FOM-Edition

Sebastian Sauer

Moderne Datenanalyse mit R

Daten einlesen, aufbereiten, visualisieren, modellieren und kommunizieren

FOM Hochschule

Springer Gabler

FOM-Edition

FOM Hochschule für Oekonomie & Management

Reihenherausgeber
FOM Hochschule für Oekonomie & Management, Essen, Deutschland

Dieses Werk erscheint in der FOM-Edition, herausgegeben von der FOM Hochschule für Oekonomie & Management.

Weitere Bände in der Reihe
http://www.springer.com/series/12753

Sebastian Sauer

Moderne Datenanalyse mit R

Daten einlesen, aufbereiten, visualisieren, modellieren und kommunizieren

Springer Gabler

FOM Hochschule

Sebastian Sauer
FOM Hochschule für Oekonomie &
Management
Nürnberg, Deutschland

FOM-Edition
ISBN 978-3-658-21586-6 ISBN 978-3-658-21587-3 (eBook)
https://doi.org/10.1007/978-3-658-21587-3

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

Springer Gabler

Springer Gabler ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Vorwort

> Wir fühlen, dass selbst, wenn alle möglichen wissenschaftlichen Fragen beantwortet sind, unsere Lebensprobleme noch gar nicht berührt sind.
>
> – *Wittgenstein, Tractatus, 6.52*

Dystopie eines Datenzeitalters

Wir leben im frühen Zeitalter der Daten und Algorithmen, dem *Algorithmozän*. Gebetsmühlenartig haben uns Unternehmensberatungen, Politiker und Google-Ingenieure dieses Mantra vorgetragen, so dass es längt zum Fundus säuberlich abgehefteter Binsenweisheiten gehört (vgl. Chui et al. 2018). Untermalt wird dieses sonore Flüstern durch ein Stakkato von Eilmeldungen wie kürzlich von *AlphaGo Zero* (D. Silver et al. 2017b). Das ist ein Programm, das sehr gut im Brettspiel *Go* ist (und wohl auch in einigen anderen Spielen laut D. Silver et al. (2017a)). Wie sein Vorgänger aus dem letzten Jahr, *AlphaGo*, basiert das Programm auf sog. neuronalen Netzen, einer wohlbekannten Methode des statistischen Modellierens (Scherer 2013). Das neue Programm spielt deutlich besser als das alte: Bei einem gemütlichen Treffen schlug der Neue den Alten vernichtend: Mit 100 zu 0 fegte der Frischling den alten Hasen vom Platz. Dabei hatte der Alte einiges vorzuweisen. Im Vorjahr hatte er einen Meister des Go-Spiels, einen Menschen, klar besiegt (D. Silver et al. 2017b). Interessant ist, dass AlphaGo Zero ohne Lernmaterial von außen auskam, im Gegensatz zu früheren Programmen wie AlphaGo. Das legt nahe, dass Maschinen grundsätzlich in der Lage sind, ohne Hilfe von Menschen zu lernen – und „übermenschliche" Leistung im Go-Spiel und vielleicht auch anderswo zu erzielen.

Ähnlich spektakulär: Eine Reihe von Fachartikeln zeigte, dass Algorithmen – ausreichend mit Daten gefüttert – die Persönlichkeit einer Person besser einschätzen können als deren Freunde (Kosinski et al. 2013; Quercia, Kosinski, Stillwell, & Crowcroft 2011; Youyou et al. 2015). Auch hier wurden moderne Modelle des statistischen Modellierens verwendet. Auf analoge Art schlägt Ihnen ein Algorithmus auf einer Webseite vor, für welche Produkte Sie sich noch interessieren könnten. Nicht immer ist der Algorithmus auch nur ansatzweise clever: Wer hat noch nicht erlebt, im Internet einen Gegenstand erworben zu haben, ein Fahrrad zum Beispiel, und danach noch wochenlang von Werbung für Fahrräder drangsaliert zu werden (mir reicht *ein* Rad, Google).

Bei aller Dystopie, die mit der Digitalisierung zusammenhängt und mit Chinas „Kreditwürdigkeitspunkten“ bisher am konsequentesten weitergedacht wurde (Botsman 2017), es gibt auch Nutzen. Moderne Daten-Technologien stecken im Smartphone, in medizinischen Anwendungen und in deutschen Autos. Wer möchte auf diesen Fortschritt verzichten? Die wenigsten offenbar.

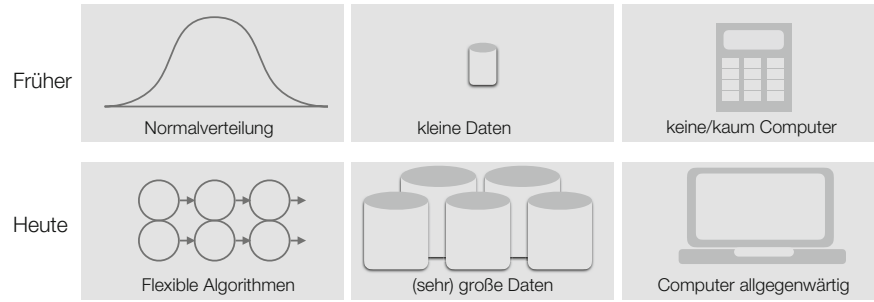
Diese Beispiele ließen sich noch länger mit Unterhaltungswert fortsetzen. Wie man den Fortschritt der Algorithmen auch einschätzt – wünschenswert, durchwachsen oder bedrohlich – man muss zum gleichen Schluss kommen: An der intensiven Beschäftigung mit dieser Technik kommen wir (jeder Einzelne) nicht vorbei. Ein Baustein, um die unheimliche Bedrohung durch panoptische, orwelleske Überwachung abzuwenden, ist das Verständnis der modernen Datentechnik. Gleichermaßen gilt: Um die offenbar gewaltigen ökonomischen Potenziale für die Unternehmen urbar zu machen, müssen wir die Technik verstehen (Brynjolfsson und McAfee 2016). Die Digitalisierung ist der bestimmende Trend des Wirtschaftslebens – wahrscheinlich (Vorsicht mit Vorhersagen); daher ist es beruflich, gesellschaftlich und politisch geboten, sich dem Algorithmozän zu stellen. Das heißt nicht, dass jeder Programmierer und Statistiker werden muss. Aber ein gewisses Grundverständnis sollte zum Bildungsstandard gehören.

