

# 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.

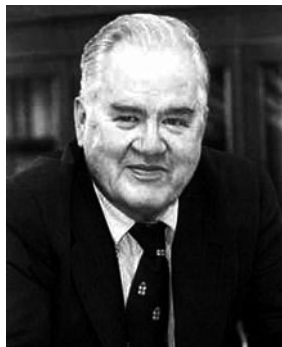


Abbildung 1-1: John Tukey, der bedeutende Statistiker, dessen vor über 50 Jahren entwickelte Ideen die Grundlage der Data Science bilden

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 kategoriale 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, während 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  $x_1, x_2, \dots, x_n$  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  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ , 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:<sup>1</sup>

```
> 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 getroffen werden.

### Schlüsselbegriffe zu Streuungsmaßen

#### *Abweichung*

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

#### *Synonyme*

Fehler, Residuen

**Varianz**

Die Summe der quadrierten Abweichungen vom Mittelwert dividiert durch  $n - 1$ , wobei  $n$  die Anzahl der Beobachtungen ist.

*Synonym*

mittlerer quadratischer Fehler

**Standardabweichung**

Die Quadratwurzel der Varianz.

**Mittlere absolute Abweichung**

Der Mittelwert der Absolutwerte der Abweichungen vom Mittelwert.

*Synonyme*

$l_1$ -Norm, Manhattan-Norm

**Mittlere absolute Abweichung vom Median**

Der Median der Absolutwerte der Abweichungen vom Median.

**Spannweite**

Die Differenz zwischen dem größten und dem kleinsten Wert in einem Datensatz (engl. *Range*).

**Ordnungsstatistik**

Eine auf den Datenwerten basierende Metrik, sortiert vom kleinsten zum größten.

*Synonym*

Rang

**Perzentil**

Der Wert, bei dem  $P$  % der Werte diesen Wert oder weniger und  $(100-P)$  % diesen Wert oder mehr annehmen.

*Synonym*

Quantil

**Interquartilsabstand**

Die Differenz zwischen dem 75%-Perzentil und dem 25%-Perzentil.

*Synonym*

IQR

So wie es verschiedene Möglichkeiten gibt, die Lage zu messen (Mittelwert, Median usw.), so gibt es auch verschiedene Möglichkeiten, das Ausmaß der Streuung zu bestimmen.

## Standardabweichung und ähnliche Maße

Die meistgenutzten Streuungsmaße basieren auf den Differenzen bzw. *Abweichungen* zwischen den Lagemaßen und den beobachteten Daten. Für eine gegebene Zahlenfolge  $\{1, 4, 4\}$  ist der Mittelwert 3 und der Median 4. Die Abweichungen vom Mittelwert entsprechen den jeweiligen Differenzen:  $1 - 3 = -2$ ,  $4 - 3 = 1$ ,  $4 - 3 = 1$ .

Diese Abweichungen geben uns Aufschluss darüber, wie weit die Daten um den Zentralwert herum gestreut sind.

Eine Möglichkeit, die Streuung zu messen, besteht darin, einen typischen Wert für diese Abweichungen zu schätzen. Eine Mittelwertbildung über die Abweichungen selbst würde uns nicht viel sagen – die negativen Abweichungen würden die positiven ausgleichen. Tatsächlich ist auch im vorliegenden Beispiel die Summe der Abweichungen vom Mittelwert genau null. Stattdessen können wir den Mittelwert der Absolutwerte der Abweichungen vom Mittelwert bilden. Im vorhergehenden Beispiel sind die Absolutwerte der Abweichungen  $\{2 \ 1 \ 1\}$ , und ihr Mittelwert ergibt  $(2 + 1 + 1) / 3 = 1,33$ . Dieses Maß wird als *mittlere absolute Abweichung* bezeichnet und mit der folgenden Formel ermittelt:

