl. *Deviation*).

Synonyme
: Fehler, Residuen