Das Statistikcurriculum ist veraltet

Die Lehrpläne der Hochschulen geben sich von dramatischen Meldungen und neuen Technologien noch weitgehend unbescholten. Zumeist gilt in Lehrplänen für Statistik: Über den *t*-Test geht nichts. Der Wirklichkeit außerhalb der Alma Mater wird das kaum gerecht. Die Gründe für diese Gemächlichkeit können darin liegen, dass sich einige Hochschullehrer[1] mit neuen Technologien schwertun und mit Daten operieren (wollen), für die die alten Methoden wie der *t*-Test geeignet sind. Frei nach Max Planck kann man behaupten, dass alte Lehrmeinungen dann erst das Zeitliche segnen, wenn das auch die Professoren tun, die die Lehrmeinungen vertreten. Vielleicht liegt es auch schlicht daran, dass unser Alltag in den meisten Belangen wenig von der Digitalisierung und von Algorithmen berührt scheint: Beim Bäcker grüßt man wie seit Altvaterzeiten; deutsche Autos rollen vom Band wie seit dem Wirtschaftswunder; Schüler und Studenten lesen in ihren Büchern, wie es ihnen in Preußen, als das deutsche Schulwesen seine Anfänge fand, eingebläut wurde (Foucault 1994). Die Revolution der Daten ist kaum spürbar; sie fühlt sich weit weg an.

Aber ein Wechsel „unter den Talaren“ zeichnet sich ab: Statistiker mit Renommee rufen dazu auf, das Datenzeitalter im Unterricht einzuläuten (Cobb 2007; Hardin et al. 2015): Datenanalyse heute ist anders als gestern (s. folgende Abbildung). Immer mehr Lehrbücher zu moderner Statistik und Datenanalysen erscheinen, auch richtig gute (Baumer et al. 2017; z. B. James et al. 2013; McElreath 2015; Wickham und Grolemund 2016). Bislang zumeist im englischen Sprachraum, aber es gibt auch zunehmend mehr deutschsprachige Bücher (z. B. Wickham und Grolemund (2017)).

[1] Aus Gründen der Lesbarkeit wird in diesem Buch das generische Maskulinum („der Leser“) verwendet; immer sind alle Geschlechter gleichermaßen gemeint.



Das vorliegende Buch versucht, einen Teil der Lücke im deutschsprachigen Raum zu schließen. Sie werden in diesem Buch die grundlegenden Ideen der modernen Datenanalyse lernen. Ziel ist es, Sie – in Grundzügen – mit moderner Statistik vertraut zu machen.[2]

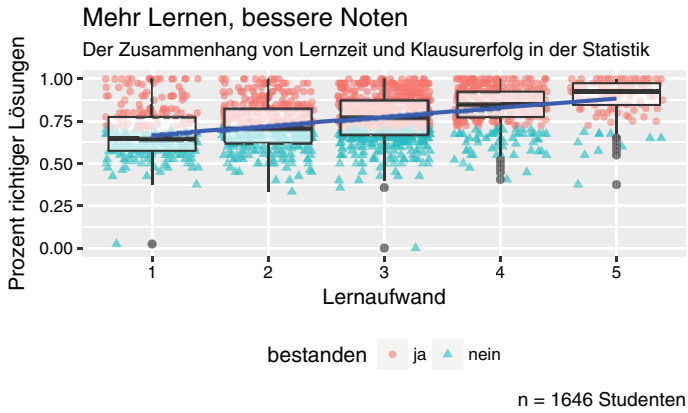
Lernziele

Zielgruppe dieses Buches sind Einsteiger; Formeln und mathematische Hintergründe sucht man meist vergebens. Das liegt zum einen daran, dass keine oder kaum Vorkenntnisse in Datenanalyse vorausgesetzt werden. Zum anderen beruht das Buch auf einem didaktischen Ansatz, der das *Tun* vor das *Wissen* setzt. Das bedeutet nicht, dass Wissen geringer geschätzt würde als Handeln. Vielmehr steht dahinter die Idee, dass es dem Verstehen und dem *statistical thinking* (C. J. Wild und Pfannkuch 1999) hilft, sich frühzeitig mit dem Ausprobieren auseinanderzusetzen. Der Werkzeug- oder Problemlöse-Charakter des Denkens steht im Mittelpunkt des Lernansatzes dieses Buches (vgl. Trilling und Fadel (2012)). Kein Inhalt dieses Buches bleibt ohne Umsetzung, ohne Anwendung; es ist ein Buch für Praktiker. Wer eine tiefere, mathematisch ansprechendere Einführung sucht, sei an das exzellente Buch von Hastie et al. (2013) verwiesen.