$$\text{Mittlere absolute Abweichung} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

wobei  $\bar{x}$  für den Stichprobenmittelwert steht.

Die bekanntesten Streuungsmaße sind die *Varianz* und die *Standardabweichung*, die auf den quadratischen Abweichungen beruhen. Die Varianz ist der Durchschnitt der quadrierten Abweichungen, und die Standardabweichung ist wiederum die Quadratwurzel der Varianz:

$$\begin{aligned} \text{Varianz} &= s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \\ \text{Standardabweichung} &= s = \sqrt{\text{Varianz}} \end{aligned}$$

Die Standardabweichung ist viel leichter zu interpretieren als die Varianz, da sie auf dieselbe Skala wie die Originaldaten bezogen ist. Dennoch mag es mit ihrer komplizierteren und weniger intuitiven Formel merkwürdig erscheinen, dass die Standardabweichung in der Statistik gegenüber der mittleren absoluten Abweichung bevorzugt wird. Sie verdankt ihre Vorrangstellung der statistischen Theorie: Mathematisch gesehen, ist es sehr viel vorteilhafter, quadrierte Werte zu verwenden – und nicht Absolutwerte –, insbesondere in statistischen Modellen.

### Die Anzahl der Freiheitsgrade und die Frage, ob $n$ oder $n - 1$ ?

In Statistikbüchern finden Sie für gewöhnlich einen Abschnitt, der erklärt, warum wir  $n - 1$  im Nenner der Formel für die Varianz anstelle von  $n$  haben, was uns zum Konzept der *Freiheitsgrade* führt. Diese Unterscheidung ist an sich nicht von großer Bedeutung, da  $n$  im Allgemeinen so groß ist, dass es keinen besonderen Unterschied macht, ob man durch  $n$  oder  $n - 1$  dividiert. Aber falls es Sie interessiert, hier folgt die Erklärung. Sie basiert auf der Prämisse, dass Sie auf Basis einer Stichprobe Schätzungen über eine Grundgesamtheit (Population) vornehmen möchten.



Wenn Sie intuitiverweise  $n$  im Nenner der Varianzformel verwenden, unterschätzen Sie den wahren Wert der Varianz und der Standardabweichung in der Grundgesamtheit. Dies wird als ein *verzerrter* Schätzer (engl. *biased*) bezeichnet. Wenn Sie jedoch  $n - 1$  anstelle von  $n$  einsetzen, ermitteln Sie einen *unverzerrten* (engl. *unbiased*) bzw. erwartungstreuen Schätzer der Varianz.

Um vollständig zu erklären, warum die Verwendung von  $n$  zu einem verzerrten Schätzer führt, müssen wir den Begriff der Freiheitsgrade heranziehen, der die Anzahl der Einschränkungen bei der Berechnung eines Schätzers berücksichtigt. In diesem Fall gibt es  $n - 1$  Freiheitsgrade, da es eine Randbedingung gibt: Die Standardabweichung hängt von der Berechnung des Stichprobenmittelwerts ab. In den meisten Anwendungsfällen müssen sich Data Scientists keine Gedanken über die Anzahl der Freiheitsgrade machen.

Weder die Varianz noch die Standardabweichung oder die mittlere absolute Abweichung ist gegenüber Ausreißern und Extremwerten robust (siehe »Median und andere robuste Lagemaße« auf Seite 11 für eine Erläuterung zu den robusten Lagemaßen). Die Varianz und die Standardabweichung sind besonders empfindlich gegenüber Ausreißern, da sie auf den quadrierten Abweichungen beruhen.