Reines Lesen dieses Buches wird dem Anfänger in etwa so viel bringen wie die Lektüre einer Schwimmfibel. Umgekehrt ist Üben die Grundlage für Fortschritt in der Kunst der Datenanalyse (s. folgende Abbildung).[3] Nutzen Sie die Übungsangebote: die reichhaltige R-Syntax, die Daten, die Aufgaben und die Verweise zu weiterführender Literatur. Wesentlich ist das Durcharbeiten der Syntax-Beispiele. Die Kapitel sind zum Teil in sich abgeschlossen; die Grundlagen (bis einschließlich Kap. 7) werden durchgängig benötigt. Eine Lektüre von vorne nach hinten ist ratsam, aber nicht zwangsläufig nötig, gerade für fortgeschrittene Leser. Vergleichsweise schwierig sind die Kap. 28 und 29.

[2] Statistiker, die diesem Buch als Vorbild Pate standen, sind: Roger D. Peng: http://www.biostat.jhsph.edu/~rpeng/, Hadley Wickham: http://hadley.nz, Jennifer Bryan: https://github.com/jennybc.

[3] Die Abbildung zeigt den Zusammenhang von Klausurerfolg und Vorbereitungsaufwand. Man sieht, dass der Klausurerfolg (Y-Achse) tendenziell steigt, wenn der Vorbereitungsaufwand (X-Achse) steigt.



Nach der Lektüre dieses Buches können Sie:

- den Ablauf eines Projekts aus der Datenanalyse in wesentlichen Schritten nachvollziehen
- Daten aufbereiten und ansprechend visualisieren
- Inferenzstatistik anwenden und kritisch hinterfragen
- klassische Vorhersagemethoden (Regression) anwenden
- moderne Methoden der angewandten Datenanalyse anwenden (z. B. Textmining)
- (wirtschaftliche) Fragestellungen mittels datengetriebener Vorhersagemodelle beantworten.

Zur Didaktik

Im Gegensatz zu anderen vergleichbaren Kursen steht hier die Umsetzung mit R (R Core Team 2018) im Vordergrund. Dies hat pragmatische Gründe: Möchte man Daten einer statistischen Analyse unterziehen, so muss man sie zumeist erst aufbereiten, und zwar oft mühselig. Selten kann man den Luxus genießen, einfach „nur“, nach Herzenslust sozusagen, ein Feuerwerk an multivariater Statistik abzubrennen. Zuvor gilt es, die Daten umzuformen, zu prüfen und zusammenzufassen. Für beide Anforderungen ist R bestens geeignet. Dem Teil des Aufbereitens der Daten ist hier ausführlich Rechnung getragen. Außerdem spielt in diesem Kurs die Visualisierung von Daten eine große Rolle. Ein Grund ist, dass Menschen bekanntlich Augentiere sind. Zum anderen bieten Diagramme bei umfangreichen Daten Einsichten, die sonst leicht wortwörtlich übersehen würden.

Lovett und Greenhouse (2000) leiten aus der kognitiven Theorie fünf Prinzipien zur Didaktik des Statistikunterrichts ab; nach diesen Prinzipien ist dieses Buch ausgerichtet. (1) *Menschen lernen am meisten durch das und von dem, was sie selber ausprobieren*: Das Selber-Tun steht im Zentrum dieses Buches. (2) *Wissen ist situiert, kontextspezifisch*: Im Unterricht bzw. in einem Buch sollte daher lebensnah und alltagsrelevantes Wissen

vermittelt werden; die Beispiele und Methoden dieses Buches sind aus typischen oder verbreiteten Fragestellungen des Wirtschaftslebens entnommen. (3) *Direktes Feedback verbessert das Lernergebnis*: Das sofortige Ausprobieren anhand der R-Syntax gibt unmittelbares Feedback, ob ein Plan aufgegangen ist. Wie beim Jonglieren: Wenn ein Ball zu Boden fällt, weiß man, es ist ein Fehler passiert. Ähnlich verhält es sich, wenn R nichts oder Kauderwelsch ausspuckt. (4) *Lernen geschieht beim Verbinden von Bekanntem mit Neuem*: Der Sprachduktus ist informell, da viele Themen gerade in Bereichen nahe der Mathematik nicht wegen des Inhalts, sondern wegen der Formalisierung kompliziert werden. Freilich setzt man mit informeller Sprache Genauigkeit und Detailtiefe aufs Spiel. Da es sich aber um ein für die avisierte Leserschaft neues Thema handelt, neigt sich die Waage hier zugunsten der informellen, intuitiven Herangehensweise. Für fortgeschrittene Leser ist dieses Buch daher weniger geeignet. (5) *Die mentale Belastung sollte ausgewogen sein*: Das Buch beginnt mit grundlegenden Themen; die vergleichsweise schwierigen warten gen Ende. Jedes Kapitel behandelt nur ein Thema, um geistige Ressourcen effektiv zu nutzen. Die „R-Philosophie“ dieses Buches orientiert sich am „Tidyverse“ (vgl. Wickham und Grolemund (2017)); alle Kapitel und alle R-Syntax sind diesem Paradigma verhaftet. Sie werden schnell den ähnlichen Aufbau der Syntax in allen Kapiteln erkennen. Erfahrenen R-Programmierern wird der ausgiebige Gebrauch der „Pfeife“ aus `magrittr` auffallen; genauso wie der ausgiebige Gebrauch von `dplyr` und anderen Figuren aus dem „Tidyverse“.

Icons

R spricht zu Ihnen; sie versucht es jedenfalls in diesem Buch, und zwar mit folgenden Icons (Fonticons 2018).

R-Pseudo-Syntax: An vielen Stellen dieses Buches findet sich R-Syntax. Neue oder kompliziertere Syntax ist Zeile für Zeile ins Deutsche übersetzt.

Achtung, aufgepasst: Schwierige, merkwürdige oder fehlerträchtige Stellen sind mit diesem Symbol markiert.

Übungsaufgaben: In jedem Kapitel finden sich Übungsaufgaben. Auf diese wird mit diesem Icon verwiesen oder die Übungen sind in einem Abschnitt mit einsichtigem Titel zu finden.

Hinweise

Kunstwerke (Bilder) sind genau wie Standard-Literatur im Text zitiert, die verwendeten R-Pakete nur im Anhang. Alle Werke (auch Daten und Software) finden sich im Literaturverzeichnis. Dieses Buch wurde mit dem R-Paket `bookdown` basierend auf R (R Core Team 2018) in RStudio (RStudio 2018) geschrieben. `bookdown` basiert wiederum u. a. auf den R-Paketen `knitr` und `rmarkdown`. Norman Markgrafs Typografie-Paket hat den

Textsatz geschliffen (Markgraf 2018). Diese Pakete stellen großartige Funktionalität zur Verfügung, kostenlos und quelloffen.

Ein Hinweis für Dozenten: Der Text enthält an vielen Stellen Abbildungen oder kurze Zusammenfassungen, die sich für Folien eignen. Getreu der Philosophie, nach der Folien keine Konkurrenz zum Buch sind, sondern nur Wesentliches illustrieren, sollten vor allem diese Bilder in einen möglichen Foliensatz für eine Lehrveranstaltung eingehen. Das hat den charmanten Vorteil, dass die Erstellung der Folien einfach von der Hand geht. Im Anhang findet sich ein Vorschlag für das Curriculum einer Einführungsveranstaltung (s. Abschn. A.2.4).

Zu diesem Buch gibt es eine Webseite: https://sebastiansauer.github.io/modar/. Dort finden sich die Abbildungen, die Syntax und mehr. Feedback, Fragen und Hinweise können Sie dort auch einstellen.

Danke

Dieses Buch ist mit der Hilfe Vieler und dank glücklicher Umstände entstanden. Von meinen Kollegen mit ihren Hinweisen, ihrem Ansporn, ihrem Lob und ihrer Kritik habe ich vielfältig profitiert; das hat dieses Buch geschliffen. Herausgreifen möchte ich Karsten Lübke, von dem ich viel gelernt habe. Norman Markgraf hat sein Wissen großzügig mit mir geteilt; für viele, gerade technische Hilfestellungen bin ich ihm dankbar. Weiter danke ich meiner Frau Sabrina, Christoph Kurz, Felix Bauer, Moritz Körber und Oliver Gansser für ihr Feedback zum Manuskript. Der Austausch mit den Kollegen vom ifes-Institut war fruchtbar und schwungvoll. Dieses Buch baut vielfach auf dieser Zusammenarbeit auf. Ich danke der Hochschulleitung der FOM für ihre Unterstützung und den Vorschuss an Vertrauen, was mein Arbeiten freudvoll machte; Gleiches gilt für den Dekan der Wirtschaftspsychologie an der FOM, Christoph Berg, der den Weg ebnete für neue Ideen, wie sie auch in diesem Buch Eingang fanden. Mein Dank gilt außerdem Kai Stumpp für die Begleitung bei der Erstellung des Buches. Nicht zuletzt danke ich meinen Studenten; ich weiß nicht, wer von wem mehr gelernt hat: sie von mir oder ich von ihnen, dank vieler Fragen und Hinweise. Vermeintlich „dumme“ Fragen zielen oft in die Mitte des Wesentlichen und die meisten komplexen Dinge kann man in einfachen Worten erklären, wenn man sie verstanden hat. Ohne umfangreiche Open-Source-Software wäre dieses Buch nicht entstanden; viele Menschen haben unentgeltlich mitgewirkt. Solcher Reichtum verblüfft mich immer wieder.

Ich hoffe, dass Sie mit diesem Buch einiges Handwerkszeug der modernen Datenanalyse lernen; dass Sie Gefallen an der „Kunst und Wissenschaft“ der Datenanalyse finden. Für Ihre Anregungen, Hinweise zu Fehlern und Ideen bin ich dankbar; am besten stellen Sie sie hier ein: https://github.com/sebastiansauer/modar/issues.

Nürnberg Sebastian Sauer
im Sommer 2018

Inhaltsverzeichnis

Teil VI Geleitetes Modellieren

Der Autor

Professor Dr. habil. Sebastian Sauer arbeitet als Hochschullehrer für Wirtschaftspsychologie an der FOM und versteht sich als Data Scientist. Datenanalyse mit R ist sein zentrales Thema im Moment. Neben dem „Wie“ der Datenanalyse beschäftigen ihn die Grenzen sowie die Gefahren, die die moderne Datenwissenschaft mit sich bringt. Außerdem engagiert er sich für das Thema Open Science und interessiert sich für die Frage, wie die Psychologie zur Klärung von Problemen mit gesellschaftlicher Relevanz beitragen kann. Sein Blog https://data-se.netlify.com/ dient ihm als Notizbuch sich entwickelnder Gedanken. Data Science für die Wirtschaft bietet er auf https://www.data-divers.com/ an.