Ein robustes Streuungsmaß ist die *mittlere absolute Abweichung vom Median* (engl. *Median Absolute Deviation from the Median*, MAD):

$$\text{Mittlere absolute Abweichung vom Median} = \text{Median}(|x_1 - m|, |x_2 - m|, \dots, |x_N - m|)$$

wobei  $m$  dem Median entspricht. Wie der Median wird auch die mittlere absolute Abweichung vom Median nicht durch Extremwerte beeinflusst. Es ist auch möglich, eine getrimmte Standardabweichung analog zum getrimmten Mittelwert zu berechnen (siehe »Mittelwert« auf Seite 9).



Die Varianz, die Standardabweichung, die mittlere absolute Abweichung und die mittlere absolute Abweichung vom Median sind keine äquivalenten Streuungsmaße – selbst dann nicht, wenn die Daten normalverteilt sind. So ist die Standardabweichung immer größer als die mittlere absolute Abweichung, die ihrerseits größer als die mittlere absolute Abweichung vom Median ist. Manchmal wird die mittlere absolute Abweichung vom Median mit einem konstanten Skalierungsfaktor multipliziert, um den Wert für den Fall, dass die Daten normalverteilt sind, genau so zu skalieren wie die Standardabweichung. Der üblicherweise verwendete Faktor von 1,4826 bedeutet, dass 50% der Normalverteilung in den Bereich  $\pm \text{MAD}$  fallen (siehe z. B. <https://oreil.ly/SfDk2>).

## Streuungsmaße auf Basis von Perzentilen

Ein anderer Ansatz zur Schätzung der Streuung basiert auf der Betrachtung der Streuung der sortierten Daten. Statistiken, die auf sortierten (d.h. in einer Rang-

folge geordneten) Daten basieren, werden als *Ordnungsstatistiken* bezeichnet. Das grundlegende Maß ist die *Spannweite*: die Differenz zwischen dem größten und dem kleinsten Wert. Die Minimal- und Maximalwerte selbst sind zwar durchaus interessant und bei der Identifizierung von Ausreißern nützlich, aber die Spannweite erweist sich als äußerst empfindlich gegenüber Ausreißern und ist als allgemeines Streuungsmaß nicht sehr hilfreich.

Um der Anfälligkeit gegenüber Ausreißern vorzubeugen, können wir vor der Ermittlung der Spannweite Werte an beiden Enden der Daten weglassen. Formal basieren diese Arten von Schätzern auf Unterschieden zwischen *Perzentilen*. In einem Datensatz ist das  $P\%$ -Perzentil so definiert, dass mindestens  $P\%$  der Werte diesen Wert oder weniger und mindestens  $(100 - P)\%$  der Werte diesen Wert oder mehr annehmen. Um zum Beispiel das 80%-Perzentil zu ermitteln, müssen Sie die Daten zunächst sortieren. Dann gehen Sie, beginnend beim kleinsten Wert, 80% der Strecke zum größten Wert weiter. Der Median ist übrigens ein und dasselbe wie das 50%-Perzentil. Ein Perzentil ist im Wesentlichen dasselbe wie ein *Quantil*, wobei Quantile durch Bruchzahlen angegeben werden (das 0,8-Quantil ist also dasselbe wie das 80%-Perzentil).