Teil I
Rahmen

Statistik heute

1

Datenanalyse, praktisch betrachtet, kann man in fünf Schritte einteilen (Wickham und Grolemund 2017), s. Abb. 1.1. Analog zu diesem Modell der Datenanalyse ist dieses Buch aufgebaut. Zuerst muss man die Daten *einlesen*, die Daten also in R (oder einer anderen Software) verfügbar machen (laden). Fügen wir hinzu: In schöner Form verfügbar machen; das man nennt auch Tidy Data (hört sich cooler an). Sobald die Daten in geeigneter Form in R geladen sind, folgt das *Aufbereiten*. Das beinhaltet das Zusammenfassen, Umformen oder Anreichern der Daten, je nach Bedarf. Ein nächster wesentlicher Schritt ist das *Visualisieren* der Daten. Ein Bild sagt bekanntlich mehr als tausend Worte. Schließlich folgt das *Modellieren* oder das Prüfen von Hypothesen: Man überlegt sich, wie sich die Daten erklären lassen könnten. Zu beachten ist, dass diese drei Schritte – Aufbereiten, Visualisieren, Modellieren – keine starre Abfolge sind, sondern eher ein munteres Hin-und-Her-Springen, ein aufeinander aufbauendes Abwechseln. Der letzte Schritt ist das *Kommunizieren* der Ergebnisse der Analyse – nicht der Daten. Niemand ist an Zahlenwüsten interessiert; es gilt, spannende Einblicke zu vermitteln. Die Datenanalyse als solche ist in einen *Rahmen* eingebettet; das beinhaltet philosophische und technische Grundlagen. Entsprechend diesen fünf Schritten sowie dem einbettenden Rahmen ist dieses Buch in Teile gegliedert. Zu Beginn jedes Teiles ist ein Diagramm analog zu 1.1 dargestellt, um einen Überblick über den jeweiligen Schritt der Datenanalyse zu geben.

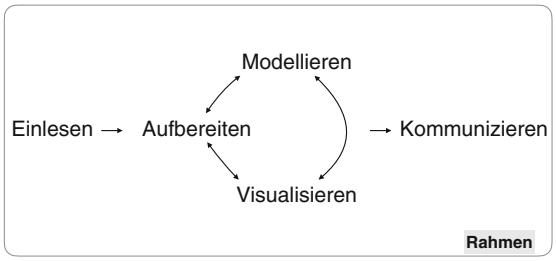


Abb. 1.1 Der Rahmen als Bestandteil der Datenanalyse

S. Sauer, *Moderne Datenanalyse mit R*, FOM-Edition,
https://doi.org/10.1007/978-3-658-21587-3_1

Lernziele

- Wissen, was Statistik ist, bzw. einige Aspekte einer Definition kennen
- Statistik zu Data Science und anderen verwandten Begriffen abgrenzen können
- Grundkonzepte wie Daten, Variable und Beobachtung definieren können
- Die Begriffe *Signal* und *Rauschen* in Verbindung bringen können
- Die Wissensgebiete der Datenanalyse aufzählen und erläutern können

1.1 Datenanalyse, Statistik, Data Science und Co.

Was ist Statistik? *Eine* Antwort dazu ist, dass Statistik die Wissenschaft von Sammlung, Analyse, Interpretation und Kommunikation von Daten mithilfe mathematischer Verfahren ist und zur Entscheidungshilfe beitragen solle (Romeijn 2016; *The Oxford Dictionary of Statistical Terms* 2006). E. N. Brown und Kass (2009) rücken die Wahrscheinlichkeitsrechnung ins Zentrum einer Definition von Statistik oder „Statistik-Denken". Cobb (2015, S. 3) spricht kurz und bündig von „thinking with and about data" als dem Wesensmerkmal von Statistik.

Wie lässt sich Statistik von Datenanalyse abgrenzen? Tukey (1962) definiert Datenanalyse als den Prozess des Erhebens von Daten, ihrer Auswertung und Interpretation. Unschwer zu sehen, dass sich diese beiden Definitionen nur um die Betonung der stochastischen Modellierung, der Anwendung der Wahrscheinlichkeitsrechnung, unterscheiden. Betrachtet man Lehrbücher der Statistik (Bortz 2013; Freedman et al. 2007), so fällt der stärkere Fokus auf mathematische Ableitung und Eigenschaften von Objekten der Statistik ins Auge; Datenanalyse scheint einen stärkeren Anwendungsfokus zu haben (im Gegensatz zu einem mathematischen Fokus). Statistik wird häufig in die zwei Gebiete *deskriptive* und *inferierende* Statistik eingeteilt (vgl. Abb. 1.2). Erstere fasst viele Zahlen zusammen (s. Kap. 8), so dass wir den Wald statt vieler Bäume sehen. Eine *Statistik* bezeichnet dabei die zusammenfassende Kenngröße; eine prototypische Statistik ist der Mittelwert. Die Inferenzstatistik verallgemeinert von den vorliegenden Daten auf eine zugrunde liegende Grundgesamtheit (Population; s. Kap. 16). So zieht man etwa eine Stichprobe von einigen College-Studenten und schließt auf dieser Basis auf alle Menschen dieser Welt. Ein abenteuerlicher Schluss, aber leider kein seltener (Henrich et al. 2010). Da *Analyse von Daten* ein allgemeiner Begriff ist, der wenig mit bestimmten Methoden aufgeladen ist, wird Datenanalyse im Folgenden als gemeinsamer Kern aller einschlägigen Disziplinen oder Begrifflichkeiten verwendet; der Fokus ist dabei als angewandt gedacht. In diesem Buch werden Statistik und Datenanalyse im Folgenden lose als Synonyme betrachtet.

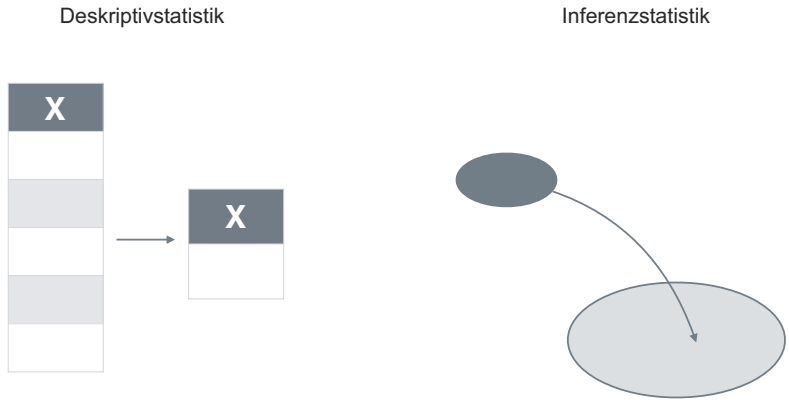


Abb. 1.2 Sinnbild für die Deskriptiv- und die Inferenzstatistik

Aufgabe der deskriptiven Statistik ist es primär, Daten prägnant zusammenzufassen. Aufgabe der Inferenzstatistik ist es, zu prüfen, ob Daten einer Stichprobe auf eine Grundgesamtheit verallgemeinert werden können.

Liegt der Schwerpunkt der Datenanalyse auf computerintensiven Methoden, so wird auch von *Data Science* gesprochen, wobei der Begriff nicht einheitlich verwendet wird (Hardin et al. 2015; Wickham und Grolemund 2017). Einige Statistiker sehen Data Science als „Statistik am Computer“, und plädieren für ein „Neudenken“ der Statistik bzw. des Statistikunterrichts, so dass Computermethoden eine zentrale Rolle spielen (Cobb 2015). Andere Statistiker wiederum grenzen Data Science von der Statistik ab, mit dem Argument, dass bei Ersterer Fragen der angewandten Informatik zentral sind, bei der Statistik nicht (Baumer et al. 2017). Die Popularität von Data Science ist dem Fortschritt in der Rechen- und Speicherkapazität der Computer zu verdanken. Heutzutage sind ganz andere Daten und Datenmengen verarbeitbar und von Interesse. Sicherlich bedingt die technische Machbarkeit auch, welche Forschungsfragen hoch im Kurs rangieren. Eine Echtzeit-Analyse von Twitter-Daten wäre bis vor einiger Zeit kaum möglich gewesen, da die Hardware nicht leistungsfähig genug war. (Wir übersehen hier geflissentlich, dass die Hardware heute immer noch an Grenzen kommt und dass es früher kein Twitter gab.) Die Datenmengen erfordern Arbeitsschritte wie z. B. Authentifizierung in die Twitter-Schnittstelle, wiederholtes Abfragen der Schnittstelle unter Beachtung der erlaubten Download-Obergrenzen, Umwandeln von einem Datenformat in ein anderes, Einlesen in eine Datenbank, Prüfung auf Programmfehler, Automatisierung des Prozesses, Verwendung mehrerer Rechenkerne, Bereinigung des Textmaterials von Artefakten, Aufbereiten der Daten und so weiter. Keine dieser Aufgaben war bei Statistikern vor 100 Jahren verbreitet. Da konnte man sich noch ganz auf die mathematischen Eigenschaften

des t-Tests konzentrieren. Die Verschiebung innerhalb der Datenanalyse spiegelt einfach die technische Entwicklung allgemein in der Gesellschaft wider. Das heißt nicht, dass heute jeder Programmierer sein muss; auch für die Datenanalyse nicht. Glücklicherweise gibt es eine Reihe von Werkzeugen, die die Handhabung der Daten einfacher macht. Diese sind hinzugekommen zum Handwerkskoffer des Datenschreiners. Nichtsdestotrotz ist ein grundlegendes Verständnis von computergestützter Datenverarbeitung wichtig und wird zunehmend wichtiger für die Datenanalyse.

Schließlich kursieren u. a. noch die Begriffe *Data Mining*, *maschinelles Lernen* und *statistisches Lernen*. Data Mining und maschinelles Lernen sind Begriffe, die eher in informatiknahen Gebieten verwendet werden; entsprechend sind die Themen mehr in Richtung Informatik verschoben. Die technische Repräsentation und technische Aspekte der Datenmanipulation (Data Warehouse, Datenbanken) werden von IT-affinen Autoren stärker betont als von Autoren, die nicht aus IT-nahen Fakultäten stammen. Beim statistischem Lernen stehen Konzepte und Algorithmen des Modellierens (s. Kap. 15) im Vordergrund. Auf der anderen Seite: Vergleicht man die Inhaltsverzeichnisse von Büchern aus allen diesen Bereichen, so stellt man eine große Überschneidung des Inhaltsverzeichnisses fest. Typische Themen in allen diesen Büchern sind (Bishop 2006; Han et al. 2011; James et al. 2013; Tan 2013) baumbasierte Verfahren (s. Kap. 21), Support Vector Machines, Dimensionsreduktion, Clusteranalyse, Regression (s. Kap. 18) oder Datenexploration (s. Kap. 11). Kurz: Die Gebiete überlappen einander beträchtlich; mal stehen IT-nahe Themen im Vordergrund, mal wird die Mathematik betont, mal die Anbindung an empirisches Forschen. Immer geht es darum, aus Daten Wissen zu generieren bzw. Entscheidungen datenbasiert und rational zu begründen.

1.2 Wissensgebiete der Datenanalyse

Egal, ob man von Datenanalyse, Statistik, Data Mining, maschinellem Lernen, Data Science oder statistischem Lernen spricht: In der Schnittmenge des Analysierens von Daten gleichen sich die Anforderungen. Wenn sich die Nuancen zwischen den Fachgebieten auch verschieben, so sind doch stets die folgenden Wissensgebiete gefragt:

1. Philosophische Grundlagen
   Dazu gehören die Annahmen, die im Alltag meist unhinterfragt für bare Münzen genommen werden. Annahmen, die das Fundament des Gebäudes der Datenanalyse stützen. Ist dieses Fundament auf Sand gebaut, so ist nicht zu erwarten, dass das Gebäude seinen Zweck erfüllt. Zu den Grundlagenfragen gehört die Frage, was Wahrscheinlichkeit, Erkenntnis, Kausalität und Unendlichkeit sind. So wird zum Beispiel kontrovers diskutiert, ob Wahrscheinlichkeit besser als Grenzwert der relativen Häufigkeit, als subjektive Angelegenheit oder als Erweiterung der Aussagenlogik zu betrachten ist (Briggs 2016; Jaynes 2003; Keynes 2013; Rucker 2004). Wichtig ist weiter die Frage, was eine Messung genau ist, woran man erkennt, ob ein Variable quantitative Aus-

sagen erlaubt und woran man die Güte eines Messinstruments festmacht (Saint-Mont 2011).