Ein gebräuchliches Streuungsmaß ist die Differenz zwischen dem 25%-Perzentil und dem 75%-Perzentil, der sogenannte *Interquartilsabstand* (engl. *Interquartile Range*, IQR). Hier ist ein einfaches Zahlenbeispiel:  $\{3, 1, 5, 3, 6, 7, 2, 9\}$ . Wir sortieren diese Zahlenfolge, um  $\{1, 2, 3, 3, 5, 6, 7, 9\}$  zu erhalten. Das 25%-Perzentil liegt bei 2,5 und das 75%-Perzentil bei 6,5, sodass der Interquartilsabstand  $6,5 - 2,5 = 4$  beträgt. Die Softwareprogramme können leicht unterschiedliche Ansätze haben, die dann unterschiedliche Ergebnisse hervorbringen (siehe folgenden Hinweis); in der Regel fallen diese Unterschiede jedoch gering aus.

Bei sehr großen Datensätzen kann die Berechnung der genauen Perzentile rechnerisch sehr aufwendig sein, da dazu alle Datenwerte sortiert werden müssen. Maschinelle Lern- und Statistikprogramme verwenden spezielle Algorithmen, wie [Zhang-Wang-2007], um einen Näherungswert für ein Perzentil zu erhalten, der sehr schnell berechnet werden kann und eine gewisse Genauigkeit gewährleistet.



#### Perzentile: Genaue Definition

Wenn wir eine gerade Anzahl an Werten haben ( $n$  ist gerade), dann ist das Perzentil im Sinne der vorhergehenden Definition mehrdeutig. Tatsächlich könnten wir jeden Wert zwischen der Ordnungsstatistik  $x_{(j)}$  und  $x_{(j+1)}$  nehmen, wobei  $j$  Folgendes erfüllt:

$$100 * \frac{j}{n} \leq P < 100 * \frac{j+1}{n}$$

In formaler Hinsicht ist das Perzentil ein gewichteter Durchschnitt:

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

für ein gegebenes Gewicht  $w$  zwischen 0 und 1. In den verschiedenen verfügbaren Statistikprogrammen gibt es leicht unterschiedliche

Ansätze für die Auswahl von  $w$ . Tatsächlich bietet die R-Funktion `quantile` neun verschiedene Alternativen zur Berechnung des Quantils. Von kleinen Datensätzen abgesehen, brauchen Sie sich in der Regel keine Gedanken darüber zu machen, wie ein Perzentil genau berechnet wird. In *Python* unterstützt das `numpy.quantile` fünf Ansätze, wobei die lineare Interpolation voreingestellt ist.

## Beispiel: Streuungsmaße für die Einwohnerzahlen der Bundesstaaten in den USA

Tabelle 1-3 (Tabelle 1-2 wird der Einfachheit halber erneut dargestellt) zeigt die ersten paar Zeilen im Datensatz, in dem die Einwohnerzahlen und Mordraten für jeden US-Bundesstaat enthalten sind.

Tabelle 1-3: Die ersten Zeilen des `data.frame` mit den Einwohnerzahlen und Mordraten nach Bundesstaaten

	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
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

Unter Verwendung der in R integrierten Funktionen für die Standardabweichung, den Interquartilsabstand (IQR) und die mittlere absolute Abweichung vom Median können wir Streuungsmaße für die Einwohnerstatistiken der Bundesstaaten berechnen:

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

Für ein Data-Frame-Objekt stehen uns in der `pandas`-Bibliothek verschiedene Methoden zur Berechnung der Standardabweichung und der Quantile zur Verfügung. Nach Ermittlung der Quantilswerte können wir den IQR berechnen. Für die robuste mittlere absolute Abweichung vom Median verwenden wir die Funktion `robust.scale.mad` aus dem `statsmodels`-Paket:

```
state['Population'].std()  
state['Population'].quantile(0.75) - state['Population'].quantile(0.25)  
robust.scale.mad(state['Population'])
```

Die Standardabweichung ist fast doppelt so groß wie die MAD (in R wird die Skalierung der mittleren absoluten Abweichung vom Median standardmäßig so angepasst, dass der Mittelwert die gleiche Skalierung besitzt). Dies ist nicht weiter verwunderlich, da die Standardabweichung gegenüber Ausreißern sensibel ist.

### Kernideen

- Die Varianz und die Standardabweichung sind die am weitesten verbreiteten und routinemäßig berichteten Streuungsmaße.
- Beide sind empfindlich gegenüber Ausreißern.
- Zu den robusteren Maßen gehören die mittlere absolute Abweichung, die mittlere absolute Abweichung vom Median und Perzentile (Quantile).

## Weiterführende Literatur

- David Lanes Online-Statistik-Ratgeber hat einen Abschnitt über Perzentile (<https://oreil.ly/o2fBI>).
- Kevin Davenport hat einen nützlichen Beitrag auf R-Bloggers (<https://oreil.ly/E7zcG>) über Abweichungen vom Median und ihre robusten Eigenschaften verfasst.

## Exploration der Datenverteilung

Alle von uns behandelten Maße fassen die Daten in einer einzigen Zahl zusammen, um die Lage oder die Streuung der Daten zu beschreiben. Es ist auch wertvoll, zu untersuchen, wie die komplette Verteilung der Daten aussieht.

### Schlüsselbegriffe zur Exploration von Verteilungen

#### **Box-Plot**

Ein von Tukey eingeführtes Diagramm zur schnellen Visualisierung der Datenverteilung.

#### *Synonyme*

Box-Whisker-Plot, Kastengrafik

#### **Häufigkeitstabelle**

Eine Übersicht über die Anzahl der numerischen Werte, die in eine Menge von Intervallen (Klassen, engl. *Bins*) fallen.

### Histogramm

Ein Diagramm der Häufigkeitstabelle mit den Intervallen auf der x-Achse und der Anzahl (oder dem relativen Anteil) auf der y-Achse. Balkendiagramme sind zwar ähnlich, sollten aber nicht mit Histogrammen verwechselt werden (siehe »Binäre und kategoriale Daten untersuchen« auf Seite 28 für eine Erläuterung des Unterschieds).

### Dichtediagramm

Eine geglättete Version des Histogramms, oft basierend auf einer *Kerndichteschätzung* (engl. *Kernel Density Estimate*).

## Perzentile und Box-Plots

In »Streuungsmaße auf Basis von Perzentilen« auf Seite 17 sind wir der Frage nachgegangen, wie Perzentile zur Messung der Streuung der Daten verwendet werden können. Perzentile sind auch nützlich, um die gesamte Verteilung zusammenfassend darzustellen. Es ist üblich, die Quartile (25 %-, 50 %- und 75 %-Perzentile) und die Dezile (10 %-, 20 %-, ..., 90 %-Perzentile) anzugeben. Perzentile sind besonders aussagekräftig, wenn man die *Enden* bzw. *Ränder* (die äußeren Bereiche) der Verteilung zusammenzufassend darstellen möchte. In der breiten Öffentlichkeit ist in diesem Zusammenhang oft von der Redewendung »das eine Prozent« die Rede, die genutzt wird, um Reiche im obersten 99 %-Perzentil der Vermögens- bzw. Einkommensverteilung zu charakterisieren.

Tabelle 1-4 stellt einige Perzentile der Mordraten in den Bundesstaaten dar. In R können wir uns die Werte mithilfe der Funktion `quantile` ausgeben lassen:

```
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
```

In *Python* können Sie für einen Data Frame die *pandas*-Methode `quantile` nutzen, um sich die Perzentile ausgeben zu lassen:

```
state['Murder.Rate'].quantile([0.05, 0.25, 0.5, 0.75, 0.95])
```

Tabelle 1-4: Perzentile der Mordraten in den Bundesstaaten

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

Der Median liegt bei vier Morden pro 100.000 Einwohner. In den Daten gibt jedoch eine beträchtliche Streuung: Das 5 %-Perzentil beträgt nur 1,6 und das 95 %-Perzentil 6,51.

Die von Tukey [Tukey-1977] eingeführten *Box-Plots* stützen sich auf Perzentile und bieten eine rasche Möglichkeit, die Verteilung Ihrer Daten zu visualisieren. Abbil-

Abbildung 1-2 zeigt einen in R erstellten Box-Plot für die Einwohnerzahlen der Bundesstaaten:

```
boxplot(state[['Population']]/1000000, ylab='Einwohnerzahl (in Millionen)')
```

Die pandas-Bibliothek bietet eine Reihe von grundlegenden informativen Diagrammen, die für Data Frames genutzt werden können; darunter auch Box-Plots:

```
ax = (state['Population']/1_000_000).plot.box()  
ax.set_ylabel('Einwohnerzahl (in Millionen)')
```

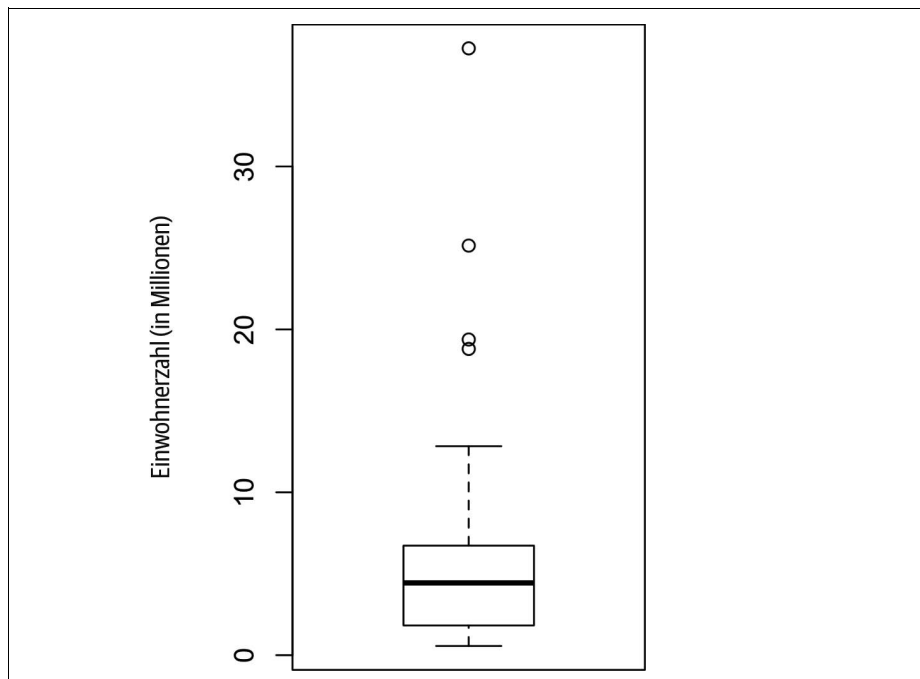


Abbildung 1-2: Box-Plot für die Einwohnerzahlen der Bundesstaaten

Bei diesem Box-Plot können wir auf einen Blick erkennen, dass die mittlere Einwohnerzahl der Bundesstaaten etwa fünf Millionen beträgt (Median), die Einwohnerzahl für die Hälfte der Staaten zwischen etwa zwei und sieben Millionen liegt und dass es einige bevölkerungsreiche Ausreißer gibt. Der obere und der untere Rand des Rechtecks (Box) kennzeichnen jeweils das 75%- bzw. 25%-Perzentil. Der Median wird durch die fett gehaltene horizontale Linie in der Box angezeigt. Die beiden vertikalen gestrichelten Linien, die als *Whisker* oder auch Antennen bezeichnet werden, erstrecken sich über den oberen und unteren Rand der Box, um den Bereich, in dem der überwiegende Teil der Daten liegt, zu kennzeichnen. Es gibt zahlreiche Varianten von Box-Plots (siehe z.B. die Dokumentation der R-Funktion `boxplot` [R-base-2015]). Standardmäßig verlängert die R-Funktion die Whisker bis zu dem Datenpunkt, der am weitesten über die Box hinausgeht – jedenfalls immer dann, wenn dieser Wert nicht mehr als das 1,5-Fache des IQR be-

trägt. Die matplotlib-Bibliothek verwendet die gleiche Implementierung. In anderen Softwareprogrammen kann eine abweichende Regel angewandt werden.

Alle Datenpunkte außerhalb der Whisker werden als einzelne Punkte oder Kreise dargestellt (die in der Regel als Ausreißer angesehen werden).

## Häufigkeitstabellen und Histogramme

Eine Häufigkeitstabelle teilt den Wertebereich einer Variablen bzw. eines Merkmals in gleich große Intervalle auf und gibt uns Auskunft darüber, wie viele Werte jeweils in jedes Intervall fallen. Tabelle 1-5 zeigt Ihnen eine in R erstellte Häufigkeitstabelle für die Einwohnerzahlen der Bundesstaaten:

```
breaks <- seq(from=min(state[['Population']]),
               to=max(state[['Population']]), length=11)
pop_freq <- cut(state[['Population']], breaks=breaks,
                right=TRUE, include.lowest=TRUE)
table(pop_freq)
```

Die Funktion `pandas.cut` erzeugt eine Zahlenfolge (Series-Objekt), die die Werte auf die einzelnen Intervalle abbildet. Mit der Methode `value_counts` erhalten wir die Häufigkeitstabelle:

```
binnedPopulation = pd.cut(state['Population'], 10)
binnedPopulation.value_counts()
```

Tabelle 1-5: Eine Häufigkeitstabelle für die Einwohnerzahlen der Bundesstaaten

Intervallnummer	Intervall	Häufigkeit	Bundesstaaten
1	563.626–4.232.658	24	WY,VT,ND,AK,SD,DE,MT,RI,NH,ME,HI,ID,NE,WV,NM,NV,UT,KS,AR,MS,IA,CT,OK,OR
2	4.232.659–7.901.691	14	KY,LA,SC,AL,CO,MN,WI,MD,MO,TN,AZ,IN,MA,WA
3	7.901.692–11.570.724	6	VA,NJ,NC,GA,MI,OH
4	11.570.725–15.239.757	2	PA,IL
5	15.239.758–18.908.790	1	FL
6	18.908.791–22.577.823	1	NY
7	22.577.824–26.246.856	1	TX
8	26.246.857–29.915.889	0	
9	29.915.890–33.584.922	0	
10	33.584.923–37.253.956	1	CA

Der Bundesstaat mit der geringsten Einwohnerzahl ist Wyoming mit 563.626 Menschen, und der bevölkerungsreichste ist Kalifornien mit 37.253.956 Einwohnern. Daraus ergibt sich ein Wertebereich von  $37.253.956 - 563.626 = 36.690.330$ , den wir in gleich große Klassen – sagen wir 10 – aufteilen müssen. Bei 10 gleich großen Klassen hat jede Klasse eine Breite von 3.669.033, sodass das erste Intervall von

563.626 bis 4.232.658 reicht. Im Gegensatz dazu liegt in der obersten Klasse für das Intervall 33.584.923 bis 37.253.956 Einwohnern nur ein einziger Bundesstaat: Kalifornien. Die beiden nächstkleineren Klassen sind unbesetzt, bis als Nächstes der Bundesstaat Texas erreicht wird. Es ist wichtig, die leeren Klassen mit einzubeziehen; die Tatsache, dass sich in diesen Intervallen keine Werte befinden, ist eine nützliche Information. Es kann auch hilfreich sein, mit verschiedenen Klassenbreiten bzw. Intervallgrößen zu experimentieren. Wenn sie zu groß sind, treten wichtige Merkmale der Verteilung gegebenenfalls nicht mehr sichtbar hervor. Werden sie zu klein gewählt, ist das Ergebnis zu feingliedrig, und die Fähigkeit, ein adäquates Gesamtbild zu liefern, geht verloren.



Sowohl Häufigkeitstabellen als auch Perzentile fassen die Daten durch die Einteilung in Klassen bzw. Intervalle übersichtlich zusammen. Im Allgemeinen haben Quartile und Dezile in jeder Klasse die gleiche Anzahl an Beobachtungen (Klassen mit gleicher Anzahl – *equal-count bins*), aber die Klassenbreite ist für gewöhnlich unterschiedlich. Bei der Häufigkeitstabelle umfassen die Klassen hingegen eine unterschiedliche Anzahl an Beobachtungen, wohingegen die Klassenbreite identisch ist (gleich breite Klassen – *equal-size bins*).

Ein Histogramm bietet die Möglichkeit, eine Häufigkeitstabelle zu visualisieren. Dabei werden die Klassen auf der x-Achse abgetragen und die Anzahl der Beobachtungen bzw. Häufigkeiten auf der y-Achse. In Abbildung 1-3 erstreckt sich die bei zehn Millionen ( $1e+07$ ) Einwohnern in der Mitte befindliche Klasse von ungefähr acht bis zwölf Millionen Einwohnern und umfasst insgesamt sechs Beobachtungen. Um ein Histogramm in R zu erstellen, das dem in Tabelle 1-5 entspricht, müssen Sie lediglich die Funktion `hist` mit dem Argument `breaks` verwenden:

```
hist(state[['Population']], breaks=breaks)
```

Die pandas-Bibliothek unterstützt die Erstellung von Histogrammen für Data-Frame-Objekte mit der Methode `DataFrame.plot.hist`. Mit dem Schlüsselwortargument `bins` können Sie die Anzahl der Klassen bestimmen. Die verfügbaren `plot`-Methoden geben jeweils ein Objekt zurück, das die Koordinaten in Bezug auf die Achsen bereithält und eine weitere Feinabstimmung des Diagramms mithilfe der `matplotlib`-Bibliothek ermöglicht:

```
ax = (state['Population'] / 1_000_000).plot.hist(figsize=(4, 4))
ax.set_xlabel('Einwohnerzahl (in Millionen)')
```

Das Histogramm wird in Abbildung 1-3 gezeigt. Im Allgemeinen werden Histogramme so erstellt:

- Auch unbesetzte Klassen werden in die Darstellung mit einbezogen.
- Die Klassen sind gleich breit.
- Die Wahl der Anzahl der Klassen (oder, äquivalent, der Klassenbreite) ist dem Anwender überlassen.