2. Mathematisch-statistische Anwendungen
   Zentrale Theorie für die Statistik oder für Wissenschaft allgemein ist die Wahrscheinlichkeitsrechnung bzw. die Theorie der Wahrscheinlichkeit (Jaynes 2003). In den meisten Lehrbüchern der Statistik, auch in den Einsteigerbüchern, findet sich eine mal schmalere, mal ausführlichere Einführung in Aspekte verschiedener Verteilungen, die gewisse Zufallsprozesse, berechenbare Zufallsprozesse, voraussetzen.[1] Häufig wird angenommen, dass sich eine bestimmte Variable nach einem bekannten stochastischen Modell verhält, so dass aus dem Modell Aussagen ableitbar sind. Besondere Berühmtheit hat die Normalverteilung erlangt, die wohl an keinem Studenten eines empirisch orientierten Faches vorbeigegangen ist.[2] Neben der Stochastik spielen aber noch weitere Felder der angewandten Mathematik eine Rolle; maßgeblich sind das die lineare Algebra und die Infinitesimalrechnung (J. D. Brown 2015). Ein eigener Zweig, der stark mit der Wahrscheinlichkeitslehre verbunden ist, ist die Bayes-Statistik (Wagenmakers et al. 2016).
3. Computerwissenschaftliche Anwendungen
   Die zunehmende Digitalisierung der Gesellschaft macht vor der Datenanalyse keinen Halt. Im Gegenteil; es liegt in der Natur der Datenanalyse, computeraffin zu sein – sind doch Daten Gegenstand sowohl der Statistik als auch der Computerwissenschaft. War es vor einigen Jahren noch ausreichend oder beeindruckend, den Knopf für den t-Test zu kennen, sind die Anforderungen bei modernen Daten meist höher. Einlesen, Aufbereiten und Speichern können leicht den größten Zeitanteil einer Datenanalyse ausmachen. In einigen Anwendungen kommt noch der Anspruch dazu, dass die Analyse schnell gehen muss: Wir haben gerade viele Kunden auf der Webseite und müssen wissen, wem wir welches Produkt und welchen Preis vorschlagen müssen. Nein, wir wollen die Antwort nicht morgen, wir brauchen sie in ein paar Sekunden; Korrektur: jetzt. Ach ja, der Datensatz ist ein paar Terabyte[3] groß.
4. Fach- und Branchenkenntnis
   Möchte man die Zufriedenheit eines Kunden vorhersagen, so ist es hilfreich, etwas über die Ursachen von Kundenzufriedenheit zu kennen – also Wissen über den Gegenstand Kundenzufriedenheit zu haben. Wenn man schon die Ursachen nicht kennt, so ist zumindest Wissen über zusammenhängende Variablen (Korrelate) sinnvoll. Wenn ein Arzt aus Erfahrung weiß, was die Risikofaktoren einer Erkrankung sind, dann sollten diese Informationen in das statistische Modell einfließen. Sach- inkl. Branchenkennt-

[1] Zufall wird hier verstanden als ein nicht näher bekannter Prozess, dessen Ergebnisse zwar nicht sicher sind, aber doch ein gewisses Muster erwarten lassen.
[2] Micceri (1989) zeigt auf, dass Normalverteilungen seltener sind als gemeinhin angenommen. McElreath (2015) bietet einen gut verständlichen Einblick in die Informationsentropie; diese Darstellung zeigt, dass eine Normalverteilung eine konservative Annahme für eine Verteilung darstellt.
[3] 1 Terabyte sind 10^{12} Byte.

nis ist zentral für gute (genaue) statistische Modelle (Shearer 2000). Wissen über den Sachgegenstand ist schon deshalb unerlässlich, weil Entscheidungen keine Frage der Statistik sind: Ob ein Kunde zufrieden ist mit einer Vorhersage durch ein statistisches Modell oder ein Patient gesund genug ist nach Aussage eines statistisches Modells, muss der Anwender entscheiden. Ob ein Betrugsversuch mit 90 %, 99 % oder 99.9 % Sicherheit erkannt werden soll, kann einen Unterschied machen – für einen bestimmten Anwender, in einer bestimmten Situation. Ein wichtiger, vielleicht der wichtigste Punkt der Datenanalyse ist es, Entscheidungen für Handlungen zu begründen. Daher muss die Datenanalyse immer wieder auf die Entscheidung und damit auf die Präferenz des Nutzers zurückgeführt werden.

1.3 Einige Grundbegriffe

Daten (die Einzahl *Datum* ist ungewöhnlich) kann man definieren als Informationen, die in einem Kontext stehen (Moore 1990), wobei eine numerische Konnotation mitschwingt. Häufig sind Daten in *Tabellen* bzw. tabellenähnlichen Strukturen gespeichert; die Excel-Tabelle ist der Prototyp davon. Tabellen, so wie sie hier verstanden werden, zeichnen sich dadurch aus, dass sie *rechteckig* sind und aus Zeilen und Spalten bestehen. Rechteckig impliziert, dass alle Zeilen gleich lang sind und alle Spalten gleich lang sind (die Tabelle muss aber nicht quadratisch sein). Knotenpunkte von Zeilen und Spalten heißen *Zellen* oder *Elemente*; die Zellen dürfen auch leer sein oder mit einem Symbol für „kein Wert vorhanden“ gefüllt sein.

Daten sind ein Produkt von *Variablen* (Merkmalen) und *Beobachtungseinheiten* (Fällen, Beobachtungen). Eine Tabelle mit ihren zwei rechtwinkligen Achsen der Zeilen und Spalten verdeutlicht das (s. Abschn. 9.3.1 und Abb. 9.6). Die Beobachtungseinheit ist das Objekt, das die untersuchten Merkmale aufweist. Oft sind es Personen, es können aber auch Firmen, Filme, Filialen oder Flüge sein. Ein Merkmal einer Beobachtungseinheit kann ihre Schuhgröße, der Umsatz, die Verspätung oder das Budget sein. Betrachten wir das Merkmal *Schuhgröße* der Beobachtungseinheit *Person S* und finden wir *46*, so ist *46* der *Wert* oder die Ausprägung dieses Merkmals dieser Beobachtungseinheit. Um nicht so viel schreiben zu müssen, wird der Wert der Beobachtungseinheit i in der Variablen k häufig als x_{ik} bezeichnet. Die Gesamtheit der verfügbaren und zusammengehörigen Daten eines Sachverhalts bezeichnen wir als *Datensatz*; meist ist es eine Stichprobe aus einer Population. Nehmen wir an, es gäbe nur zwei *verschiedene* Schuhgrößen: 36 und 46. Dann hat die Variable Schuhgröße zwei *Ausprägungen*.

Natürlich gibt es auch Daten, die sich nicht (so einfach) in das enge Korsett einer Tabelle pressen lassen; Textdokumente, Sprachdaten oder Bilder zum Beispiel. Nicht tabellarisierte Daten werden auch als *unstrukturiert*, Daten in Tabellenform als *strukturiert* bezeichnet. Rein mengenmäßig überwiegen unstrukturierte Daten in der Welt. Allerdings sind strukturierte Daten einfacher zu verarbeiten; wir werden uns auf diese konzentrieren.

1.4 Signal und Rauschen

Die Aufgabe der Wissenschaft – oder sogar jeglichen Erkennens – kann man als zweistufigen Prozess betrachten: erstens Signale (Phänomene) erkennen und zweitens diese dann erklären (Bogen und Woodward 1988; Silver 2012). *Signale erkennen* ist der Versuch, aus Daten ein Muster, Regularitäten, herauszulesen. Das impliziert, dass Daten dieses Muster nicht direkt offenbaren und nicht identisch mit dem Muster sind. Mit dem alten Bild, im Rauschen ein (leises) Geräusch, das Signal, herauszuhören, ist der Sachverhalt gut beschrieben (Haig 2014). Daten sind vergänglich, veränderlich, vorübergehend – und für den erkennenden Verstand ohne Interesse. Das Muster ist es, welches von alleinigem Interesse ist. Schält sich ein Muster heraus, ist es über die Zeit, Erhebungsmethode und Situation hinweg stabil, so hat man ein *Phänomen* identifiziert. Im Gegensatz zu Daten ist ein Phänomen unbeobachtbar; es ist die Abstrahierung der Gemeinsamkeit aus der Konkretheit der Daten. Ein Phänomen könnte sein, dass „Männer, die Windeln kaufen, auch Bier kaufen". Ist ein Phänomen identifiziert, so ist die wichtigste Frage zumeist, was die Ursache des Phänomens ist. Wissenschaftliche und Alltagstheorien untersuchen Phänomene, nicht Daten. Warum ist es wichtig, Ursachen von Phänomen zu kennen? Ein Grund ist, dass man das Auftreten eines Phänomens beeinflussen kann, wenn man seine Ursache kennt. Überspitzt formuliert: Die Ursache der Entzündung sind Bakterien? Entferne die Bakterien, und die Entzündung klingt ab.

Abb. 1.3 stellt den Unterschied (und den Zusammenhang) von Rauschen und Signal dar. Der linke Teil des Diagramms zeigt die Körpergröße einer Reihe von Frauen und Männern (und damit von zwei Variablen). Die Beobachtungseinheit, als schwarzer Punkt dargestellt, ist eine Person. Wie man sieht, unterscheiden sich die Körpergrößen der Personen; einige sind größer, andere kleiner. Auch zwischen den *Geschlechtern* gibt es Unterschiede. Unser Auge ist schnell mit der Erkenntnis des Musters, dass „Männer größer sind als Frauen". Nicht alle Männer sind größer als alle Frauen; einige Frauen sind größer als einige Männer. Aber in den meisten Fällen gilt: Der Mann ist größer als die Frau. Anders betrachtet: Der „mittlere" Mann ist größer als die „mittlere" Frau (als Quadrat bzw. Linie dargestellt im rechten Teil der Abb. 1.4). Als Beobachtung in den Daten existiert aber weder die mittlere Frau noch der mittlere Mann; wir erkennen dies als Phänomen in den Daten bzw. aus den Daten heraus.

Das Rauschen in den Daten kann vielerlei Ursachen haben: Messfehler, Besonderheiten der ausgewählten Merkmalsträger oder der Situation. Zufall ist keine Ursache im strengen Sinne des Wortes; in der Regel wird dieser Begriff verwendet, wenn man die wahre Ursache nicht kennt.[4] Experimentieren ist nichts anderes als die Kunst, Rauschen *vor der Messung* zu verringern. Genauer gesagt solches Rauschen zu verringern, welches das Signal sonst überlagert. Analog kann man sagen, dass Datenanalyse das Ziel hat, Rau-

[4] Diesem Gedanken hinterliegt ein deterministisches Weltbild; strittige quantentheoretische Phänomene (Jaynes 2003) und der (meiner Meinung nach) freie menschliche Wille sind davon ausgeklammert.